The Undergraduate Companion to Theoretical Physics

Andrea Kouta Dagnino[‡]

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

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

Contents

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

A1 Numbers

1.1   The set $\mathbb{N}$ of Natural Numbers

Peano Axioms

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

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

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

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

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

N5. Let $A \subset \mathbb{N}$ which contains 1 and contains $\mathrm{succ}(n)$ 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 $S = \{n \in \mathbb{N} | \, n \notin A\}$. 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 $P_1, P_2...$ be a list of propositions, then the principle of mathematical induction asserts that they are true provided:

$\mathbf{I}_1$ P_1 is true (basis of induction)

$\mathbf{I}_2$ $(P_n$ is true$) \implies (P_{n+1}$ is true$)$ (inductive step)

> **Proposition 5.1** *(Sum of natural numbers)*

The sum of the first n natural numbers is:

$$\sum_{i=1}^{n} = \frac{n(n+1)}{2}. \tag{1.1.1}$$

Proof. We define the nth proposition to be:

$$P_n : \sum_{i=1}^{n} = \frac{n(n+1)}{2} \tag{1.1.2}$$

$\mathbf{I_1}$ 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.

$\mathbf{I_2}$ For the inductive step, suppose P_n is true, so we assume:

$$\sum_{i=1}^{n} = \frac{n(n+1)}{2} \tag{1.1.3}$$

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 nth proposition is:

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

$\mathbf{I_1}$ The basis for induction is true, since $5^1 - 4 - 1 = 0$ which is divisible by 16.

$\mathbf{I_2}$ For the inductive step, suppose P_n is true, we wish to verify P_{n+1}. To do so, we write:

$$5^{n+1} - 4(n+1) - 1 = 5(5^n - 4n - 1) + 16n = 5 \cdot 16m + 16n = 16(5m + n) \tag{1.1.4}$$

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 nth proposition is:

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

$\mathbf{I}_1$ The basis for induction is clearly true, since $|\sin x| \leq |\sin x|$.

$\mathbf{I}_2$ 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:

$$(1+x)^n \geq 1 + nx, \quad \text{when } x \geq -1 \tag{1.1.5}$$

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

$$P(n) : (1+x)^n \geq 1 + nx \tag{1.1.6}$$

$\mathbf{I}_1$ $P(1)$ is obviously true, since the LHS reads $(1+x)^1 = (1+x)$ which is equal to the RHS.

$\mathbf{I}_2$ Now let $P(k)$ be true for some $k \geq 1$, so that:

$$(1+x)^k \geq 1 + kx \tag{1.1.7}$$

It follows that:

$$
\begin{aligned}
(1+x)^{k+1} &\geq (1 + kx)(1 + x) \tag{1.1.8} \\
&\geq 1 + (k+1)x + kx^2 \tag{1.1.9} \\
&\geq 1 + (k+1)x \tag{1.1.10}
\end{aligned}
$$

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

$$(1+x)^n \geq 1 + nx, \quad \text{when } x \geq -1 \tag{1.1.11}$$

as desired.

■

1.2 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:

$$c_n x^n + c_{n-1} x^{n-1} + ... + c_1 x + c_0 = 0 \tag{1.2.1}$$

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:

$$nx - m = 0. \tag{1.2.2}$$

$\blacksquare$

Theorem 5.4 *(Rational Zeros Theorem)*

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

$$c_n x^n + c_{n-1} x^{n-1} + ... + c_1 x + c_0 = 0 \tag{1.2.3}$$

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:

$$c_n \left(\frac{c}{d} \right)^n + c_{n-1} \left(\frac{c}{d} \right)^{n-1} + ... + c_1 \left(\frac{c}{d} \right) + c_0 = 0$$
$$c_n c^n + c_{n-1} c^{n-1} d + ... + c_1 c d^{n-1} + c_0 d^n = 0$$

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

$$c_0 d^n = -c[c_n c^{n-1} + c_{n-1} c^{n-1} d + ... + c_2 c d^{n-2} + c_1 d^{n-1}] \tag{1.2.4}$$

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 $c_n c^n$:

$$c_n c^n = -d[c_{n-1} c^{n-1} + c_{n-2} c^{n-2} d + ... + c_1 c d^{n-2} + c_0 d^{n-1}] \tag{1.2.5}$$

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:

$$x^n + c_{n-1}x^{n-1} + \ldots + c_1 x + c_0 = 0. \tag{1.2.6}$$

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:

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

which gives the polynomial equation:

$$x^6 - 6x^4 + 12a^2 - 13 = 0 \tag{1.2.7}$$

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

◀

1.3   The Set 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. $\forall a, b,\ a \leq b \lor b \leq a$ (multiplication inverse)

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

O3. $(a \leq b \land b \leq C) \implies a \leq c$

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:

$$b^p - a^p = (b - a) \underbrace{(b^{p-1} + b^{p-2}a + \ldots + ba^{p-2} + a^{p-1})}_{>0} \tag{1.3.1}$$

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

$$b - a > 0 \iff b^p - a^p > 0 \tag{1.3.2}$$

∎

1.4 Absolute Value

Definition 5.8 *(Absolute value and distance)*
We define the *absolute value* of a as:

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a \leq 0 \end{cases} \tag{1.4.1}$$

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 $a \geq 0 \land b \geq 0$, then $ab \geq 0 \implies |a| \cdot |b| = ab = |ab|$. If $a \leq 0 \land b \leq 0$, then $ab \geq 0 \implies |a| \cdot |b| = (-a)(-b) = ab = |ab|$. If $a \geq 0 \land b \leq 0$, then $ab \leq 0 \implies |a| \cdot |b| = a(-b) = -ab = |ab|$. If $a \leq 0 \land b \geq 0$, 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:

$$-|a| - |b| \leq a + b \leq |a| + |b| \tag{1.4.2}$$

so that

$$- (|a| + |b|) \leq a + b \leq |a| + |b| \tag{1.4.3}$$

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:

$$|a - b| \geq ||a| - |b|| \iff (a - b)^2 \geq (|a| - |b|)^2 \tag{1.4.4}$$
$$\iff a^2 - 2ab + b^2 \geq a^2 - 2|a||b| + b^2 \tag{1.4.5}$$
$$\iff -2ab \geq -|2ab| \tag{1.4.6}$$
$$\iff ab \leq |ab| \tag{1.4.7}$$

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

∎

1.5 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 $s = \max S$.
  (i) 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:

$$\{1, 2, \frac{1}{3}, 4, \frac{1}{5}, 6, \frac{1}{7}...\} \tag{1.5.1}$$

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) $\forall M_1 < M, \exists s_1 \in S$ 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 $\sqrt{7} \notin \mathbb{Q}$. 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, $\exists m \in \mathbb{R}$ 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. ∎

> **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. ∎

> **Theorem 5.15** *(Denseness 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:

$$a < r = \frac{m}{n} < b \implies an < m < bn \tag{1.5.2}$$

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

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:
>
> $$\sup(A + B) = \sup A + \sup B, \ \inf(A + B) = \inf A + \inf B \tag{1.5.3}$$

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 $\sup(A+B) \geq \sup A + \sup B$. ∎

A2 Sequences

2.1 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 nth 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 $(1, -1, 1, -1...)$, 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 $n = 1, 2, 3...$
>   (ii) **increasing (decreasing)** if $a_{n+1} \geq a_n$ $(a_{n+1} \leq a_n)$ for $n = 1, 2, 3...$
>   (iii) **strictly increasing** if $a_{n+1} > a_n$ $(a_{n+1} < a_n)$ for $n = 1, 2, 3...$
> 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 $a_n = (n-1)(n-2)$, $n = 1, 2...$, which is monotonic increasing. Indeed, note that:
>
> $$a_{n+1} - a_n = 2n - 2 \geq 0 \tag{2.1.1}$$
>
> 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:

$$\frac{a_{n+1}}{a_n} = \frac{n + 1 + \frac{1}{n+1}}{n + \frac{1}{n}} = \frac{(n+1)^2}{n^2} \cdot \frac{n}{n+1} = \frac{n+1}{n} > 1 \tag{2.1.2}$$

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:

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)^4}{n^4} \frac{4^n}{4^{n+1}} = \frac{1}{4}\left(\frac{n+1}{n}\right)^4 < 1 \implies 1 + \frac{1}{n} < \sqrt{2} \implies \frac{1}{\sqrt{2}-1} < n \tag{2.1.3}$$

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

2.2 Convergence of sequences

Definition 6.1 *(Sequence convergence)*
A sequence (s_n) of real numbers converges to s (i.e. $\lim_{n \to \infty} s_n = s$) provided that:

$$\forall \epsilon > 0, \; \exists N \in \mathbb{R} \text{ s.t. } \; n > N \implies |s_n - s| < \epsilon \tag{2.2.1}$$

A sequence that does not converge to a real number is divergent.
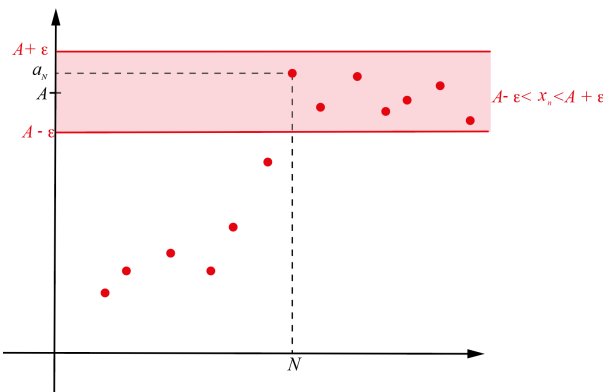
Figure 2.1. Geometrical interpretation of the epsilon-delta definition

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, $\lim_{n\to\infty} s_n = \frac{3}{7}$ means that:

$$\forall \epsilon > 0, \exists N \text{ s.t. } n > N \implies \left| \frac{3n+1}{7n-4} - \frac{3}{7} \right| < \epsilon \tag{2.2.2}$$

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.

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

◀

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

$$\lim_{n\to\infty} s_n = s \wedge \lim_{n\to\infty} s_n = t \implies s = t \tag{2.2.3}$$

Indeed, the first implies that:

$$\exists N_1 \text{ s.t. } n > N_1 \implies |s_n - s| < \frac{\epsilon}{2} \tag{2.2.4}$$

and the second implies that:

$$\exists N_2 \text{ s.t. } n > N_2 \implies |s_n - t| < \frac{\epsilon}{2} \tag{2.2.5}$$

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

$$|s - t| = |(s - s_n) + (s_n - t)| \le |s - s_n| + |s_n - t| \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \tag{2.2.6}$$

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. ∎

2.3 Formal Proofs of Limit Theorems

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

<u>Discussion</u>: 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:

$$\left| \frac{19}{7(7n-4)} \right| < \epsilon \tag{2.3.1}$$
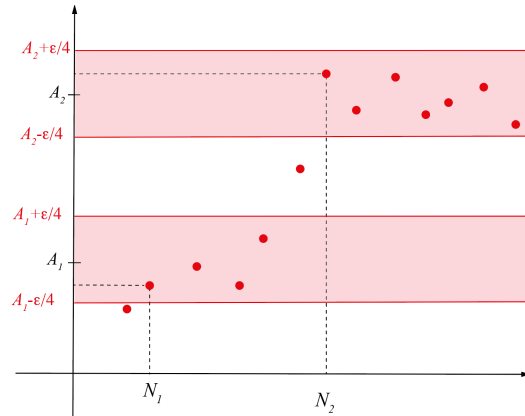
Figure 2.2. Uniqueness of Limits

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

$$\frac{19}{7(7n-4)} < \epsilon \implies \frac{19}{7\epsilon} < 7n - 4 \implies \frac{19}{49\epsilon} + \frac{4}{7} < n \tag{2.3.2}$$

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

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

$$n > N \implies n > \frac{19}{49\epsilon} + \frac{4}{7} \implies 7n - 4 > \frac{19}{7\epsilon} \implies \epsilon > \frac{19}{7(7n-4)} \tag{2.3.3}$$

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

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

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

$$|\frac{4n^3 + 3n}{n^3 - 6} - 4| < \epsilon \implies |\frac{3n + 24}{n^3 - 6}| < \epsilon \tag{2.3.4}$$

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 $3n + 24 \le 27n$ for $n > 1$ and $n^3 - 6 \ge \frac{n^3}{2}$ for $n > 2$. So:

$$|\frac{3n + 24}{n^3 - 6}| \le \frac{27n}{\frac{1}{2}n^3} < \epsilon \implies n > \sqrt{\frac{54}{\epsilon}} \tag{2.3.5}$$

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

$$n > N \implies n > \sqrt{\frac{54}{\epsilon}} \implies \frac{27n}{n^3/2} < \epsilon \tag{2.3.6}$$

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

$$|\frac{3n + 24}{n^3 - 6}| \leq \frac{27n}{n^3/2} < \epsilon \tag{2.3.7}$$

and hence

$$|\frac{4n^3 + 3n}{n^3 - 6} - 4| < \epsilon \tag{2.3.8}$$

as required. ◀

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

Proof.

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

$$\exists N \text{ s.t. } n > N \implies |s_n - s| < \sqrt{s}\epsilon \tag{2.3.9}$$

so:

$$\exists N \text{ s.t. } n > N \implies |\sqrt{s_n} - \sqrt{s}| = \frac{|s_n - s|}{\sqrt{s_n} + \sqrt{s}} \leq \frac{|s_n - s|}{\sqrt{s}} < \frac{\sqrt{s}}{\sqrt{s}}\epsilon = \epsilon \tag{2.3.10}$$

as desired. ◀

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

$$\exists N \text{ s.t. } n > N \implies |s_n| < \epsilon^2 \tag{2.3.11}$$

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

$$|\sqrt{s_n} - 0| < \epsilon \implies \lim_{n \to \infty} \sqrt{s_n} = s = 0 \tag{2.3.12}$$

as desired. ◀

Example. Let (s_n) be a convergent sequence of non-zero real numbers such that $\lim_{n \to \infty} s_n = s \neq 0$. 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 $\lim_{n \to \infty} s_n = s$:

$$\exists N \text{ s.t. } n > N \implies |s_n - s| \leq \frac{|s|}{2} \tag{2.3.13}$$

We must have that:

$$\exists N \text{ s.t. } n > N \implies |s_n| \geq \frac{|s|}{2} \tag{2.3.14}$$

since:

$$|s| = |s - s_n + s_n| \leq |s - s_n| + |s_n| \leq \frac{|s|}{2} + |s_n| \implies |s_n| \geq \frac{|s|}{2}. \tag{2.3.15}$$

Setting:

$$m = \min\{\frac{|s|}{2}, |s_1|, ..., |s_n|\} \tag{2.3.16}$$

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

◀

2.4   Null sequences

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

$$\forall \epsilon > 0, \exists N \in \mathbb{R} \text{ s.t. } n > N \implies |a_n| < \epsilon \tag{2.4.1}$$

Example.   Consider the sequence

$$a_n = \frac{(-1)^n}{n^4 + 1}, \ n = 1, 2, ... \tag{2.4.2}$$

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

$$|a_n| = \frac{1}{n^4 + 1} < \epsilon \iff n^4 > \frac{1}{\epsilon} - 1 \iff n > \left(\frac{1}{\epsilon} - 1\right)^{\frac{1}{4}} = N \tag{2.4.3}$$

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:

$$|a_n| = \frac{1}{n^4 + 1} < \epsilon \iff n^4 > \frac{1}{\epsilon} - 1 \tag{2.4.4}$$

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

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

$$a_n < \epsilon^{1/p}, \ \forall n > N \tag{2.4.5}$$

so that for all positive ϵ:

$$a_n^p < \epsilon, \ \forall n > N \tag{2.4.6}$$

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

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

$$|a_n| = \frac{1}{n} < \epsilon \iff \epsilon < n, \tag{2.4.7}$$

so that $|a_n| \leq \epsilon$ for all $n > N = \epsilon$. Hence $\lim_{n \to \infty} a_n = 0$. 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:

$$c = \frac{1}{a+1}, \ a > 0 \tag{2.4.8}$$

We may apply bernoulli's inequality:

$$c^n = \frac{1}{(a+1)^n} \leq \frac{1}{na+1} \leq \frac{1}{na} \tag{2.4.9}$$

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 $a > 0$:

$$c = \frac{1}{a+1} \tag{2.4.10}$$

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

$$(1+a)^n \geq 1 + na + \frac{1}{2}n(n-1)a^2 \geq \frac{1}{2}n(n-1)a^2, \; n = 2, 3, \ldots \tag{2.4.11}$$

it follows that

$$nc^n = \frac{n}{(a+1)^n} \leq \frac{n}{\frac{1}{2}n(n-1)a^2} = \frac{2}{a^2(n-1)}, \; n = 2, 3, \ldots \tag{2.4.12}$$

The sequence defined by:

$$b_n = \frac{2}{a^2(n-1)}, \; n = 2, 3\ldots \tag{2.4.13}$$

is equivalent to:

$$b_n = \frac{2}{a^2 n}, \; n = 1, 2\ldots \tag{2.4.14}$$

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:

$$n^p c^n = (nd^n)^p, \; n = 1, 2\ldots \tag{2.4.15}$$

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 $n > m + 1$:

$$\frac{c^n}{n!} = \prod_{k=1}^{n} \frac{c}{k} \leq \frac{c}{n} \prod_{k=1}^{m} \frac{c}{k} = \frac{c^m}{m!} \cdot \frac{c}{n} \tag{2.4.16}$$

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:

$$\frac{n^p}{n!} = \frac{n^p}{2^n} \cdot \frac{2^n}{n!} \tag{2.4.17}$$

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

∎

Example. Consider the sequence described by

$$a_n = \frac{1}{3n^4(2n-1)^{1/3}}, \; n = 1, 2, \ldots \tag{2.4.18}$$

We can rewrite the terms of this sequence as:

$$a_n = \frac{1}{3} \cdot \frac{1}{n^4} \cdot \left(\frac{(-1)^n}{n^4+1}\right)^{1/3} \tag{2.4.19}$$

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:

$$|a_n| \leq b_n, \ \forall n = 1, 2, ... \tag{2.4.20}$$

then (a_n) is null.

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

$$-b_n \leq a_n \leq b_n, \ \forall n = 1, 2, ... \tag{2.4.21}$$

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

Example. Consider the sequence:

$$a_n = \frac{\sin\left(n^2\right)}{n^2 + 2^n}, \ n = 1, 2, ... \tag{2.4.22}$$

Then:

$$\left|\frac{\sin\left(n^2\right)}{n^2 + 2^n}\right| \leq \left|\frac{1}{n^2 + 2^n}\right| \tag{2.4.23}$$

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

$$\frac{1}{n^2 + 2^n} < \frac{1}{n^2} \tag{2.4.24}$$

so:

$$\left|\frac{\sin\left(n^2\right)}{n^2 + 2^n}\right| \leq \frac{1}{n^2} \tag{2.4.25}$$

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.

◀

2.5 Limit theorems for Convergent sequences

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

$$|a_n| \leq M, \ n = 1, 2, ... \tag{2.5.1}$$

and is **unbounded otherwise**

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

$$a_n = \frac{2n+1}{n} = 2 + \frac{1}{n} \leq 2 + 1 = 3 \tag{2.5.2}$$

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

◀

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

$$|b_n| = |n| = n < M \tag{2.5.3}$$

for all $n = 1, 2, ...$, 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 $\lim_{n \to \infty} s_n = s$. Then we can apply the Definition 6.1 with $\epsilon = 1$ to find:

$$n > N \implies |s_n - s| < 1 \implies |s_n| < 1 + |s| \tag{2.5.4}$$

where we use the triangle inequality. Let us now define $M = \max\{|s| + 1, |s_1|, ..., |s_N|\}$ 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:

$$|a_n| = n^{(-1)^n} = n \tag{2.5.5}$$

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

$$|a_n| = \frac{1}{n} < 1 \tag{2.5.6}$$

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

$$\frac{n^2 + n}{n^2 + 1} \le \frac{n^2 + n^2}{n^2} = 2 \tag{2.5.7}$$

so the sequence is bounded. Moreover,

$$\lim_{n\to\infty} \frac{n^2 + n}{n^2 + 1} = \lim_{n\to\infty} \frac{1 + \frac{1}{n}}{1 + \frac{1}{n^2}} = 1 \tag{2.5.8}$$

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

Proof.

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

$$n > N \implies |s_n - s| < \frac{\epsilon}{|k|} \tag{2.5.9}$$

  Then:

$$n > N \implies |ks_n - ks| < \epsilon \tag{2.5.10}$$

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

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

$$|s_n + t_n - (s + t)| < \epsilon \tag{2.5.11}$$

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

$$n > N_1 \implies |s_n - s| < \frac{\epsilon}{2} \tag{2.5.12}$$

and similarly for t_n:

$$n > N_2 \implies |t_n - t| < \frac{\epsilon}{2} \tag{2.5.13}$$

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

$$n > N \implies |s_n - s| + |t_n - t| < \epsilon \tag{2.5.14}$$

Finally, we make use of the triangle inequality:

$$n > N \implies |s_n + t_n - (s + t)| \leq |s_n - s| + |t_n - t| < \epsilon \tag{2.5.15}$$

as desired.

(iii) Firstly note that:

$$|s_n t_n - st| = |s_n t_n - s_n t + s_n t - st| \leq |s_n| \cdot |t_n - t| + |t| \cdot |s_n - s| \tag{2.5.16}$$

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:

$$n > N_1 \implies |t_n - t| < \frac{\epsilon}{2M} \tag{2.5.17}$$

and similarly for s_n:

$$n > N_2 \implies |s_n - s| < \frac{\epsilon}{2(|t| + 1)} \tag{2.5.18}$$

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

$$n > N \implies |s_n t_n - st| \leq M \frac{\epsilon}{2M} + |t| \frac{\epsilon}{2(|t| + 1)} < \epsilon \tag{2.5.19}$$

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

(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:

$$n > N \implies |s - s_n| < \epsilon m |s| \tag{2.5.20}$$

Then $n > N$ implies:

$$\left| \frac{1}{s_n} - \frac{1}{s} \right| = \frac{|s - s_n|}{|s_n s|} \leq \frac{|s - s_n|}{m|s|} \leq \epsilon \tag{2.5.21}$$

as desired.

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

$$\lim_{n \to \infty} \frac{t_n}{s_n} = \lim_{n \to \infty} \frac{1}{s_n} t_n = \frac{t}{s} \tag{2.5.22}$$

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:

$$a_n = \frac{n^2 + 2^n}{3^n + n^3} \tag{2.5.23}$$

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

$$a_n = \frac{2^n}{3^n} \cdot \frac{\frac{n^2}{2^n} + 1}{\frac{n^3}{3^n} + 1} = \left(\frac{2}{3}\right)^n \cdot \frac{\frac{n^2}{2^n} + 1}{\frac{n^3}{3^n} + 1} \tag{2.5.24}$$

Hence:

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \left[\left(\frac{2}{3}\right)^n \cdot \right] \frac{\frac{n^2}{2^n} + 1}{\frac{n^3}{3^n} + 1} \tag{2.5.25}$$

$$= \lim_{n\to\infty} \left(\frac{2}{3}\right)^n \cdot \frac{\lim_{n\to\infty} \frac{n^2}{2^n} + 1}{\lim_{n\to\infty} \frac{n^3}{3^n} + 1} \tag{2.5.26}$$

$$= \lim_{n\to\infty} \left(\frac{2}{3}\right)^n \tag{2.5.27}$$

$$= 0 \tag{2.5.28}$$

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 $n = 1, 2, ...$
  (ii) $\lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = l$
Then $\lim_{n\to\infty} a_n = l$.

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

$$l - \epsilon < b_n < l + \epsilon, \ \forall n > N_1 \tag{2.5.29}$$

$$l - \epsilon < c_n < l + \epsilon, \ \forall n > N_2 \tag{2.5.30}$$

Then, it follows from condition (i) that:

$$l - \epsilon < b_n \leq a_n \leq c_n < l + \epsilon, \ \forall n > N \tag{2.5.31}$$

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

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

$$(1 + x)^n \geq \frac{n(n-1)}{2} x^2 \qquad (2.5.32)$$

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

$$\left(1 + \sqrt{\frac{2}{n-1}}\right)^n \geq \frac{n(n+1)}{2} \frac{2}{n-1} = n \qquad (2.5.33)$$

implying that

$$n^{1/n} \leq 1 + \sqrt{\frac{2}{n-1}}, \ n = 2, 3, ... \qquad (2.5.34)$$

If we define b_n to be:

$$b_1 = 1 \qquad (2.5.35)$$

$$b_n = 1 + \sqrt{\frac{2}{n-1}}, \ n = 2, 3, ... \qquad (2.5.36)$$

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

$$|b_n - 1| < \epsilon \iff n > \frac{2}{\epsilon^2} + 1 = N > 1 \qquad (2.5.37)$$

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:

$$\lim_{n\to\infty} n^{1/n} = 1 \qquad (2.5.38)$$

◄

Proposition *(Limit inequality)*
If $\lim_{n\to\infty} a_n = l, \lim_{n\to\infty} b_n = m$ and:

$$a_n \leq b_n, \ \forall n = 1, 2, ... \qquad (2.5.39)$$

then $l \leq m$.

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

$$\lim_{n\to\infty} (a_n - b_n) = l - m > 0 \qquad (2.5.40)$$

so that:

$$a_n - b_n > \frac{1}{2}(l - m) \qquad (2.5.41)$$

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

$$|(a_n - b_n) - (l - m)| < \frac{l}{2} \implies -\frac{l - m}{2} < (a_n - b_n) - (l - m) < \frac{l - m}{2} \tag{2.5.42}$$

$$\implies \frac{l - m}{2} < a_n, \; \forall n > N \tag{2.5.43}$$

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

$$l - m \leq 0 \implies l \leq m \tag{2.5.44}$$

a contradiction.

> **Proposition** *(Limit uniqueness)*
> The limit of a sequence is unique.

Suppose $\lim_{n\to\infty} a_n = l$ and $\lim_{n\to\infty} a_n = m$, 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 $\lim_{n\to\infty} a_n = l$ then $\lim_{n\to\infty} |a_n| = |l|$.

Proof. Using the triangle inequality:

$$||a_n| - |l|| \leq |a_n - l| \tag{2.5.45}$$

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

$$|a_n - l| \leq \epsilon, \; \forall n > N \tag{2.5.46}$$

implying that:

$$||a_n| - |l|| \leq \epsilon, \; \forall n > N \tag{2.5.47}$$

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

2.6 Divergent sequences

> **Definition** *(Infinite limit)*
> A sequence (a_n) tends to infinity if:
>
> $$\forall M > 0, \; \exists N \; s.t. \; a_n > M, \forall n > N \tag{2.6.1}$$
>
> We write that $a_n \to \infty$.
> The sequence (a_n) tends to minus infinity if:
>
> $$-a_n \to \infty \tag{2.6.2}$$

> **Theorem** *(Reciprocal rule)*
> If (a_n) is eventually positive, and $\left(\frac{1}{a_n}\right)$ is a null sequence then $a_n \to \infty$.

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

$$a_n > 0, \ \forall n > N_1 \tag{2.6.3}$$

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

$$\left|\frac{1}{a_n}\right| < \frac{1}{M*}, \ \forall n > N_2 \tag{2.6.4}$$

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

$$0 < \frac{1}{a_n} < \frac{1}{M}, \ \forall n > N \tag{2.6.5}$$

implying:

$$a_n > M, \ \forall n > N \tag{2.6.6}$$

as desired. ∎

> **Example.** Consider the sequence $a_n = n! - 10^n$ for $n = 1, 2, ...$ The dominant term is $n!$, so let us write:
>
> $$a_n = n!\left(1 - \frac{10^n}{n!}\right), \ n = 1, 2, ... \tag{2.6.7}$$
>
> Therefore, a_n is indeed eventually positive, since $10^n < n!$ is true for $n \geq 25$. Then:
>
> $$\lim_{n\to\infty} \frac{1}{a_n} = \lim_{n\to\infty} \frac{1}{n!} \frac{1}{1 - \frac{10^n}{n!}} = 0 \cdot \frac{0}{1 - 0} = 0 \tag{2.6.8}$$
>
> By the reciprocal rule we see that $a_n \to \infty$.
>
> ◄

> **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 $n = 1, 2, ...$ then (a_n) tends to infinity.

> **Example.** Consider the sequence $a_n = \frac{2^n}{n} + 5n^9, \ n = 1, 2, ...$ Then let us define $b_n = \frac{2^n}{n}$ and $c_n = n^9$ so that:
>
> $$a_n = b_n + 5c_n \tag{2.6.9}$$

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

$$\lim_{n \to \infty} \frac{1}{b_n} = \lim_{n \to \infty} n\left(\frac{1}{2}\right)^n = 0 \tag{2.6.10}$$

where we used the standard null sequence (iii) with $p = 1$.
Similarly, $c_n \to \infty$ since:

$$\lim_{n \to \infty} \frac{1}{c_n} = \lim_{n \to \infty} \frac{1}{n^9} = 0 \tag{2.6.11}$$

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

$$a_n \geq \frac{2^n}{n} \tag{2.6.12}$$

but $\frac{2^n}{n} \to \infty$ so $a_n \to \infty$ 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:

$$a_{2n+1} = \frac{1}{2n+1} \tag{2.6.13}$$

whereas the even subsequence is given by:

$$a_{2n} = 2n \tag{2.6.14}$$

◄

Theorem *(Subsequence divergence)*
For any subsequence (a_{n_k}) of a sequence (a_n):
  (i) if $\lim_{n \to \infty} a_n = l$ then $\lim_{n \to \infty} a_{n_k} = l$.
  (ii) if $a_n \to \infty$ then $a_{n_k} \to \infty$

Proof.

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

$$|a_n - l| < \epsilon, \ \forall n > N \tag{2.6.15}$$

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

$$n_k \geq n_K \geq N, \ \forall k > K \tag{2.6.16}$$

so that:

$$|a_{n_k} - l| < \epsilon, \ \forall n_k > N \tag{2.6.17}$$

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

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

$$a_n > M, \ \forall n > N \tag{2.6.18}$$

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

$$n_k \geq n_K \geq N, \ \forall k > K \tag{2.6.19}$$

so that:

$$a_{n_k} > M, \ \forall n_k > N \tag{2.6.20}$$

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

$$\blacksquare$$

Proposition *(Subsequence rules)*

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

A3 Series

3.1   Introduction to Series

Definition *(Series)*

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

$$\sum_{i=m}^{\infty} a_i \tag{3.1.1}$$

an infinite series, and define by nth partial sum:

$$s_n = \sum_{i=m}^{n} a_i \tag{3.1.2}$$

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

$$\lim_{n \to \infty} \left(\sum_{i=m}^{n} a_i \right) = s \tag{3.1.3}$$

Otherwise it is said to be divergent.

Proposition *(Geometric series)*

The series of the form:

$$\sum_{k=0}^{\infty} ar^n \tag{3.1.4}$$

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

$$\sum_{k=0}^{\infty} ar^n = \frac{a}{1-r} \quad \text{if } |r| < 1 \tag{3.1.5}$$

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:

$$s_n = \sum_{k=0}^{n} ar^k = a\frac{1 - r^{n+1}}{1 - r} \tag{3.1.6}$$

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

$$(1-r)\sum_{k=0}^{n} ar^k = \sum_{k=0}^{n} ar^k - \sum_{k=0}^{n} ar^{k+1} \tag{3.1.7}$$

$$= (a + ar + ar^2 + ... + ar^n) - (ar + ar^2 + ... + ar^n + ar^{n+1}) \tag{3.1.8}$$

$$= a + ar^{n+1} \tag{3.1.9}$$

$$\implies \sum_{k=0}^{n} ar^k = a\frac{1 - r^{n+1}}{1 - r} \tag{3.1.10}$$

as desired.

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

> **Definition** *(Cauchy criterion)*
> A series $\sum a_n$ satisfies the Cauchy criterion if its partial sums form a Cauchy sequence:
>
> $$\forall \epsilon > 0, \ \exists N \text{ such that } m, n > N \implies |s_n - s_m| < \epsilon \tag{3.1.11}$$

3.2 Telescoping series

Telescoping series are series with terms of the form:

$$a_n = b_n - b_{n+i} \tag{3.2.1}$$

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

$$s_n = \sum_{k=1}^{n} (b_k - b_{n\pm i}) \tag{3.2.2}$$

$$= (b_1 + ... + b_{i+1} + b_{i+2} + ... + b_n) - (b_{i+1} + b_{i+2} + ... + b_{i+n}) \tag{3.2.3}$$

$$= (b_1 + b_2 + ... + b_i + b_n + b_{n+1} + ... + b_{n+i}) \tag{3.2.4}$$

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

> **Example.** Consider the series:
>
> $$\sum_{n=1}^{\infty} \frac{1}{2n(2n+2)} \tag{3.2.5}$$
>
> We can expand the partial fraction as:
>
> $$\frac{1}{2n(2n+2)} = \frac{1}{2}\left(\frac{1}{2n} - \frac{1}{2n+2}\right) \tag{3.2.6}$$

so that the partial sums turn into:

$$s_n = \frac{1}{2} \sum_{k=1}^{n} \left(\frac{1}{2k} - \frac{1}{2k+2} \right) \tag{3.2.7}$$

$$= \frac{1}{2} \left[\left(\frac{1}{2} - \frac{1}{4} \right) + \left(\frac{1}{4} - \frac{1}{6} \right) + ... + \left(\frac{1}{2n-2} - \frac{1}{2n} \right) + \left(\frac{1}{2n} - \frac{1}{2n+2} \right) \right] \tag{3.2.8}$$

$$= \frac{1}{2} \left(\frac{1}{2} - \frac{1}{2n+2} \right) = \frac{n}{4(n+1)} \tag{3.2.9}$$

We quickly see that the series does indeed converge:

$$\lim_{n \to \infty} s_n = \frac{1}{4} \tag{3.2.10}$$

◄

3.3 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:

$$\sum_{n=1}^{\infty} (\lambda a_n + b_n) = \lambda s + t \; \forall \lambda \in \mathbb{R} \tag{3.3.1}$$

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

$$s_n = \sum_{k=1}^{n} a_k \; \text{ and } \; t_n = \sum_{k=1}^{n} b_k \tag{3.3.2}$$

Then:

$$\sum_{k=1}^{\infty} (\lambda a_n + b_n) = (\lambda a_1 + b_1) + (\lambda a_2 + b_2) + ... + (\lambda a_n + b_n) \tag{3.3.3}$$

$$= \lambda(a_1 + a_2 + ... + a_n) + (b_1 + b_2 + ... + b_n) \tag{3.3.4}$$

$$= \lambda s_n + t_n \tag{3.3.5}$$

By theorem 6.3:
$$\lim_{n \to \infty} (\lambda s_n + t_n) = \lambda \lim_{n \to \infty} s_n + \lim_{n \to \infty} t_n = \lambda s + t \tag{3.3.6}$$

as desired, the series is convergent to:

$$\sum_{n=1}^{\infty} (\lambda a_n + b_n) = \lambda s + t \tag{3.3.7}$$

■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \left(\left(-\frac{3}{4} \right)^n - \frac{2}{n(n+1)} \right) = \sum_{n=1}^{\infty} \left(-\frac{3}{4} \right)^n - 2 \sum_{n=1}^{\infty} \frac{1}{n(n+1)} \tag{3.3.8}$$

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

$$\sum_{n=1}^{\infty} \left(-\frac{3}{4} \right)^n = \frac{-\frac{3}{4}}{1 + \frac{3}{4}} = -\frac{3}{7} \tag{3.3.9}$$

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

$$s_n = \sum_{n=1}^{n} \frac{1}{n(n+1)} = \sum_{n=1}^{n} \left(\frac{1}{n} - \frac{1}{n+1} \right) \tag{3.3.10}$$

$$= \left(1 - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \dots + \left(\frac{1}{n-1} - \frac{1}{n} \right) + \left(\frac{1}{n} - \frac{1}{n+1} \right) \tag{3.3.11}$$

$$= 1 - \frac{1}{n+1} = \frac{n}{n+1} \tag{3.3.12}$$

so that:

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_{n\to\infty} s_n = 1 \tag{3.3.13}$$

Hence the original series converges to:

$$\sum_{n=1}^{\infty} \left(\left(-\frac{3}{4} \right)^n - \frac{2}{n(n+1)} \right) = -\frac{3}{7} - 2 = -\frac{17}{7} \tag{3.3.14}$$

◄

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 nth partial sum of a_n. Since $\sum a_n$ converges, s_n must also converge to some limit s. Note that:

$$a_n = s_n - s_{n-1} \implies \lim_{n\to\infty} a_n = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = s - s = 0 \tag{3.3.15}$$

so that a_n is indeed a null sequence. ∎

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n^2}{2n^2 + 1} \tag{3.3.16}$$

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

$$a_n = \frac{(-1)^{n+1}n^2}{2n^2 + 1} \tag{3.3.17}$$

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

$$a_{2n} = -\frac{n^2}{2n^2 + 1} \tag{3.3.18}$$

which is convergent to:

$$\lim_{n\to\infty} a_{2n} = -\lim_{n\to\infty} \frac{1}{2 + \frac{1}{n^2}} = \frac{1}{2 + \lim_{n\to\infty}\frac{1}{n^2}} = \frac{1}{2} \tag{3.3.19}$$

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.

◀

3.4   Non-negative series

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

Example.   Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n} \tag{3.4.1}$$

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

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + ... \tag{3.4.2}$$

$$= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + ... \tag{3.4.3}$$

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

$$s_2 = 1 + \frac{1}{2} \tag{3.4.4}$$

$$s_4 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) \geq 1 + \frac{1}{2} \tag{3.4.5}$$

$$s_8 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) \tag{3.4.6}$$

$$s_{2^k} = \sum_{n=1}^{2^k} \frac{1}{n} \geq 1 + \frac{1}{2} + \frac{1}{2} + ... + \frac{1}{2} = 1 + \frac{1}{2}k \tag{3.4.7}$$

Since $1 + \frac{1}{2}k \to \infty$ as $k \to \infty$ 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:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \tag{3.4.8}$$

Consider the term:

$$\frac{1}{k^2} < \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k} \quad k > 1 \tag{3.4.9}$$

Therefore the kth partial sum of the series is:

$$s_n = \sum_{k=1}^{n} \frac{1}{k^2} < 1 + \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2 - \frac{1}{n} < 2 \tag{3.4.10}$$

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

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n \to \infty} s_n \text{ converges} \tag{3.4.11}$$

◀

Theorem *(Comparison test)*
If $0 \le a_n \le b_n$ 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:

$$\sum_{n=1}^{\infty} \frac{1}{n^3} \tag{3.4.12}$$

Then, we may write that:

$$0 \le n^2 \le n^3 \tag{3.4.13}$$

so that

$$0 \le \frac{1}{n^3} \le \frac{1}{n^2} \tag{3.4.14}$$

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:

$$\sum_{n=1}^{\infty} \frac{\cos^2 2n}{n^3} \tag{3.4.15}$$

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

$$0 \leq \frac{\cos^2 2n}{n^3} \leq \frac{1}{n^3} \tag{3.4.16}$$

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

$$\lim_{n \to \infty} \frac{a_n}{b_n} = L \neq 0 \tag{3.4.17}$$

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:

$$\frac{a_n}{b_n} \leq K \implies a_n \leq Kb_n, \; n = 1, 2, \dots \tag{3.4.18}$$

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:

$$\lim_{n \to \infty} \frac{b_n}{a_n} = \frac{1}{L} \neq 0 \tag{3.4.19}$$

since $L \neq 0$. Then:

$$\frac{b_n}{a_n} \leq K \implies b_n \leq Ka_n, \; n = 1, 2, \dots \tag{3.4.20}$$

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:

$$\sum_{n=1}^{\infty} a_n \text{ converges} \implies \sum_{n=1}^{\infty} b_n \text{ converges} \tag{3.4.21}$$

which is equivalent to its converse:

$$\sum_{n=1}^{\infty} b_n \text{ diverges} \implies \sum_{n=1}^{\infty} a_n \text{ diverges} \tag{3.4.22}$$

as desired. ■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{1}{n^3 + n} = \sum_{n=1}^{\infty} a_n \tag{3.4.23}$$

Then, define:

$$b_n = \frac{1}{n^3} \tag{3.4.24}$$

so that:

$$\lim_{n \to \infty} \frac{a_n}{b_n} = \lim_{n \to \infty} \frac{n^3}{n^3 + n} = \lim_{n \to \infty} \frac{1}{1 + \frac{1}{n^2}} = 1 \neq 0 \tag{3.4.25}$$

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:

$$\sum_{n=1}^{\infty} \frac{n+4}{2n^3 - n + 1} = \sum_{n=1}^{\infty} a_n \tag{3.4.26}$$

Then we may define

$$b_n = \frac{1}{n^2} \tag{3.4.27}$$

so that:

$$\lim_{n \to \infty} \frac{a_n}{b_n} = \lim_{n \to \infty} \frac{n^3 + 4n^2}{2n^3 - n + 1} = \lim_{n \to \infty} \frac{1 + \frac{4}{n}}{2 - \frac{1}{n^3} + \frac{1}{n^2}} = \frac{1}{2} \neq 0 \tag{3.4.28}$$

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

Proof.

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

$$l + \epsilon < 1 \tag{3.4.29}$$

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

$$\frac{a_{n+1}}{a_n} \leq r, \ \forall n \geq N \tag{3.4.30}$$

Then:

$$\frac{a_n}{a_N} = \prod_{k=n-1}^{N} \frac{a_{k+1}}{a_k} \leq \prod_{k=n-1}^{N} r = r^{n-N} \tag{3.4.31}$$

Hence:

$$a_n \leq a_N r^{n-N} \tag{3.4.32}$$

Note however that:

$$\sum_{n=1}^{\infty} a_N r^{n-N} \tag{3.4.33}$$

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

(ii) and (iii) Suppose that:

$$\frac{a_{n+1}}{a_n} \to \infty \;\; \text{or} \;\; \frac{a_{n+1}}{a_n} \to l \tag{3.4.34}$$

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

$$\frac{a_{n+1}}{a_n} \geq 1, \; \forall n \geq N \tag{3.4.35}$$

Therefore:

$$\frac{a_n}{a_N} = \prod_{k=n-1}^{N} \frac{a_{k+1}}{a_k} \leq 1 \tag{3.4.36}$$

implying that:

$$a_n \geq a_N > 0, \; \forall n \geq N \tag{3.4.37}$$

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

■

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{(2n)!}{n^n} = \sum_{n=1}^{\infty} a_n \tag{3.4.38}$$

Then:

$$\frac{a_{n+1}}{a_n} = \frac{(2n+2)!}{(n+1)^{n+1}} \cdot \frac{n^n}{(2n)!} \tag{3.4.39}$$

$$= (2n+2)(2n+1) \frac{n^n}{(n+1)^{n+1}} \tag{3.4.40}$$

$$= \frac{(2n+2)(2n+1)}{n+1} \frac{1}{(1+\frac{1}{n})^n} \tag{3.4.41}$$

Therefore:

$$\lim_{n \to \infty} \frac{a_{n+1}}{a_n} = \frac{2}{e} \lim_{n \to \infty} (2n+1) = \infty \tag{3.4.42}$$

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

> **Proposition** *(Standard series)*
> The following series converge:
>
> (i) $\sum\limits_{n=1}^{\infty} \frac{1}{n^p}$ for $p \geq 2$
>
> (ii) $\sum\limits_{n=1}^{\infty} c^n$ for $0 \leq c < 1$
>
> (iii) $\sum\limits_{n=1}^{\infty} n^p c^n$ for $p > 0$ and $0 \leq c < 1$
>
> (iv) $\sum\limits_{n=1}^{\infty} \frac{c^n}{n!}$ for $c \geq 0$
>
> The following series is divergent:
>
> (v) $\sum\limits_{n=1}^{\infty} \frac{1}{n^p}$ for $0 < p \leq 1$

Proof.

(i) Note that if $p \geq 2$ then:
$$\frac{1}{n^p} \leq \frac{1}{n^2}, \; n = 1, 2, \ldots \tag{3.4.43}$$

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:
$$a_n = (n^p b^n) b^n, \; n = 1, 2, \ldots \tag{3.4.44}$$

Since $0 \leq b < 1$, $n^p b^n$ is a standard null sequence. Setting $\epsilon = 1$, there exists N such that we have:
$$n^p b^n < 1, \; \forall n > N \tag{3.4.45}$$

and thus:
$$a_b < b^n, \forall n > N \tag{3.4.46}$$

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:
$$\frac{a_{n+1}}{a_n} = \frac{c^{n+1}}{(n+1)!} \cdot \frac{c^n}{n!} = \frac{c}{n+1} \tag{3.4.47}$$

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

(v) Note that if $p \leq 1$ then:
$$\frac{1}{n^p} \geq \frac{1}{n^2}, \; n = 1, 2, \ldots \tag{3.4.48}$$

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.

∎

3.5 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:

$$b_n = \begin{cases} a_n, & \text{if } a_n \geq 0 \\ 0, & \text{if } a_n < 0 \end{cases} \quad , \text{ and } c_n = \begin{cases} 0, & \text{if } a_n \geq 0 \\ -a_n, & \text{if } a_n < 0 \end{cases} \tag{3.5.1}$$

then both b_n and c_n are non-negative. Moreover, $b_n \leq |a_n|$ and $c_n \leq |a_n|$ for $n = 1, 2, ...$, 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:

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} (b_n - c_n) = \sum_{n=1}^{\infty} b_n - \sum_{n=1}^{\infty} c_n \tag{3.5.2}$$

also converges. ∎

Example. Consider the series:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n}{n^3 + 1} \tag{3.5.3}$$

and its absolute analogue:

$$\sum_{n=1}^{\infty} \frac{n}{n^3 + 1} \tag{3.5.4}$$

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

$$\lim_{n \to \infty} \frac{a_n}{b_n} = \lim_{n \to \infty} \frac{n^3}{n^3 + 1} = 1 \neq 0 \tag{3.5.5}$$

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. ◀

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

$$\left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n| \tag{3.5.6}$$

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:

$$b_n = \begin{cases} a_n, & \text{if } a_n \geq 0 \\ 0, & \text{if } a_n < 0 \end{cases} \quad , \text{ and } c_n = \begin{cases} 0, & \text{if } a_n \geq 0 \\ -a_n, & \text{if } a_n < 0 \end{cases} \tag{3.5.7}$$

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

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n - \sum_{n=1}^{\infty} c_n \tag{3.5.8}$$

we find that

$$-\sum_{n=1}^{\infty} c_n \leq \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} b_n \tag{3.5.9}$$

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

$$-\sum_{n=1}^{\infty} |a_n| \leq \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} |a_n| \implies \left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n| \tag{3.5.10}$$

as required. ∎

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

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2^n} \tag{3.5.11}$$

converges. Note that:

$$\left| \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2^n} \right| \leq \sum_{n=1}^{\infty} \frac{1}{2^n} = 1 \tag{3.5.12}$$

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

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

$$\sum_{n=1}^{\infty} a_n \text{ converges} \tag{3.5.13}$$

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

$$s_{2k} = (b_1 - b_2) + (b_3 - b_4) + \ldots + (b_{2k-1} + b_{2k}) \tag{3.5.14}$$

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

$$s_{2k} = b_1 - (b_2 - b_3) - (b_4 - b_5) - \ldots - (b_{2k-2} - b_{2k-1}) - b_{2k} \leq b_1 \tag{3.5.15}$$

By the monotone convergence theorem it follows that

$$\lim_{n \to \infty} s_{2n} = s \tag{3.5.16}$$

for some s. Also:

$$s_{2k-1} = b_1 - (b_2 - b_3) - (b_4 - b_5) - \ldots - (b_{2k-2} - b_{2k-1}) = s_{2k} + b_{2k} \tag{3.5.17}$$

so that:

$$\lim_{n\to\infty} s_{2k+1} = \lim_{n\to\infty} s_{2k} + \lim_{n\to\infty} b_{2k} = s \tag{3.5.18}$$

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

Example. Consider the sequence:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n + \sqrt{n}} \tag{3.5.19}$$

We can write its terms as:

$$a_n = \frac{(-1)^{n+1}}{n + \sqrt{n}} = (-1)^{n+1} \frac{1}{n + \sqrt{n}} = (-1)^{n+1} b_n \tag{3.5.20}$$

Now:

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

$$\frac{1}{n + \sqrt{n}} \leq \frac{1}{n} \tag{3.5.21}$$

   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:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n + \sqrt{n}} \tag{3.5.22}$$

   converges.

◀

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 $n = 1, 2, ...$ Then the subsequence:

$$a_{3n} = n - \lfloor n \rfloor = 0 \tag{3.5.23}$$

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

$$a_{3n+1} = n + \frac{1}{3} - n = \frac{1}{3} \tag{3.5.24}$$

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

> **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:
>
> $$\lim_{n \to \infty} a_n = l \tag{3.5.25}$$

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

$$|a_{m_k} - l| < \epsilon, \ \forall k > K_1 \tag{3.5.26}$$

$$|a_{n_k} - l| < \epsilon, \ \forall k > K_2 \tag{3.5.27}$$

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

$$|a_n - l| < \epsilon, \ \forall k > N \tag{3.5.28}$$

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 $N \geq K_2$. ∎

3.6 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 $\max\{a_n : n = 1, 2, ...\} = l$, and let $\epsilon > 0$. Then there exists an integer N such that

$$a_N > l - \epsilon, \ \forall n > N \tag{3.6.1}$$

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:

$$a_n > l - \epsilon \iff l - a_n < \epsilon, \ \forall n > N \tag{3.6.2}$$

Hence:

$$|a_n - l| = l - a_n < \epsilon, \ \forall n > N \tag{3.6.3}$$

proving that (a_n) converges to l. ∎

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

$$|a_n - l| < \epsilon, \ \forall n > N \tag{3.6.4}$$

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

$$|a_n| < \epsilon + |l|, \ \forall n > n \tag{3.6.5}$$

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

$$|a_n| \leq M \tag{3.6.6}$$

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 ∞.

A4 Functions and Continuity

4.1   Real functions

> **Definition** *(Real function)*
> Let $A \subseteq \mathbb{R}$ and $f : A \to \mathbb{R}$ is a bijective function. Then the inverse function $f^{-1} : \operatorname{Im}(A) \to A$ and has rule:
> $$f^{-1}(f(x)) = x, \ \forall x \in A \tag{4.1.1}$$

Example. Consider the function:

$$f(x) = \frac{1}{1-x}, \ \forall x \in (-\infty, 1) \tag{4.1.2}$$

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

$$y = \frac{1}{1-x} \iff \frac{1}{y} = 1 - x \iff x = 1 - \frac{1}{y} \tag{4.1.3}$$

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

$$f^{-1}(x) = 1 - \frac{1}{x} \tag{4.1.4}$$

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

$$\operatorname{Im}(f) = f(A) = \left\{ \frac{1}{1-x} : \forall x < -1 \right\} = (0, \infty) \tag{4.1.5}$$

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

◀

Unfortunately, often times it is hard to solve $y = f(x9)$ 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:

$$f(x) = x^5 + x - 1, \ \forall x \in \mathbb{R} \tag{4.1.6}$$

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

$$x_1^5 + x_1 < x_2^5 + x_2 \implies x_1^5 + x_1 - 1 < x_2^5 + x_2 \implies f(x_1) < f(x_2) \tag{4.1.7}$$

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

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

$$f(x) = x^3 - 2x^2 \tag{4.2.1}$$

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

$$f(x_n) = x_n^3 - 2x_n^2 \to 2^3 - 2 \cdot 2^2 = 0 \tag{4.2.2}$$

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

Example. Consider the function:

$$f(x) = \lfloor x \rfloor \tag{4.2.3}$$

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

$$f(x_n) = \left\lfloor 1 - \frac{1}{n} \right\rfloor = 0 \tag{4.2.4}$$

since for any $n \geq 1$, $1 - \frac{1}{n} < 1$ and thus $\lfloor 1 - \frac{1}{n} \rfloor = 0$. However, $f(1) = 1 \neq 0$, thus the

function is discontinuous at $a = 1$. ◄

Example. Consider the function:

$$f(x) = \begin{cases} \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{4.2.5}$$

Note that:

$$\sin\left(2n + \frac{1}{2}\right)\pi = 1 \tag{4.2.6}$$

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

$$\sin x_n \to 1 \neq 0 = f(0) \tag{4.2.7}$$

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

Example. Consider the function:

$$f(x) = |x|, \ \forall x \in \mathbb{R} \tag{4.2.8}$$

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

$$|x_n - a| \geq ||x_n| - |a||, \ n = 1, 2, \dots \tag{4.2.9}$$

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

Example. Consider the function:

$$f(x) = \sqrt{x}, \ \forall x \in [0, \infty) \tag{4.2.10}$$

Consider any $a \in [0, \infty)$ and let (x_n) be any sequence in $\mathbb{R}$ such that $x_n \to a$ as $n \to \infty$. 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:

$$\sqrt{|x_n - a|} \geq |\sqrt{x_n} - \sqrt{a}| \tag{4.2.11}$$

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

4.3 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:

$$C = \{x \in A : f(x) \in B\} \ni a \tag{4.3.1}$$

We know that $f(x_n) \to a$ 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 $g(f(x_n)) \to g(f(a))$, as desired. ∎

The following theorem results immediately

Theorem *(Continuity of polynomials and their rationals)*
The following are continuous:
  (i) any polynomial $p(x) = a_0 + a_1 x + ... + a_n x^n$
  (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:

$$f(x) = \sqrt{x^2 + 2x + 2} - \frac{3x}{x^4 + 4}, \ \forall x \in \mathbb{R} \tag{4.3.2}$$

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. ◀

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 $x_n \to a$, there must exist N such that:

$$x_n \in (a - \epsilon, a + \epsilon) \subseteq I, \ \forall n > N \tag{4.3.3}$$

Hence by condition 1:

$$g(x_n) \leq f(x_n) \leq h(x_n) \tag{4.3.4}$$

Condition 2,3 imply that:

$$\lim_{n \to \infty} g(x_n) = \lim_{n \to \infty} h(x_n) = f(a) \tag{4.3.5}$$

so by the Squeeze rule of sequences, $\lim_{n \to \infty} f(x_n) = f(a)$, and f is continuous at a. ∎

Example. Consider the function:

$$f(x) = \begin{cases} x^2 \cos \frac{1}{x^2}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{4.3.6}$$

Then, we know that:

$$-1 \leq \cos \frac{1}{x^2} \leq 1, \ x \neq 0 \tag{4.3.7}$$

so that:

$$-x^2 \leq x^2 \cos \frac{1}{x^2} \leq x^2, \ x \neq 0 \tag{4.3.8}$$

Now, since $-x^2 = 0 \leq f(0) = 0 \leq 0 = x^2$, we may assert that:

$$g(x) \leq f(x) \leq h(x) \tag{4.3.9}$$

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 $x = 0$. ◀

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:

$$x_n \in (a - \epsilon, a + \epsilon) \subseteq I, \ \forall n > N \tag{4.3.10}$$

since $x_n \to a$. We define $(x_n)_N^\infty$, consists of two subsequences (x_{m_k}) and (x_{n_k}) satisfying:

$$x_{m_k} < a, \ \text{ and } \ x_{n_k} \geq a \tag{4.3.11}$$

The conditions give:

$$g(x_{m_k}) \to g(a) = f(a), \ \text{ and } \ h(x_{n_k}) \to h(a) = g(a) \tag{4.3.12}$$

so that:

$$f(x_{m_k}) \to f(a), \text{ and } f(x_{n_k}) \to f(a) \tag{4.3.13}$$

Therefore, $f(x_n)$ consists of two subsequences convergent to $f(a)$, and thus $f(x_n) \to f(a)$, as desired. ■

Example. Consider:

$$f(x) = \begin{cases} x^3 - 3x + 5, \ x < 1 \\ \frac{2x+1}{3x-2}, \ x \geq 1 \end{cases} \tag{4.3.14}$$

Let's define $g(x) = x^3 - 3x + 5$ and $h(x) = \frac{2x+1}{3x-2}$, then:

$$f(x) = g(x), \text{ for } x < 1 \tag{4.3.15}$$

and

$$f(x) = h(x), \text{ for } x > 1 \tag{4.3.16}$$

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. ◄

4.4 Trigonometric and exponential functions

We will now prove that the function $\sin, \cos, \tan$ and $\exp$ are continuous.

Proposition *(Sine inequality)*
We have that:

$$\sin x \leq x, \text{ for } 0 \leq x \leq \frac{\pi}{2} \tag{4.4.1}$$

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:

$$f'(x) = 1 - \cos x \tag{4.4.2}$$

and since $0 \leq \cos x \leq 1$ over this interval, we have that:

$$0 \leq f'(x) \leq 1 \tag{4.4.3}$$

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. ■

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 $x > \frac{\pi}{2}$:

$$|\sin x| \leq 1 < \frac{\pi}{2} < x = |x| \tag{4.4.4}$$

Finally, if $x < 0$ then:

$$|\sin(x)| = |\sin(-x)| \leq |-x| = |x| \tag{4.4.5}$$

as desired. ∎

> **Theorem** *(Continuity of trigonometric functions)*
> The trigonometric functions $\sin, \cos, \tan$ are continuous.

Proof. We need to show that:

$$\sin x_n \to \sin a \tag{4.4.6}$$

for all null sequences $x_n - a$.

We can use the property:

$$\sin x - \sin a = 2\cos\left(\frac{1}{2}(x+a)\right)2\sin\left(\frac{1}{2}(x-a)\right) \tag{4.4.7}$$

So:

$$|\sin x_n - \sin a| = 2\left|\cos\left(\frac{1}{2}(x+a)\right)2\sin\left(\frac{1}{2}(x-a)\right)\right| \tag{4.4.8}$$

$$\leq 2\left|\sin\left(\frac{1}{2}(x-a)\right)\right| \tag{4.4.9}$$

$$\leq |x_n - a| \tag{4.4.10}$$

Since $x_n - a$ is a null sequence, we must have that $\sin x_n \to \sin a$ as desired.

Since $\cos x = \sin\left(x + \frac{\pi}{2}\right)$, we can use the continuity of composite functions to state that it too must be continuous. The same goes for $\tan x = \frac{\sin x}{\tan x}$. ∎

> **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:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \leq 1 + x \tag{4.4.11}$$

for $x \geq 0$. Similarly:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \leq 1 + x + x^2 + x^3 + \ldots = \frac{1}{1-x} \tag{4.4.12}$$

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:

$$1 - x \leq e^{-x} \leq \frac{1}{1 + x} \tag{4.4.13}$$

Since all terms in the inequality are non-zero on $0 < -x < 1$, we can take the reciprocal:

$$1 + x \leq e^x \leq 1 - x \tag{4.4.14}$$

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 $x_n \to a$ we have $e^{x_n} \to e^a$.

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:

$$1 + (x_n + a) \leq e^{x_n - a} \leq \frac{1}{1 - (x_n - a)}, \; \forall n > N \tag{4.4.15}$$

By the squeeze rule, we find that $e^{x_n - a} \to 1$ and thus $e^{x_n} \to e^a$ 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) $f(x) = |x|$
> (ii) $f(x) = \sqrt{x}$
> (iv) the trigonometric functions
>   (v) the exponential function

A5 Limits

5.1 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:
>
> $$N_\varepsilon(c) = (c - \varepsilon, c) \cup (c, c + \varepsilon) \tag{5.1.1}$$

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 $x_n \to c$, we have that $f(x_n) \to l$. We write this as:
>
> $$\lim_{x \to c} f(x) = l \tag{5.1.2}$$

Example. Let us try to prove that $\lim\limits_{x \to 0} \frac{\sin x}{x} = 1$.
Firstly, in the previous unit we established the inequality:

$$\sin x \leq x, \ \ \text{for } 0 < x \leq \frac{\pi}{2} \tag{5.1.3}$$

We may also deduce that:

$$x \leq \tan x, \ \ \text{for } 0 < x \leq \frac{\pi}{2} \tag{5.1.4}$$

as can be seen from the figures below:

shaded area $= \frac{1}{2}x$ shaded area $= \frac{1}{2}\tan x$
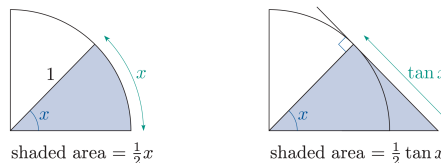
The first inequality (5.1.3) may be rearranged into the more useful form:

$$\frac{\sin x}{x} \leq 1, \ \ \text{for } 0 < x \leq \frac{\pi}{2} \tag{5.1.5}$$

and similarly for the second inequality (5.1.4) may be written as:

$$\cos x \leq \frac{\sin x}{x}, \ \text{ for } 0 < x \leq \frac{\pi}{2} \tag{5.1.6}$$

Hence, these two may be combined into a single inequality providing upper and lower bounds for $\frac{\sin x}{x}$:

$$\cos x \leq \frac{\sin x}{x} \leq 1, \ \text{ for } 0 < x \leq \frac{\pi}{2} \tag{5.1.7}$$

Now since both $\cos x$ and $\frac{\sin x}{x}$ are both even functions, we may substitute $x' = -x$ into (5.1.7) to find that:

$$\cos x \leq \frac{\sin x}{x} \leq 1, \ \text{ for } 0 < |x| \leq \frac{\pi}{2} \tag{5.1.8}$$

Suppose (x_n) is a null sequence in the neighborhood $N_{\frac{\pi}{2}}(0)$ so that:

$$\cos x_n \leq \frac{\sin x_n}{x_n} \leq 1, \ n = 1, 2, \dots \tag{5.1.9}$$

Using the squeeze rule, we see that since $\cos x_n \to 1$ it must be that $\frac{\sin x_n}{x_n} \to 1$. This proves that:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \tag{5.1.10}$$

◀

Example. Consider $\lim_{x \to 0} \lfloor x \rfloor$, 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:

$$\lim_{n \to \infty} f(x_n) = \lim_{n \to \infty} \left\lfloor 1 - \left(\frac{1}{2}\right)^n \right\rfloor = 0 \tag{5.1.11}$$

whereas:

$$\lim_{n \to \infty} f(x_n) = \lim_{n \to \infty} \left\lfloor 1 + \left(\frac{1}{2}\right)^n \right\rfloor = 1 \tag{5.1.12}$$

Therefore, we have found two different sequences on $N_{\frac{1}{2}}(0)$ which converge to different values, showing that $f(x) = \lfloor x \rfloor$ does not tend to a limit as $x \to 1$. ◀

Theorem 9.3 *(Continuity $\Longleftrightarrow$ limit)*
Let f be a function defined on an open interval I and let $c \in I$. Then:

$$f \text{ is continuous at } c \iff \lim_{x \to c} f(x) = f(c) \tag{5.1.13}$$

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 $\lim\limits_{x \to c} f(x) = l$ and $\lim\limits_{x \to l} g(x) = L$, then $\lim\limits_{x \to c} g(f(x)) = L$ provided:

$$g \text{ is defined and continuous at } l \tag{5.1.14}$$

or

$$f(x) \neq l, \forall x \in N_\varepsilon(c) \tag{5.1.15}$$

Example. Let us try to evaluate $\lim\limits_{x \to 0} \sqrt{\frac{x}{\sin x}}$. Let us define $f(x) = \frac{\sin x}{x}$ and $g(x) = \frac{1}{\sqrt{x}}$. Then, we have that:

$$\lim_{x \to 0} f(x) = 1 \tag{5.1.16}$$

as was found previously. Moreover

$$\lim_{x \to 1} g(x) = 1 \tag{5.1.17}$$

so that, since g is defined and continuous at 1 (using the composition rules of continuous functions):

$$\lim_{x \to 0} g(f(x)) = \lim_{x \to 0} \sqrt{\frac{x}{\sin x}} = 1 \tag{5.1.18}$$

◀

Example. We consider the limit:

$$\lim_{x \to 0} \frac{1 - \cos x}{x} \tag{5.1.19}$$

We use the identity:

$$\cos x = 1 - 2\sin^2 \frac{x}{2} \tag{5.1.20}$$

to find that:

$$\lim_{x \to 0} \frac{2\sin^2 \frac{x}{2}}{x} = \lim_{x \to 0} \frac{\sin \frac{x}{2}}{\frac{x}{2}} \cdot \lim_{x \to 0} \sin \frac{x}{2} = 0 \tag{5.1.21}$$

where we used the substitution $u = \frac{x}{2}$ to evaluate $\lim\limits_{x \to 0} \frac{\sin \frac{x}{2}}{\frac{x}{2}}$. ◀

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) $\lim\limits_{x \to c} g(x) = \lim_{x \to c} h(x) = l$ then $\lim\limits_{x \to c} f(x) = l$.

Example. We have shown previously that:

$$1 + x \leq e^x \leq \frac{1}{1 - x}, \quad \text{for } |x| < 1 \tag{5.1.22}$$

This may rearranged into:

$$1 + x \le e^x \le \frac{1}{1-x}, \quad \text{for } |x| < 1 \tag{5.1.23}$$

$$1 \le \frac{e^x - 1}{x} \le \frac{e^x - 1}{x} \le \frac{1}{1-x}, \quad \text{for } 0 < |x| < 1 \tag{5.1.24}$$

Now we have that:

$$1 - \frac{|x|}{1-x} = \frac{x - |x|}{1-x} \le 1 \tag{5.1.25}$$

since if $x < 0$ then $|x| - x = 2|x| > 0$ whereas if $x > 0$ then $|x| - x = 0$. Similarly:

$$\frac{1}{1-x} \le 1 + \frac{|x|}{1-x} = \frac{1 + (|x| - x)}{1-x} \tag{5.1.26}$$

Consequently, we have the following inequality:

$$1 - \frac{|x|}{1-x} \le \frac{e^x - 1}{x} \le 1 + \frac{|x|}{1-x} \tag{5.1.27}$$

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:

$$g(x) \le f(x) \le h(x) \tag{5.1.28}$$

Next, we also have that:

$$\lim_{x \to 0} g(x) = 1 = \lim_{x \to 0} h(x) \tag{5.1.29}$$

Therefore we may conclude that

$$\lim_{x \to 0} \frac{e^x - 1}{x} = 1 \tag{5.1.30}$$

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:

$$\lim_{x \to c^+} f(x) = l \tag{5.1.31}$$

provided that for each sequence (x_n) in $(c, c + r)$ such that $x_n \to c$, $f(x_n) \to l$.
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:

$$\lim_{x \to c^-} f(x) = l \tag{5.1.32}$$

provided that for each sequence (x_n) in $(c - r, c)$ such that $x_n \to c$, $f(x_n) \to l$.

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:

$$\lim_{x \to c} f(x) = l \iff \lim_{x \to c^+} f(x) = \lim_{x \to c^-} f(x) = l \qquad (5.1.33)$$

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:

$$f \text{ is continuous at } c \iff \lim_{x \to c^\pm} f(x) = f(c) \qquad (5.1.34)$$

Example. Let us evaluate $\lim_{x \to 0^+} \left(\frac{\sin x}{x} + \sqrt{x} \right)$.
We have from typical combination rules that:

$$\lim_{x \to 0^+} \left(\frac{\sin x}{x} + \sqrt{x} \right) = \lim_{x \to 0^+} \frac{\sin x}{x} + \lim_{x \to 0^+} \sqrt{x} = 1 + 0 = 0 \qquad (5.1.35)$$

where we used Theorem 9.6 to get:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \implies \lim_{x \to 0^+} \frac{\sin x}{x} = 1 \qquad (5.1.36)$$

and Proposition 9.7 to get

$$\lim_{x \to 0} \sqrt{x} = 1 \implies \lim_{x \to 0^+} \sqrt{x} = 1 \qquad (5.1.37)$$

◀

5.2 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 $x_n \to c$, $f(x_n) \to \infty$. We write that:

$$f(x) \to \infty \text{ as } x \to c \qquad (5.2.1)$$

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) $f(x) \to 0$ as $x \to c$
then $\frac{1}{f(x)} \to \infty$ as $x \to c$.

Example. Consider the asymptotic behaviour of $\frac{1}{x^3-1}$ as $x \to 1^+$. Define $f(x) = x^3 - 1$, then we have that:

$$f(x) = x^3 - 1 > 0 \; \forall x \in (1, \infty) = (1, 1 + r) \tag{5.2.2}$$

for some $r > 0$. Moreover

$$\lim_{x \to 1^+} f(x) = 0 \tag{5.2.3}$$

Therefore:

$$\frac{1}{x^3 - 1} \to \infty \text{ as } x \to 1^+ \tag{5.2.4}$$

◄

Theorem 9.10 *(Squeeze rule for $x \to \infty$)*
Let f, g, h be defined on (R, ∞). Then:
  (a) if
        (i) $g(x) \le f(x) \le h(x), \; \forall x \in (R, \infty)$
        (ii) $\lim\limits_{x \to \infty} g(x) = \lim\limits_{x \to \infty} h(x) = l$
        then $\lim\limits_{x \to \infty} f(x) = l$.
  (b) if
        (i) $g(x) \le f(x), \; \forall x \in (R, \infty)$
        (ii) $g(x) \to \infty$ as $x \to \infty$
        then $f(x) \to \infty$ as $x \to \infty$.

Example. Let's examine the asymptotic behaviour as $x \to \infty$ of:

$$f(x) = \frac{\sin(1/x)}{x} \tag{5.2.5}$$

We have that for $x \ne 0$:

$$-1 \le \sin(1/x) \le 1 \tag{5.2.6}$$

so that:

$$-\frac{1}{x} \le \frac{\sin(1/x)}{x} \le \frac{1}{x}, \; \forall x \ne 0 \tag{5.2.7}$$

Then, since:

$$\lim_{x \to \infty} -\frac{1}{x} = \lim_{x \to \infty} \frac{1}{x} = 0 \tag{5.2.8}$$

we must have that

$$\lim_{x \to \infty} \frac{\sin(1/x)}{x} = 0 \tag{5.2.9}$$

◄

Proposition 9.11 *(Asymptotic behaviour of standard functions)*
  (a) let $a_0, a_1, ..., a_{n-1} \in \mathbb{R}$, and let

$$p(x) = x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0 \tag{5.2.10}$$

Then

$$p(x) \to \infty, \text{ and } \frac{1}{p(x) \to 0} \text{ as } x \to \infty \tag{5.2.11}$$

(b) For $n \in \mathbb{N}$, then:

$$\frac{e^x}{x^n} \to \infty, \text{ and } \frac{x^n}{e^x} \to 0 \text{ as } x \to \infty \tag{5.2.12}$$

(c) we have $\log x \to \infty$ as $x \to \infty$, and for $a > 0$ we have:

$$\lim_{x \to \infty} \frac{\log x}{x^a} = 0 \tag{5.2.13}$$

Proof. (a) Firstly, we have that the zeros of the polynomial p must lie in some interval $(-M, M)$, so we have that:

$$p(x) > 0, \ \forall x \in (M, \infty) \tag{5.2.14}$$

Now if $x \neq 0$ then:

$$p(x) = x^n \left(1 + \frac{a_{n-1}}{x} + ... + \frac{a_0}{x^n} \right) \tag{5.2.15}$$

Now we have that:

$$1 + \frac{a_{n-1}}{x} + ... + \frac{a_0}{x^n} \to 1 + 0 + ... + 0 = 1 \text{ as } x \to \infty \tag{5.2.16}$$

Therefore

$$\frac{1}{p(x)} = \frac{\frac{1}{x^n}}{1 + \frac{a_{n-1}}{x} + ... + \frac{a_0}{x^n}} \to 0 \text{ as } x \to \infty \tag{5.2.17}$$

It follows from the reciprocal rule that, since $p(x) \to \infty$ as $x \to \infty$.

(b) We use the series expansion:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \geq \frac{x^{n+1}}{(n+1)!}, \ \forall x \geq 0 \tag{5.2.18}$$

Hence, if $x > 0$ then:

$$\frac{e^x}{x^n} \geq \frac{x}{(n+1)!}, \text{ and } 0 \leq \frac{x^n}{e^x} \leq \frac{(n+1)!}{x} \tag{5.2.19}$$

Now $\frac{x}{(n+1)!} \to \infty$ so that:

$$\frac{e^x}{x^n} \to \infty, \text{ as } x \to \infty \tag{5.2.20}$$

Similarly, using the Squeeze rule

$$\frac{x^n}{e^x} \to 0 \text{ as } x \to \infty \tag{5.2.21}$$

(c) We know that $\log x$ is the strictly increasing inverse of the exponential, so that $\log x \to \infty$.

Now let $a > 0$, if we define $t(x) = a \log x$ then $x^a = e^{a \log x} = e^t$, giving:

$$\frac{\log x}{x^a} = \frac{t}{ae^t} \tag{5.2.22}$$

We have shown that $t \to \infty$ since a is positive. Hence, using part (b)

$$\frac{t}{ae^t} \to 0 \tag{5.2.23}$$

Using the composition rule of limits:

$$\lim_{x \to \infty} \frac{\log x}{x^a} = 0 \tag{5.2.24}$$

as desired.

∎

Example. Let us examine the bahaviour of

$$f(x) = \frac{2e^x - x^2}{e^x + \log x} \tag{5.2.25}$$

Then:

$$f(x) = \frac{\frac{2e^x}{x^2} - 1}{\frac{e^x}{x^2} + \frac{\log x}{x^2}} \tag{5.2.26}$$

Using the combination rules of limits:

$$\lim_{x \to \infty} f(x) = \frac{\lim_{x \to \infty} \frac{2e^x}{x^2} - 1}{\lim_{x \to \infty} \left(\frac{e^x}{x^2} + \frac{\log x}{x^2} \right)} \tag{5.2.27}$$

$$= \frac{\lim_{x \to \infty} \frac{2e^x}{x^2} - 1}{\lim_{x \to \infty} \frac{e^x}{x^2}} \tag{5.2.28}$$

$$= \lim_{x \to \infty} \left(2 - \frac{x^2}{e^x} \right) \tag{5.2.29}$$

$$= 2 \tag{5.2.30}$$

◀

Example. Let's prove that $\lim_{x \to \infty} x \sin \frac{1}{x} = 1$. We use the substitution $u(x) = \frac{1}{x}$ so that:

$$\lim_{x \to \infty} x \sin \frac{1}{x} = \lim_{u \to 0} \frac{\sin u}{u} = 1 \tag{5.2.31}$$

as desired.

◀

5.3 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:

$$|f(x) - f(c)| < \epsilon, \ \forall x \in A \text{ with } |x - c| < \delta \tag{5.3.1}$$

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

$$|f(x) - f(c)| < \epsilon, \ \forall x \in A \text{ with } |x - c| < \delta \tag{5.3.2}$$

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:

$$|f(x) - f(c)| < \epsilon, \ \forall x \in A \text{ with } |x - c| < \delta \tag{5.3.3}$$

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:
$$|x^3 - 1| < \epsilon, \ \forall x \text{ with } |x - 1| < \delta \tag{5.3.4}$$

We can write that:
$$x^3 - 1 = (x - 1)(x^2 + x + 1) \tag{5.3.5}$$

Now we find an upper bound for $|x^2 + x + 1|$. When $|x - 1| \leq 1$ then $x \in [0, 2]$ so that:

$$|x^2 + x + 1| \leq |x^2| + |x + 1| \leq 4 + 3 = 7, \ \forall |x - 1| \leq 1 \tag{5.3.6}$$

Therefore:
$$|f(x) - f(1)| \leq 7|x - 1|, \ \forall |x - 1| \leq 1 \tag{5.3.7}$$

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:

$$|f(x) - f(1)| < \epsilon, \ \forall x \text{ with } |x - 1| < \delta \tag{5.3.8}$$

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:

$$|f(x) - f(c)| < \epsilon, \ \text{ for } |x - c| < \delta \tag{5.3.9}$$

Now consider a sequence $x_n \in A$ such that $x_n \to c$. Then, there exists N such that:

$$|x_n - c| < \delta, \ \forall n > N \tag{5.3.10}$$

so that:

$$|f(x_n) - f(c)| < \epsilon, \ \forall n > N \tag{5.3.11}$$

Consequently, $f(x_n) \to f(c)$, 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

$$|f(x) - f(c)| < \epsilon, \ \text{ for } |x - c| < \delta \tag{5.3.12}$$

Hence, for all n, $\exists x_n \in A$ with $|x_n - 1| < \frac{1}{n}$ such that

$$|f(x) - f(c)| \geq \epsilon \tag{5.3.13}$$

By the sequential definition $\lim_{n \to \infty} f(x_n) = f(c)$ contradicting the above inequality. Hence the $\epsilon - \delta$ definition must also be satisfied. ∎

5.4 Unusual function continuity

> **Proposition 9.14** *(Dirichlet function)*
> The Dirichlet function defined as:
>
> $$f(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ 0, & \text{if } x \text{ is irrational} \end{cases} \tag{5.4.1}$$
>
> is discontinous everywhere on $\mathbb{R}$.

Proof. Let $c \in \mathbb{R}$. By the density of $\mathbb{R}$, each interval $\left(c - \frac{1}{n}, c + \frac{1}{n}\right)$ with n natural contains a rational x_n and irrational y_n. Then $x_n \to c$ and $y_n \to c$, 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:
>
> $$s(x) = \begin{cases} x - \lfloor x \rfloor, & \text{if } 0 \leq x - \lceil x \rceil \leq \frac{1}{2} \\ 1 - (x - \lfloor x \rfloor), & \text{if } \frac{1}{2} < x - \lfloor x \rfloor < 1 \end{cases} \tag{5.4.2}$$
>
> and the Blancmange function B is defined as:
>
> $$B(x) = \sum_{n=0}^{\infty} \frac{1}{2^n} s(2^n x) \tag{5.4.3}$$
>
> is continous everywhere on $\mathbb{R}$.

Proof. Let $c \in \mathbb{R}$, and let $\epsilon > 0$. Then:

$$B(x) - B(c) = \sum_{n=0}^{\infty} \frac{1}{2^n} (s(2^n x) - s(2^n c)) \tag{5.4.4}$$

Using the triangle ineqality:

$$|B(x) - B(c)| = \left| \sum_{n=0}^{\infty} \frac{1}{2^n} (s(2^n x) - s(2^n c)) \right| \leq \sum_{n=0}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \tag{5.4.5}$$

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:

$$|s(2^n x) - s(2^n c)| \leq \frac{1}{2} \implies \sum_{n=N}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \frac{1}{2} \sum_{n=N}^{\infty} \frac{1}{2}^n \tag{5.4.6}$$

and using our standadrd results for geoemtric series:

$$\sum_{n=N}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \frac{1}{2^N} \tag{5.4.7}$$

Now consider:

$$x \mapsto s(2^n x), \ n = 0, 1, 2... \tag{5.4.8}$$

which is a continuous function. Consequently, for all n there is a positive δ_n such that:

$$|s(2^n x) - s(2^n c)| < \frac{\epsilon}{4}, \ \forall |x - c| < \delta_n \tag{5.4.9}$$

Choosing $\delta = \min_{n \in \mathbb{N}} \delta_n$ we get that for $|x - c| < \delta$

$$\sum_{n=0}^{N-1} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \sum_{n=0}^{N-1} \frac{1}{2^N} \frac{\epsilon}{4} < 2 \cdot \frac{\epsilon}{4} = \frac{\epsilon}{2} \tag{5.4.10}$$

Consequently

$$\sum_{n=0}^{\infty} \frac{1}{2^n} |s(2^n x) - s(2^n c)| \leq \frac{\epsilon}{2} + \frac{1}{2^N} \tag{5.4.11}$$

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
> $$|f(x) - l| < \epsilon, \ \forall x \text{ such that } 0 < |x - c| < \delta \tag{5.4.12}$$
>
> We then write that:
> $$\lim_{x \to c} f(x) = l \tag{5.4.13}$$

Example. We evaluate

$$\lim_{x \to 1} \frac{2x^3 + 3x - 5}{x - 1} \tag{5.4.14}$$

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:

$$|f(x) - 9| < \epsilon, \ \forall x \text{ with } 0 < |x - 1| < \delta \tag{5.4.15}$$

But we have that for $0 < |x - 1| < 1$ then $x \in (-1, 2)$ so that:

$$|2x^2 + 2x + 5| < 17, \ \forall 0 < |x - 1| < 1 \tag{5.4.16}$$

so that

$$|f(x) - 9| < 17|x - 1|, \ \forall 0 < |x - 1| < 1 \tag{5.4.17}$$

Consequently if $|x - 1| < \delta$ then:

$$|f(x) - 9| < 17\delta, \ \forall 0 < |x - 1| < \delta \tag{5.4.18}$$

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. ◄

5.5 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:

$$|f(x) - f(y)| < \epsilon, \ \forall x, y \in I \text{ such that } |x - y| < \delta \tag{5.5.1}$$

Example. Let us prove that $f(x) = x^2$ is uniformly continuous on $I = [-4, 4]$.
Let $\epsilon > 0$, we have:

$$f(x) - f(y) = x^2 - y^2 = (x + y)(x - y) \tag{5.5.2}$$

Therefore $x, y \in [-4, 4]$ implies that $|x| \leq 4$ and $|y| \leq 4$ so that:

$$|f(x) - f(y)| = |x + y||x - y| \tag{5.5.3}$$
$$\leq (|x| + |y|)|x - y| \tag{5.5.4}$$
$$\leq 8|x - y| \tag{5.5.5}$$

Thus if we choose $\delta = \frac{\epsilon}{8}$ and $|x - y| < \delta$ then:

$$|f(x) - f(y)| < \epsilon \tag{5.5.6}$$

as desired. ◄

> **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) $|x_n - y_n| \to 0$ as $n \to \infty$
> (ii) $|f(x_n) - f(y_n)| \geq \epsilon$, for $n = 1, 2...$

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:

$$|x - y| < \delta \text{ and } |f(x) - f(y)| \geq \epsilon \tag{5.5.7}$$

Setting $\delta = 1, \frac{1}{2}, \frac{1}{3}...$ we obtain

$$|x_n - y_n| < \frac{1}{n} \text{ and } |f(x_n) - f(y_n)| \geq \epsilon, \ n = 1, 2, ... \tag{5.5.8}$$

Therefore $|x_n - y_n| \to 0$ and $|f(x_n) - f(y_n)| \geq \epsilon$ as desired.

Now suppose that $|x_n - y_n| \to 0$ 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:

$$|f(x) - f(y)| < \epsilon, \ \forall x, y \in I \text{ with } |x - y| < \delta \tag{5.5.9}$$

But by statement 1, $|x_n - y_n| < \delta$ for $n > N$, so

$$|f(x_n) - f(y_n)| < \epsilon, \ \forall n > N \tag{5.5.10}$$

contradicting the second statement. Therefore f is not uniformly continuous on I. ∎

> **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:
>
> $$|x_n - y_n| = \left| \frac{1}{n} \right| = \frac{1}{n} \to 0 \tag{5.5.11}$$
>
> Moreover:
>
> $$|f(x_n) - f(y_n)| = \left| \left(n + \frac{1}{n} \right)^2 - n^2 \right| \tag{5.5.12}$$
>
> $$= 2 + \frac{1}{n^2} > \epsilon \tag{5.5.13}$$
>
> where $\epsilon = 2$. ◀

A6 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
>
> $$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} = \lim_{x \to c} Q(x) \qquad (6.0.1)$$
>
> 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:

$$Q(x) = \frac{f(x) - f(c)}{x - c} = \frac{1/x - 1/c}{x - c} = -\frac{1}{cx} \qquad (6.0.2)$$

so that:

$$\lim_{x \to c} Q(x) = \lim_{x \to c} \frac{1}{cx} = \frac{1}{c^2} \qquad (6.0.3)$$

where $c \in \mathbb{R}^*$. We may therefore conclude that:

$$f'(x) = -\frac{1}{x^2} \qquad (6.0.4)$$

◀

Example. Let us prove that the function:

$$f(x) = \begin{cases} x^2 \cos \frac{1}{x}, \ x \neq 0 \\ 0, \ x = 0 \end{cases} \qquad (6.0.5)$$

is differentiable at $x = 0$. Indeed:

$$Q(x) = \frac{f(x) - f(0)}{x} = x \cos \frac{1}{x} \qquad (6.0.6)$$

so we must prove that the following limit exists:

$$\lim_{x \to 0} x \cos \frac{1}{x} \qquad (6.0.7)$$

Let us define
$$f_{new}(x) = \begin{cases} x \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{6.0.8}$$

To do so, we know that:
$$-x \leq x \cos \frac{1}{x} \leq x \tag{6.0.9}$$

Now by the squeeze rule, we see that since $h(x) = x$ and $g(x) = -x$ are continuous at 0 and $h(0) = g(0) = f_{new}(0) =$ then
$$\lim_{x \to 0} x \cos \frac{1}{x} = 0 \tag{6.0.10}$$

so that f is indeed differentiable at $x = 0$ with $f'(0) = 0$. ◄

Example. Let us examine the differentiability of
$$f(x) = \begin{cases} \sqrt{|x|} \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{6.0.11}$$

The difference quotient is:
$$Q(x) = \frac{\sqrt{|x|} \sin \frac{1}{x}}{x} \tag{6.0.12}$$

Now consider the sequence $h_n = \frac{1}{n\pi}$ so that
$$Q(h_n) = n\pi \sqrt{\frac{1}{n\pi}} \sin n\pi \tag{6.0.13}$$

which does not exist. Hence we must have that $\lim_{x \to 0} Q(x)$ 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:
$$Q(h) = \frac{f(h + c) - f(h)}{h} \tag{6.0.14}$$

so that differentiability at c is given when the following limit exists:
$$\lim_{h \to 0} Q(h) \tag{6.0.15}$$

Proof. (a) we have that
$$Q(x) = \frac{k - k}{x - c} = 0 \implies f'(x) = 0 \tag{6.0.16}$$

(b) we have that:

$$Q(h) = \frac{(c+h)^n - c^n}{h} = nc^{n-1} + \frac{n(n-1)}{2}c^{n-2}h + \dots + h^{n-1} \qquad (6.0.17)$$

so that $Q(h) \to nc^{n-1}$ as $h \to 0$.

(c) The difference quotient is:

$$Q(h) = \frac{\sin(c+h) - \sin c}{h} = \frac{\sin c \cos h + \sin h \cos c - \sin c}{h} \qquad (6.0.18)$$

$$= \cos c \frac{\sin h}{h} + \sin c \left(\frac{\cos h - 1}{h} \right) \qquad (6.0.19)$$

so using some standard trigonometric limits:

$$\lim_{h \to 0} Q(h) = \cos c \implies f'(x) = \cos x \qquad (6.0.20)$$

(d) The difference quotient is:

$$Q(h) = \frac{\cos(c+h) - \cos c}{h} = \frac{\cos c \cos h - \sin h \sin c - \cos c}{h} \qquad (6.0.21)$$

$$= -\sin c \frac{\sin h}{h} + \cos c \left(\frac{\cos h - 1}{h} \right) \qquad (6.0.22)$$

so using some standard trigonometric limits:

$$\lim_{h \to 0} Q(h) = -\sin c \implies f'(x) = -\sin x \qquad (6.0.23)$$

(e) The difference quotient is:

$$Q(h) = \frac{e^c e^h - e^c}{h} = e^c \frac{e^h - 1}{h} \qquad (6.0.24)$$

so $Q(h) \to e^c$ as $h \to 0$ 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:

$$f'_L(c) = \lim_{x \to c^-} \frac{f(x) - f(c)}{x - c} = \lim_{h \to 0^-} Q(h) \qquad (6.0.25)$$

If this limit exists, then f is left-differentiable at c. Similarly for the right derivative:

$$f'_R(c) = \lim_{x \to c^+} \frac{f(x) - f(c)}{x - c} = \lim_{h \to 0^+} Q(h) \qquad (6.0.26)$$

Proposition 10.4 *(Differentiability and one-sided differentiability)*

Let f be defined on I and let $c \in I$.

$$f \text{ is differentiable at } c \iff f'_R(c) = f'_L(c) = f'(c). \tag{6.0.27}$$

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:

$$f(x) = \begin{cases} g(x), \ x > c \\ h(x), \ x \leq c \end{cases} \tag{6.0.28}$$

then we will have that:

$$f'(x) = \begin{cases} g'(x), \ x > c \\ h'(x), \ x < c \end{cases} \tag{6.0.29}$$

Example. Let us prove that:

$$f(x) = \begin{cases} -x^2, \ x < 0 \\ x^2, \ x \geq 0 \end{cases} \tag{6.0.30}$$

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:

$$f(0) = g(0) = h(0) = 0 \tag{6.0.31}$$

Finally, we also know that $g'(0) = 0$ and $h'(0) = 0$, so that by the Glue rule $f'(0) = 0$.

$$f'(x) = \begin{cases} g'(x) = -2x, \ x < 0 \\ h'(x) = 2x, \ x > 0 \\ 0, \ x = 0 \end{cases} = 2|x|, \ x \in \mathbb{R} \tag{6.0.32}$$

as desired. ◄

6.1 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

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} \tag{6.1.1}$$

For $x \in I$ and $x \neq c$ we have that:

$$f(x) - f(c) = f'(c)(x - c) \tag{6.1.2}$$

Hence

$$\lim_{x \to c} f(x) - f(c) = f'(c) \cdot 0 = 0 \tag{6.1.3}$$

so that $f(x) \to f(c)$ implying continuity. ∎

6.2 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:

$$\frac{F(x) - F(c)}{x - c} = \frac{f(x) - f(c)}{x - c} + \frac{g(x) - g(c)}{x - c} \tag{6.2.1}$$

$$\to f'(c) + g'(c) \tag{6.2.2}$$

as required.

 (ii) Use product rule with $f = \lambda$.

(iii) Let $F = fg$ then:

$$\frac{F(x) - F(c)}{x - c} = \frac{f(x)g(x) - f(c)g(c)}{x - c} + \frac{g(x) - g(c)}{x - c} \tag{6.2.3}$$

$$= \frac{f(x) - f(c)}{x - c} g(c) + f(c) \frac{g(x) - g(c)}{x - c} \tag{6.2.4}$$

$$\to f'(c)g(c) + f(c)g'(c) \tag{6.2.5}$$

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:

$$|g(x) - g(c)| < \frac{1}{2}|g(c)|, \ \forall x \text{ with } |x - c| < \delta \tag{6.2.6}$$

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:

$$\frac{F(x) - F(c)}{x - c} = \frac{f(x)/g(x) - f(c)/g(c)}{x - c} + \frac{g(x) - g(c)}{x - c} \tag{6.2.7}$$

$$= \frac{f(x)g(c) - f(c)g(x)}{(x - c)g(x)g(c)} \tag{6.2.8}$$

$$= \frac{g(c)(f(x) - f(c)) - f(c)(g(x) - g(c))}{(x - c)g(x)g(c)} \tag{6.2.9}$$

$$= \frac{1}{g(x)g(c)}\left(g(c)\frac{f(x) - f(c)}{x - c} - f(c)\frac{g(x) - g(c)}{x - c}\right) \tag{6.2.10}$$

$$\rightarrow \frac{g(c)f'(c) - f(c)g'(c)}{(g(c))^2} \tag{6.2.11}$$

as desired.

∎

> **Theorem 10.8** *(Composition rule)* Let f be defiend 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
>
> $$(g \circ f)'(c) = g'(f(c))f'(c) \tag{6.2.12}$$

Proof. Let $F = g \circ f$ then:

$$\frac{F(x) - F(c)}{x - c} = \frac{g(f(x)) - g(f(c))}{x - c} \tag{6.2.13}$$

Let $y = f(x)$ with $x \in I$ and let $d = f(c)$. Then the right hand side of the above equation is:

$$\left(\frac{g(y) - g(d)}{y - d}\right)\left(\frac{f(x) - f(c)}{x - c}\right),\ y \neq d \tag{6.2.14}$$

We may introduce the function $h(y)$ to deal with the discontinuity at $y \neq d$:

$$h(y) = \begin{cases} \frac{g(y) - g(d)}{y - d}, & y \neq d \\ g'(f(c)), & f(x) = f(c) \end{cases} \tag{6.2.15}$$

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:

$$(h \circ f)(x) = \begin{cases} \frac{g(f(x)) - g(f(c))}{f(x) - f(c)}, & f(x) \neq f(c) \\ g'(f(c)), & f(x) = f(c) \end{cases} \tag{6.2.16}$$

is continuous at c.

Therefore:

$$\frac{F(x) - F(c)}{x - c} = (h \circ f)(x)\left(\frac{f(x) - f(c)}{x - c}\right) \tag{6.2.17}$$

so that as $x \rightarrow c$ then:

$$\frac{F(x) - F(c)}{x - c} \rightarrow g'(f(c))f'(c) \tag{6.2.18}$$

as desired. ∎

Example. Let's find the derivative of $f(x) = \cos\left(\frac{\cos 2x}{x^2}\right)$.

Let us define
$$g(x) = \cos x, x \in I \tag{6.2.19}$$

and
$$h(x) = \frac{\cos 2x}{x^2}, x \in I = (0, \infty) \tag{6.2.20}$$

Then $h(I) \subseteq (0, \infty) = I$ so we may apply the composition rule:

$$(g \circ f)'(x) = -\sin\left(\frac{\cos 2x}{x^2}\right) \cdot \frac{-2x^2 \sin 2x - 2x \cos 2x}{x^4} \tag{6.2.21}$$

$$= 2\sin\left(\frac{\cos 2x}{x^2}\right) \frac{x \sin 2x + \cos 2x}{x^3} \tag{6.2.22}$$

◀

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:

$$(f^{-1})'(d) = \frac{1}{f'(c)} \tag{6.2.23}$$

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:

$$\frac{f^{-1}(y) - f^{-1}(d)}{y - d} = \frac{x - c}{f(x) - f(c)} = \frac{1}{\frac{f(x) - f(c)}{x - c}} \tag{6.2.24}$$

Taking the limit as $y \to d \implies x = f^{-1}(y) \to c$ due to the continuity of f^{-1}. Hence:

$$\lim_{y \to d} \frac{f^{-1}(y) - f^{-1}(d)}{y - d} = \frac{1}{f'(c)} \tag{6.2.25}$$

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:

$$(f^{-1})'(y) = \frac{1}{\sec^2 x} = \frac{1}{1 + \tan^2 x} = \frac{1}{1 + y^2} \tag{6.2.26}$$

implying that:

$$(\tan^{-1})'(x) = \frac{1}{1 + x^2}, \ \forall x \in \mathbb{R} \tag{6.2.27}$$

◄

6.3   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:
$$f(x) \leq f(c), \ \forall x \in I \tag{6.3.1}$$

2. **local minimum** $f(c)$ at $x = c$ if there exists $I = (c - r, c + r) \subseteq J$ where $r > 0$ such that:
$$f(c) \leq f(x), \ \forall x \in I \tag{6.3.2}$$

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:

$$f(x) \leq f(c), \ \text{ for } c - r < x < c + r \tag{6.3.3}$$

Let $x_n = c + \frac{r}{n}$ and $x'_n = c - \frac{r}{n}$ for $n = 2, 3, ...$ Then $c < x_n < c + r$ so that $f(x_n) \leq f(c)$ and $x_n > c$. Hence:
$$\frac{f(x_n) - f(c)}{x_n - c} \leq 0 \implies f'(c) \leq 0 \tag{6.3.4}$$

We also have that $c - r < x'_n < c$ so that $f(x'_n) \leq f(c)$ and $x'_n < c$ and hence:

$$\frac{f(x'_n) - f(c)}{x_n - c} \geq 0 \implies f'(c) \geq 0 \tag{6.3.5}$$

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:

$$f'(x) = 2 \sin x \cos x - \sin x \tag{6.3.6}$$

which vanishes when:

$$2 \sin x \cos x = \sin x \implies \sin x = 0 \text{ or } \cos x = \frac{1}{2} \tag{6.3.7}$$

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 $f(x) = 1$. ◀

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.

∎

6.4 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:

$$f'(c) = \frac{f(b) - f(a)}{b - a} \tag{6.4.1}$$

Proof. The gradient of the chord joining the points $(a, f(a))$ and $(b, f(b))$ is:

$$m = \frac{f(b) - f(a)}{b - a} \tag{6.4.2}$$

Consequently its equation will be $y = \frac{f(b)-f(a)}{b-a}(x - a) + f(a)$. Let us then define:

$$h(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a) - f(a) \tag{6.4.3}$$

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:

$$h'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} = 0 \implies f'(c) = \frac{f(b) - f(a)}{b - a} \tag{6.4.4}$$

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:

$$m = \frac{2e^2}{2} = e^2 \implies \exists c \in (0, 2) \text{ s.t.} f'(c) = e^2 \tag{6.4.5}$$

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:

$$f'(c) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \tag{6.4.6}$$

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

$$g(x) \geq h(x), \ \forall x \in [a, b] \tag{6.4.7}$$

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:

$$f(a) \geq 0 \ \text{ and } \ f'(x) \geq 0 \forall x \in (a, b) \tag{6.4.8}$$

which shows that f is smallest at a in $[a, b]$, or:

$$f(b) \geq 0 \ \text{ and } \ f'(x) \leq 0 \forall x \in (a, b) \tag{6.4.9}$$

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 $x \geq -1$:

$$(1 + x)^\alpha \geq 1 + \alpha x \tag{6.4.10}$$

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:

$$f'(x) = \alpha(1 + x)^{\alpha - 1} - \alpha = \alpha((1 + x)^{\alpha - 1} - 1) \tag{6.4.11}$$

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. ◄

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:

$$|f''(x) - f''(c)| < \epsilon = \frac{1}{2}f''(c), \ \forall x \in (c - \delta, c + \delta) \tag{6.4.12}$$

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:

$$f'(x) < 0, \ \forall x \in (c - \delta, c) f'(x) > 0, \ \forall x \in (c, c + \delta) \tag{6.4.13}$$

This implies that

$$f(x) \text{ is decreasing}, \ \forall x \in (c - \delta, c) f(x) \text{ is increasing}, \ \forall x \in (c, c + \delta) \tag{6.4.14}$$

proving that f has a local minimum at c. ∎

6.5 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:

$$f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a)) \tag{6.5.1}$$

Proof. Let us define:
$$h(x) = f(x)(g(b) - g(a)) - g(x)(f(b) - f(a)) \tag{6.5.2}$$

Clearly by the combination rules of continuity and differentiability, h must be continuous on $[a, b]$ and differentiable on (a, b). Note also that:

$$h(a) = f(a)g(b) - g(a)f(b) = h(b) \tag{6.5.3}$$

We may therefore apply Rolle's Theorem:

$$\exists c \in (a, b) \text{ s.t.} h'(c) = f'(c)(g(b) - g(a)) - g'(c)(f(b) - f(a)) = 0 \tag{6.5.4}$$

thus implying that:

$$f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a)) \tag{6.5.5}$$

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:
>
> $$\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)} \tag{6.5.6}$$
>
> provided the latter limit exists.

Proof. Assume that:

$$\lim_{x \to c} \frac{f'(x)}{g'(x)} = l \tag{6.5.7}$$

Let $\epsilon > 0$, then it follows from the existence of the above limit that $\exists \delta > 0$ such that:

$$\left| \frac{f'(x)}{g'(x)} - l \right| < \epsilon, \ 0 < |x - c| < \delta \tag{6.5.8}$$

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:

$$f'(d)(g(x) - g(c)) = g'(d)(f(x) - f(c)) \implies \frac{f'(d)}{g'(d)} = \frac{f(x) - f(c)}{g(x) - g(c)} \tag{6.5.9}$$

and since $f(c) = g(c) = 0$ we get that:

$$\frac{f'(d)}{g'(d)} = \frac{f(x)}{g(x)} \tag{6.5.10}$$

Therefore:

$$\left| \frac{f(x)}{g(x)} - l \right| = \left| \frac{f'(d)}{g'(d)} - l \right| < \epsilon, \ 0 < |x - c| < \delta \tag{6.5.11}$$

proving that

$$\lim_{x \to c} \frac{f(x)}{g(x)} = l \tag{6.5.12}$$

as desired. ∎

> **Example.** Consider
>
> $$\lim_{x \to 0} \frac{\sin x - x \cos x}{x^3} \tag{6.5.13}$$
>
> 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:
>
> $$\lim_{x \to 0} \frac{\sin x - x \cos x}{x^3} = \lim_{x \to 0} \frac{\cos x - \cos x + x \sin x}{3x^2} = \lim_{x \to 0} \frac{\sin x}{3x} = \frac{1}{3} \tag{6.5.14}$$

A7 Integration

7.1 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 $[a, b]$:

$$P = \{[x_0, x_1], [x_1, x_2], ..., [x_{i-1}, x_i], ..., [x_{n-1}, x_n]\} \tag{7.1.1}$$

with:

$$a = x_0 < x_1 < ... < x_i < ... < x_n = b \tag{7.1.2}$$

The points x_i are known as partition points, and the ith 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 $[0, 1]$:

$$P = \left\{ \left[0, \frac{1}{2}\right], \left[\frac{1}{2}, \frac{3}{5}\right], \left[\frac{3}{5}, \frac{3}{4}\right], \left[\frac{3}{4}, 1\right] \right\} \tag{7.1.3}$$

We find that:

$$\delta x_1 = \frac{1}{2}, \ \delta x_2 = \frac{1}{10}, \ \delta x_3 = \frac{3}{20}, \ \delta x_4 = \frac{1}{4} \tag{7.1.4}$$

so that:

$$\|P\| = \max\{\delta x_1, \delta x_2, \delta x_3, \delta x_4\} = \frac{1}{2} \tag{7.1.5}$$

◀

Definition *(Riemann sums)*

Let f be bounded on $[a, b]$ and let $P = \{[a, x_1], [x_1, x_2], ..., [x_{i-1}, x_i], ..., [x_{n-1}, b]\}$. Define:

$$m_i = \inf_{x \in I_i} f, \ M_i = \sup_{x \in I_i} f \tag{7.1.6}$$

Then the lower Riemann sum for f on $[a, b]$ is:

$$L(f, P) = \sum_{i=1}^{n} m_i \delta x_i \tag{7.1.7}$$

while the upper Riemann sum for f on $[a, b]$ is:

$$U(f, P) = \sum_{i=1}^{n} M_i \delta x_i \tag{7.1.8}$$

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 ith vertical strip, while $M_i \delta x_2$ will overestimate it.

Example. Consider the function

$$f(x) = \begin{cases} 2x, \ 0 < x < 1 \\ 1, \ x = 0, 1 \end{cases} \tag{7.1.9}$$

and the partition $P = \left\{ \left[0, \frac{1}{4}\right], \left[\frac{1}{4}, \frac{1}{2}\right], \left[\frac{1}{2}, \frac{3}{4}\right], \left[\frac{3}{4}, 1\right] \right\}$. Then we see that:

$$m_1 = 0, \ M_1 = 1, \ \delta x_1 = \frac{1}{4} \tag{7.1.10}$$

$$m_2 = \frac{1}{2}, \ M_2 = 1, \ \delta x_2 = \frac{1}{4} \tag{7.1.11}$$

$$m_3 = 1, M_3 = \frac{3}{2}, \ \delta x_3 = \frac{1}{4} \tag{7.1.12}$$

$$m_4 = 1, M_4 = 2, \ \delta x_4 = \frac{1}{4} \tag{7.1.13}$$

$$\tag{7.1.14}$$

and therefore:

$$L(f, P) = \frac{1}{4}\left(0 + \frac{1}{2} + 1 + 1\right) = \frac{5}{8} \tag{7.1.15}$$

$$U(f, P) = \frac{1}{4}\left(1 + 1 + \frac{3}{2} + 2\right) = \frac{11}{8} \tag{7.1.16}$$

◀

Example. Let

$$f(x) = \begin{cases} x^2, \ 0 \le x \le 1 \\ 2, \ 1 < x \le 2 \end{cases} \tag{7.1.17}$$

and the partition $P = \left\{ \left[0, \frac{1}{n}\right], \left[\frac{1}{n}, \frac{2}{n}\right], ..., \left[2 - \frac{1}{n}, 2\right] \right\}$.
Now we see that the ith interval is $\left[\frac{i-1}{n}, \frac{i}{n}\right]$, 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 ith inter-

val, while M_i will be the value of f at the right endpoint of the ith 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:

$$L(f, P) = \sum_{i=1}^{n+1} \frac{(i-1)^2}{n^2} \frac{1}{n} + \sum_{i=n+2}^{2n} 2\frac{1}{n} \tag{7.1.18}$$

$$= \frac{1}{n^3} \sum_{i=0}^{n} i^2 + \sum_{i=n+2}^{2n} 2\frac{1}{n} \tag{7.1.19}$$

$$= \frac{1}{n^3} \frac{n(n+1)(2n+1)}{6} + \frac{2}{n}(2n - n - 1) \tag{7.1.20}$$

$$= \frac{n(n+1)(2n+1)}{6n^3} + 2 - \frac{2}{n} \tag{7.1.21}$$

Instead:

$$U(f, P) = \sum_{i=1}^{n} \frac{i^2}{n^2} \frac{1}{n} + \sum_{n+1}^{2n} 2\frac{1}{n} \tag{7.1.22}$$

$$= \frac{n(n+1)(2n+1)}{6n^3} + 2\frac{1}{n}(2n - n) \qquad = \frac{n(n+1)(2n+1)}{6n^3} + 2 \tag{7.1.23}$$

Note that taking the limit as $n \to \infty$:

$$\lim_{n \to \infty} L(f, P) = \frac{1}{3} + 2 = \frac{7}{3} \tag{7.1.24}$$

and similarly:

$$\lim_{n \to \infty} U(f, P) = \frac{1}{3} + 2 = \frac{7}{3} \tag{7.1.25}$$

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 $P = \{[x_0, x_1], ..., [x_{n-1}, x_n]\}$. 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:

$$L(f, P) \le L(f, P'') \le U(f, P'') \le U(f, P') \tag{7.1.26}$$

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 $L(g, P) \le U(g, P')$ and thus $L(f, P) \le U(f, P')$ 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:

$$\underline{\int_a^b} f = \sup_P L(f, P) \tag{7.1.27}$$

while the upper integral of f is:

$$\overline{\int_a^b} f = \inf_P U(f, P) \tag{7.1.28}$$

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:

$$\underline{\int_0^2} f = \overline{\int_0^2} f = \frac{7}{3} \implies \int_0^2 f = \frac{7}{3} \tag{7.1.29}$$

Example. Let f be the Dirichlet function defined by

$$f(x) = \begin{cases} 1, & 0 \le x \le 1, \ x \text{ is irrational} \\ 0, & 0 \le x \le 1, \ x \text{ is rational} \end{cases} \tag{7.1.30}$$

with a partition $P = \{[0, x_1], [x_1, x_2], ..., [x_{n-1}, 1]\}$. 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:

$$L(f, P) = \sum_i m_i \delta x_i = 0 \tag{7.1.31}$$

while

$$U(f, P) = \sum_i M_i \delta x_i = \sum_i \delta x_i = 1 \tag{7.1.32}$$

Therefore:

$$\underline{\int_0^1} f = 0 \ne \overline{\int_0^1} f = 1 \tag{7.1.33}$$

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 $\|P_n\| \to 0$ then:
$$\lim_{n \to \infty} L(f, P_n) = \lim_{n \to \infty} U(f, P_n) = A \in \mathbb{R} \tag{7.1.34}$$
then $\int_a^b f = A$.

Proof. Let $\epsilon > 0$, then there exists n such that:
$$|L(f, P_n) - A| < \frac{1}{2}\varepsilon \implies L(f, P_n) > A - \frac{1}{2}\varepsilon \tag{7.1.35}$$

and similarly:
$$U(f, P_n) < A + \frac{1}{2}\varepsilon \tag{7.1.36}$$

Also, by definition we must have that:
$$L(f, P_n) \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq U(f, P_n) \tag{7.1.37}$$

so that:
$$A - \frac{1}{2}\varepsilon \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq A + \frac{1}{2}\varepsilon \tag{7.1.38}$$

Therefore, we find that f is integrable on $[a, b]$ with:
$$\int_a^b f = A \tag{7.1.39}$$

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:
$$P_n = \{[a, x_1], [x_1, x_2], ..., [x_{n-1}, n]\} \tag{7.1.40}$$

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:
$$U(f, P_n) - L(f, P_n) = \sum_i (f(x_i) - f(x_{i-1}))\frac{b-a}{n} = (f(b) - f(a))\frac{b-a}{n} \tag{7.1.41}$$

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:

$$\int_a^c f + \int_c^b f = \int_a^b f \tag{7.1.42}$$

Suppose that f is integrable on $[a, b]$, then $|f|$ is also integrable on $[a, b]$. Also:

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g \tag{7.1.43}$$

and

$$\int_a^b \lambda f = \lambda \int_a^b f \tag{7.1.44}$$

Finally, both fg and f/g are integrable, provided that $\frac{1}{g}$ is bounded on $[a, b]$ in the latter case.

7.2 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:

$$\int_a^b f \leq \int_a^b g \tag{7.2.1}$$

(ii) if $m \leq f(x) \leq M$, $\forall x \in [a, b]$ then:

$$m(b - a) \leq \int_a^b f \leq M(b - a) \tag{7.2.2}$$

(iii)

$$\left| \int_a^b f \right| \leq \int_a^b |f| \tag{7.2.3}$$

Proof. (i) Suppose that $f(x) \leq g(x)$, $\forall x \in [a, b]$, and let P be any partition of $[a, b]$. Then:

$$\inf_{[x_i, x_{i+1}]} f \leq \inf_{[x_i, x_{i+1}]} \tag{7.2.4}$$

implying that:

$$\int_a^b f = \sup_P L(f, p) \leq \sup_P L(g, P) = \int_a^b g \tag{7.2.5}$$

as desired.

(ii) Suppose $m \leq f(x) \leq M$ over $[a, b]$. Then from the results of part (a)

$$\int_a^b m \leq \int_a^b f \leq \int_a^b M \implies m(b - a) \leq \int_a^b f \leq M(b - a) \tag{7.2.6}$$

(iii) Note that $-f(x) \le |f(x)| \le f(x)$ for all $x \in [a, b]$, so that:

$$-\int_a^b |f| \le \int_a^b f \le \int_a^b |f| \implies \left| \int_a^b f \right| \le \int_a^b |f| \qquad (7.2.7)$$

as desired.

∎

Example.

(a) Let us prove that:

$$\int_1^3 x \sin \frac{1}{x^{10}} dx \le 4 \qquad (7.2.8)$$

We have that for all $x \in [1, 3]$:

$$-1 \le \sin \frac{1}{x^{10}} \le 1 \qquad (7.2.9)$$

implying that:

$$\int_1^3 x \sin \frac{1}{x^{10}} dx \le \int_1^3 x = \frac{1}{2}[x^2]_0^3 = 4 \qquad (7.2.10)$$

as desired.

(b) Let us prove that:

$$\frac{1}{2} \le \int_0^{\frac{1}{2}} e^{x^2} dx \le \frac{1}{2} e^{1/4} \qquad (7.2.11)$$

Indeed, note that:

$$\frac{d}{dx}(e^{x^2}) = 2xe^{x^2} \ge 0, \ \forall x \in \left[0, \frac{1}{2}\right] \qquad (7.2.12)$$

showing that e^{x^2} is increasing on $\left[0, \frac{1}{2}\right]$. It then follows that:

$$1 \le e^{x^2} \le e^{\frac{1}{4}}, \ \forall x \in \left[0, \frac{1}{2}\right] \qquad (7.2.13)$$

and hence:

$$\frac{1}{2} \le \int_0^{\frac{1}{2}} e^{x^2} dx \le \frac{1}{2} e^{1/4} \qquad (7.2.14)$$

as desired.

(c) Finally, let us prove that:

$$\left| \int_0^{\frac{\pi}{4}} \frac{\tan x}{3 - \sin x^2} dx \right| \le \frac{1}{4} \log 2 \qquad (7.2.15)$$

Indeed, we have that:

$$-1 \le \sin(x^2) \le 1 \implies 2 \le 3 - \sin x^2 \le 4 \qquad (7.2.16)$$

Thus:

$$\frac{1}{4} \tan x \le \frac{\tan x}{3 - \sin x^2} \le \frac{1}{2} \tan x \qquad (7.2.17)$$

implying that:

$$\left| \frac{\tan x}{3 - \sin x^2} \right| \leq \frac{1}{2} |\tan x| \tag{7.2.18}$$

Also $\tan x \geq 0$, $\forall x \in \left[0, \frac{\pi}{4}\right]$, so that:

$$\left| \frac{\tan x}{3 - \sin x^2} \right| \leq \frac{1}{2} \tan x \tag{7.2.19}$$

Using the limit inequality:

$$\left| \int_0^{\frac{\pi}{4}} \frac{\tan x}{3 - \sin x^2} dx \right| \leq \int_0^{\frac{\pi}{4}} \frac{1}{2} \tan x \, dx = \frac{1}{2} [\ln |\sec x|]_0^{\frac{\pi}{4}} \tag{7.2.20}$$

$$= \frac{1}{2} \ln \left| \sqrt{2} \right| = \frac{1}{4} \ln 2 \tag{7.2.21}$$

as desired.

◀

Wallis' formula

Lemma. Let

$$I_n = \int_0^{\frac{\pi}{2}} \sin^n x \, dx, \; n = 0, 1, 2... \tag{7.2.22}$$

Then $I_n = \frac{n-1}{n} I_{n-2}$ for $n \geq 2$.

Proof. It is easy to see that:

$$I_0 = \frac{\pi}{2}, \; I_1 = 1 \tag{7.2.23}$$

Also, for $n \geq 2$:

$$I_n = \int_0^{\pi/2} \sin^n x \, dx = \int_0^{\pi/2} \sin x \sin^{n-1} x \, dx \tag{7.2.24}$$

$$= [-\cos x \sin^{n-1} x]_0^{\pi/2} + \int_0^{\pi/2} (n-1) \cos^2 x \sin^{n-2} x \, dx \tag{7.2.25}$$

$$= (n-1) \left(\int_0^{\pi/2} \sin^{n-2} x \, dx - \int_0^{\pi/2} \sin^n x \, dx \right) \tag{7.2.26}$$

$$\implies n I_n = (n-1) I_{n-2} \implies I_n = \frac{n-1}{n} I_{n-2} \tag{7.2.27}$$

as desired. ∎

For example, we have that even powers of $\sin x$ integrate to:

$$I_4 = \frac{3}{4} \frac{1}{2} \frac{\pi}{2}, \; I_6 = \frac{5}{6} \frac{3}{4} \frac{1}{2} \frac{\pi}{2} \tag{7.2.28}$$

or more generally:

$$I_{2n} = \frac{2n-1}{2n} \frac{2n-3}{2n-2} ... \frac{3}{4} \frac{1}{2} \frac{\pi}{2} \tag{7.2.29}$$

Similarly, we have that odd powers of $\sin x$ integrate to:

$$I_5 = \frac{4}{5}\frac{2}{3}, \; I_7 = \frac{6}{7}\frac{4}{5}\frac{2}{3} \tag{7.2.30}$$

or more generally:

$$I_{2n+1} = \frac{2n}{2n+1}\frac{2n-2}{2n-1}...\frac{4}{5}\frac{2}{3} \tag{7.2.31}$$

Lemma. Let:

$$a_n = \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}...\frac{2n}{2n-1}\frac{2n}{2n+1} \tag{7.2.32}$$

$$b_n = \frac{(n!)^2 2^{2n}}{(2n)!\sqrt{n}} \tag{7.2.33}$$

then:

$$b_n^2 = \frac{2n+1}{n}a_n, \; n = 1, 2, ... \tag{7.2.34}$$

Proof. Firstly note that:

$$b_1^2 = \frac{2^4}{4} = 4, \; a_1 = \frac{2}{1}\frac{2}{3} = \frac{4}{3} \implies b_1^2 = 3a_1 \tag{7.2.35}$$

More generally, suppose that for some n:

$$b_n^2 = \frac{2n+1}{n}a_n \tag{7.2.36}$$

then:

$$b_{n+1}^2 = \frac{((n+1)!)^4 2^{4(n+1)}}{((2n+2)!)^2(n+1)} \tag{7.2.37}$$

$$= \frac{(n!)^4(n+1)^4 2^{4n}2^4}{(2n!)^2(2n+2)^2(2n+1)^2(n+1)} \tag{7.2.38}$$

$$= \frac{(n!)^4 2^{4n}}{((2n)!)^2 n}\frac{n}{n+1}\frac{(n+1)^4 2^4}{2^2(n+1)^2(2n+1)^2} \tag{7.2.39}$$

$$= \frac{2n+1}{n}\frac{n}{n+1}\frac{(n+1)^2 2^2}{(2n+1)^2}a_n \tag{7.2.40}$$

$$= \frac{4(n+1)}{2n+1}a_n \tag{7.2.41}$$

Now note that:

$$a_{n+1} = \frac{2n+2}{2n+1}\frac{2n+2}{2n+3}a_n \implies \frac{4(n+1)}{2n+1}a_n = \frac{2n+3}{n+1}a_{n+1} \tag{7.2.42}$$

and hence:

$$b_{n+1}^2 = \frac{2n+3}{n+1}a_{n+1} \tag{7.2.43}$$

as desired. ∎

> **Theorem** *(Wallis' Formula)* Wallis' formula:
>
> $$\lim_{n\to\infty}\frac{(n!)^2 2^{2n}}{(2n)!\sqrt{n}} = \sqrt{\pi} \tag{7.2.44}$$

Proof. We begin by proving that:

$$\lim_{n \to \infty} a_n = \frac{\pi}{2} \tag{7.2.45}$$

Indeed, note that:

$$\frac{\pi}{2} a_n = \frac{I_{2n+1}}{I_{2n}} \tag{7.2.46}$$

so we need to prove that:

$$\lim_{n \to \infty} \frac{I_{2n+1}}{I_{2n}} = 1 \tag{7.2.47}$$

Furthermore, for $x \in [0, \pi/2]$ then:

$$\sin^{2n+2} x \le \sin^{2n+1} x \le \sin^{2n} x \implies I_{2n+2} \le I_{2n+1} \le I_{2n} \tag{7.2.48}$$

using the inequality rules for integrals. Therefore:

$$\frac{I_{2n+2}}{I_{2n}} = \frac{2n+1}{2n+2} \le \frac{I_{2n+1}}{I_{2n}} \le 1 \tag{7.2.49}$$

Hence, using the squeeze rule:

$$\lim_{n \to \infty} \frac{I_{2n+1}}{I_{2n}} = 1 \implies \lim_{n \to \infty} a_n = \frac{\pi}{2} \tag{7.2.50}$$

as desired.

We then have that:

$$\lim_{n \to \infty} b_n^2 = \lim_{n \to \infty} \frac{2n+1}{n} a_n = \pi \implies \lim_{n \to \infty} b_n = \sqrt{\pi} \tag{7.2.51}$$

$\blacksquare$

7.3 Series

> **Theorem** *(Integral test)*
> Let f be positive and decreasing on $[1, \infty)$, and suppose $\lim_{x \to \infty} f(x) = 0$. Define:
>
> $$I_n = \int_1^n f \tag{7.3.1}$$
>
> 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 nth partial sum of $\sum_{n=1}^{\infty} f(n)$, and let P_{n-1} be the partition of $[1, n]$:

$$P_n = \{[1,2], ..., [i, i+1], ..., [n-1, n]\} \tag{7.3.2}$$

Now since $f(x)$ is decreasing on $[0, \infty)$, we must have that:

$$m_i = f(i+1) \implies L(f, P_{n-1}) = f(2) + f(3) + ... + f(n) = s_n - f(1) \qquad (7.3.3)$$

$$M_i = f(i) \implies U(f, P_{n-1}) = f(1) + f(2) + ... + f(n-1) = s_n - f(n) \qquad (7.3.4)$$

Hence:

$$s_n - f(1) \leq \int_1^n f \leq s_n - f(n) \qquad (7.3.5)$$

(i) Suppose $I_n = \int_1^n f$ is bounded above by some M:

$$s_n - f(1) \leq I_n \leq M \implies s_n \leq M + f(1) \qquad (7.3.6)$$

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:

$$I_{n+1} - I_n = \int_n^{n+1} f \geq 0 \implies I_n \text{ is increasing} \qquad (7.3.7)$$

So we have that I_n diverges. Note also that:

$$s_n \geq I_n \qquad (7.3.8)$$

so using the Squeeze rule, s_n diverges, and hence so does $\sum_{n=1}^{\infty} f(n)$.

$\blacksquare$

Example. Consider:

$$\int \frac{dx}{x(\log x)^2} = \int \frac{du}{u^2} = -\frac{1}{\log x} \qquad (7.3.9)$$

where we used $u = \log x$, $du = \frac{1}{x}$. Hence:

$$I_n = \int_2^n \frac{dx}{x(\log x)^2} = \left[\frac{1}{\log x} \right]_n^2 = \frac{1}{\log 2} - \frac{1}{\log n} \qquad (7.3.10)$$

Note that $x(\log x)^2$ is positive and increasing on $[2, \infty)$, implying that $f = \frac{1}{x(\log x)^2}$ is positive an decreasing on $[2, \infty)$. Furthermore:

$$I_n = \frac{1}{\log 2} - \frac{1}{\log n} \leq \frac{1}{\log 2} \qquad (7.3.11)$$

so I_n is bounded above. It follows that:

$$\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2} \text{ is convergent} \qquad (7.3.12)$$

$\blacktriangleleft$

A8 Power series

8.1 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:
>
> $$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}(x-a)^k = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + ... + \frac{f^{(n)}(x)}{n!}(x-a)^n$$
>
> $$(8.1.1)$$

Example. Consider $f(x) = \cos x$, let us find its nth order Taylor polynomial about $a \in \mathbb{R}$. Firstly we evaluate the derivatives:

$$
\begin{aligned}
f(x) &= \cos x &\implies f(a) &= \cos a \\
f'(x) &= -\sin x &\implies f'(a) &= -\sin a \\
f''(x) &= -\cos x &\implies f''(a) &= -\cos a \\
f^{(3)}(x) &= \sin x &\implies f^{(3)}(a) &= \sin a \\
f^{(4)}(x) &= \cos x &\implies f^{(4)}(a) &= \cos a \\
&\qquad\qquad \vdots \\
f^{(2n)}(x) &= (-1)^n \cos x &\implies f^{(2n)}(a) &= (-1)^n \cos a \\
f^{(2n+1)}(x) &= (-1)^{n+1} \sin x &\implies f^{(2n+1)} &= (-1)^{n+1} \sin a
\end{aligned}
$$

Consequently, even order taylor polynomials are:

$$T_{2n}(a) = \sum_{k=0}^{n} \frac{(-1)^k \cos a}{(2k)!}(x-a)^{2k} + \sum_{k=0}^{n-1} \frac{(-1)^{k+1} \sin a}{(2k+1)!}(x-a)^{2k+1} \qquad (8.1.2)$$

while odd order taylor polynomials are:

$$T_{2n+1}(a) = \sum_{k=0}^{n} \frac{(-1)^k \cos a}{(2k)!}(x-a)^{2k} + \sum_{k=0}^{n} \frac{(-1)^{k+1} \sin a}{(2k+1)!}(x-a)^{2k+1} \qquad (8.1.3)$$

Taking $a = 0$ we find that:

$$T_{2n}(0) = T_{2n+1}(0) = \sum_{k=0}^{n} \frac{(-1)^k}{(2k)!}(x-a)^{2k} \tag{8.1.4}$$

◀

Example. Consider $f(x) = \sin x$, let us find its nth order Taylor polynomial about $a \in \mathbb{R}$. Firstly we evaluate the derivatives:

$$f(x) = \sin x \implies f(a) = \sin a$$
$$f'(x) = \cos x \implies f'(a) = \cos a$$
$$f''(x) = -\sin x \implies f''(a) = -\sin a$$
$$f^{(3)}(x) = -\cos x \implies f^{(3)}(a) = -\cos a$$
$$f^{(4)}(x) = \sin x \implies f^{(4)}(a) = \sin a$$
$$\vdots$$
$$f^{(2n)}(x) = (-1)^n \sin x \implies f^{(2n)}(a) = (-1)^n \sin a$$
$$f^{(2n+1)}(x) = (-1)^n \cos x \implies f^{(2n+1)} = (-1)^n \cos a$$

Consequently, even order taylor polynomials are:

$$T_{2n+2}(a) = \sum_{k=0}^{n+1} \frac{(-1)^k \sin a}{(2k)!}(x-a)^{2k} + \sum_{k=0}^{n} \frac{(-1)^k \cos a}{(2k+1)!}(x-a)^{2k+1} \tag{8.1.5}$$

while odd order taylor polynomials are:

$$T_{2n+1}(a) = \sum_{k=0}^{n} \frac{(-1)^k \sin a}{(2k)!}(x-a)^{2k} + \sum_{k=0}^{n} \frac{(-1)^k \cos a}{(2k+1)!}(x-a)^{2k+1} \tag{8.1.6}$$

Taking $a = 0$ we find that:

$$T_{2n+1}(0) = T_{2n+2}(0) = \sum_{k=0}^{n} \frac{(-1)^k}{(2k+1)!}(x-a)^{2k+1} \tag{8.1.7}$$

◀

Example. Consider $f(x) = e^x$, let us find its nth order Taylor polynomial about $a \in \mathbb{R}$. First we evaluate the derivatives:

$$f^{(n)}(a) = e^a \tag{8.1.8}$$

so that:

$$T_n(a) = \sum_{k=1}^{n} \frac{e^a}{k!}(x-a)^k \implies T_n(0) = \sum_{k=1}^{n} \frac{x^k}{k!} \tag{8.1.9}$$

◀

Theorem *(Taylor's theorem)*
Let $f \in \mathcal{C}^{n+1}(I)$ and $a, x \in I$. Then:

$$f(x) = T_n(x) + R_n(x) \tag{8.1.10}$$

where $T_n(x)$ is the nth order Taylor about a, and:

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \tag{8.1.11}$$

for some $c \in (a, x)$ is known as the error term.

Proof. Let's consider:

$$h(t) = f(t) - T_n(t) - A(t-a)^{n+1} \tag{8.1.12}$$

where A is chosen so that $h(x) = 0$. Note also that:

$$f^{(k)}(a) = T_n^{(k)}(a) \implies h^{(k)}(a) = 0, \ k = 0, 1, ..., n \tag{8.1.13}$$

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 $h'', h^{(3)}, ..., h^{(n)}$ on the intervals:

$$[a, c_2], [a, c_3], ..., [a, c_n], \ c_2 > c_3 > ... > c_n > a \tag{8.1.14}$$

we find that there must be some $c \in [a, c_n]$ such that:

$$h^{(n+1)}(c) = f^{(n+1)}(c) - A(n+1)! = 0 \implies A = \frac{f^{(n+1)}(c)}{(n+1)!} \tag{8.1.15}$$

Substituting this into our original expression for $h(t)$, and setting $t = x$ with $h(x) = 0$:

$$f(x) = T_n(x) + \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \tag{8.1.16}$$

as desired.

Example. The Taylor expansion of $f(x) = \log x$ about $a = 0$ is:

$$T_n(x) = \sum_{k=1}^{n} \frac{(-1)^{k+1}x^k}{k} \tag{8.1.17}$$

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:

$$T_2(x) = x - \frac{x^2}{2}, \ x \in I \tag{8.1.18}$$

implying that:

$$|R_2(x)| = \left| \frac{f^{(3)}(c)}{3!} x^3 \right| \le |f^{(3)}(c)| \frac{r^3}{3!} \tag{8.1.19}$$

for some $c \in I$. Now we have that:

$$|f^{(3)}(c)| = \left| \frac{2}{(1+c)^3} \right| \le 2 \tag{8.1.20}$$

so that:

$$|R_2(x)| \le 2 \cdot \frac{(0.02)^3}{3!} = 2.66 \times 10^{-6} \tag{8.1.21}$$

so we know that the error will be negligible up to 5 decimal places. Consequently:

$$\log(1.02) \approx (0.02) - \frac{(0.02)^2}{2} \approx 0.01980 (5 \text{ d.p.}) \tag{8.1.22}$$

◀

Example. The fourth order Taylor polynomial $T_4(x)$ at π for $f(x) = \cos x$ is:

$$T_4(x) = \sum_{k=0}^{4} \frac{(-1)^{k+1}}{(2k)!} (x - \pi)^{2k} = -1 + \frac{1}{2} (x - \pi)^2 \tag{8.1.23}$$

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 and upper bound for the remainder $|R_4(x)| = \left| \frac{f^{(5)}(c)}{5!} (x - a)^5 \right|$. Now $|x - a| \le r = \frac{\pi}{4}$, so:

$$|R_4(x)| \le |f^{(5)}(c)| \frac{r^5}{5!} \tag{8.1.24}$$

for some $c \in [3\pi/4, 5\pi/4]$. We also have that:

$$f^{(5)}(x) = -\sin x \implies |f^{(5)}(c)| = |\sin c| \le 1 \tag{8.1.25}$$

and thus:

$$|R_4(x)| \le \frac{(\pi/4)^5}{5!} = 0.0025 < 0.01 \tag{8.1.26}$$

as desired. ◀

■

> **Theorem** *(Taylor series)*
> Let f be a class $\mathcal{C}^\infty$ on an open interval I at points a and x. If $R_n(x) \to 0$ as $n \to \infty$ then
>
> $$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n \tag{8.1.27}$$
>
> 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)*
>
> $$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \ |x| < 1 \tag{8.1.28}$$
>
> $$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}, \ x \in \mathbb{R} \tag{8.1.29}$$
>
> $$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}, \ x \in \mathbb{R} \tag{8.1.30}$$
>
> $$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \ x \in \mathbb{R} \tag{8.1.31}$$
>
> $$\log(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}, \ -1 < x \le 1 \tag{8.1.32}$$

Proof. (a) It can easily be seen that the nth order taylor polynomial of $f(x) = \frac{1}{1-x}$ is:

$$T_n(x) = \sum_{k=0}^{n} x^k \tag{8.1.33}$$

which is a geometric series. It converges to $\frac{1}{1-x}$ only for $|x| \le 1$, as desired.

(b) The nth order taylor polynomial of $f(x) = \sin x$ is:

$$T_n(x) \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \tag{8.1.34}$$

implying that the error term $R_n(x)$ may be expressed as:

$$|R_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \le \frac{x^{n+1}}{(n+1)!} \tag{8.1.35}$$

since $f^{(n+1)}(c) = \pm \sin c$ or $\pm \cos c$. Therefore:

$$|R_n(x)| \le \frac{|x|^{n+1}}{(n+1)!} \to 0 \text{ as } n \to \infty \tag{8.1.36}$$

where we have used the squeeze rule for null sequences. This is true for all x, and the result follows.

(c) The nth order taylor polynomial of $f(x) = \cos x$ is:

$$T_n(x) \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \tag{8.1.37}$$

implying that the error term $R_n(x)$ may be expressed as:

$$|R_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \leq \frac{x^{n+1}}{(n+1)!} \tag{8.1.38}$$

since $f^{(n+1)}(c) = \pm \sin c$ or $\pm \cos c$. Therefore:

$$|R_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \to 0 \text{ as } n \to \infty \tag{8.1.39}$$

where we have used the squeeze rule for null sequences. This is true for all x, and the result follows.

(d) The nth order taylor polynomial of $f(x) = e^x$ is:

$$T_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!} \tag{8.1.40}$$

The error term may be written as:

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} = e^c \frac{x^{n+1}}{(n+1)!} \tag{8.1.41}$$

and since c lies between 0 and x:

$$|R_n(x)| \leq e^x \frac{|x^{n+1}|}{(n+1)!} \to 0 \text{ as } n \to \infty \tag{8.1.42}$$

as desired.

(e) The nth order Taylor polynomial of $f(x) = \log(1+x)$ is:

$$T_n(x) = \sum_{k=1}^{n} \frac{(-1)^{k+1} x^k}{k} \tag{8.1.43}$$

Consequently, we have that for $0 < x \leq 1$ then the error term reads:

$$|R_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| \tag{8.1.44}$$

$$= \left| \frac{n!}{(1+c)^{n+1}} \frac{x^{n+1}}{(n+1)!} \right| \tag{8.1.45}$$

$$= \frac{|x|^{n+1}}{(n+1)(c+1)^{n+1}} \leq \frac{1}{n} \to 0 \text{ as } n \to \infty \tag{8.1.46}$$

where we noted that $|x|^{n+1} \leq 1$ and $1 + c > 1$. Using the sequeeze rule we readily find the desired result.

$\blacksquare$

8.2 Convergence

> **Definition** *(Power series)* Let $a, a_n, x \in \mathbb{R}$ for $n = 0, 1, 2....$ Then a power series at a in x is a series of the form:
> $$\sum_{n=0}^{\infty} a_n(x-a)^n \tag{8.2.1}$$

Lemma. If the power series $\sum\limits_{n=0}^{\infty} a_n x^n$ 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 $\sum\limits_{n=0}^{\infty} a_n x_0^n$ implies that $(a_n x_0^n)$ is a null sequence, and hence there exists some K such that:

$$|a_n| r^n = |a_n x_0^n| \leq K, \ n = 0, 1, 2... \tag{8.2.2}$$

Let $|x| < r$, then clearly:

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} a_n r^n \frac{x^n}{r^n} \implies |a_n x^n| \leq K \frac{|x|^n}{r^n} \tag{8.2.3}$$

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 $\sum\limits_{n=0}^{\infty} a_n x^n$ converges for $|x| < |x_0|$. ∎

> **Theorem** *(Radius of convergence)*
> The power series $\sum\limits_{n=0}^{\infty} a_n(x-a)^n$ 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:

$$E = \left\{ x \in \mathbb{R} : \sum_{n=0}^{\infty} a_n(x-a)^n \text{ converges} \right\} \tag{8.2.4}$$

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 $\sum\limits_{n=0}^{\infty} a_n(x-a)^n$ 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 sastisfy 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 $\sum\limits_{n=0}^{\infty} a_n(x - a)^n$ were to converge then we would find that $\sum\limits_{n=0}^{\infty} a_n x_2^n$ converges, a contradiction.

So here condition (c) is satisfied. ∎

Theorem *(Ratio test)*

Suppose that $\sum\limits_{n=0}^{\infty} a_n(x - a)^n$ is a power series with radius of convergence R.

(a) If $\left|\frac{a_{n+1}}{a_n}\right| \to \infty$ as $n \to \infty$ then $R = 0$.

(b) If $\left|\frac{a_{n+1}}{a_n}\right| \to 0$ as $n \to \infty$ then $R = \infty$.

(c) If $\left|\frac{a_{n+1}}{a_n}\right| \to L$ as $n \to \infty$ then $R = \frac{1}{L}$ provided $L > 0$.

Proof. (a) Suppose that $\left|\frac{a_{n+1}}{a_n}\right| \to \infty$ as $n \to \infty$. If $x \neq a$ then:

$$\frac{|a_{n+1}(x - a)^{n+1}|}{|a_n(x - a)^n|} = \left|\frac{a_{n+1}}{a_n}\right| |x - a| \to \infty \text{ as } n \to \infty \tag{8.2.5}$$

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 $\left|\frac{a_{n+1}}{a_n}\right| \to 0$ as $n \to \infty$. If $x \neq a$ then:

$$\frac{|a_{n+1}(x - a)^{n+1}|}{|a_n(x - a)^n|} = \left|\frac{a_{n+1}}{a_n}\right| |x - a| \to 0 \text{ as } n \to \infty \tag{8.2.6}$$

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 $\left|\frac{a_{n+1}}{a_n}\right| \to L$ as $n \to \infty$. If $x \neq a$ then:

$$\frac{|a_{n+1}(x - a)^{n+1}|}{|a_n(x - a)^n|} = \left|\frac{a_{n+1}}{a_n}\right| |x - a| \to L|x - a| \text{ as } n \to \infty \tag{8.2.7}$$

If $|x - a| > \frac{1}{L}$ then we find that:

$$\frac{|a_{n+1}(x - a)^{n+1}|}{|a_n(x - a)^n|} \to L|x - a| > 1 \text{ as } n \to \infty \tag{8.2.8}$$

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:

$$\frac{|a_{n+1}(x - a)^{n+1}|}{|a_n(x - a)^n|} \to L|x - a| < 1 \text{ as } n \to \infty \tag{8.2.9}$$

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):

$$\sum_{n=1}^{\infty} \frac{(n!)^2}{(2n)!} x^n \tag{8.2.10}$$

We see that $a_n = \frac{(n!)^2}{(2n)!}$, and hence:

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{((n+1)!)^2}{(n!)^2} \tag{8.2.11}$$

$$\frac{(2n)!}{(2n+2)!} = \frac{(n+1)^2}{(2n+1)(2n+2)} \tag{8.2.12}$$

$$= \frac{n+1}{2(2n+1)} \to \frac{1}{4} \text{ as } n \to \infty \tag{8.2.13}$$

Using the ratio test for power series we see that $R = 4$.
Consider the power series (about 0):

$$\sum_{n=1}^{\infty} n^n x^n \tag{8.2.14}$$

We see that $a_n = n^n$, and hence:

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{(n+1)^{n+1}}{n^n} = \left(1 + \frac{1}{n}\right)^n (n+1) \to \infty \text{ as } n \to \infty \tag{8.2.15}$$

Using the ratio test for power series we see that $R = 0$.
Consider the power series (about 0):

$$\sum_{n=1}^{\infty} (n + 2^{-n})(x-1)^n \tag{8.2.16}$$

We see that $a_n = (n + 2^{-n})$, and hence:

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{n+1+2^{-n-1}}{n+2^{-n}} \tag{8.2.17}$$

$$1 + \frac{2^{-n-1} - 2^{-n}}{n+2^{-n}} + \frac{1}{n+2^{-n}} \tag{8.2.18}$$

$$= 1 + \frac{1}{2(n2^n+1)} + \frac{1}{n+2^{-n}} \to 1 \text{ as } n \to \infty \tag{8.2.19}$$

Using the ratio test for power series we see that $R = 1$. ◀

Example. Let's consider the following power series:

$$\sum_{n=0}^{\infty} \frac{\alpha(\alpha - 1)...(\alpha - n + 1)}{n!} x^n \tag{8.2.20}$$

where α is not an integer. Then this is a power series about 0 with coefficients:

$$a_n = \frac{\alpha(\alpha - 1)...(\alpha - n + 1)}{n!} \tag{8.2.21}$$

We see that:

$$\left| \frac{a_{n+1}}{a_n} \right| = \left| \frac{\alpha(\alpha - 1)...(\alpha - n)}{\alpha(\alpha - 1)...(\alpha - n - 1)} \frac{n!}{(n+1)!} \right| \tag{8.2.22}$$

$$= \left| \frac{\alpha - n}{n + 1} \right| \to 1 \text{ as } n \to \infty \tag{8.2.23}$$

◀

8.3 The combination rules

Proposition *(Combination rules)*
Let f, g be a function represented by a Taylor series at a:

$$f(x) = \sum_{n=0}^{\infty} a_n (x - a)^n, \ |x - a| < R \tag{8.3.1}$$

$$g(x) = \sum_{n=0}^{\infty} b_n (x - a)^n, \ |x - a| < R' \tag{8.3.2}$$

then for $r = \min\{R, R'\}$ and $\lambda \in \mathbb{R}$:

$$(f + g)(x) = \sum_{n=0}^{\infty} (a_n + b_n)(x - a)^n, \ |x - a| < r \tag{8.3.3}$$

$$\lambda f(x) = \sum_{n=0}^{\infty} (a_n + b_n)(x - a)^n, \ |x - a| < R \tag{8.3.4}$$

$$\tag{8.3.5}$$

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:

$$f(x) = \cosh x = \frac{e^x + e^{-x}}{2} = \frac{1}{2} \sum_{n=0}^{\infty} (1 + (-1)^n) \frac{x^n}{n} = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \tag{8.3.6}$$

with infinite radius of convergence.

◀

Proposition *(Product rule)*

Let f, g be a function represented by a Taylor series at a:

$$f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n, \ |x-a| < R \tag{8.3.7}$$

$$g(x) = \sum_{n=0}^{\infty} b_n(x-a)^n, \ |x-a| < R' \tag{8.3.8}$$

then for $r = \min\{R, R'\}$:

$$(fg)(x) = \sum_{n=0}^{\infty} c_n(x-a)^n, \ |x-a| < r \tag{8.3.9}$$

where

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} \tag{8.3.10}$$

Example. Let's consider the following function:

$$f(x) = (1+x)\log(1+x) \tag{8.3.11}$$

The taylor series for $\log(1+x)$ at 0 is:

$$\log(1+x) = \sum_{n=0}^{\infty} \frac{(-1)^{n+1} x^n}{n}, \ -1 < x \leq 1 \tag{8.3.12}$$

implying that:

$$(1+x)\log(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}(x^n + x^{n+1}) \tag{8.3.13}$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^{n+1} \tag{8.3.14}$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n + \sum_{n=2}^{\infty} \frac{(-1)^n}{n-1} x^n \tag{8.3.15}$$

$$= x + \sum_{n=2}^{\infty} \left(\frac{(-1)^{n+1}}{n} + \frac{(-1)^n}{n-1} \right) x^n \tag{8.3.16}$$

$$= x + \sum_{n=2}^{\infty} (-1)^n \frac{x^n}{n(n-1)} \tag{8.3.17}$$

for $-1 < x \leq 1$.

Let's consider the following function:

$$f(x) = \frac{1}{(1-x)^2} \tag{8.3.18}$$

The taylor series for $\frac{1}{1-x}$ is:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \ |x| < 1 \tag{8.3.19}$$

Therefore:

$$\frac{1}{(1-x)^2} = \sum_{n=0}^{\infty} c_n x^{2n}, \ |x| < 1 \tag{8.3.20}$$

where

$$c_n = \sum_{k=0}^{n} 1 = n+1 \implies \frac{1}{(1-x)^2} = \sum_{n=0}^{\infty} (n+1) x^n, \ |x| < 1 \tag{8.3.21}$$

Consider the function:

$$f(x) = \frac{1}{1+2x^2} \tag{8.3.22}$$

The taylor series for $\frac{1}{1+x}$ is:

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n, \ |x| < 1 \tag{8.3.23}$$

Consequently:

$$\frac{1}{1+2x^2} = \sum_{n=0}^{\infty} (-1)^n (2x^2)^n = \sum_{n=0}^{\infty} (-2)^n x^{2n}, \ |2x^2| < 1 \tag{8.3.24}$$

so the range of validity for this expansion is $2x^2 < 1 \implies |x| \leq \frac{1}{\sqrt{2}}$.
Consider the function

$$f(x) = \frac{e^x}{(1-x)^2} \tag{8.3.25}$$

The taylor series reads:

$$\sum_{n=0}^{\infty} c_n x^n, \ |x| < 1 \tag{8.3.26}$$

where

$$c_n = \sum_{k=0}^{n} \frac{k+1}{(n-k)!} \tag{8.3.27}$$

so:

$$c_0 = 1, \ c_1 = 1 + 2 = 3, \ c_2 = \frac{1}{2} + 2 + 3 = \frac{11}{2} \tag{8.3.28}$$

◀

Theorem *(Differentiation rule)* The following taylor series:

$$f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n, \tag{8.3.29}$$

$$g(x) = \sum_{n=1}^{\infty} n a_n (x-a)^{n-1} \tag{8.3.30}$$

have the same radius of convergence, and $f'(x) = g(x)$ for $|x - a| < R$.

Theorem *(Integration rule)* The following taylor series:

$$f(x) = \sum_{n=0}^{\infty} a_n (x - a)^n, \tag{8.3.31}$$

$$F(x) = \sum_{n=0}^{\infty} \frac{a_n}{n+1} (x - a)^{n+1} \tag{8.3.32}$$

have the same radius of convergence R, and if $R > 0$ then:

$$\int f(x) dx = F(x), \ |x - a| < R \tag{8.3.33}$$

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:

$$f'(x) = \frac{1}{1 - x^2} = \sum_{n=0}^{\infty} x^{2n}, \ |x| < 1 \tag{8.3.34}$$

Consequently:

$$f(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{2n+1}, \ |x| < 1 \tag{8.3.35}$$

Let's find the taylor series of e^{-x^2}:

$$e^{-x^2} = \sum_{n=0}^{\infty} \frac{(-x^2)^n}{n!} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} x^{2n}, \ x \in \mathbb{R} \tag{8.3.36}$$

implying that:

$$\int_0^1 e^{-x^2} = \sum_{n=0}^{\infty} \int_0^1 frac(-1)^n n! x^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{1}{2n+1} \tag{8.3.37}$$

We define the error function as:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n!} \frac{1}{2n+1} \implies \int_0^1 e^{-x^2} = \frac{\sqrt{\pi}}{2} \text{erf}(1) \tag{8.3.38}$$

Finally, let's find the taylor series of $f(x) = \log(1 + x)$. We know that:

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n, \ |x| < 1 \tag{8.3.39}$$

implying that:

$$\log(1+x) = \int \frac{1}{1+x} dx \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} x^{n+1} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n \tag{8.3.40}$$

for $|x| < 1$. ◀

Theorem *(General binomial theorem)* Let $\alpha \in \mathbb{R}$, then:

$$(1+x)^{\alpha} = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n, \ |x| < 1 \tag{8.3.41}$$

Proof. Let us define:

$$f(x) = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n, \ g(x) = f(x)(1+x)^{-\alpha} \tag{8.3.42}$$

Differentiating g we find that:

$$g'(x) = f'(x)(1+x)^{-\alpha} - \alpha f(x)(1+x)^{-\alpha-1} \tag{8.3.43}$$

$$= (1+x)^{-\alpha-1}((1+x)f'(x) - \alpha f(x)) \tag{8.3.44}$$

$$= (1+x)^{-\alpha-1}\left((1+x) \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^{n-1} - \alpha \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \right) \tag{8.3.45}$$

We can simplify the expression in brackets:

$$(1+x) \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^{n-1} - \alpha \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \tag{8.3.46}$$

$$= \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^{n-1} + \sum_{n=1}^{\infty} \binom{\alpha}{n} nx^n - \alpha \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \tag{8.3.47}$$

$$= \sum_{n=0}^{\infty} \binom{\alpha}{n+1} (n+1)x^n + \sum_{n=1}^{\infty} \binom{\alpha}{n} (n-\alpha)x^n \tag{8.3.48}$$

We find that:

$$\binom{\alpha}{n+1}(n+1) = \frac{\alpha!}{(n+1)!(\alpha-n-1)!}(n+1) = \frac{\alpha!}{(n)!(\alpha-n-1)!} \tag{8.3.49}$$

and

$$\binom{\alpha}{n}(n-\alpha) = \frac{\alpha!}{n!(\alpha-n)!}(n-\alpha) = -\frac{\alpha!}{n!(\alpha-n-1)!} \tag{8.3.50}$$

implying that:

$$\binom{\alpha}{n+1}(n+1) + \binom{\alpha}{n}(n-\alpha) = \frac{\alpha!}{(n)!(\alpha-n-1)!} - \frac{\alpha!}{n!(\alpha-n-1)!} = 0 \tag{8.3.51}$$

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

GT1 Symmetry and groups

9.1   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:
>
> $$f : \mathbb{R}^2 \to \mathbb{R}^2 \qquad\qquad (9.1.1)$$
> $$\mathcal{F} \mapsto \mathcal{F} \qquad\qquad (9.1.2)$$

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.

(a)   (b)
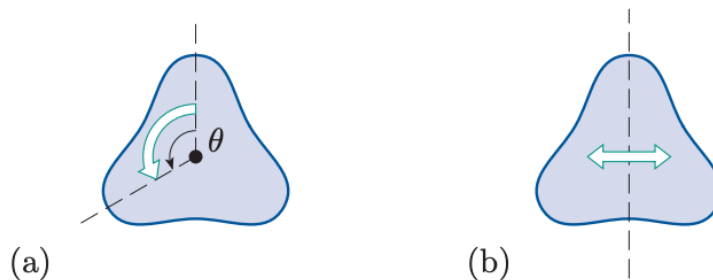
The axes of symmetry of a figure all pass through a point, called the centre of the bounded figure, is also the center of all rotational symmetries.

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.

| initial position | identity symmetry | rotation through $\pi/2$ | rotation through π | rotation through $3\pi/2$ |
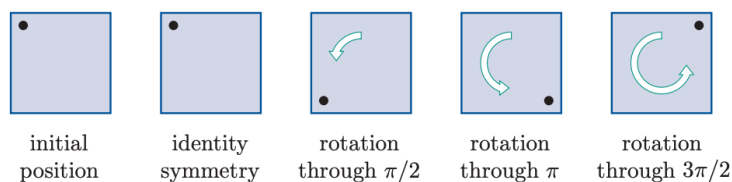
Figure 9.1. Rotational symmetry of a square

The square also has 4 reflection symmetries, 1 horizontal, 1 vertical and 2 diagonal:

| initial position | reflection in vertical | reflection in diagonal | reflection in horizontal | reflection in diagonal |
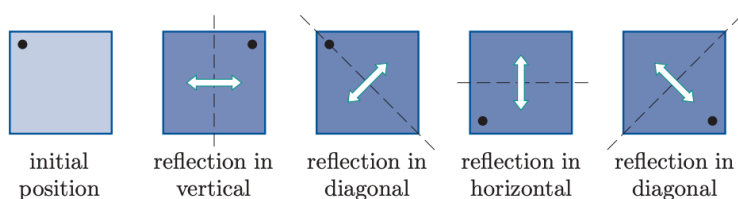
Figure 9.2. Reflection symmetries of a square

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:

Then clearly, if we apply $t \circ u$ as shown below, it is equivalent to applying c, so $t \circ u = c \in S(\mathcal{F})$.
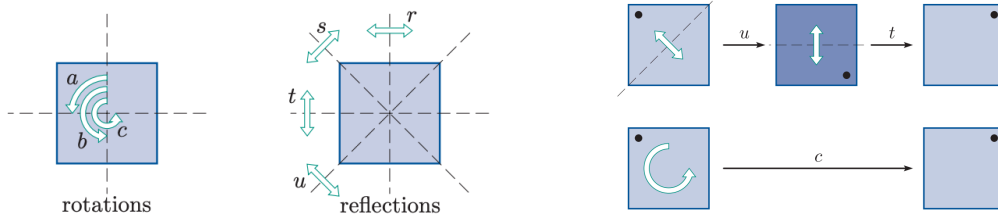
Figure 9.3. Symmetries of a square

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:

$$S^+(\square) = \{e, a, b, c\} \tag{9.1.3}$$

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 $S(\square)$: ∎
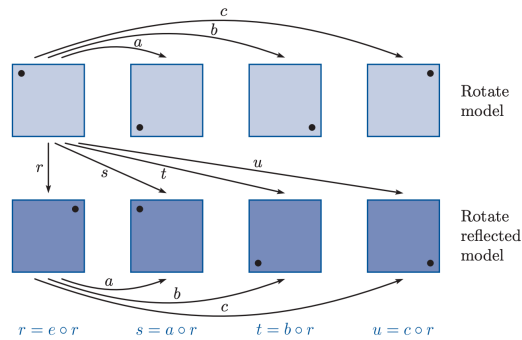
Figure 9.4. Deriving indirect symmetries from direct symmetries

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.

9.2 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
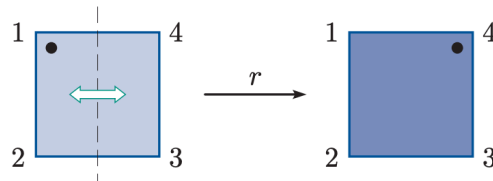
Figure 9.5. Two line symmetry notation

can then represent this transformation as:

$$r \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \tag{9.2.1}$$

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 $1, 2, 3...n$ is:
>
> $$f \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ f(1) & f(2) & f(3) & \dots & f(n) \end{pmatrix} \tag{9.2.2}$$
>
> The inverse f^{-1} is then clearly represented by:
>
> $$f^{-1} \leftrightarrow \begin{pmatrix} f(1) & f(2) & f(3) & \dots & f(n) \\ 1 & 2 & 3 & \dots & n \end{pmatrix} \tag{9.2.3}$$

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 xth row and yth column.

The Cayley table for the symmetries of a rectangle is shown below:

$\circ$	e	a	r	s
e	e	a	r	s
a	a	e	s	r
r	r	s	e	a
s	s	r	a	e
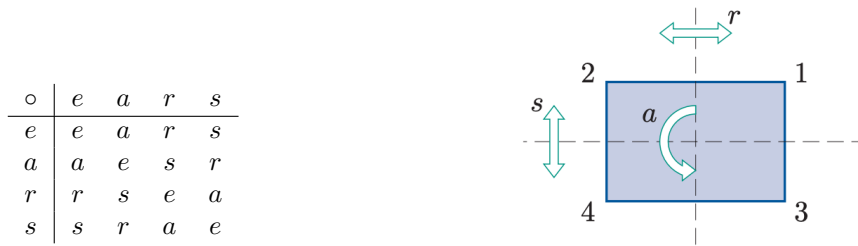
Table 9.1. Rectangle symmetries and its Cayley table

Since we chose to list the direct symmetries and indirect symmetries separately, we formed four blocks each containing only direct or indirect symmetries.

The same occurs with the Cayley table for a square:

$\circ$	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

$\circ$	direct	indirect
direct	direct	indirect
indirect	indirect	direct
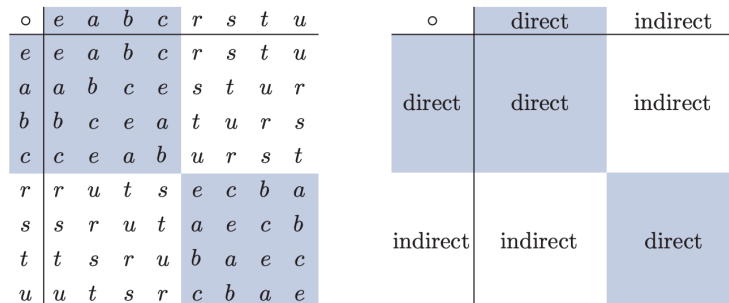
Figure 9.6. Block patterns in Cayley table

9.3   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:
>
> $$* : G \times G \longrightarrow G \tag{9.3.1}$$
> $$(f, g) \longmapsto h \tag{9.3.2}$$
>
> 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:

$$(a, b) * (c, d) = (ac, ad + b) \in \mathbb{R} \tag{9.3.3}$$

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:

$$((a, b) * (c, d)) * (e, f) = (ac, ad + b) * (e, f) = (ace, acf + ad + b) \tag{9.3.4}$$

and

$$(a, b) * ((c, d) * (e, f)) = (a, b) * (ce, cf + d) = (ace, acf + ad + b) \tag{9.3.5}$$

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 $\forall (a, b) \in X$:

$$(a, b) * (e_1, e_2) = (ae_1, ae_2 + b) = (a, b) \tag{9.3.6}$$
$$(e_1, e_2) * (a, b) = (ae_1, be_1 + e_2) = (a, b) \tag{9.3.7}$$

which upon equating terms component-wise gives:

$$\begin{cases} ae_1 = a \\ ae_2 + b = b \\ be_1 + e_2 = b \end{cases} \implies \begin{cases} e_1 = 1 \\ e_2 = 0 \end{cases} \tag{9.3.8}$$

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, $\forall (a,b) \in X$:

$$(a,b) * (1,0) = (a \cdot 1, a \cdot 0 + b) = (a,b) \tag{9.3.9}$$

$$(1,0) * (a,b) = (1 \cdot a, 1 \cdot b + 0) = (a,b) \tag{9.3.10}$$

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:

$$(a,b) * (c,d) = (ac, ad + b) = (1,0) \tag{9.3.11}$$

$$(c,d) * (a,b) = (ac, bc + d) = (1,0) \tag{9.3.12}$$

which implies:

$$\begin{cases} ac = 1 \\ ad + b = 0 \\ bc + d = 0 \end{cases} \implies \begin{cases} c = \frac{1}{a} \\ d = -\frac{b}{a} \end{cases} \tag{9.3.13}$$

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:

$$(a,b) * \left(\frac{1}{a}, -\frac{b}{a}\right) = \left(a\frac{1}{a}, -\frac{b}{a}a + b\right) = (1,0) \tag{9.3.14}$$

$$(c,d) * \left(\frac{1}{a}, -\frac{b}{a}\right) = \left(a\frac{1}{a}, b\frac{1}{a} - \frac{b}{a}\right) = (1,0) \tag{9.3.15}$$

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. ∎
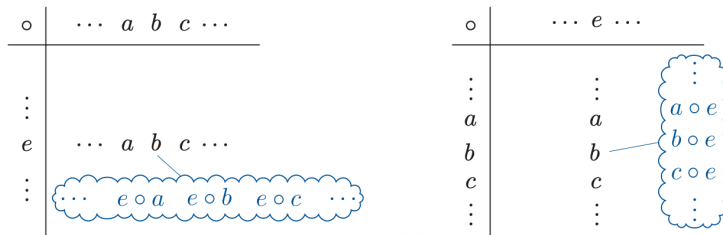
Figure 9.7. Row and column of the identity element e

> **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:

$$(\mathbb{Z}, +), \ (\mathbb{Q}, +), \ (\mathbb{R}, +), \ (\mathbb{C}, +), \ (\mathbb{Q}^*, \times), \ (\mathbb{R}^*, \times), \ (\mathbb{C}^*, \times) \tag{9.3.16}$$

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 $\times_n$:

$$(c \times_n d) \times_n (a \times_n b) = (c \times_n a) \times_n (d \times_n b) = 1 \times_n 1 = 1 \tag{9.3.17}$$

and similarly:

$$(a \times_n b) \times_n (c \times_n d) = 1 \tag{9.3.18}$$

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 $\forall a \in \mathbb{Z}_n$:

$$a \times_n 1 = 1 \times_n a = a \tag{9.3.19}$$

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:

$$a \times_n b = b \times_N a = 1 \tag{9.3.20}$$

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.

9.4 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 $(g * h) * (h^{-1} * g^{-1}) = e$:

$$(g * h) * (h^{-1} * g^{-1}) = g * (h * h^{-1}) * g^{-1} \tag{9.4.1}$$
$$= g * e * g^{-1} \tag{9.4.2}$$
$$= g * g^{-1} \tag{9.4.3}$$
$$= e \tag{9.4.4}$$

  We now show that $(h^{-1} * g^{-1}) * (g * h) = e$:

$$(h^{-1} * g^{-1}) * (g * h) = h^{-1} * (g^{-1} * g) * h \tag{9.4.5}$$
$$= h^{-1} * e * h \tag{9.4.6}$$
$$= h^{-1} * h \tag{9.4.7}$$
$$= e \tag{9.4.8}$$

  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) $(g^m)^n = \underbrace{\underbrace{x * x * \ldots * x}_{m \text{ times}} * \ldots * \underbrace{x * x * \ldots * x}_{m \text{ times}}}_{n \text{ times}} = \underbrace{g * g * g \ldots * g * g}_{mn \text{ times}} = g^{mn}$

■

9.5 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:
>
> $$f : \mathbb{R}^3 \to \mathbb{R}^3 \qquad (9.5.1)$$
> $$\mathcal{F} \mapsto \mathcal{F} \qquad (9.5.2)$$
>
> 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:
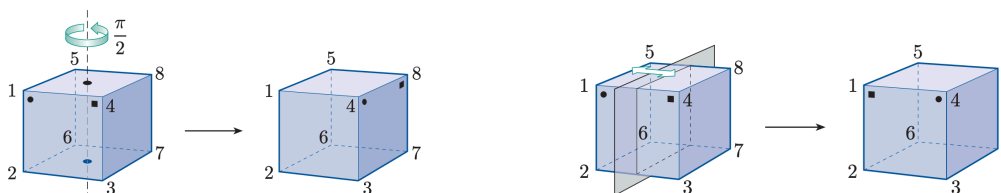
(a) Rotation of a cube through $\pi/2$ about its vertical axis

(b) Reflection of a cube in the vertical plane

Figure 9.8.

Using two-line symbols we can write:

$$g \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 3 & 7 & 8 & 1 & 2 & 6 & 5 \end{pmatrix} \qquad (9.5.3)$$

Similarly a reflection in the vertical plane is represented by:

$$f \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 3 & 2 & 1 & 8 & 7 & 6 & 5 \end{pmatrix} \tag{9.5.4}$$

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.

$$f \circ g \leftrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 6 & 5 & 4 & 3 & 7 & 8 \end{pmatrix} \tag{9.5.5}$$

which is a reflection in the diagonal plane as shown below:
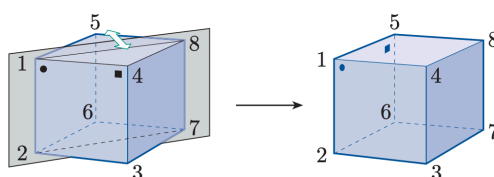
Figure 9.9. Reflection in diagonal plane of a cube

> **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.
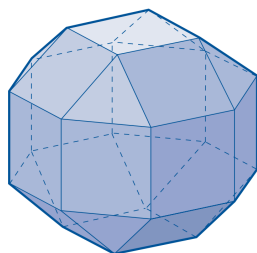
Figure 9.10. Rhombicuboctahedron

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 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.
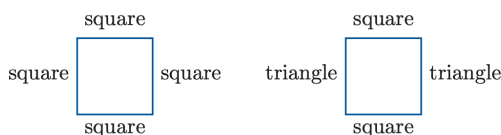
Figure 9.11. Square faces of a rhombicuboctahedron

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.

Had we chosen 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:

$$e = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} \tag{9.5.6}$$

For each base, there is a rotational symmetry about a fixed axis through the opposite vertex.

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:
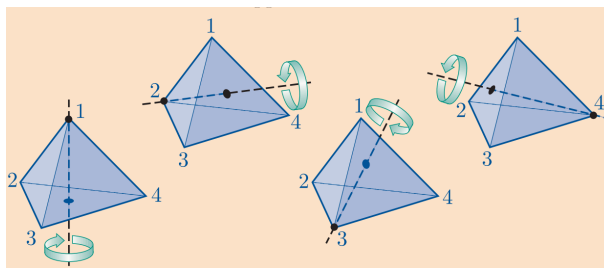
Figure 9.12. Rotational symmetries of the tetrahedron about axes through bases
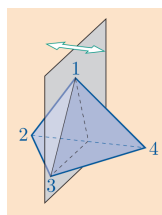
Figure 9.13. Reflectional symmetry of tetrahedron

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.

9.6 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:

$$r : \mathbb{C} \longrightarrow \mathbb{C} \tag{9.6.1}$$

$$z \mapsto z \cdot e^{\frac{2i\pi}{n}} \tag{9.6.2}$$

which is a rotation about the center of the polygon by $\frac{2\pi}{n}$. This is a symmetry, since it preserves distances $\forall z, w \in \mathbb{C}$:

$$|r(z) - r(w)| = |(z - w) \cdot e^{\frac{2i\pi}{n}}| = |z - w| \tag{9.6.3}$$

We define the reflection in the x-axis by the following map:

$$s : \mathbb{C} \longrightarrow \mathbb{C} \tag{9.6.4}$$

$$z \mapsto \bar{z} \tag{9.6.5}$$

which is again another symmetry since $\forall z, w \in \mathbb{C}$:

$$|s(z) - s(w)| = |\bar{z} - \bar{w}| = \sqrt{(\bar{z} - \bar{w})(z - w)} = |z - w| \tag{9.6.6}$$

We will show that:

$$D_{2n} = \{\underbrace{e, r, r^2, r^3, .., r^{n-1}}_{\text{rotations}}, \underbrace{s, rs, r^2s, ..., r^{n-1}s}_{\text{reflections}}\} \tag{9.6.7}$$

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 $e^{2\pi i k}n$ for some $0 \geq k < n$. however, r^k also maps $1 \mapsto e^{2\pi i k}n$ 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:

$$d((f \circ g)(z), (f \circ g)(w)) = |(f \circ g)(z) - (f \circ g)(w)| \tag{9.6.8}$$
$$= |f(g(z)) - f(g(w))| \tag{9.6.9}$$
$$= |g(z) - g(w)| = |z - w| \tag{9.6.10}$$
$$= d(z, w) \tag{9.6.11}$$

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 $e^{2\pi i k}n$ with $k \leq m - 1$ for some $0 < m < n$ have been mapped to themselves by g. Then, $e^{2\pi i m}n$ 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, $g(e^{2\pi i m}n) = e^{2\pi i(m-2)}n$, 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 reflections[1], 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 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 $0 \leq i < n$

---

[1] they are also all distinct. Clearly, if $r^m = r^k$, then 1 gets mapped to $e^{\frac{2\pi i m}{n}}$ and $e^{\frac{2\pi i k}{n}}$, 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

(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 $e^{\frac{2\pi i m}{n}}$. s sends it to $e^{\frac{-2\pi i m}{n}}$, followed by r^k which sends it to $e^{\frac{2\pi i(k-m)}{n}}$. Similarly, sr^{-k} maps $e^{\frac{2\pi i m}{n}}$ 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 $e^{\frac{2\pi i m}{n}}$ to the same vertex, and are equivalent. $sr^{-k} = r^k s \implies r^{-k} = sr^k s$.

(ii) Since $e, r, r^2, ..., r^{n-1}$ are all distinct, and $r^n = e$, it follows that $\mathrm{ord}(r) = n$, 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 $\mathrm{ord}(r^i s) = 2$.

(iii) We have that $rs = sr^{-1}$, and suppose $rs = sr$, then $sr^{-1} = sr \implies r^2 = e$, which is a contradiction since $\mathrm{ord}(r) = n > 2$ by assumption.

(iv) Consider the rotation r^i for some $0 \leq i < n$. Then, it may be written as $r^i = (r^{i+1}s) \circ (sr)$

■

GT2 Subgroups and isomorphisms

10.1 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) $x * y \in H$
> (SG2) $e_G \in H$
> (SG3) $x_G^{-1} \in H$

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:

$$(1, b) * (1, d) = (1, d + b) \in A \qquad (10.1.1)$$

  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:
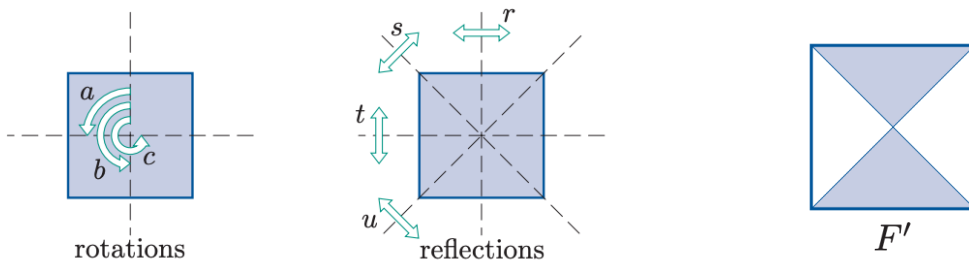
Figure 10.1. Symmetries of $\square$, and the modified square $\mathcal{F}'$ with symmetries $\{e, b, r, t\}$

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 $\mathcal{F}F$ 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).
>
> ◄
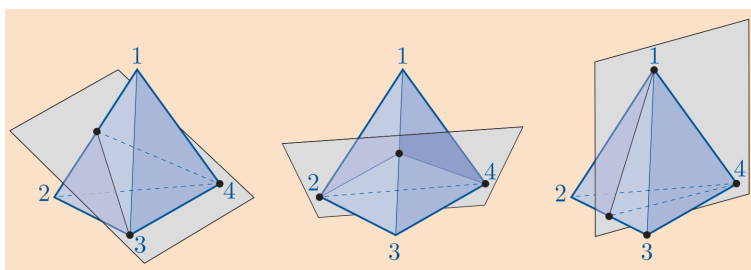
Figure 10.2. Indirect symmetries of the tetrahedron fixing a vertex

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 $..., x^{-3}, x^{-2}, x^{-1}, x^0, x, x^2, x^3...$ 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:

$$x^s = x^t \implies x^s * (x^s)^{-1} = x^t * (x^s)^{-1} \implies e = x^{t-s} \tag{10.1.2}$$

  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 $n \in \mathbb{Z}$:

$$x^n = e \implies (x^n)^{-1} * x^n = (x^n)^{-1} \implies (x^{-1})^n = e \tag{10.1.3}$$

  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 $(x^{-1})^n$, 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:
> $$e, x, x^2, ..., x^{n-1} \tag{10.1.4}$$
> 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 $e, x, x^2, ..., x^{n-1}$ are not distinct, so that $x^u = x^t$ for some $0 \le t < u \le n - 1$. 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:

$$x^{kn} = (x^n)^k = e^k = e \tag{10.1.5}$$

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 $0 \le t < u \le n - 1$, $x^u = x^t$. 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:
>
> $$1 +_6 1 +_6 1 +_6 1 +_6 1 +_6 1 = 0 \tag{10.1.6}$$
>
> so 1 has order 6, and $1^{-1} = 5$ has order 6 as well.
> For the element 2:
> $$2 +_6 2 +_6 2 = 0 \tag{10.1.7}$$
>
> 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. ◄

10.2 Cyclic groups and subgroups

We can consider the powers of a group element as forming a cycle that repeats itself after n operations.

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}.
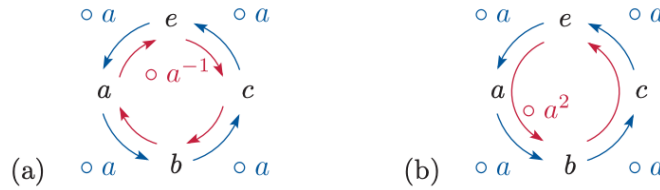
Figure 10.3. Cycle of powers of a in $S(\square)$

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:

$$g * h = x^{s+t} \in \langle x \rangle \tag{10.2.1}$$

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 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:

$$f * g = a^s * a^t = a^{s+t} = a^{t+s} = a^t * a^s = g * f \tag{10.2.2}$$

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:

$$a^r = a^{k-qm} = a^k * (a^m)^{-q} = h * (a^m)^{-q} \tag{10.2.3}$$

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:

$$a^k = (a^m)^q \tag{10.2.4}$$

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:
>
> $$\langle 1 \rangle = \{1\} \tag{10.2.5}$$
> $$\langle 2 \rangle = \{2, 4, 3, 1\} = \mathbb{Z}_5^* \tag{10.2.6}$$
> $$\langle 3 \rangle = \{3, 4, 2, 1\} = \mathbb{Z}_5^* \tag{10.2.7}$$
> $$\langle 4 \rangle = \{4, 1\} \tag{10.2.8}$$

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:

$$\{1\}, \{1, 4\}, \mathbb{Z}_5^* \tag{10.2.9}$$

◀

10.3 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 $d = \mathrm{GCD}(m, n)$.

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 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 $d = \mathrm{GCD}(m, n)$. 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 $\frac{n}{d} \leq \mathrm{ord}(m) \leq \frac{n}{d}$.

To show that $\mathrm{ord}(m) \leq \frac{n}{d}$, consider the cycle of multiples of 1:
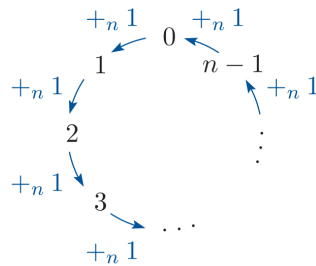
Figure 10.4. Cycle of 1 in $\mathbb{Z}_n, +_n)$

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, $\mathrm{ord}(m)$ is indeed at most $\frac{n}{d}$.

Now let us show that $\frac{n}{d} \leq \mathrm{ord}(m)$ by contradiction. Indeed, suppose that $1 \leq \mathrm{ord}(m) = r < \frac{n}{d}$. 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 $k \in \mathbb{N}$:

$$rm = kn \implies r\frac{m}{d} = k\frac{n}{d} \tag{10.3.1}$$

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:

$$rd = \frac{kd}{m}n \tag{10.3.2}$$

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 $1 \le \text{ord}(m) = r < \frac{n}{d}$. Thus the order or m cannot be less than $\frac{n}{d}$, and we may conclude that:

$$\text{ord}(m) = \frac{n}{d} \tag{10.3.3}$$

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 $m \in \langle d \rangle \implies \langle m \rangle \le \langle d \rangle$ 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. ∎

10.4 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
e	e	a	b	c
a	a	b	c	e
b	b	c	e	a
c	c	e	a	b

$(S^+(\square), \circ)$

$+_4$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

$(\mathbb{Z}_4, +_4)$

Figure 10.5. 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 10.6. 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
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

original table

$\longrightarrow$ swap columns 3, 4

$\times_5$	1	2	4	3
1	1	2	4	3
2	2	4	3	1
3	3	1	2	4
4	4	3	1	2

intermediate table

$\longrightarrow$ swap rows 3, 4

$\times_5$	1	2	4	3
1	1	2	4	3
2	2	4	3	1
4	4	3	1	2
3	3	1	2	4

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:

$$\phi : (S^+(\square), \circ) \longrightarrow \mathbb{Z}_4, +_4) \tag{10.4.1}$$

$$\{e, a, b, c\} \mapsto \{1, 2, 3, 4\} \tag{10.4.2}$$

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.

$$
\begin{array}{c|cccc}
\circ & e & a & b & c \\
\hline
e & e & a & b & c \\
a & a & b & c & e \\
b & b & c & e & a \\
c & c & e & a & b
\end{array}
\qquad \longrightarrow \qquad
\begin{array}{c|cccc}
 & 2 & 3 & 4 & 1 \\
\hline
2 & 2 & 3 & 4 & 1 \\
3 & 3 & 4 & 1 & 2 \\
4 & 4 & 1 & 2 & 3 \\
1 & 1 & 2 & 3 & 4
\end{array}
$$

$$(S^+(\Box), \circ)$$

Figure 10.7. 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 $\phi(x * y) = \phi(x) * \phi(y), \ \forall x, y \in G$.

$$
\begin{array}{c|ccc}
\circ & \cdots & y & \cdots \\
\hline
\vdots & & \vdots & \\
x & \cdots & x \circ y & \cdots \\
\vdots & & \vdots &
\end{array}
\qquad \longrightarrow \qquad
\begin{array}{c|ccc}
* & \cdots & \phi(y) & \cdots \\
\hline
\vdots & & \vdots & \\
\phi(x) & \cdots & \phi(x \circ y) & \cdots \\
\vdots & & \vdots &
\end{array}
$$

$$(G, \circ) \qquad\qquad\qquad (H, *)$$

Definition 15.18 *(Isomorphic groups)*
Two groups $(G, \circ), (H, *)$ are isomorphic if there exists a mapping $\phi : G \longrightarrow H$ called **isomorphism** such that:
   (i) ϕ is bijective
   (ii) $\forall x, y \in G \ \phi(x \circ y) = \phi(x) * \phi(y)$
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:

$$\psi : G \longrightarrow \mathbb{Z} \tag{10.4.3}$$
$$2^k \mapsto k \tag{10.4.4}$$

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 $\text{Im}(\phi) = \{k \in \mathbb{Z} : 2^k \in \mathbb{G}\} = \mathbb{Z}$ and thus surjectivity is satisfied.
Finally, for $2^j, 2^k \in G$:

$$\phi(2^j \times 2^k) = \phi(2^{j+k}) = j + k = \phi(2^j) + \phi(2^k) \tag{10.4.5}$$

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 $\phi : (G, \circ) \longrightarrow (H, *)$ 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 $\phi : a^k \mapsto b^k$, $k = 0, 1, 2...n - 1$ for finite ordered groups and $k \in \mathbb{Z}$ for infinite ordered groups.

Proof. It is trivial to see that ϕ is bijective. Also $\forall a^j, a^k \in G$:

$$\phi(a^j \circ a^k) = \phi(a^{j+k}) = b^{j+k} = b^j * b^k = \phi(a^j) * \phi(a^k) \tag{10.4.6}$$

as required.

∎

10.5   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:

$$\phi : \mathbb{Z}_4, +_4) \longrightarrow (\mathbb{Z}_5, \times_5) \tag{10.5.1}$$
$$0 \mapsto 1 \tag{10.5.2}$$
$$1 \mapsto 2 \tag{10.5.3}$$
$$2 \mapsto 4 \tag{10.5.4}$$
$$3 \mapsto 3 \tag{10.5.5}$$

◄

> **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 $(U_8, \times_8)$:

$\circ$	e	a	r	s
e	e	a	r	s
a	a	e	s	r
r	r	s	e	a
s	s	r	a	e

$$(S(\square), \circ)$$

$\times_8$	1	3	5	7
1	1	3	5	7
3	3	1	7	5
5	5	7	1	3
7	7	5	3	1

$$(U_8, \times_8)$$

We can see that they both share the same structure, and therefore are isomorphic to each other.

10.6 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:
> $$G \times H = \{(g, h) : g \in G, h \in H\} \tag{10.6.1}$$
>
> is the **cartesian product** of G, H, equipped with the operation
>
> $$(g_1, h_1) * (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2) \tag{10.6.2}$$
>
> 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 $(g_1, h_1), (g_2, h_2) \in G \times H$:

$$(g_1, h_1) * (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2) \in G \times H \tag{10.6.3}$$

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 $\forall (g, h) \in G \times H$:

$$(e_G, e_H) * (g, h) = (e_G * g, e_H *_H h) = (g, h) \tag{10.6.4}$$

and

$$(g, h) * (e_G, e_H) = (g *_G e_G, h *_H e_H) = (g, h) \tag{10.6.5}$$

as desired.

Finally, the inverse of (g, h) is (g^{-1}, h^{-1}), since:

$$(g, h) * (g^{-1}, h^{-1}) = (g * g^{-1}, h *_H h^{-1}) = (e_G, e_H) \tag{10.6.6}$$

and

$$(g^{-1}, h^{-1}) * (g, h) = (g^{-1} *_G g, h^{-1} *_H h) = (e_G, e_H) \tag{10.6.7}$$

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)*
> $C_n \times C_m \cong C_{nm}$ *iff* $\mathrm{GCD}(m, n) = 1$.

Proof. Suppose that $\mathrm{GCD}(m, n) = 1$, and let $C_n = \langle a \rangle$, $C_m = \langle b \rangle$, and $\mathrm{ord}((a, b)) = k$ so that:

$$(a, b)^k = (a^k, b^k) = e \tag{10.6.8}$$

which can only be the case if k is a common multiple of $m = \mathrm{ord}(a)$ and $n = \mathrm{ord}(b)$. Since k must be the minimum such integer, we require $\mathrm{LCM}(m, n) = k$. Using the well known relation that:

$$\mathrm{LCM}(m, n) = \frac{n \cdot m}{\mathrm{GCD}(m, n)} = n \cdot m = k \tag{10.6.9}$$

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 $|\langle (a, b) \rangle| = \mathrm{ord}((a, b)) = n \cdot m$. However, $|C_n \times C_n| = n \cdot m$ 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 $\mathrm{GCD}(m, n) \neq 1$, 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:

$$\phi : H \times F \to G \tag{10.6.10}$$
$$(h, g) \mapsto h * g \tag{10.6.11}$$

is an isomorphism. Indeed,

$$\phi((h_1, f_1) * (h_2, f_2)) = \phi(h_1 * h_2, f_1 * f_2) \tag{10.6.12}$$
$$= h_1 * h_2 * f_1 * f_2 \tag{10.6.13}$$
$$= (h_1 * f_1) * (h_2 * f_2) \tag{10.6.14}$$
$$= \phi(h_1, f_1) * \phi(h_2, f_2) \tag{10.6.15}$$

where we used the commutativity of h, f in the third line.

Also, ϕ is injective. Indeed:

$$\phi(h_1, f_1) = \phi(h_2, f_2) \implies h_1 * f_1 = h_2 * f_2 \implies h_1 * h_2^{-1} = f_2 * f_1^{-1} \tag{10.6.16}$$

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. ∎

GT3 Permutations

11.1 Permutations

> **Definition 16.1** *(Permutation)*
> A **permutation** of a finite set S is a bijective map $\sigma : S \to S$. The set of all permutations of this set is denoted $\mathrm{Sym}(S)$.

We use the typical two-line notation to denote a permutation:

$$\sigma \leftrightarrow \begin{pmatrix} 1\ 2\ 3\ 4 \\ 4\ 1\ 2\ 3 \end{pmatrix} \tag{11.1.1}$$

A more convenient and effective notation is the **cycle form**. Indeed starting from 1 and looking at where each element gets mapped we find:

$$1 \xrightarrow{\ \sigma\ } 4 \xrightarrow{\ \sigma\ } 3 \xrightarrow{\ \sigma\ } 2 \xrightarrow{\ \sigma\ } 1 \tag{11.1.2}$$

or more concisely as:

$$\sigma = (1\ 4\ 3\ 2) \tag{11.1.3}$$

However it is not always possible to write a permutation in one single cycle. Indeed some permutations map only some symbols in each cycle:

$$\sigma_g = \begin{pmatrix} 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8 \\ 4\ 6\ 8\ 3\ 1\ 2\ 7\ 5 \end{pmatrix} \tag{11.1.4}$$

then we have three **disjoint** cycles (disjoint because every member appears only in one cycle):
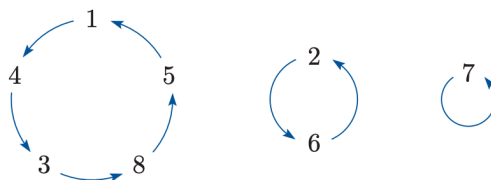
Figure 11.1. Cycles of σ_g

We can then write:
$$\sigma_g = (1\ 4\ 3\ 8\ 5)(2\ 6)(7) \tag{11.1.5}$$
and we say that it is a product of three cycles.

> **Definition 16.2** *(Permutation cycles)*
> A permutation σ of a set S is said to be cyclic if there exist $a_i...a_k \in \{1, 2, ..., n\}$ such that:
> $$a_i\sigma = a_{i+1} \quad \text{for } 1 \le i < k \tag{11.1.6}$$
> $$a_k\sigma = a_1 \tag{11.1.7}$$
> and is denoted in **cycle form** as $(a_1\ \sigma a_1\ \sigma^2 a_1\ ...\ \sigma^{k-1} a_1)$. 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 $1 \le i < k$, 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:
$$a_1\sigma = 4 = a_2, \tag{11.1.8}$$
$$a_2\sigma = 3 = a_3, \tag{11.1.9}$$
$$a_3\sigma = 8 = a_4, \tag{11.1.10}$$
$$a_4\sigma = 5 = a_5 \tag{11.1.11}$$
$$a_5\sigma = 1 = a_1 \tag{11.1.12}$$

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 $\sigma_a = (a_1\ ...\ a_k)$ and $\sigma_b = (b_1\ ...\ b_l)$ so that $a_i \ne b_i$. Then applying $\sigma_b \sigma_a$:
$$a_i\sigma_a\sigma_b = a_{i+1}\sigma_b = a_{i+1} \qquad \text{for } i < k \tag{11.1.13}$$
$$a_k\sigma_a\sigma_b = a_1\sigma_b = a_1 \tag{11.1.14}$$
$$b_i\sigma_a\sigma_b = b_i\sigma_b = b_{i+1} \qquad \text{for } i < l \tag{11.1.15}$$
$$b_l\sigma_a\sigma_b = b_l\sigma_b = b_1 \tag{11.1.16}$$

Similarly applying $\sigma_b \sigma_a$ one finds:
$$a_i\sigma_b\sigma_a = a_i\sigma_a = a_{i+1} \qquad \text{for } i < k \tag{11.1.17}$$
$$a_k\sigma_b\sigma_a = a_k\sigma_a = a_1 \tag{11.1.18}$$
$$b_i\sigma_b\sigma_a = b_{i+1}\sigma_a = b_{i+1} \qquad \text{for } i < l \tag{11.1.19}$$
$$b_l\sigma_b\sigma_a = b_1\sigma_a = b_1 \tag{11.1.20}$$

∎

An immediate consequence of Proposition 16.3 is that, if a permutation σ can be expressed as a product of disjoint cycles σ_i:

$$\sigma = \prod_{i=1}^{n} \sigma_i \implies \sigma^k = \prod_{i=1}^{n} \sigma_i^k \tag{11.1.21}$$

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 $\sigma \in \text{Sym}(\{1, 2, ..., n\})$ and the infinite sequence:

$$a_1, a_1\sigma, a_1\sigma^2, a_1\sigma^3... \tag{11.1.22}$$

where $a_1 \in \{1, 2, ..., n\}$. Because $\{1, 2, ..., n\}$ 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 $\{a_1, a_1\sigma...a_1\sigma^{k_1-1}\}$ 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:

$$\sigma = (a_1 \ a_1\sigma... \ a_1\sigma^{k_1-1})(a_2 \ a_2\sigma... \ a_2\sigma^{k_2-1})...(a_r \ a_R\sigma... \ a_r\sigma^{k_r-1}) \tag{11.1.23}$$

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:

$$\sigma = \rho_1\rho_2...\rho_k = \tau_1\tau_2...\tau_l \tag{11.1.24}$$

Then $a_1 \in \{1, 2, ..., n\}$ 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:

$$\rho_1 = (1 \ 1\sigma \ 1\sigma^2...1\sigma^{k-1}) = \tau_1 \tag{11.1.25}$$

an repeating this for all other ρ and τ we get the desired result.

$\blacksquare$

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 $\sigma \in \mathrm{Sym}(S)$ is the set $\{a_1,\ \sigma a_1 ... \sigma^{k_1-1} a_1\}$ 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 $\sigma_g \circ \sigma_f = (1\ x...)$.
>   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:
>
> $$(1\ 4\ 6)(3\ 5) \circ (1\ 5\ 3\ 2\ 4) \circ (1\ 2)(3\ 5)(4\ 6) = (1\ 6\ 4)(2\ 3\ 5) \tag{11.1.26}$$
>
> ◀

11.2 Permutation groups

> **Theorem 16.6** *(Symmetric group)*
> The set S_n of all permutations of $\{1, 2, 3, ..., n\}$ 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 $\{1, 2, ..., n\}$ 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 $\{1, 2, ..., n\}$ 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 $\{1, 2, ..., n\}$ 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:

$$n \times (n-1) \times (n-2)... \times 2 \times 1 = n! \tag{11.2.1}$$

are all the possible permutations. Now if $x_1, x_2, x_3 \in \{1, 2, ..., n\}$ are distinct then we can define:

$$\sigma_1 : x_1 \mapsto x_1, \ x_2 \mapsto x_3, \ x_3 \mapsto x_2 \tag{11.2.2}$$
$$\sigma_2 : x_1 \mapsto x_2, \ x_2 \mapsto x_1, \ x_3 \mapsto x_3 \tag{11.2.3}$$

We then find that:

$$\sigma_2 \circ \sigma_1 : x_1 \mapsto x_2, \ x_2 \mapsto x_3, \ x_3 \mapsto x_1 \tag{11.2.4}$$
$$\sigma_1 \circ \sigma_2 : x_1 \mapsto x_3, \ x_2 \mapsto x_1, \ x_3 \mapsto x_2 \tag{11.2.5}$$

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 $\sigma = \rho_1 \rho_2 ... \rho_k$ where ρ_i are disjoint cycles of order l_i is:
>
> $$\text{ord}(\sigma) = \text{LCM}(l_1, l_2...l_k) \tag{11.2.6}$$
>
> 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:

$$\sigma^n = \rho_1^n \rho_2^n \rho_3^n ... \rho_k^n = e \tag{11.2.7}$$

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:

$$\rho_i^n = e, \forall i \tag{11.2.8}$$

so that $n | l_i$. Since n must be the smallest such integer, it follows that $n = \text{LCM}(l_1, l_2, ..., l_k)$ as desired. ∎

We can use cycle form to denote symmetries of figures as well. For example the symmetry:

$$\begin{pmatrix} 1 \ 2 \ 3 \ 4 \\ 1 \ 4 \ 3 \ 2 \end{pmatrix} \tag{11.2.9}$$

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:
>
> $$e, \ (1\ 2\ 3), \ (1\ 3\ 2), \ (1\ 2), \ (1\ 3), \ (2\ 3) \tag{11.2.10}$$
>
> We can then compose with the reflection $(4\ 5)$ to find:
>
> $$(4\ 5), \ (1\ 2\ 3)(4\ 5), \ (1\ 3\ 2)(4\ 5), \ (1\ 3)(4\ 5), \ (2\ 3)(4\ 5) \tag{11.2.11}$$
>
> These are twelve distinct symmetries, and therefore we have found all the symmetries. ◄

11.3 Even and Odd symmetries

> **Strategy.** *(Expressing cycles as composite of transpositions)*
> Consider a cycle $(a_1 \ a_2 \ a_3 ... a_r)$, then we can express:
>
> $$(a_1 \ a_2 \ a_3 ... a_r) = (a_1 \ a_r) \circ (a_1 \ a_{r-1}) \circ ... \circ (a_1 \ a_2) \tag{11.3.1}$$

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.

Now consider the a_s where $2 \leq s \leq r - 1$. We then see that $(a_1 \ a_2), ...(a_1 \ a_{s-1})$ 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}.
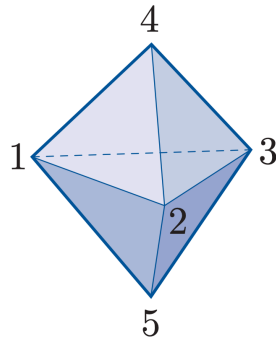
Figure 11.2. Octahedron

Finally, all the successive transpositions $(a_1\ a_{s+2})...(a_1\ a_r)$ 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 $(a_1\ a_2),...(a_1\ a_{r-1})$ 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 $(a_1\ a_2\ a_3...a_r)$ 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 transposition. Indeed in the previous example we could have written:

$$(2\ 4\ 3\ 5) = (4\ 3\ 5\ 2) = (4\ 2) \circ (4\ 5) \circ (4\ 3) \tag{11.3.2}$$

$$= (2\ 4) \circ (5\ 4) \circ (3\ 4) \tag{11.3.3}$$

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 $x_1, \ldots, x_n$ and consider the Vandermonde polynomial [1]

$$P = \prod_{1 \le i < j \le n} (x_j - x_i) \tag{11.3.4}$$

We now choose a permutation $\sigma \in S_n$, and define the function f_σ as:

$$f_\sigma(P) = \prod_{1 \le i < j \le n} (x_{\sigma(j)} - x_{\sigma(i)}) \tag{11.3.5}$$

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 $j, i \ne a, b$ 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:

$$\sigma = \tau_1 \cdots \tau_r = \rho_1 \cdots \rho_s \tag{11.3.6}$$

Then

$$f_\sigma(P) = f_{\tau_1 \cdots \tau_r}(P) = (f_{\tau_1} \circ \cdots \circ f_{\tau_r})(P) = (-1)^r P \tag{11.3.7}$$
$$f_\sigma(P) = f_{\rho_1 \cdots \rho_s}(P) = (f_{\rho_1} \circ \cdots \circ f_{\rho_s})(P) = (-1)^s P \tag{11.3.8}$$

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:

$$\sigma \text{ is } \begin{cases} \text{even permutation, if } k \text{ is odd} \\ \text{odd permutation, if } k \text{ is even} \end{cases} \tag{11.3.9}$$

---

[1] For $n = 4$, we would have $P = (x_2 - x_1)(x_3 - x_1)(x_4 - x_1)(x_3 - x_2)(x_4 - x_2)(x_4 - x_3)$.

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
>
> $$\begin{cases} \text{even permutation, if } k \text{ is odd} \\ \text{odd permutation, if } k \text{ is even} \end{cases} \tag{11.3.10}$$
>
> 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 $\{1, 2, ...n\}$ 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:

$$e, \ (\_\_)(\_\_), (\_\_\_\_) \tag{11.3.11}$$

and we must fill the gaps with $\{1, 2, 3, 4\}$.

Note that there are only 3 cycle structures of the form $(\_\_)(\_\_)$. 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 (_ _ _ _). 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 $f_1, ..., f_r \in A_n$ then consider:

$$(1\ 2) \circ f_1,\ (1\ 2) \circ f_2,\ ...\ ,(1\ 2) \circ f_r \tag{11.3.12}$$

these are all distinct odd permutations. So there are at least r odd permutations in S_n.

A similar argument uses $g_1...g_r$ 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 $|A_n| = \frac{1}{2}n!$. ∎

11.4 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:
>
> $$\sigma = \tau \circ \rho \circ \tau^{-1} \tag{11.4.1}$$
>
> 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:

$$\left.\begin{array}{l} x = (* \ * \ \ldots \ *)(* \ * \ \ldots \ *)\ldots(*)(*) \\ y = (* \ * \ \ldots \ *)(* \ * \ \ldots \ *)\ldots(*)(*) \end{array}\right\}\tau \qquad (11.4.2)$$

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:

$$\left.\begin{array}{l} x = (1\,2\,4)(3\,5) \\ y = (2\,5\,3)(1\,4) \end{array}\right\}\tau \qquad (11.4.3)$$

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:

$$\left.\begin{array}{l} x = (1\,2\,4)(3\,5) \\ y = (3\,2\,5)(1\,4) \end{array}\right\}\tau \qquad (11.4.4)$$

whose corresponding conjugate permutation is $\tau = (1\,3)(4\,5)$.

Finally we can rewrite $(2\,5\,3)$ as $(3\,5\,2)$ and find:

$$\left.\begin{array}{l} x = (1\,2\,4)(3\,5) \\ y = (3\,5\,2)(1\,4) \end{array}\right\}\tau \qquad (11.4.5)$$

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:

$$g \circ H \circ g^{-1} = \{g \circ h \circ g^{-1} \ : \ h \in H\} \qquad (11.4.6)$$

So it suffices to substitute every element in H using g.

If we let for example $H = \langle(1\,2\,4\,5)\rangle$ then:

$$H = \{e, (1\,2\,4\,5), (1\,2\,4\,5)^2, (1\,2\,4\,5)^3\} \qquad (11.4.7)$$
$$= \{e, (1\,2\,4\,5), (1\,4)(2\,5), (1\,5\,4\,2)\} \qquad (11.4.8)$$

Then, if we let $g = (3\,5)$ we find that:

$$g \circ H \circ g^{-1} = \{e, (1\,2\,4\,3), (1\,4)(2\,3), (1\,3\,4\,2)\} \qquad (11.4.9)$$

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:

$$h \circ k = (g \circ h' \circ g^{-1}) \circ (g \circ k' \circ g^{-1}) \tag{11.4.10}$$
$$= g \circ h' \circ g^{-1} \circ g \circ k' \circ g^{-1} \tag{11.4.11}$$
$$= g \circ (h' \circ k) \circ g^{-1} \tag{11.4.12}$$
$$= g \circ l \circ g^{-1} \tag{11.4.13}$$

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:

$$e_H = g \circ g^{-1} = g \circ e_H \circ g^{-1} \in g \circ H \circ g^{-1]} \tag{11.4.14}$$

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:

$$g \circ h \circ g^{-1} \circ g \circ h^{-1} \circ g^{-1} = g \circ h \circ h^{-1} \circ g^{-1} = g \circ g^{-1} = e_H \tag{11.4.15}$$

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.

∎

11.5 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 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:

$$\{e, (1\,2)\}, \{e, (1\,3)\}, \{e, (1\,4)\}, \{e, (2\,3)\}, \{e, (2\,4)\}, \{e, (3\,4)\} \tag{11.5.1}$$

and:

$$\{e, (1\,2)(3\,4)\}, \{e, (1\,3)(2\,4)\}, \{e, (1\,4)(2\,3)\} \tag{11.5.2}$$

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:

$$\{\langle (1\,2\,3) \rangle, \langle (1\,3\,4) \rangle, \langle (1\,4\,2) \rangle, \langle (2\,3\,4) \rangle\} \tag{11.5.3}$$

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 11.3. Cycle structures of elements in S_4

The cyclic subgroups of order 4 can be found similarly. (1 2 3 4) generates the following set:

$$\langle (1\ 2\ 3\ 4) \rangle = \{e, (1\ 2\ 3\ 4), (1\ 3)(2\ 4), (1\ 4\ 3\ 2)\} \tag{11.5.4}$$

We now choose a permutation of order 4 that is not in this list, and finds its generator:

$$\langle (1\ 2\ 4\ 3) \rangle = \{e, (1\ 2\ 4\ 3), (1\ 4)(2\ 3), (1\ 3\ 4\ 2)\} \tag{11.5.5}$$

Repeat this process one final time:

$$\langle (1\ 3\ 2\ 4) \rangle = \{e, (1\ 3\ 2\ 4), (1\ 2)(3\ 4), (1\ 4\ 2\ 3)\} \tag{11.5.6}$$

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:
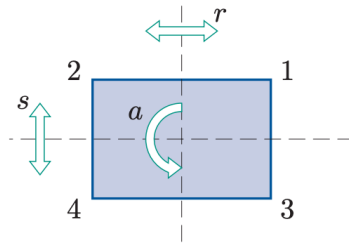
Then the symmetry group is:

$$\{e, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\} \tag{11.5.7}$$

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.

11.6 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 $G = \{g_1, ..., g_n\}$ so that for each element g_i the table shows:

$*$	g_1	g_2	g_3	$\cdots$	g_n
g_1					
$\vdots$					
x	$x * g_1$	$x * g_2$	$x * g_3$	$\cdots$	$x * g_n$
$\vdots$					
g_n					

so that:

$$p_x = \begin{pmatrix} g_1 & g_2 & \cdots & g_n \\ x * g_1 & x * g_2 & \cdots & x * g_n \end{pmatrix} \tag{11.6.1}$$

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:

$$p_x \circ p_y = \begin{pmatrix} g_1 & g_2 & \cdots & g_n \\ x*y*g_1 & x*y*g_2 & \cdots & x*y*g_n \end{pmatrix} = p_{x*y} \tag{11.6.2}$$

so that:

$*$	$\cdots$	y	$\cdots$
$\vdots$		$\vdots$	
x	$\cdots$	$x*y$	$\cdots$
$\vdots$		$\vdots$	

$(G, *)$

$\circ$	$\cdots$	p_y	$\cdots$
$\vdots$		$\vdots$	
p_x	$\cdots$	p_{x*y}	$\cdots$
$\vdots$		$\vdots$	

$(P, \circ)$

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:

$$p : G \to P \tag{11.6.3}$$
$$x \mapsto p_x \tag{11.6.4}$$

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. $\blacksquare$

For example, consider the group $(\mathbb{Z}_6, +_6)$. For each $x \in \mathbb{Z}_6$, we associate a permutation p_x whose

$+_6$	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 11.4. Cayley table for $(\mathbb{Z}_6, +_6)$

two line form has its first line as the column heading and its second line as the row labelled x. So:

$$p_2 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 0 & 1 \end{pmatrix} = (0\ 2\ 4)(1\ 3\ 5) \tag{11.6.5}$$

This gives us the permutations $p_0, p_1...p_6$ for each element of $\mathbb{Z}_6$, obtaining:

$$p_0 = e \tag{11.6.6}$$
$$p_1 = (0\ 1\ 2\ 3\ 4\ 5) \tag{11.6.7}$$
$$p_2 = (0\ 2\ 4)(1\ 3\ 5) \tag{11.6.8}$$
$$p_3 = (0\ 3)(1\ 4)(2\ 5) \tag{11.6.9}$$
$$p_4 = (0\ 4\ 2)(1\ 5\ 3) \tag{11.6.10}$$
$$p_5 = (0\ 5\ 4\ 3\ 2\ 1) \tag{11.6.11}$$

Let $P = \{p_0, p_1, p_2...p_5\}$. 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:

$$\phi : \mathbb{Z}_6 \to P \tag{11.6.12}$$
$$x \mapsto p_x \tag{11.6.13}$$

is an isomorphism.

GT4 Lagrange's Theorem and small groups

12.1 Lagrange's Theorem

> **Theorem 17.1** *(Lagrange's Theorem)*
> Let G be a finite group, and let $H \leq G$. Then $\mathrm{ord}(H)|\mathrm{ord}(G)$, 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 $\mathrm{ord}(G) = s$ and $\mathrm{ord}(H) = r$.

We begin by writing down all the elements of H:

$$\begin{pmatrix} h_1 & h_2 & \ldots & h_r \end{pmatrix} \tag{12.1.1}$$

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.

$$\begin{pmatrix} h_1 & h_2 & \ldots & h_r \\ g_2 \circ h_1 & g_2 \circ h_2 & \ldots & g_2 \circ h_r \end{pmatrix} \tag{12.1.2}$$

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:

$$\begin{pmatrix} h_1 & h_2 & \ldots & h_r \\ g_2 \circ h_1 & g_2 \circ h_2 & \ldots & g_2 \circ h_r \\ g_3 \circ h_1 & g_3 \circ h_2 & \ldots & g_3 \circ h_r \end{pmatrix} \tag{12.1.3}$$

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:

$$
\begin{pmatrix}
h_1 & h_2 & \dots & h_r \\
g_2 \circ h_1 & g_2 \circ h_2 & \dots & g_2 \circ h_r \\
g_3 \circ h_1 & g_3 \circ h_2 & \dots & g_3 \circ h_r \\
\vdots & \vdots & & \vdots \\
g_k \circ h_1 & g_k \circ h_2 & \dots & g_k \circ h_r
\end{pmatrix}
\tag{12.1.4}
$$

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 kth row, we have that if for some $h_i, h_j \in H$ distinct:

$$
g_k \circ h_i = g_k \circ h_j
\tag{12.1.5}
$$

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 $g_k \circ h_j$ another row k:

$$
g_l \circ h_i = g_k \circ h_j \implies g_l = g_k \circ h_j \circ h_i^{-1}
\tag{12.1.6}
$$

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 kth row. This is a contradiction, since we assumed that the rows l and k are different.
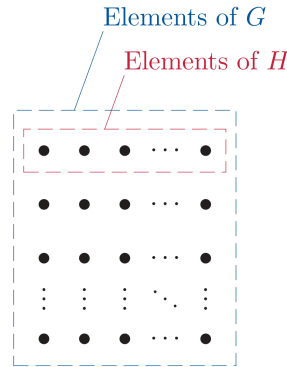
Figure 12.1. Visualization of Lagrange's proof

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, $\mathrm{ord}(G) = k \cdot r = k \cdot \mathrm{ord}(H)$. It follows immediately that $\mathrm{ord}(H) | \mathrm{ord}(G)$. ∎

Corollary.

(i) let $g \in G$, then the $\mathrm{ord}(g) | \mathrm{ord}(G)$.

(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 $\mathrm{ord}(g) = \mathrm{ord}(\langle g \rangle)$ so that $\mathrm{ord}(g)|\mathrm{ord}(G)$.

(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.

∎

12.2 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	$(\mathbb{Z}_5^*, \times_5)$, $S^+(\square)$, $S(\maltese)$, $(\{0, 3, 6, 9\}, +_{12})$, $(\{1, -1, i, -i\}, \times)$
	V, $S(\square)$	abelian, non-cyclic	$(U_8, \times_8)$, $(U_{12}, \times_{12})$, $(\{1, 7, 9, 15\}, \times_{16})$, $(\{1, 9, 11, 19\}, \times_{20})$
5	C_5, $(\mathbb{Z}_5, +_5)$	cyclic	$S^+(\bigcirc)$
6	C_6, $(\mathbb{Z}_6, +_6)$	cyclic	$S^+(\bigcirc)$, $(\mathbb{Z}_7^*, \times_7)$, $(U_9, \times_9)$, $(\{0, 2, 4, 6, 8, 10\}, +_{12})$, $(U_{14}, \times_{14})$
	$S(\triangle)$	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 $\mathrm{ord}(G) > 2$ and each element except e has order 2, then $4|\mathrm{ord}(G)$.
> (iii) if $\mathrm{ord}(G)$ 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:

$$(xy)^2 = e \implies xyxy = e \implies xey = x^2 yxy^2 \implies xy = yx \qquad (12.2.1)$$

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

$$z = e \implies xy = e \implies y = x \qquad (12.2.2)$$
$$z = x \implies xy = x \implies y = e \qquad (12.2.3)$$
$$z = y \implies xy = y \implies x = e \qquad (12.2.4)$$

It now remains to prove that $\{e, x, y, z\} \le G$. 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 $\{e, x, y, z\} \le G$, and by Lagrange's theorem, $4 | \mathrm{ord}(G)$.

(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 od 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 my the same logic as in the proof of (ii) $z = xy$ and we retrieve the following Cayley table: which is the table of the Klein four-group V, so

	e	x	y	z
e	e	x	y	z
x	x	e	z	y
y	y	z	e	x
z	z	y	x	e

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 $4|\mathrm{ord}(G)$ which clearly isn't the case.

So let $g, h \in G$ be some elements of order 2 and 3 respectively. We define:

$$H = \langle h \rangle = \{e, h, h^2\} \tag{12.2.5}$$

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:

$$\begin{pmatrix} e & h & h^2 \\ g & gh & gh^2 \end{pmatrix} \tag{12.2.6}$$

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:

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:

$$hg = gh \neq e \implies (hg)^3 = (hg)(gh)(hg) = g \neq e \tag{12.2.7}$$

so that hg has order greater than 3. The only possible value for $\mathrm{ord}(hg)$ is then 6, a contradiction.

	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			

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.

GT5 Cosets and normal subgroups

13.1 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:
$$\det(AB) = \det A \cdot \det B \neq 0 \implies AB \in GL(n, \mathbb{R}) \tag{13.1.1}$$

Associativity matrix multiplication is associative.

Identity The identity matrix $\mathbb{I}$ is the identity of $GL(n, \mathbb{R})$. Indeed $\forall A \in GL(n, \mathbb{R})$:
$$\mathbb{I}A = A\mathbb{I} = A \tag{13.1.2}$$

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:
$$AA^{-1} = A^{-1}A = \mathbb{I} \tag{13.1.3}$$

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, $\dim GL(n, \mathbb{R}) = n^2$. Indeed, the restriction that $\det A \neq 0$ only restrict the values that the matrix elements cannot take, but it doe snot force some matrix elements to take a specific value. Hence they are spanned by the standard basis of $\text{Mat}_n(\mathbb{R})$.

> **Definition 18.2** *(Special linear group $SL(n, \mathbb{R})$)*
> The subset of $GL(n, \mathbb{R})$ 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 $A, B \in SL(n, \mathbb{R})$, and let A^{-1} and B^{-1} be their inverses. Then, since $\det A = 1$ and $\det B = 1$, we find that:
$$\det(AB) = \det A \cdot \det B = 1 \implies AB \in SL(n, \mathbb{R}) \tag{13.1.4}$$

Identity The identity matrix $\mathbb{I} \in SL(n, \mathbb{R})$ since $\det \mathbb{I} = 1$.

Inverses Let $A \in SL(n, \mathbb{R})$, then it must have an inverse A^{-1}, and:
$$\det A^{-1} = \frac{1}{\det A} = 1 \tag{13.1.5}$$

so $A^{-1} \in SL(n, \mathbb{R})$.

Therefore all the subgroup axioms are satisfied, and $SL(n, \mathbb{R})$ forms a sgroup 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 $A \in SL(n, \mathbb{R})$:
$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix} \tag{13.1.6}$$

The condition $\det A = 1$ can be expanded as:
$$\det A = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{\sigma(1),1} a_{\sigma(2),2} a_{\sigma(3),3} \dots a_{\sigma(n),n} \tag{13.1.7}$$

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 A, giving:
$$\dim SL(n, \mathbb{R}) = n^2 - 1 \tag{13.1.8}$$

The extension of these results for $SL(n, \mathbb{C})$ is immediate:
$$\dim SL(n, \mathbb{C}) = 2 \cdot \dim SL(n, \mathbb{R}) = 2n^2 - 2 \tag{13.1.9}$$

> **Definition 18.3** *(Orthogonal group $O(n, \mathbb{R})$)*
> The subset of $GL(n, \mathbb{R})$ 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 $n\mathbb{R}$.

Proof. Firstly note that if $A \in O(n, \mathbb{R})$, then:

$$\det AA^T = (\det\{A\})^2 = \det \mathbb{I} = 1 \implies \det\{A\} = \pm 1 \tag{13.1.10}$$

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:

$$(AB)(AB)^T = ABB^T A^T = \mathbb{I} \tag{13.1.11}$$

so $AB \in O(n, \mathbb{R})$.

Identity The identity matrix $\mathbb{I} \in O(n, \mathbb{R})$ since $\mathbb{I}\mathbb{I}^T = \mathbb{I}$.

Inverses Let $A \in O(n, \mathbb{R})$. Then its inverse A^{-1} in $GL(n, \mathbb{R})$ satisfies

$$(A^{-1})(A^{-1})^T = (A^{-1})(A^T)^{-1} = (A^T A)^{-1} = \mathbb{I}^{-1} = \mathbb{I} \tag{13.1.12}$$

so that $A \in O(n, \mathbb{R})$ as desired. So $A^{-1} \in SL(n, \mathbb{R})$.

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:

$$AA^T - \mathbb{I} = 0 \tag{13.1.13}$$

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 $\frac{1}{2}(n^2 - n)$. Hence can be expressed as:

$$\begin{pmatrix} b_{11} & \ldots & \ldots & B \\ \vdots & b_{22} & & \vdots \\ \vdots & & & \vdots \\ B & \ldots & & b_{nn} \end{pmatrix} = 0 \tag{13.1.14}$$

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 $\mathbb{A}$. Hence:

$$\dim O(n, \mathbb{R}) = n^2 - \frac{1}{2}n(n+1) = \frac{1}{2}n(n-1) \tag{13.1.15}$$

> **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 $U(n, \mathbb{R})$)*
> The subset of $\mathrm{GL}(n, \mathbb{C})$ comprising of all unitary $n \times n$ matrices ($AA^\dagger = \mathbb{1}$ forms a subgroup of the general linear group, called the **unitary group**, and denoted $U(n, \mathbb{R})$.

Proof. We need to show that the subgroup axioms are satisfied.

Closure Let $A, B \in U(n, \mathbb{C})$. Then we find that:

$$(AB)(AB)^\dagger = ABB^\dagger A^\dagger = \mathbb{1} \tag{13.1.16}$$

so $AB \in O(n, \mathbb{C})$.

Identity The identity matrix $\mathbb{1} \in U(n, \mathbb{C})$ since $\mathbb{1}\mathbb{1}^\dagger = \mathbb{1}$.

Inverses Let $A \in U(n, \mathbb{C})$. Then its inverse A^{-1} in $\mathrm{GL}(n, \mathbb{C})$ satisfies

$$(A^{-1})(A^{-1})^\dagger = (A^{-1})(A^\dagger)^{-1} = (A^\dagger A)^{-1} = \mathbb{1}^{-1} = \mathbb{1} \tag{13.1.17}$$

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 $\mathrm{Mat}_n(\mathbb{C})$ has dimension $2n^2$, n^2 free real parameters, and n^2 free complex parameters. Now let's see how many constraints unitarity ($AA^\dagger = \mathbb{1}$) imposes.

Note that $AA^\dagger$ is hermitian, since $(AA^\dagger)^\dagger = (A^\dagger)^\dagger A^\dagger = AA^\dagger$. 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:

$$\dim U(n, \mathbb{C}) = 2n^2 - n^2 = n^2 \tag{13.1.18}$$

> **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 $\mathrm{SU}(n, \mathbb{C})$.

Note that the determinant of $A \in U(n, \mathbb{R})$ is such that:

$$\det AA^\dagger = (\det A)(\det A)^* = |\det A|^2 = 1 \implies \det A = e^{i\theta}, \ \forall \theta \in [0, 2\pi) \tag{13.1.19}$$

whereas we are restricting $\det A = 1 \implies \theta = 0$. 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:
$$\dim \mathrm{SU}(n, \mathbb{C}) = n^2 - 1 \tag{13.1.20}$$

We summarize the dimensionalities of these matrix groups below:

> **Theorem 18.5** *(Dimensions of matrix groups)*
> We have that:
>   (i) $\dim \mathrm{GL}(n, \mathbb{R}) = n^2$
>   (ii) $\dim \mathrm{SL}(n, \mathbb{R}) = n^2 - 1$
>   (iii) $\dim O(n, \mathbb{R}) = \frac{1}{2}n(n-1)$
>   (iv) $\dim \mathrm{SU}(n, \mathbb{R}) = \frac{1}{2}n(n-1)$
>   (v) $\dim U(n, \mathbb{C}) = n^2$
>   (vi) $\dim \mathrm{SU}(n, \mathbb{C}) = n^2 - 1$

13.2 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:
> $$gH = \{gh : h \in H\} \tag{13.2.1}$$
> 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 $S(\Delta)$:

$$eH = e\{e, s\} = \{e, s\} \tag{13.2.2}$$
$$aH = a\{e, s\} = \{a, r\} \tag{13.2.3}$$
$$bH = b\{e, s\} = \{b, t\} \tag{13.2.4}$$
$$rH = r\{e, s\} = \{r, a\} = aH \tag{13.2.5}$$
$$sH = s\{e, s\} = \{s, e\} = eH \tag{13.2.6}$$
$$tH = t\{e, s\} = \{t, b\} = bH \tag{13.2.7}$$

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:

$$1H = 1\{1, 2, 4\} = \{1, 2, 4\} \tag{13.2.8}$$
$$2H = 2\{1, 2, 4\} = \{2, 4, 1\} \tag{13.2.9}$$
$$3H = 3\{1, 2, 4\} = \{3, 6, 5\} \tag{13.2.10}$$
$$4H = 4\{1, 2, 4\} = \{4, 1, 2\} \tag{13.2.11}$$
$$5H = 5\{1, 2, 4\} = \{5, 3, 6\} \tag{13.2.12}$$
$$6H = 6\{1, 2, 4\} = \{6, 5, 3\} \tag{13.2.13}$$

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:

$$x \sim y \text{ if } x \in yH \tag{13.2.14}$$

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 $y = zH$. Therefore, $\exists h_1, h_2 \in H$ such that:

$$x = yh_1 \text{ and } y = zh_2 \implies x = zh_1h_2 \tag{13.2.15}$$

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:

$$[x] = \{y \in G : y \sim x\} = \{y \in G : y \in xH\} = xH \tag{13.2.16}$$

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 $g_1 H$ and $g_2 H$ 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 $H = \{h_1, h_2, ...h_{,m}\}$, and let $g \in G$. Then:

$$gH = \{gh_1, gh_2, ..., gh_m\} \tag{13.2.17}$$

Suppose $gh_i = gh_j$, then by the left cancellation law $h_1 = h_j$, 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:
> $$3H = 3\{1, 19\} = \{3, 17\} \tag{13.2.18}$$
>
> An element missing from both H and $3H$ is 7, and its corresponding left coset is:
>
> $$7H = 7\{1, 19\} = \{7, 13\} \tag{13.2.19}$$
>
> The final missing element from all these left cosets is 9:
>
> $$9H = 9\{1, 19\} = \{9, 11\} \tag{13.2.20}$$
>
> 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:

$$A_4 = \{e, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3), (1\ 2\ 3), (1\ 3\ 2), (1\ 2\ 4), (1\ 4\ 2), \qquad (13.2.21)$$

$$(1\ 3\ 4), (1\ 4\ 3), (2\ 3\ 4), (2\ 4\ 3)\} \qquad (13.2.22)$$

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:

$$(1\ 2\ 3)H = \{(1\ 2\ 3), (1\ 3\ 4), (2\ 4\ 3), (1\ 4\ 2)\} \qquad (13.2.23)$$

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:

$$H \cup \{(1\ 3\ 2), (1\ 2\ 4), (1\ 4\ 3), (2\ 3\ 4)\} = A_4 \qquad (13.2.24)$$

◄

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 $|G|$. ∎

13.3 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:
> $$Hg = \{hg : h \in H\} \qquad (13.3.1)$$
> 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:

$$He = \{e, r\}e = \{e, r\} \tag{13.3.2}$$
$$Ha = \{e, r\}a = \{a, u\} \tag{13.3.3}$$
$$Hb = \{e, r\}b = \{b, t\} \tag{13.3.4}$$
$$Hc = \{e, r\}c = \{c, s\} \tag{13.3.5}$$
$$Hr = \{e, r\}r = \{r, e\} \tag{13.3.6}$$
$$Hs = \{e, r\}s = \{s, c\} \tag{13.3.7}$$
$$Ht = \{e, r\}t = \{t, b\} \tag{13.3.8}$$
$$Hu = \{e, r\}u = \{u, a\} \tag{13.3.9}$$

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:

$$Ha = \{e, s\}a = \{a, t\} \tag{13.3.10}$$

Now another element in G that does not belong to these two cosets is b, so:

$$Hb = \{e, s\}b = \{b, r\} \tag{13.3.11}$$

We have exhausted all elements of $S(\Delta)$, so we may write the partition of the latter as:

$$S(\Delta) = \{e, s\} \cup \{a, t\} \cup \{b, r\} \tag{13.3.12}$$

◀

Note that if G is abelian and $H < G$, then H is also abelian and thus:

$$Hg = \{hg : \forall h \in H\} = \{gh : \forall h \in H\} = gH \tag{13.3.13}$$

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 $x^{-1} \in Hg'$ for some $g' \in G$. Indeed:

$$gH = \{gh : h \in H\} \text{ and } \{(gh)^{-1} : h \in H\} = \{h^{-1}g^{-1} : h \in H\} = \{hg^{-1} : h \in H\} = Hg^{-1} \tag{13.3.14}$$

due to the inverse axiom of subgroups[1]. 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.

> **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 $\frac{|G|}{|H|}$. ■

> **Example.** Let's partition $(2\mathbb{Z}, +)$ into cosets of $6\mathbb{Z}$.
> Note that $2\mathbb{Z} = \{..., -6, -4, -2, 0, 2, 4, 6, ...\}$, and $6\mathbb{Z} = \{..., -18, -12, -6, 0, 6, 12, 18, ...\}$. 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 $2\mathbb{Z}$.

---

[1]indeed suppose

$$H = \{h_1, h_2, ..., h_n, h_1^{-1}, h_2^{-1}, ..., h_n^{-1}\}$$

then

$$\{h^{-1} : h \in H\} = \{h_1^{-1}, h_2^{-1}, ..., h_n^{-1}, h_1, h_2, ..., h_n\} = H$$

Now note that $6\mathbb{Z}$ is itself a coset so:

$$0 + 6\mathbb{Z} = \{6k : k \in \mathbb{Z}\} \tag{13.3.15}$$

An element that was not included is 2:

$$2 + 6\mathbb{Z} = \{..., -16, -10, -4, 2, 8, 14, 20, ...\} = \{2 + 6k : k \in \mathbb{Z}\} \tag{13.3.16}$$

An element that was not included is 4, so:

$$4 + 6\mathbb{Z} = \{..., -14, -8, -2, 4, 10, 16, 22, ...\} = \{4 + 6k : k \in \mathbb{Z}\} \tag{13.3.17}$$

Note that:

$$\{2(3k+2) : k \in \mathbb{Z}\} \cup \{2(3k+1) : k \in \mathbb{Z}\} \cup \{2(3k) : k \in \mathbb{Z}\} = \{2k : k \in \mathbb{Z}\} = 2\mathbb{Z} \tag{13.3.18}$$

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. ◀

13.4 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:
>
> $$\{e, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\} \cup \{(1\ 3\ 2), (1\ 2\ 4), (1\ 4\ 3), (2\ 3\ 4)\} = A_4 \tag{13.4.1}$$

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:

$$\{e, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\} \cup \{(1\,2\,3), (1\,4\,2), (1\,3\,4), (2\,4\,3)\} = A_4 \qquad (13.4.2)$$

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:

$$gH = \{gh : h \in H\} = \{hg : h \in H\} = Hg \qquad (13.4.3)$$

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 H. ∎

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_nI. f instead $n = 1$, then $A_n = S_n$, and from proposition 18.18 we see that A_1 is normal to S_1. ◀

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 $S(\square)$. ◀

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.

∎

GT6 Quotient groups and conjugacy

14.1 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:
>
> $$X \cdot Y = \{xy : x \in X, y \in Y\} \tag{14.1.1}$$
>
> 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 compositon. 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:
>
> $$xN \cdot yN = (xy)N, \ \forall x, y \in G \tag{14.1.2}$$

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:

$$z = xn_1yn_2 \tag{14.1.3}$$

Since N is a normal subgroup, $n_1y \in Ny \implies n_1y \in yN$ so

$$n_1y = yn_3, \ n_3 \in N \tag{14.1.4}$$

Then:

$$z = xyn_3n_2 = xyn, \; n = n_3n_2 \in N \tag{14.1.5}$$

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:

$$z = xyn_1 \tag{14.1.6}$$

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:

$$xN \cdot (yN \cdot zN) = xN \cdot (yz)N = (x(yz))N = (xyz)N \tag{14.1.7}$$

since G is a group, and therefore its operation is associative. Similarly

$$(xN \cdot yN) \cdot zN = ((xy)z)N = (xyz)N \tag{14.1.8}$$

Identity We prove that eN is the identity element. Indeed:

$$eN \cdot xN = (ex)N = xN, \; \forall x \in G \tag{14.1.9}$$

and similarly:

$$xN \cdot eN = (xe)N = xN, \; \forall x \in G \tag{14.1.10}$$

as desired.

Inverses Suppose xN is a coset of N in G. Then:

$$x^{-1}N \cdot xN = (x^{-1}x)N = eN, \; \forall x \in G \tag{14.1.11}$$

and similarly:

$$xN \cdot x^{-1}N = (xx^{-1})N = eN, \; \forall x \in G \tag{14.1.12}$$

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:

$\cdot$	$S^+(\mathcal{F})$	$S^-(\mathcal{F})$
$S^+(\mathcal{F})$	$S^+(\mathcal{F})$	$S^-(\mathcal{F})$
$S^-(\mathcal{F})$	$S^-(\mathcal{F})$	$S^+(\mathcal{F})$

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
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

$\circ$	direct	indirect
direct	direct	indirect
indirect	indirect	direct

Figure 14.1. Block effect in $S(\square)$

Example. Consider the subgroup $H = \langle 6 \rangle$ of $\mathbb{Z}_{12}$. The elements of H are:

$$\langle 6 \rangle = \{0, 6\} \implies \text{ord}(6) = 2 \tag{14.1.13}$$

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:

$$H = \{0, 6\} \tag{14.1.14}$$
$$1 + H = \{1, 7\} \tag{14.1.15}$$
$$2 + H = \{2, 8\} \tag{14.1.16}$$
$$3 + H = \{3, 9\} \qquad 4 + H = \{4, 10\} \tag{14.1.17}$$
$$5 + H = \{5, 11\} \tag{14.1.18}$$

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. ◀

14.2 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:

$$4\mathbb{Z} \tag{14.2.1}$$
$$1 + 4\mathbb{Z} = \{1 + 4n : n \in \mathbb{Z}\} \tag{14.2.2}$$
$$2 + 4\mathbb{Z} = \{2 + 4n : n \in \mathbb{Z}\} \tag{14.2.3}$$
$$3 + 4\mathbb{Z} = \{3 + 4n : n \in \mathbb{Z}\} \tag{14.2.4}$$
$$\tag{14.2.5}$$

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:

$$\phi : \mathbb{Z}/4\mathbb{Z} \to \mathbb{Z}_4 \tag{14.2.6}$$
$$a + 4\mathbb{Z} \mapsto a \forall a \in \mathbb{Z}_4 \tag{14.2.7}$$

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:

$$\phi : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}_n \tag{14.2.8}$$
$$a + n\mathbb{Z} \mapsto a, \ \forall a \in \mathbb{Z}_n \tag{14.2.9}$$

Proof. Let us firstly find the distinct cosets of $n\mathbb{Z}$. We have that:

$$a + n\mathbb{Z} = b + n\mathbb{Z} \iff a \in b + n\mathbb{Z} \iff a \equiv b(\text{mod}n) \tag{14.2.10}$$

so the distinct cosets must be:

$$n\mathbb{Z}, 1 + n\mathbb{Z}, ..., (n-1) + n\mathbb{Z} \tag{14.2.11}$$

It then follows that ϕ is bijective. Indeed, it is injective since

$$\phi(a + n\mathbb{Z}) = \phi(b + n\mathbb{Z}) \implies a \equiv b(\text{mod}n) \implies a + n\mathbb{Z} = b + n\mathbb{Z} \tag{14.2.12}$$

It is also surjective since:

$$a \in \mathbb{Z}_n \implies a \leq n - 1 \implies \phi(a + n\mathbb{Z}) = a \tag{14.2.13}$$

Finally, for $a, b \in \mathbb{Z}_n$ we find that:

$$\phi((a + n\mathbb{Z}) + (b + n\mathbb{Z})) = \phi((a +_n b) + n\mathbb{Z}) \tag{14.2.14}$$
$$= \phi(c + n\mathbb{Z}) \tag{14.2.15}$$
$$= c \tag{14.2.16}$$

where $c \equiv a + b(\text{mod}n)$. Moreover:

$$\phi(a + n\mathbb{Z}) +_n \phi(b + n\mathbb{Z}) = a +_n b = c \tag{14.2.17}$$

so that:

$$\phi((a + n\mathbb{Z}) + (b + n\mathbb{Z})) = \phi(a + n\mathbb{Z}) +_n \phi(b + n\mathbb{Z}) \tag{14.2.18}$$

as desired. ∎

> **Example.** Consider the group $\mathbb{Z}/6\mathbb{Z}$. We have established that:
>
> $$\phi : \mathbb{Z}/6\mathbb{Z} \to \mathbb{Z}_6 \tag{14.2.19}$$
> $$a + n\mathbb{Z} \mapsto a, \forall a \in \mathbb{Z}_6 \tag{14.2.20}$$
>
> 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 $\mathbb{Z}/6\mathbb{Z}$. ◄

14.3   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:
>
> $$y = gxg^{-1} \tag{14.3.1}$$

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:

$$y^k = gx^k g^{-1} \implies y^{k+1} = gx^k g^{-1} y \tag{14.3.2}$$
$$= gx^k g^{-1} gxg^{-1} \tag{14.3.3}$$
$$= gx^k xg^{-1} = gx^{k+1} g^{-1} \tag{14.3.4}$$

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:

$$y^n = gx^n g^{-1} = geg^{-1} = e \tag{14.3.5}$$

Similarly, suppose that $y^n = e$, then:

$$x^n = g^{-1} y^n g = g^{-1} eg = e \tag{14.3.6}$$

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:
$$\{gxg^{-1} : g \in G\} \tag{14.3.7}$$

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.

Proof. Reflexive: 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 $g \in G$:

$$x = gyg^{-1} \implies y = g^{-1} xg = g^{-1} x(g^{-1})^{-1} \tag{14.3.8}$$

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:

$$x = g_1 y g_1^{-1}, \ y = g_2 z g_2^{-1} \tag{14.3.9}$$

for some $g_1, g_2 \in G$. It follows that:

$$x = g_1 g_2 z g_2^{-1} g_1^{-1} = g_3 z g_3^{-1} \tag{14.3.10}$$

where $g_3 = g_1 g_2 \in G$. Hence x is conjugate to z.

Since the equivalence classes of an equivalence class 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:
> $$\{e\}, \ \{r, s, t\}, \ \{a, b\} \tag{14.3.11}$$
> 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:
>
> $$ara^{-1} = arb = at = s \tag{14.3.12}$$
> $$brb^{-1} = bra = bs = t \tag{14.3.13}$$
>
> so we see that $\{r, s, t\}$ is indeed a conjugacy class. Since ther eare 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:
>
> $$rar^{-1} = rar = rt = b \tag{14.3.14}$$
>
> 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 $y = x$:
>
> $$g \in G, y = gxg^{-1} = gg^{-1}x = x \tag{14.3.15}$$
>
> Hence, each conjugacy class can only contain one element:
>
> $$\{1\}, \ \{2\}, \ \{3\}, \ \{4\}, \ \{5\}, \ \{6\} \tag{14.3.16}$$

◀

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 $g \in G$:

$$gxg^{-1} = gg^{-1}x = ex = x \tag{14.3.17}$$

Therefore, x is only conjugate to itself. Hence the conjugacy class of x is $\{x\}$, as desired. ∎

It is important when talking abou 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_n.
> ◀

14.4 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:

$$(a) \iff (c), \ (b) \implies (c), \ (c) \implies (d), \ (d) \implies (b) \tag{14.4.1}$$

$(a) \implies (c)$ Suppose that $gH = Hg$, $\forall g \in G$. Now suppose $x \in gHg^{-1}$, then $\exists h \in H$ such that:

$$x \in gHg^{-1} \tag{14.4.2}$$
$$\iff x = ghg^{-1} \tag{14.4.3}$$
$$\iff x = h'gg^{-1} = h', \ h' \in H \implies x \in H \tag{14.4.4}$$

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:

$$x \in gH \tag{14.4.5}$$
$$\iff x = gh \tag{14.4.6}$$
$$\iff x = ghg^{-1}g \tag{14.4.7}$$
$$\iff x = h_1 g \implies x \in Hg \tag{14.4.8}$$

Therefore $gH = Hg$ as desired.

$(b) \implies (c)$ Suppose that $ghg^{-1} \in H$, for each $h \in H, g \in G$. Then:

$$h = gg^{-1}hgg^{-1} = g \underbrace{\left(g^{-1}h(gg^{-1})^{-1}\right)}_{\in H} g^{-1} \tag{14.4.9}$$

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, $ghg^{-1} \in H$.

$\blacksquare$

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:
$$gxg^{-1} = ghg^{-1} = h'gg^{-1} = h \tag{14.4.10}$$

for some $h, h' \in H$. Similarly:

$$gxg^{-1} = gkg^{-1} = k'gg^{-1} = k' \tag{14.4.11}$$

for some $k, k' \in K$. It follows then that $gxg^{-1} \in H \cap K$. ◀

Example. Consider the group $X = \{(a,b) \in \mathbb{R}^2 : a \neq 0\}$ equipped with the binary operation:
$$(a,b) * (c,d) = (ac, ad + b) \tag{14.4.12}$$

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:

$$k_1 * k_2 = (1, n_1) * (1, n_2) = (1, n_1 + n_2) \in K \tag{14.4.13}$$

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:

$$xkx^{-1} = (a, b) * (1, n) * \left(\frac{1}{a}, -\frac{b}{a}\right) = (a, b) * \left(\frac{1}{a}, n - \frac{b}{a}\right) \tag{14.4.14}$$

$$= (1, an) \tag{14.4.15}$$

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. ◀

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:

$$(ghg^{-1})(gkg^{-1}) = ghkg^{-1} = gxg^{-1} \tag{14.4.16}$$

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:

$$aHa^{-1} = aHc = a\{e, s\}c = a\{c, t\} = \{e, u\} \tag{14.4.17}$$

◀

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:

$$\phi : H \to K \tag{14.4.18}$$
$$h \mapsto ghg^{-1} \tag{14.4.19}$$

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:

$$\phi(xy) = gxyg^{-1} = (gxg^{-1})(gyg^{-1}) = \phi(x)\phi(y) \tag{14.4.20}$$

as desired. ∎

Example. Consider the following subgroup of A_4:

$$K = \{e, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\} \tag{14.4.21}$$

Let's find the following conjugate subgroups:

$$(1\,2\,4)K(1\,2\,4)^{-1} = (1\,2\,4)K(1\,4\,2), \text{ and } (2\,4\,3)K(2\,4\,3)^{-1} = (2\,4\,3)K(2\,3\,4) \tag{14.4.22}$$

Firstly, using the fact that conjugacy does not affect the cycle structure:

$$(1\,2\,4)e(1\,4\,2) = (1\,2\,4)(1\,4\,2) = e \tag{14.4.23}$$
$$(1\,2\,4)(1\,2)(3\,4)(1\,4\,2) = (1\,3)(2\,4) \tag{14.4.24}$$
$$(1\,2\,4)(1\,3)(2\,4)(1\,4\,2) = (1\,4)(2\,3) \tag{14.4.25}$$
$$(1\,2\,4)(1\,4)(2\,3)(1\,4\,2) = (1\,2)(3\,4) \tag{14.4.26}$$

Similarly:

$$(2\,4\,3)e(2\,3\,4) = (2\,4\,3)(2\,3\,4) = e \tag{14.4.27}$$
$$(2\,4\,3)(1\,2)(3\,4)(2\,3\,4) = (1\,4)(2\,3) \tag{14.4.28}$$
$$(2\,4\,3)(1\,3)(2\,4)(2\,3\,4) = (1\,2)(3\,4) \tag{14.4.29}$$
$$(2\,4\,3)(1\,4)(2\,3)(2\,3\,4) = (1\,3)(2\,4) \tag{14.4.30}$$

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. ◀

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:

$$\{e\}, \ \{a,b\}, \ \{r,s,t\} \tag{14.4.31}$$

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:

$$\{e\} = \{e\} \tag{14.4.32}$$
$$\{e,a,b\} = \{e\} \cup \{a,b\} \tag{14.4.33}$$
$$S(\Delta) = \{e\} \cup \{a,b\} \cup \{r,s,t\} \tag{14.4.34}$$

◀

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. ◄

14.5 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.
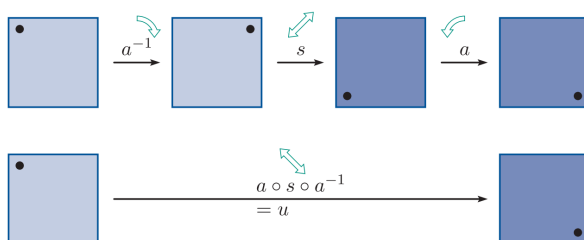
Figure 14.2. Application of rar^{-1}

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 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.
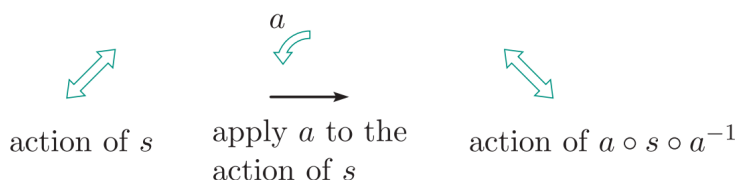
Figure 14.3. Effect of conjugacy on symmetries

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:
>
> $$\text{Fix } f = \{P \in \mathcal{F} : f(P) = P\} \tag{14.5.1}$$
>
> 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 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 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:

$$g^{(-1)}(P) \in \text{Fix } f \implies (fg^{-1})(P) = g^{-1}(P) \implies (gfg^{-1})(P) = (P) \tag{14.5.2}$$

In other words, $P \in \text{Fix } gfg^1$.

Now, suppose that $P \in \text{Fix } gfg^1$. Then:

$$(gfg^{-1})(P) = P \implies (fg^{-1})(P) = g^{-1}(P) \implies f(g^{-1})(P) = (g^{-1}(P)) \tag{14.5.3}$$

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 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:
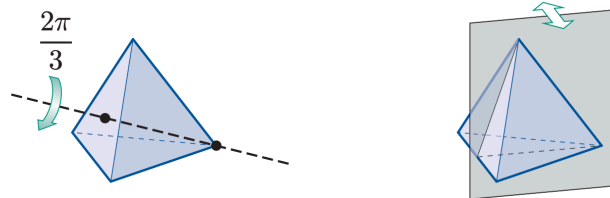
Figure 14.4. Two non-conjugate symmetries of a tetrahedron

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:
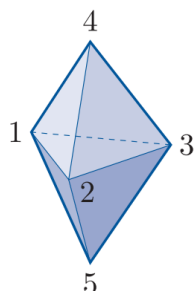
Firstly, since $|S(\mathcal{F})| = 12$, we see that the conjugacy classes can only contain 1,2,3,4,6, or 12 elements.
We can see that its symmetries may be categorized in terms of their cycle structure as follows:

$$\{e\} \tag{14.5.4}$$
$$\{(1\ 2), (1\ 3), (2\ 3), (4\ 5)\} \tag{14.5.5}$$
$$\{(1\ 2\ 3), (1\ 3\ 2)\} \tag{14.5.6}$$
$$\{(1\ 2)(4\ 5), (1\ 3)(4\ 5), (2\ 3)(4\ 5)\} \tag{14.5.7}$$
$$\{(1\ 2\ 3)(4\ 5), (1\ 3\ 2)(4\ 5)\} \tag{14.5.8}$$

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:

$$(2\,3)(1\,2)(2\,3)^{-1} = (1\,3) \tag{14.5.9}$$

$$(1\,2)(1\,3)(1\,2)^{-1} = (2\,3) \tag{14.5.10}$$

and since conjugacy is an equivalence relation, transitivty 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:

$$(-\,-)(4\,5)(-\,-) = (1\,2) \tag{14.5.11}$$

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 $(2\,3) \in S(\mathcal{F})$:

$$(2\,3)(1\,2\,3)(2\,3)^{-1} = (1\,3\,2) \tag{14.5.12}$$

Similarly, we also find that $(1\,2\,3)(4\,5)$ and $(1\,3\,2)(4\,5)$ are conjugate through $(2\,3) \in S(\mathcal{F})$:

$$(2\,3)(1\,2\,3)(4\,5)(2\,3)^{-1} = (1\,3\,2)(4\,5) \tag{14.5.13}$$

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.
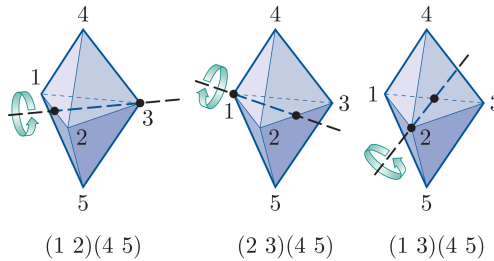
$(1\,2)(4\,5)$ $(2\,3)(4\,5)$ $(1\,3)(4\,5)$

More rigorously:

$$(2\,3)(1\,2)(4\,5)(2\,3)^{-1} = (1\,3)(4\,5) \tag{14.5.14}$$

$$(1\,2)(1\,3)(4\,5)(1\,2)^{-1} = (2\,3)(4\,5) \tag{14.5.15}$$

as desired. Hence the conjugacy classes of $S(\mathcal{F})$ are:

$$\{e\} \tag{14.5.16}$$
$$\{(1\ 2), (1\ 3), (2\ 3)\} \tag{14.5.17}$$
$$\{(4\ 5)\} \tag{14.5.18}$$
$$\{(1\ 2\ 3), (1\ 3\ 2)\} \tag{14.5.19}$$
$$\{(1\ 2)(4\ 5), (1\ 3)(4\ 5), (2\ 3)(4\ 5)\} \tag{14.5.20}$$
$$\{(1\ 2\ 3)(4\ 5), (1\ 3\ 2)(4\ 5)\} \tag{14.5.21}$$

◀

Example. We label the hexagon as shown:
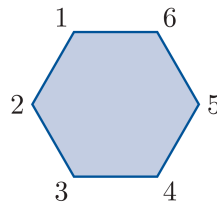
Consider the symmetries of a hexagon:

$$\{e\} \tag{14.5.22}$$
$$\{(1\ 3)(4\ 6), (2\ 6)(3\ 5), (1\ 5)(2\ 4)\} \tag{14.5.23}$$
$$\{(1\ 4)(2\ 5)(3\ 6), (1\ 6)(2\ 5)(3\ 4), (1\ 2)(3\ 6)(4\ 5), (1\ 4)(2\ 3)(5\ 6)\} \tag{14.5.24}$$
$$\{(1\ 3\ 5)(2\ 4\ 6), (1\ 5\ 3)(2\ 6\ 4)\} \tag{14.5.25}$$
$$\{(1\ 2\ 3\ 4\ 5\ 6), (1\ 6\ 5\ 4\ 3\ 2)\} \tag{14.5.26}$$

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:

$$\{e\} \tag{14.5.27}$$
$$\{(1\ 4)(2\ 5)(3\ 6)\} \tag{14.5.28}$$
$$\{(1\ 3)(4\ 6), (2\ 6)(3\ 5), (1\ 5)(2\ 4)\} \tag{14.5.29}$$
$$\{(1\ 6)(2\ 5)(3\ 4), (1\ 2)(3\ 6)(4\ 5), (1\ 4)(2\ 3)(5\ 6)\} \tag{14.5.30}$$
$$\{(1\ 3\ 5)(2\ 4\ 6), (1\ 5\ 3)(2\ 6\ 4)\} \tag{14.5.31}$$
$$\{(1\ 2\ 3\ 4\ 5\ 6), (1\ 6\ 5\ 4\ 3\ 2)\} \tag{14.5.32}$$

If we now modify the hexagon as shown below:
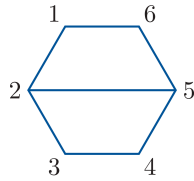
The symmetries of the new modified picture form a subgroup of the symmetries of a hexagon. This subgroup can be partitioned into permutations of the same cyclic structure as:

$$\{e\} \tag{14.5.33}$$

$$\{(1\,3)(4\,6)\} \tag{14.5.34}$$

$$\{(1\,6)(2\,5)(3\,4)\} \tag{14.5.35}$$

◀

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.

GT7 Homomorphisms

> **Definition 20.1** *(Homomorphism)*
> Let $(G, \circ)$ and $(H, *)$ be groups. A mapping $\phi : (G, \circ) \to (H, *)$ is a **homomorphism** if it satisfied the property:
> $$\phi(x \circ y) = \phi(x) * \phi(y), \ \forall x, y \in G \tag{15.0.1}$$

It follows that isomorphisms are homomorphisms that are also bijective.
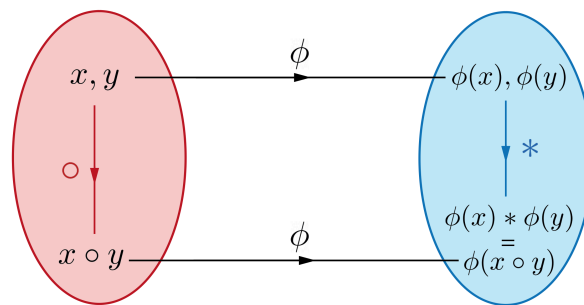
Figure 15.1. Why the homomorphism property must be satisfied for "sensible" maps

Example. Consider the following mapping:

$$\phi : (S_n, \circ) \longrightarrow (\mathbb{Z}_2, +_2) \tag{15.0.2}$$

$$\sigma \mapsto \begin{cases} 0 \text{ if } \sigma \text{ is an even permutation} \\ 1 \text{if } \sigma \text{ is an odd permutation} \end{cases} \tag{15.0.3}$$

Lets show that it is a homomorphism, thus satisfying $\phi(x \circ y) = \phi(x) +_2 \phi(y), \ \forall x, y \in S_n$. Suppose x, y are both of even permutations, then $x \circ y$ is also an even permutation, so that:

$$\phi(x \circ y) = 0 = 0 +_2 0 = \phi(x) +_2 \phi(y) \tag{15.0.4}$$

If instead x, y are both odd permutations, then $x \circ y$ must instead be an even permutation, so that:

$$\phi(x \circ y) = 0 = 1 +_2 1 = \phi(x) +_2 \phi(y) \tag{15.0.5}$$

Finally, if x, y have opposite parity, then $x \circ y$ is an odd permutation, so that:

$$\phi(x \circ y) = 1 = 0 +_2 1 = \phi(x) +_2 \phi(y) \tag{15.0.6}$$

Therefore, ϕ is a homomorphism. ◀

Example. Let us show that the following mapping is not a homomorphism:

$$\phi : (\mathrm{GL}(n, \mathbb{R}), \times) \longrightarrow (\mathrm{GL}(n, \mathbb{R}), \times) \tag{15.0.7}$$
$$\mathsf{A} \longmapsto \mathsf{A}^{-1} \tag{15.0.8}$$

Consider $\mathsf{A}, \mathsf{B} \in \mathrm{GL}(n, \mathbb{R})$ then:

$$\phi(\mathsf{A}) \times \phi(\mathsf{B}) = \mathsf{A}^{-1} \times \mathsf{B}^{-1} \tag{15.0.9}$$

whereas

$$\phi(\mathsf{A} \times \mathsf{B}) = (\mathsf{A} \times \mathsf{B})^{-1} = \mathsf{B}^{-1} \times \mathsf{A}^{-1} \tag{15.0.10}$$

However note that $\mathsf{B}^{-1} \times \mathsf{A}^{-1} \neq \mathsf{A}^{-1} \times \mathsf{B}^{-1}$ generally. ◀

Proposition 20.2 *($\mathbb{Z}$ homomorphic to $\mathbb{Z}_n$)*
For $n \geq 2$, the following is a homomorphism to:

$$\phi : (\mathbb{Z}, +) \longrightarrow (\mathbb{Z}_n, +_n) \tag{15.0.11}$$
$$k \longmapsto k_{(\mathrm{mod}\ n)} \tag{15.0.12}$$

where $k_{(\mathrm{mod}\ n)}$ 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:

$$\phi(r + s) = (r + s)_{(\mathrm{mod}\ n)} \tag{15.0.13}$$
$$\equiv r + s \pmod{n} \tag{15.0.14}$$
$$\equiv r_{(\mathrm{mod}\ n)} + s_{(\mathrm{mod}\ n)n} \pmod{n} \tag{15.0.15}$$
$$\equiv r_{(\mathrm{mod}\ n)} +_n s_{(\mathrm{mod}\ n)n} \pmod{n} \tag{15.0.16}$$
$$= \phi(r) +_n \phi(s) \tag{15.0.17}$$

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. ∎

Example. Let us prove that:

$$\phi : (G, \circ) \longrightarrow (G, \circ) \tag{15.0.18}$$
$$x \longmapsto x \circ x \tag{15.0.19}$$

is a homomorphism *iff* $(G, \circ)$ is abelian.

Suppose ϕ is indeed a homomorphism, so that:

$$\phi(x \circ y) = (x \circ y) \circ (x \circ y) = (x \circ x) \circ (y \circ y) = \phi(x) \circ \phi(y) \tag{15.0.20}$$

$$\Longleftrightarrow \ x \circ y \circ x \circ y = x \circ x \circ y \circ y \tag{15.0.21}$$

$$\Longleftrightarrow \ y \circ x = x \circ y \tag{15.0.22}$$

This implies that $(G, \circ)$ is indeed an abelian group, as desired. ◀

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:

$$\phi : (G, \circ) \longrightarrow (H, *) \tag{15.0.23}$$

$$x \longmapsto e_H \tag{15.0.24}$$

Proof. Let $x, y \in G$, then:
$$\phi(x \circ y) = e_H = e_H * e_H = \phi(x) * \phi(y) \tag{15.0.25}$$

as desired. ∎

Proposition 20.4 *(Properties of homomorphisms)*
Let $\phi : (G, \circ) \longrightarrow (H, \odot)$ be a homomorphism, then
  (i) let $x_1, x_2, ..., x_n \in G$, then $\phi\left(\underset{k=1}{\overset{n}{\bigcirc}} x_k \right) = \underset{k=1}{\overset{n}{\bigodot}} \phi(x_k)$
  (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:

$$\phi\left(\underset{k=1}{\overset{n}{\bigcirc}} x_k \right) = \underset{k=1}{\overset{n}{\bigodot}} \phi(x_k) \tag{15.0.26}$$

  Then:

$$\phi\left(\underset{k=1}{\overset{n+1}{\bigcirc}} x_k \right) = \phi\left(\left(\underset{k=1}{\overset{n}{\bigcirc}} x_k \right) \circ x_{n+1} \right) \tag{15.0.27}$$

$$= \left(\phi\left(\underset{k=1}{\overset{n}{\bigcirc}} x_k \right) \right) \odot \phi(x_{n+1}) \tag{15.0.28}$$

$$= \underset{k=1}{\overset{n}{\bigodot}} \phi(x_k) \odot \phi(x_{n+1}) \tag{15.0.29}$$

$$= \underset{k=1}{\overset{n+1}{\bigodot}} \phi(x_k) \tag{15.0.30}$$

as desired. Hence by the principle of mathematical induction, we have that $\phi\left(\bigcirc_{k=1}^{n} x_k\right) = \bigodot_{k=1}^{n} \phi(x_k)$.

(ii) We have that $e_G \circ e_G = e_G$, then:

$$\phi(e_G \circ e_G) = \phi(e_G) \tag{15.0.31}$$

giving:

$$\phi(e_G) \odot \phi(e_G) = \phi(e_G) = \phi(e_G) \odot e_H \tag{15.0.32}$$

from which we find that $\phi(e_G) = e_H$.

(ii) Let $x \in G$. Then:

$$x \circ x^{-1} = x^{-1} \circ x = e_G \tag{15.0.33}$$

Then, applying ϕ:

$$\phi(x \circ x^{-1}) = \phi(x^{-1} \circ x) = \phi(e_G) = e_H \tag{15.0.34}$$

Therefore:

$$\phi(x) \odot \phi(x^{-1}) = \phi(x) \odot \phi(x^{-1}) = e_G \tag{15.0.35}$$

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:

$$\phi(x^0) = \phi(e_G) = (\phi(x))^0 = e_H \tag{15.0.36}$$

which was proven previously. Now suppose that for some integer $k \geq 0$

$$\phi(x^k) = (\phi(x))^k \tag{15.0.37}$$

Then:

$$\phi(x^{k+1}) = \phi(x^k \circ x) = \phi(x^k) \odot \phi(x) \tag{15.0.38}$$
$$= (\phi(x))^k \odot \phi(x) \tag{15.0.39}$$
$$= (\phi(x))^{k+1} \tag{15.0.40}$$

as desired.

Now suppose $n < 0$, and let $x \in G$. We can write $n = -m$ where $m > 0$. Them:

$$\phi(x^n) = \phi(x^{-m}) \tag{15.0.41}$$
$$= \phi((x^{-1})^m) \tag{15.0.42}$$
$$= (\phi(x^{-1}))^m \tag{15.0.43}$$
$$= (\phi(x))^{-m} \tag{15.0.44}$$
$$= (\phi(x))^n \tag{15.0.45}$$

Therefore, both from cases 1 and 2, we find that:

$$\phi(x^n) = (\phi(x))^n \tag{15.0.46}$$

as desired.

∎

> **Theorem 20.5** *(Order of element and homomorphism image)*
> Let $\phi : (G, \circ) \longrightarrow (H, *)$ 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:

$$e = x^r \tag{15.0.47}$$
$$= x^{as+b} \tag{15.0.48}$$
$$= (x^s)^a \circ x^b \tag{15.0.49}$$
$$= e^a \circ x^b \tag{15.0.50}$$
$$= x^b \tag{15.0.51}$$

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:

$$(\phi(x))^s = \phi(x^s) = \phi(e_G) = e_H \tag{15.0.52}$$

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 $\phi : (G, \circ) \longrightarrow (H, *)$ 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:

$$y = g \circ x \circ g^{-1} \tag{15.0.53}$$

Therefore:

$$\phi(y) = \phi(g \circ x \circ g^{-1}) = \phi(g) * \phi(x) * \phi(g^{-1}) = \phi(g) * \phi(x) * (\phi(g))^{-1} \tag{15.0.54}$$

so that $\phi(x), \phi(y)$ are also conjugate. ∎

15.1 Image and kernels

Definition 20.7 *(Image and kernel of homomorphism)*
Let $\phi(G, \circ) \longrightarrow (H, *)$ be a homomorphism. Then the **image** of ϕ is:

$$\text{Im} \, \phi = \{\phi(g) : g \in G\} \tag{15.1.1}$$

and the **kernel** of ϕ is:

$$\ker \phi = \{g \in G : \phi(g) = e_H\} \tag{15.1.2}$$

Theorem 20.8 *(Image subgroup of homomorphism)*
Let $\phi : (G, \circ) \longrightarrow (H, *)$ 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:

$$h_1 * h_2 = \phi(g_1) * \phi(g_2) = \phi(g_1 \circ g_2) \tag{15.1.3}$$

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:

$$h = \phi(g) \implies h^{-1} = (\phi(g))^{-1} = \phi(g^{-1}) \tag{15.1.4}$$

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 $\phi : (G, \circ) \longrightarrow (H, *)$ 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 odomain of this mapping. Consequently, φ is an isomorphism $(G, \circ) \longrightarrow \text{Im} \, \phi$, as desired. ∎

Theorem 20.10 *(Preserved structures under homomorphism)*
Let $\phi : (G, \circ) \longrightarrow (H, *)$ 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 $h_1, h_2 \in \operatorname{Im} \phi$, so that $\exists g_1, g_2$ such that $h_1 = \phi(g_1)$ and $h_2 = \phi(g_2)$. Therefore:

$$\phi(g_1 \circ g_2) = h_1 * h_2 = \phi(g_2 \circ g_1) = h_2 * h_2 \tag{15.1.5}$$

proving that H is also abelian.

(ii) Suppose that $\langle a \rangle = G$, and let $h \in \operatorname{Im} \phi$, so that $h = \phi(g)$ for some $g \in G$. Now since G is generated by a we find that:

$$g = a^k \implies \phi(g) = \phi(a^k) = (\phi(a))^k \tag{15.1.6}$$

Consequently:

$$h = (\phi(a))^k \tag{15.1.7}$$

proving that $\operatorname{Im} \phi$ is generated by $\phi(a)$, and thus also cyclic.

∎

> **Theorem 20.11** *(Kernel normal subgroup)*
> Let $\phi : (G, \circ) \longrightarrow (H, *)$ be a homomorphism. Then $\operatorname{Ker}(\phi) \trianglelefteq (G, \circ)$.

Proof. We firstly need to prove that $\operatorname{Ker}(\phi)$ is a subgroup of $(G, \circ)$ by checking the subgroup properties.

Closure: let $k_1, k_2 \in \operatorname{Ker}(\phi)$. Then $\phi(k_1) = e_H$ and $\phi(k_2) = e_H$, so that:

$$\phi(k_1 \circ k_2) = \phi(k_1) * \phi(k_2) = e_H * e_H = e_H \tag{15.1.8}$$

Hence $k_1 \circ k_2 \in \operatorname{Ker}(\phi)$.

Identity: we have $\phi(e_G) = e_H$, so that $e_G \in \operatorname{Ker}(\phi)$.

Inverses: let $k \in \operatorname{Ker}(\phi)$, then $\phi(k) = e_H$ so that:

$$\phi(k^{-1}) = (\phi(k))^{-1} = e_H^{-1} = e_H \tag{15.1.9}$$

so that $k^{-1} \in \operatorname{Ker}(\phi)$.

So, we have that $(\operatorname{Ker}(\phi), \circ) \leq (G, \circ)$.

To prove normality, we must prove that $g \circ k \circ g^{-1} \in \operatorname{Ker}(\phi)$ for $k \in \operatorname{Ker}(\phi)$ and $g \in G$. Indeed:

$$\phi(g \circ k \circ g^{-1}) = \phi(g) * \phi(k) * \phi(g^{-1}) \tag{15.1.10}$$
$$= \phi(g) * e_H * (\phi(g))^{-1} \tag{15.1.11}$$
$$= e_H \tag{15.1.12}$$

as desired, we find that $g \circ k \circ g^{-1} \in \operatorname{Ker}(\phi)$. Hence $(\operatorname{Ker}(\phi), \circ) \trianglelefteq (G, \circ)$ ∎

> **Theorem 20.12** *(Injectivity of homomorphism)*
> Let $\phi : (G, \circ) \longrightarrow (H, *)$ be a homomorphism. Then ϕ is injective *iff* $\operatorname{Ker}(\phi) = \{e_G\}$.

Proof. We begin by proving $\implies$. 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 $\impliedby$ part. Suppose that $\text{Ker}(\phi) = \{e_G\}$, and let $\phi(x) = \phi(y)$ for some $x, y \in G$. Then:

$$e_H = \phi(x) * (\phi(y))^{-1} \tag{15.1.13}$$
$$= \phi(x) * \phi(y^{-1}) \tag{15.1.14}$$
$$= \phi(x \circ y^{-1}) \tag{15.1.15}$$

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. ∎

> **Theorem 20.13** *(Normality $\iff$ kernel)*
> Let $K \leq G$, then $K \unlhd G \iff K = \text{Ker}(\phi)$ 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 $\impliedby$. Suppose that K is a normal subgroup of G. Then it is the kernel of the map ϕ defined by:

$$\phi : (G, \circ) \longrightarrow (G \setminus K, \cdot) \tag{15.1.16}$$
$$x \mapsto xK \tag{15.1.17}$$

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 $x, y \in G$:

$$\phi(x \circ y) = (x \circ y)K \tag{15.1.18}$$
$$= (xK) \cdot (yK) \tag{15.1.19}$$
$$= \phi(x) \cdot \phi(y) \tag{15.1.20}$$

Also:
$$\text{Ker}(\phi) = \{x \in G : \phi(x) = K\} = \{x \in G : xK = K\} = K \tag{15.1.21}$$

∎

15.2   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 $g \in G$:
>
> $$\phi(x) = \phi(y) \iff x, y \in g\text{Ker}(\phi) \tag{15.2.1}$$

Proof. Firstly, suppose that $\phi(x) = \phi(y)$ for any $x, y \in G$. Then:

$$\phi(x) * \phi(y^{-1}) = \phi(x \circ y^{-1}) = e_H \tag{15.2.2}$$

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 $k_1, k_2 \in \text{Ker}(\phi)$:

$$x, y \in g\text{Ker}(\phi) \implies x = g \circ k_1, y = g \circ k_2 \implies x = y \circ k_2^{-1} \circ k_1 \tag{15.2.3}$$

Then:

$$\phi(x) = \phi(y \circ k_2^{-1} \circ k_1) = \phi(y) * (\phi(k_2))^{-1} * \phi(k_1) = \phi(y) \tag{15.2.4}$$

as desired. ∎
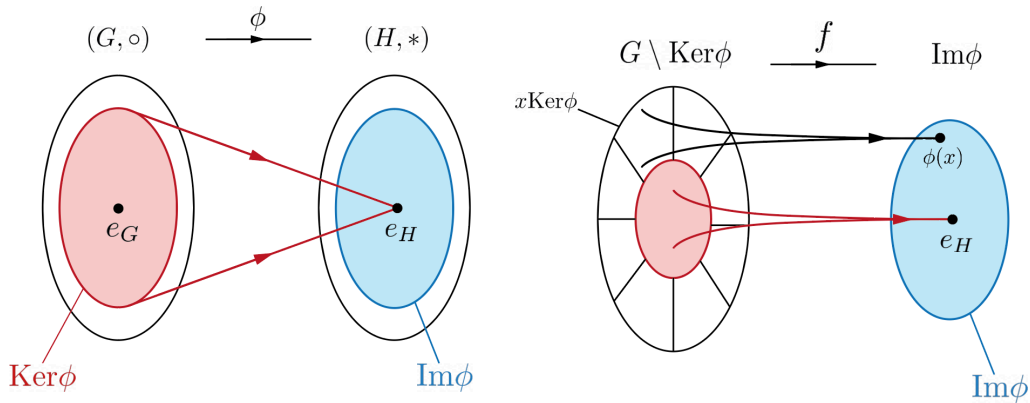
Figure 15.2. Set diagram of first isomorphism theorem

> **Theorem 20.15** *(First isomorphism theorem)*
> Let $\phi(G, \circ) \longrightarrow (H, *)$ be a homomorphism. Then:
>
> $$f : G \setminus \text{Ker}(\phi) \longrightarrow \text{Im}(\phi) \tag{15.2.5}$$
> $$x\text{Ker}(\phi) \longmapsto \phi(x) \tag{15.2.6}$$
>
> 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:

$$x \in y\text{Ker}(\phi) = \{yk_1, yk_2, ...\} \tag{15.2.7}$$
$$y \in x\text{Ker}(\phi) = \{xk_1, xk_2, ...\} \tag{15.2.8}$$

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 $xk_n = yk_ik_n \in y\text{Ker}(\phi)$, 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:

$$f(xK \cdot yK) = f((x \circ y)K) \tag{15.2.9}$$
$$= \phi(x \circ y) \tag{15.2.10}$$
$$= \phi(x) * \phi(y) \tag{15.2.11}$$
$$= f(xK) * f(yK) \tag{15.2.12}$$

as desired. It follows that f is a bijective homomorphism, hence an isomorphism, so that $G \setminus \mathrm{Ker}(\phi) \cong \mathrm{Im}(\phi)$. ∎

Example. Consider the following mapping ϕ:

$$\phi : (L, \times) \longrightarrow (\mathbb{R}^*, \times) \tag{15.2.13}$$

$$\begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \longmapsto ac \tag{15.2.14}$$

where L is the group of lower triangular 2×2 matrices.
This is clearly a homomorphism, since for any $\mathsf{A}, \mathsf{B} \in L$:

$$\mathsf{A} = \begin{pmatrix} a_1 & 0 \\ b_1 & c_1 \end{pmatrix}, \ \mathsf{B} = \begin{pmatrix} a_2 & 0 \\ b_2 & c_2 \end{pmatrix} \tag{15.2.15}$$

then:

$$\mathsf{AB} = \begin{pmatrix} a_1 a_2 & 0 \\ a_2 b_1 + c_1 b_2 & c_1 c_2 \end{pmatrix} \tag{15.2.16}$$

Hence:

$$\phi(\mathsf{AB}) = \phi \begin{pmatrix} a_1 a_2 & 0 \\ a_2 b_1 + c_1 b_2 & c_1 c_2 \end{pmatrix} \tag{15.2.17}$$

$$= (a_1 a_2)(c_1 c_2) \tag{15.2.18}$$

$$= \phi(\mathsf{A})\phi(\mathsf{B}) \tag{15.2.19}$$

as desired.
The image of ϕ is:

$$\mathrm{Im}(\phi) = \{\phi(\mathsf{A}) : \mathsf{A} \in L\} = \{ac : a, c \in \mathbb{R}^*\} = \mathbb{R}^* \tag{15.2.20}$$

The kernel of ϕ is:

$$\mathrm{Ker}(\phi) = \{\mathsf{A} \in L : \det\{\mathsf{A}\} = 1\} = \left\{ \begin{pmatrix} a & 0 \\ b & \frac{1}{a} \end{pmatrix} : a \in \mathbb{R}^*, b \in \mathbb{R} \right\} \tag{15.2.21}$$

By the first isomorphism theorem:

$$L \setminus \mathrm{Ker}(\phi) \cong \mathrm{Im}(\phi) = (\mathbb{R}^*, \times) \tag{15.2.22}$$

◀

Example. Consider the following map:

$$\phi : (L, \times) \longrightarrow (L, \times) \tag{15.2.23}$$

$$\begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \longmapsto \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} \tag{15.2.24}$$

This is clearly a homomorphism, since for any $A, B \in L$:

$$A = \begin{pmatrix} a_1 & 0 \\ b_1 & c_1 \end{pmatrix}, \ B = \begin{pmatrix} a_2 & 0 \\ b_2 & c_2 \end{pmatrix} \tag{15.2.25}$$

then:

$$AB = \begin{pmatrix} a_1 a_2 & 0 \\ a_2 b_1 + c_1 b_2 & c_1 c_2 \end{pmatrix} \tag{15.2.26}$$

Hence:

$$\phi(AB) = \begin{pmatrix} \frac{1}{a_1 a_2} & 0 \\ 0 & 1 \end{pmatrix} \tag{15.2.27}$$

$$= \begin{pmatrix} \frac{1}{a_1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{a_2} & 0 \\ 0 & 1 \end{pmatrix} \tag{15.2.28}$$

$$= \phi(A)\phi(B) \tag{15.2.29}$$

Moreover it is easy to see that:

$$\mathrm{Im}(\phi) = \left\{ \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} : a \in \mathbb{R}^* \right\} \tag{15.2.30}$$

and:

$$\mathrm{Ker}(\phi) = \left\{ \begin{pmatrix} 1 & 0 \\ b & c \end{pmatrix} : b \in \mathbb{R}, c \in \mathbb{R}^* \right\} \tag{15.2.31}$$

Now, by the first isomorphism theorem, we have that:

$$L \setminus \mathrm{Ker}(\phi) \cong \left\{ \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} : a \in \mathbb{R}^* \right\} \tag{15.2.32}$$

We now prove that $\mathrm{Im}(\phi) \in (\mathbb{R}^*, \times)$. An example of an isomorphism between them is:

$$\varphi : \mathrm{Im}(\phi) \longrightarrow (\mathbb{R}^*, \times) \tag{15.2.33}$$

$$\begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix} \mapsto a \tag{15.2.34}$$

This is a homomorphism since for $A, B \in \mathrm{Im}(\phi)$:

$$A = \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & 1 \end{pmatrix}, \ B = \begin{pmatrix} \frac{1}{b} & 0 \\ 0 & 1 \end{pmatrix} \tag{15.2.35}$$

with $a \neq 0$, $b \neq 0$, we have that

$$\varphi(\mathsf{AB}) = ab = \varphi(\mathsf{A})\varphi(\mathsf{B}) \tag{15.2.36}$$

Moreover, φ is injective, since $\varphi(\mathsf{A}) = \varphi(\mathsf{B}) \implies a = b \implies \mathsf{A} = \mathsf{B}$. Finally, φ is surjective, since every $x \in \mathbb{R}^*$ is the map of the matrix:

$$\mathsf{X} = \begin{pmatrix} \frac{1}{x} & 0 \\ 0 & 1 \end{pmatrix} \tag{15.2.37}$$

Hence, we may conclude that $L \setminus \mathrm{Ker}(\phi) \cong (\mathbb{R}^*, \times)$.   ◀

Proposition 20.16 *(Order of kernel, image and group)* Let $\phi : (G, \circ) \longrightarrow (H, *)$ be a homomorphism with a finite group domain, then:

$$|\mathrm{Ker}(\phi)| \cdot |\mathrm{Im}(\phi)| = |G| \tag{15.2.38}$$

Proof. From the first isomorphism theorem, since ϕ is a homomorphism:

$$G \setminus \mathrm{Ker}(\phi) \cong \mathrm{Im}(\phi) \tag{15.2.39}$$

and since isomorphic finite groups must have same order:

$$|G \setminus \mathrm{Ker}(\phi)| = |\mathrm{Im}(\phi)| \tag{15.2.40}$$

Now since G is finite dimensional, we must have that $|G \setminus \mathrm{Ker}(\phi)| = \frac{|G|}{|\mathrm{Ker}(\phi)|}$ so that:

$$|\mathrm{Ker}(\phi)| \cdot |\mathrm{Im}(\phi)| = |G| \tag{15.2.41}$$

as desired.   ■

To summarize, we have that for finite groups G, H, and any homomorphism ϕ between them:

(i) $|\mathrm{Ker}(\phi)|$ divides $|G|$ by Lagrange's theorem, since $\mathrm{Ker}(\phi) \leq G$

(ii) $|\mathrm{Im}(\phi)|$ divides $|H|$ by Lagrange's theorem, since $\mathrm{Im}(\phi) \leq H$

(iii) $|\mathrm{Im}(\phi)|$ 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 $|\mathrm{Ker}(\phi)|$ and $|\mathrm{Im}(\phi)|$ divide $|S(\triangle)| = 6$. Hence they can have values of 1,2,3,6. Moreover, we must have that $|\mathrm{Im}(\phi)|$ divides 3, hence it can only take values 1,3, for which $|\mathrm{Ker}(\phi)|$ takes the values of 6,2 respectively.
Now we know that $|\mathrm{Ker}(\phi)|$ is a normal subgroup of $S(\Delta)$, and we found no normal subgroups of order 2. Hence $|\mathrm{Ker}(\phi)| = 6 = |S(\Delta)|$ and $\mathrm{Im}(\phi) = 1$. Consequently, since $\mathrm{Ker}(\phi) \leq S(\Delta)$ and $\mathrm{Im}(\phi) \leq \mathbb{Z}_3$, we have that $\mathrm{Ker}(\phi) = |S(\Delta)|$ and $\mathrm{Im}(\phi) = \{0\}$. The only possible ϕ with such structures is the trivial homomorphism.   ◀

GT8 Group actions

16.1   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**.
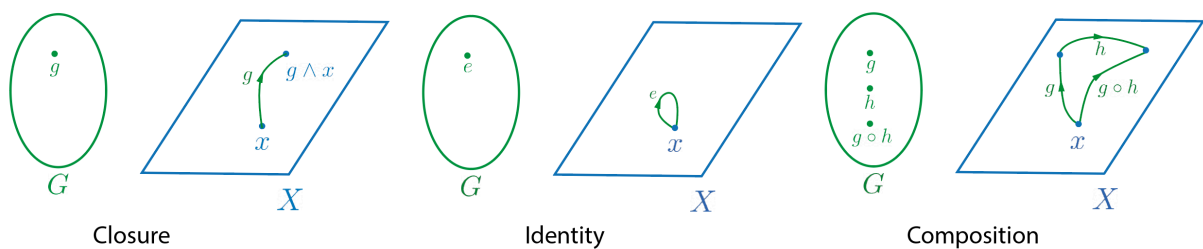
Figure 16.1. Visual illustration of group action axioms.

> **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:

$$g \wedge x = g(x) \tag{16.1.1}$$

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 $\mathrm{id}(x) = \mathrm{id} \wedge x = x$ for all $x \in X$.

GA3 For $g, h \in G$ and $x \in X$ we have that:

$$g \wedge (h \wedge x) = g \wedge h(x) \tag{16.1.2}$$
$$= g(h(x)) = (g \circ h)(x) \tag{16.1.3}$$
$$= (g \circ x) \wedge x \tag{16.1.4}$$

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:

$$g \wedge (x, y) = (x + g, y + g) \tag{16.1.5}$$

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:

$$g \wedge (h \wedge x) = g \wedge (x + h, y + x) = (x + h + g, y + h + g) \tag{16.1.6}$$

whereas:

$$(g \times h) \wedge x = (gh) \wedge x = (x + gh, y + gh) \neq (x + h + g, y + h + g) \tag{16.1.7}$$

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 $g^{-1} \in G$:

$$g^{-1} \wedge (g \wedge x) = g^{-1} \wedge (g \wedge y) \implies x = y \tag{16.1.8}$$

using the composition axiom.

Next, let $g \in G$ and $y \in X$. Then:

$$e \wedge y = y \implies g \wedge (g^{-1} \wedge y) = y \tag{16.1.9}$$

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. ∎

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:

$$g \wedge A = g(A), \forall g \in G, A \in X \tag{16.1.10}$$

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:

$$g \wedge (h \wedge A) = g(h(A)) = (g \circ h)A \tag{16.1.11}$$

by definition of the composition operation. ∎

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:

$$R \qquad S \qquad T \qquad U$$

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	$(R\,T)$

◀

16.2 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:
$$\text{Orb } x = \{g \wedge x : g \in G\} \subseteq X \tag{16.2.1}$$

that is, the set of elements in X which are the image of x under g.
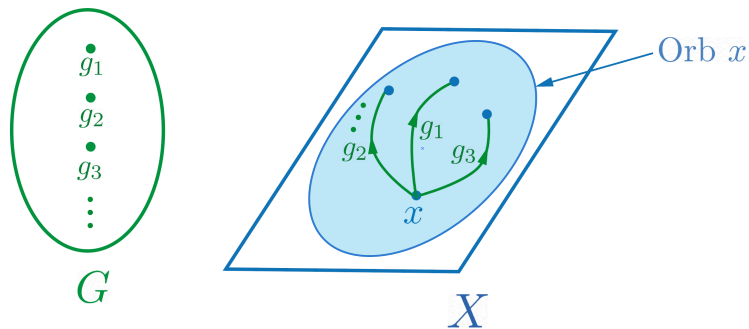
Figure 16.2. Visualization of the orbit of some x under $\wedge$.

Example. We consider the action of $S(\square)$ on the set $\{1, 2, 3, 4\}$ of labelled vertices of a square. Then:

$$\text{Orb } 1 = \{a \wedge 1, b \wedge 1, c \wedge 1, r \wedge 1, s \wedge 1, t \wedge 1, u \wedge 1\} \tag{16.2.2}$$

$$= \{2, 3, 4, 4, 1, 2, 3\} = \{1, 2, 3, 4\} \tag{16.2.3}$$

Similarly:

$$\text{Orb } 2 = \{a \wedge 2, b \wedge 2, c \wedge 2, r \wedge 2, s \wedge 2, t \wedge 2, u \wedge 2\} \tag{16.2.4}$$

$$= \{3, 4, 1, 3, 4, 1, 2\} = \{1, 2, 3, 4\} \tag{16.2.5}$$

It is easy to verify that Orb $3 =$ 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 Orb P definitely contains the circle $\mathcal{C}_P$ 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 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:

$$x \sim y \text{ if } y \in \text{Orb } x \tag{16.2.6}$$

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:

$$g^{-1} \wedge y = (g^{-1} \circ g) \wedge x = x \implies x \in \text{Orb } y \tag{16.2.7}$$

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:

$$y = g_1 \wedge x, \text{ and } z = g_2 \wedge y \tag{16.2.8}$$

for some $g_1, g_2 \in G$. Then:

$$z = g_2 \wedge (g_1 \wedge x) = (g_2 \circ g_1) \wedge x \implies z \in \text{Orb } x \tag{16.2.9}$$

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:

$$[x] = \{y : y \in \text{Orb } x\} = \text{Orb } x \tag{16.2.10}$$

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:

$$A_1 \qquad A_2 \qquad A_3$$

$$A_4 \qquad A_5 \qquad A_6$$

$$A_7 \qquad A_8 \qquad A_9$$
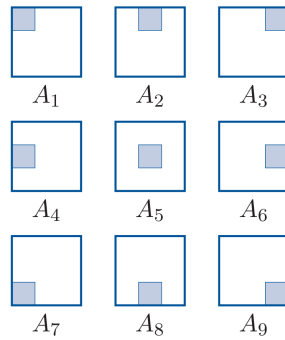
We begin by writing down Orb A_1:

$$\text{Orb } A_1 = \{A_1, A_7, A_9, A_3\} \tag{16.2.11}$$

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, :

$$\text{Orb } A_2 = \{A_2, A_4, A_8, A_6\} \tag{16.2.12}$$

The only remaining element of X is A_5, so its orbit must only contain itself:

$$\text{Orb } A_5 = \{A_5\} \tag{16.2.13}$$

Hence the orbit partition of X is:

$$X = \{A_1, A_3, A_7, A_9\} \cup \{A_2, A_4, A_6, A_8\} \cup \{A_5\} \tag{16.2.14}$$

◀

Example. Consider the matrix group

$$G = \left\{ \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} : a, b \in \mathbb{R}^+ \right\} \tag{16.2.15}$$

and the group action $\wedge$ on $\mathbb{R}^2$ defined by:

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \wedge (x, y) = (ax, by) \tag{16.2.16}$$

for all $(x, y) \in \mathbb{R}^2$. It follows that:

$$\text{Orb}\,(x, y) = \{(ax, by) : a, b \in \mathbb{R}^+\} \tag{16.2.17}$$

For example:

$$\text{Orb}\,(1, 0) = \{(a, 0) : a \in \mathbb{R}^+\} \tag{16.2.18}$$

which is the positive x-axis excluding the origin. By symmetry, Orb $(-1, 0)$ must be the negative x-axis excluding the origin.
Similarly:

$$\text{Orb}\,(0, -1) = \{(0, -b) : b \in \mathbb{R}^+\} \tag{16.2.19}$$

which is the negative y-axis excluding the origin. By symmetry, 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:

$$\text{Orb}\,(0, 0) = \{(0, 0)\} \tag{16.2.20}$$

We are missing the four quadrants of $\mathbb{R}^2$. Note that

$$\text{Orb}\,(1, 1) = \{(a, b) \in \mathbb{R}^+\} \tag{16.2.21}$$

which is the upper right quadrant excluding the origin and the axes. By symmetry, Orb $(1, -1)$ must be the lower right quadrant, Orb $(-1, 1)$ must be the upper left quadrant and Orb $(-1, -1)$ must be the lower left quadrant. ◀

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:
$$\text{Stab}\,x = \{g \in G : g \wedge x = x\} \tag{16.2.22}$$

We may interpret the stabiliser geometrically as shown below:

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 Stab $P = S(\bigcirc)$
   (ii) P is not the origin: then Stab $P = \{e, q\}$ contains e and the reflection in the line containing P and O, which we call q.

◀
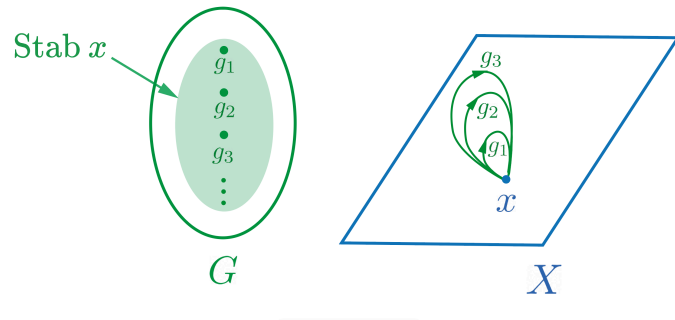
Figure 16.3. Geometrical interpretation of the stabiliser

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 $x \in X$, Stab $x \leq G$.

Proof. We show that the subgroup axioms are satisfied:

Closure: let $g, h \in$ Stab x. Then $(g \circ h) \wedge x = g \wedge (h \wedge x) = g \wedge x = x$ so $g \circ h \in$ Stab x as desired.

Identity: let e be the identity of G. Then $e \wedge x = x$ by the group action axioms, hence $e \in$ Stab x.

Inverses: let $g \in$ Stab x, then $g \wedge x = x \implies x = g^{-1} \wedge x$ so that $g^{-1} \in$ Stab x.

$\blacksquare$
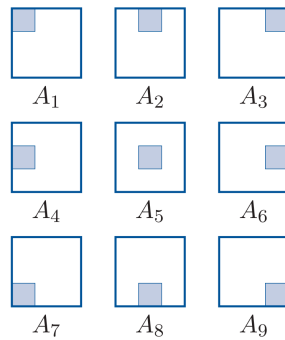
> **Example.** Consider the action of $S(\square)$ on the set $X = \{A_i : 1 \leq i \leq 9\}$ of squares shown below:
>
>
>
> Then:
>
> $$\text{Stab } A_1 = \{e, s\} = \text{Stab } A_9 \qquad (16.2.23)$$
> $$\text{Stab } A_2 = \{e, r\} = \text{Stab } A_8 \qquad (16.2.24)$$
> $$\text{Stab } A_3 = \{e, u\} = \text{Stab } A_7 \qquad (16.2.25)$$
> $$\text{Stab } A_4 = \{e, t\} = \text{Stab } A_6 \qquad (16.2.26)$$
> $$\text{Stab } A_5 = S(\square) \qquad (16.2.27)$$

These are all subgroups of $S(\square)$, since the first four lines contain e and a reflection (which are self inverse). ◀

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:

$$\begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \wedge (x, y) = (ax, ay), \ \forall (x, y) \in \mathbb{R}^2 \tag{16.2.28}$$

Then, we have that if (x, y) is not the origin:

$$\text{Stab }(x, y) = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \in G : (ax, ay) = (x, y) \right\} = \left\{ \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} : b \in \mathbb{R} \right\} \tag{16.2.29}$$

If instead $(x, y) = (0, 0)$ then:

$$\text{Stab }(0, 0) = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \in G : (a \cdot 0, a \cdot 0) = (0, 0) \right\} = G \tag{16.2.30}$$

◀

16.3 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:

$$g \wedge x = h \wedge x \iff g, h \text{ lie in the same left coset of Stab } x \tag{16.3.1}$$

Proof. We firstly prove $\implies$. Indeed, suppose g, h lie in the same left coset of Stab x, so that $h \in$ Stab x. Then $\exists k \in$ Stab x such that $h = g \circ k$ and thus:

$$h \wedge x = g \wedge k \wedge x = g \wedge x \tag{16.3.2}$$

as desired.

Now suppose that $h \wedge x = g \wedge x$. Then:

$$(g^{-1} \circ h) \wedge x = g^{-1} \wedge (h \wedge x) \tag{16.3.3}$$
$$= g^{-1} \wedge (g \wedge x) \tag{16.3.4}$$
$$= (g^{-1} \circ g) \wedge x \tag{16.3.5}$$
$$= e \wedge x \tag{16.3.6}$$
$$= x \tag{16.3.7}$$

so that $g^{-1} \circ h \in$ Stab x, and consequently $h \in g$Stab x, so that h, g both lie in the same left coset. ∎
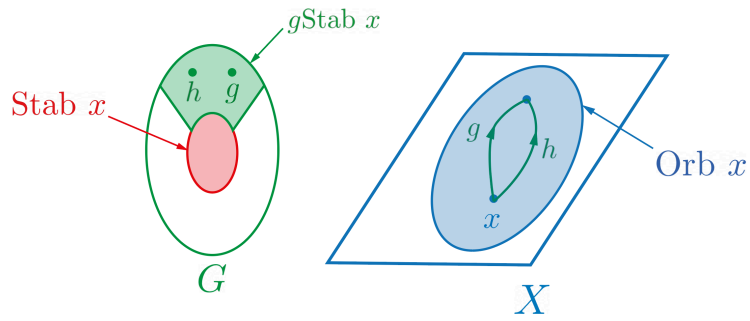
Figure 16.4. Visual interpretation of the left coset gStab x and its action on X

Example. We consider the action of S_3 on $\{1, 2, 3\}$, and find that:

$$\text{Stab } 1 = \{\sigma \in S_3 : \sigma(1) = 1\} = \{e, (2\ 3)\} \tag{16.3.8}$$

Then, the left cosets of this stabiliser are:

$$e\text{Stab } 1 = \text{Stab } 1 \tag{16.3.9}$$
$$(1\ 2)\text{Stab } 1 = \{(1\ 2), (1\ 2\ 3)\} \tag{16.3.10}$$
$$(1\ 3)\text{Stab } 1 = \{(1\ 3), (1\ 3\ 2)\} \tag{16.3.11}$$
$$(2\ 3)\text{Stab } 1 = \text{Stab } 1 \tag{16.3.12}$$
$$(1\ 2\ 3)\text{Stab } 1 = \{(1\ 2), (1\ 2\ 3)\} \tag{16.3.13}$$
$$(1\ 3\ 2)\text{Stab } 1 = \{(1\ 3), (1\ 3\ 2)\} \tag{16.3.14}$$

Now we partition S_3 according to where its elements map 1:

$$\{e, (2\ 3)\} \text{ map 1 to 1} \tag{16.3.15}$$
$$\{(1\ 2), (1\ 2\ 3)\} \text{ map 1 to 2} \tag{16.3.16}$$
$$\{(1\ 3), (1\ 3\ 2)\} \text{ map 1 to 3} \tag{16.3.17}$$

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:

$$\phi : G\text{Stab } x \longrightarrow \text{Orb } x \tag{16.3.18}$$
$$g\text{Stab } x \mapsto g \wedge x \tag{16.3.19}$$

is a bijective map.

Proof. Since elements of different cosets of Stab x have different images under ϕ, we find that ϕ

must be injective. Indeed, suppose that $\phi(g\text{Stab } x) = \phi(h \wedge x)$, so that $g \wedge x = h \wedge x$:

$$g \in h\text{Stab } x = \{hs_1, hs_2, ...\} \tag{16.3.20}$$

$$h \in g\text{Stab } x = \{gs_1, gs_2, ...\} \tag{16.3.21}$$

from which it follows that $g = hs_i$ and $h = gs_i^{-1} = gs_j$ for some $k_i, k_j \in \text{Stab } x$. 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:

$$\forall x \in X, \ |\text{Orb } x| \times |\text{Stab } x| = |G| \tag{16.3.22}$$

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:

$$|\text{Orb } x| \cdot |\text{Stab } x| = |G| \tag{16.3.23}$$

∎

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:

$$g \wedge x = gxg^{-1} \tag{16.3.24}$$

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:

$$g \wedge (h \wedge x) = g \wedge (hxh^{-1}) \tag{16.3.25}$$

$$= g(hxh^{-1})g^{-1} \tag{16.3.26}$$

$$= (gh)x(gh)^{-1} \tag{16.3.27}$$

$$= (gh) \wedge x \tag{16.3.28}$$

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:
>
> $$f \wedge (h \wedge g) = f \wedge (hx) \tag{16.3.29}$$
> $$= fhx \tag{16.3.30}$$
> $$= (fh)x \tag{16.3.31}$$
> $$= (fh) \wedge x \tag{16.3.32}$$
>
> 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:

$$\text{Orb } x = \{gxg^{-1} : g \in G\} = [x] \tag{16.3.33}$$

so the orbit of x is the conjugacy class of x. We also have that:

$$\text{Orb } x \text{ divides } |G| \tag{16.3.34}$$

giving the desired result. ∎

> **Proposition 21.13** *(Homomorphism group action)*
> Let $\phi : (G, \circ) \longrightarrow (H, *)$ be a homomorphism, and let $\wedge$ be defined as:
>
> $$g \wedge h = \phi(g) * h \tag{16.3.35}$$
>
> 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:

$$g \wedge (f \wedge h) = g \wedge (\phi(f) * h) \tag{16.3.36}$$
$$= \phi(g) * (\phi(f) * h) \tag{16.3.37}$$
$$= \phi(g \circ f) * h \tag{16.3.38}$$
$$= (g \circ f) \wedge h \tag{16.3.39}$$

as desired.

∎

Notice that for the homomorphism group action of G on H:

$$\text{Orb } e_H = \{\phi(g) * e_H : g \in G\} = \{\phi(g) : g \in G\} = \text{Im}(\phi) \tag{16.3.40}$$

and:

$$\text{Stab } e_H = \{g : \phi(g) * e_H = e_H\} = \{g : \phi(g) = e_H\} = \text{Ker}(\phi) \tag{16.3.41}$$

Applying the orbit stabiliser theorem:

$$|\text{Orb } e_H| \cdot |\text{Stab } e_H| = |\text{Im}(\phi)| \cdot |\text{Ker}(\phi)| = |G| \tag{16.3.42}$$

which is precisely the result proven in Proposition 20.16.

16.4   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:
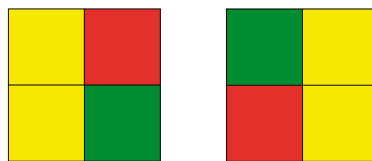
Figure 16.5. Two identical patterns which were double counted.

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 symemtry 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:
> $$\text{Fix } g = \{x \in X : g \wedge x = x\} \tag{16.4.1}$$
> 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:

$$\begin{array}{cccc} A & B & C & D \end{array}$$
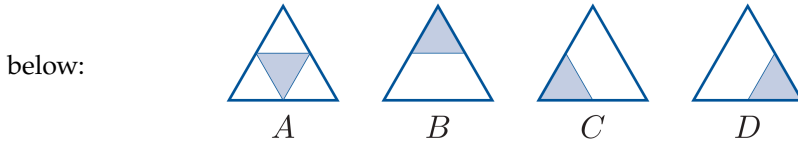
Then, we see that:

$$\text{Fix } e = X \tag{16.4.2}$$
$$\text{Fix } a = \text{Fix } b = \{A\} \tag{16.4.3}$$
$$\text{Fix } r = \{A, B\} \tag{16.4.4}$$
$$\text{Fix } s = \{A, C\} \tag{16.4.5}$$
$$\text{Fix } t = \{A, D\} \tag{16.4.6}$$

◀

Example. Let

$$G = \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} : a, b \in \mathbb{R}, \ a \neq 0 \right\} \tag{16.4.7}$$

and consider the action of G on $\mathbb{R}^2$ defined by:

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \wedge (x, y) = (ax, y) \tag{16.4.8}$$

Then, since $a \neq 0$ it follows that

$$\text{Fix } \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} = \{(x, y) \in \mathbb{R}^2 : (x, y) = (ax, y)\} = \{(0, y) : y \in \mathbb{R}\} \tag{16.4.9}$$

if $a \neq 1$ and

$$\text{Fix } \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} = \mathbb{R}^2 \tag{16.4.10}$$

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, Fix $e = X$, so that $|\text{Fix } e| = 5^4$.
> Moreover, 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 Fix a by c_1, c_2, c_3, c_4 representing their colors. If we rotate this pattern, we find that:
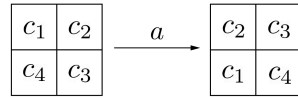
so that $c_1 = c_2, c_4 = c_1, c_3 = c_4, c_2 = c_3$ implying that all colors must be the same. There are 5 such squares.

By similar arguments, we can see that 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, 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:

$$\frac{1}{|G|} \sum_{g \in G} |\text{Fix } g| \tag{16.4.11}$$

Proof. Let t be the number of orbits, and let B be one such orbit. Then:

$$\sum_{x \in B} |\text{Stab } x| = \sum_{x \in B} \frac{|G|}{|\text{Orb } x|} \tag{16.4.12}$$

$$= |G| \sum_{x \in B} \frac{1}{|\text{Orb } x|} \tag{16.4.13}$$

$$= |G| \sum_{x \in B} \frac{1}{|B|} \tag{16.4.14}$$

$$= |G| \cdot |B| \cdot \frac{1}{|B|} \tag{16.4.15}$$

$$= |G| \tag{16.4.16}$$

Therefore, since orbits partition X:

$$\sum_{x \in X} |\text{Stab } x| = t|G| \iff t = \frac{1}{|G|} \sum_{x \in X} |\text{Stab } x| \qquad (16.4.17)$$

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 $\sum_{g \in G} |\text{Fix } g|$.

■

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 $S(\square)$:

$$\frac{1}{8}(5^4 + 5 + 5^2 + 5 + 5^2 + 5^3 + 5^2 + 5^3) = 120 \qquad (16.4.18)$$

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:

$$\frac{1}{24}(3^6 + 6 \cdot 3^3 + 3 \cdot 3^4 + 8 \cdot 3^2 + 6 \cdot 3^3) = 57 \qquad (16.4.19)$$

hence there are 57 different colored cubes. ◄

Part III

Representation Theory and Lie Algebra

Part IV

Differential Equations

ODE1 Fundamentals

17.1 Definitions

Definition 23.1 *(nth order ODE)*
An nth order ordinary differential equation (ODE) in $\mathbb{K}^N$ is an equation:

$$y^{(n)} = f(t, y, y'...y^{(n-1)}) \tag{17.1.1}$$

for $(t, y, y'...y^{(n-1)}) \in \Omega, t \in \mathbb{R}, y \in \mathbb{K}^n$ where $\Omega \subseteq \mathbb{R} \times (\mathbb{K}^\mathbb{N})^n$ and $f : \Omega \to \mathbb{K}^n$, with $n, N \in \mathbb{N}^*$.

We highlight the fundamental case where $n = 1$, in which case:

$$y' = f(t, y), \ (t, y) \in \Omega, \ t \in \mathbb{R}, \ y \in \mathbb{K}^N \tag{17.1.2}$$

Definition 23.2 *(Solution)*
A solution to an nth order ODE consists of a function $y : I \to \mathbb{K}^N$ n times differentiable such that:
  (i) $\forall t \in I, (t, y, y'...y^{(n-1)}) \in \Omega$
  (ii) $\forall t \in I, y^{(n)} = f(t, y, y'...y^{(n-1)})$

Definition 23.3 *(Linear and Homogeneous)*
An ODE is said to be **linear** if f is a polynomial function, that is:

$$f(t, y, y'...y^{(n-1)}) = \sum_{i=0}^{n} A_i(t) y^{(i)} \tag{17.1.3}$$

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 $(t, y, y'...y^{(n-1)}) \to (\lambda t, \lambda y, \lambda y'...\lambda y^{(n-1)})$ for $\lambda \in \mathbb{K}$.

Proposition 23.4 *(Smoothness of Solutions)*
If $f : \Omega \to \mathbb{K}^N$ 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}. ∎

17.2 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 $f : [a,b] \to \mathbb{R}$ and $\mathbb{K} : [a,b]^2 \to K$ are continuous, with $t \in [a,b]$ then:

$$\text{Volterra non-homogeneous} : y(t) = f(t) + \int_a^t K(t,s)y(s)ds \tag{17.2.1}$$

$$\text{Fredholm non-homogeneous} : y(t) = f(t) + \lambda \int_a^b K(t,s)y(s)ds \tag{17.2.2}$$

and the two corresponding homogeneous equations. We call $K(t,s)$ the **kernel** of the integral equation.

Note that for the Fredholm homogeneous equation:

$$y(t) = \lambda \int_a^b K(t,s)y(s)ds \tag{17.2.3}$$

we may consider this as an eigenfunction equation, with λ as an eigenvalue and y as an eigenfunction.

Lemma 1 Suppose that $f : [a,b] \to \mathbb{R}$ is continuous. Then:

$$\int_a^x \int_a^{x'} f(t)dtdx' = \int_a^x (x-t)f(t)dt \tag{17.2.4}$$

Proof. Define the integral transform $G : [a,b] \to \mathbb{R}$ by:

$$F(x) = \int_a^x (x-t)f(t)dt \tag{17.2.5}$$

then because $f(t)$ and $x-t$ are continuous we may use the Leibniz integral rule to find:

$$F'(x) = [(x-t)f(t)]_{t=x}\overset{0}{\frac{d}{dx}}(x) + \int_a^x \frac{\partial}{\partial x}[(x-t)f(t)]dt = \int_a^x f(t)dt \tag{17.2.6}$$

We then deduce from the fundamental theorem of Calculus that:

$$\int_a^{x'} F'(x)dx = F(x') - \overset{0}{F(a)} = \int_a^{x'} \int_a^x f(t)dtdx \tag{17.2.7}$$

Substituting $x \to x'$ then:

$$F(x) = \int_a^x \int_a^{x'} f(t)dtdx' \tag{17.2.8}$$

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):

$$y'(t) - y'(0) + \lambda \int_0^t y(s)ds = \int_0^t g(s)ds \tag{17.2.9}$$

Further integration gives:

$$y(t) - y(0) - ty'(0) + \lambda \int_0^t (t - s)f(s)ds = \int_0^t (t - s)g(s)ds \tag{17.2.10}$$

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 $(t, y, y'...y^{(n)})$ 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:

$$y(t) = At + \int_0^t (t - s)g(s)ds - \lambda \int_0^t (t - s)y(s)ds \tag{17.2.11}$$

   which is a Volterra non-homogeneous integral equation with $K(t, s) = \lambda(t - s)$ and $f(t) = At + \int_0^x (t - s)g(s)ds$.

2. **Boundary condition:** suppose $y(0) = 0$ and $y(L) = B$. Then we find upon inserting $t = L$ that:

$$y'(0) = \frac{1}{L}\left(\lambda \int_0^L (L - s)y(s)ds - \int_0^L (L - s)g(s)ds + B\right) \tag{17.2.12}$$

   and substituting back into the original integral equation we find:

$$y = \frac{Bt}{L} - \int_0^L \frac{t}{L}(L - s)g(s)ds + \int_0^t (t - s)g(s)ds + \lambda\left(\int_0^L \frac{t}{L}(L - s)g(s)ds - \int_0^t (t - s)g(s)ds\right) \tag{17.2.13}$$

   If we now define a function $K(s, t)$ such that:

$$\int_0^L K(s, t)f(s)ds = \int_0^L \frac{t}{L}(L - s)f(s)ds - \int_0^t (t - s)f(s)ds \tag{17.2.14}$$

   then we may write:

$$y = \underbrace{\frac{Bt}{L} - \int_0^L K(s, t)g(s)ds}_{f(x)} + \lambda\left(\int_0^L K(s, t)y(s)ds\right) \tag{17.2.15}$$

It turns out that the kernel is[1] :

$$K(s,t) = \begin{cases} \frac{s}{L}(L-t) & \text{when } 0 \le s \le t \le L \\ \frac{t}{L}(L-s) & \text{when } 0 \le t \le s \le L \end{cases} \tag{17.2.20}$$

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.

17.3 Picard iteration

Consider the Volterra integral equation:

$$y(t) = f(t) + \int_a^t K(s,t)y(s)ds \tag{17.3.1}$$

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:

$$\begin{cases} y_0 = f(t) \\ y_k(t) = f(t) + \int_a^t K(s,t)y_{k-1}(s)ds \end{cases} \tag{17.3.2}$$

Because $f(t)$ is continuous by hypothesis, so are all y_i for $i = 0, 1, 2...$. We now conjecture that:

> **Proposition 23.6** *(Picard iteration convergence)*
> We have that:
> $$|u_n(t)| = |y_n(t) - y_{n-1}(t)| \le M_n \tag{17.3.3}$$
> with $\sum_{n=1}^{\infty} M_n$ converging.

Proof. Since K, f are continuous over $[a,b]$, they must be bounded:

$$|K(s,t)| \le L, \ |f(t)| \le M \ \forall s, t \in [a,b] \tag{17.3.4}$$

---

[1]Indeed:

$$\int_0^L K(s,t)f(s)ds = \int_0^t K(s,t)f(s)ds + \int_t^L K(s,t)f(s)ds \tag{17.2.16}$$

$$= \int_0^t \frac{s}{L}(L-t)f(s)ds + \int_t^L \frac{t}{L}(L-s)f(s)ds \tag{17.2.17}$$

$$= \int_0^t \frac{s}{L}(L-t)f(s)ds + \int_0^L \frac{t}{L}(L-s)f(s)ds - \int_0^t \frac{t}{L}(L-s)f(s)ds \tag{17.2.18}$$

$$= \int_0^t (s-t)f(s)ds + \int_0^L \frac{t}{L}(L-s)f(s)ds \tag{17.2.19}$$

as required.

We can therefore write:

$$|y_1(t) - y_0(t)| = \left| \int_a^t K(s,t) y_0(s) ds \right| = \int_a^x |K(s,t)||f(s)| ds \le LM(x-a) \tag{17.3.5}$$

Let us now suppose that for some $n \ge 2$:

$$|y_{n-1}(t) - y_{n-2}(t)| \le L^{n-1} M \frac{(t-a)^{n-1}}{(n-1)!} \tag{17.3.6}$$

We then find:

$$|y_n(s) - y_{n-1}(s)| = \left| \int_a^t K(s,t)(y_{n-1}(s) - y_{n-2}(s)) ds \right| \tag{17.3.7}$$

$$\le \int_a^t |K(s,t)||y_{n-1}(s) - y_{n-2}(s)| ds \tag{17.3.8}$$

$$\le \int_a^t L^n M \frac{(s-a)^{n-1}}{(n-1)!} ds \tag{17.3.9}$$

$$= \le L^n M \frac{(t-a)^n}{n!} \tag{17.3.10}$$

as required.

We can therefore define M_n as:

$$|y_n(t) - y_{n-1}(t)| \le L^n M \frac{(t-a)^n}{n!} \le L^n M \frac{(b-a)^n}{n!} \equiv M_n \tag{17.3.11}$$

Consequently:

$$\sum_{n=1}^{\infty} M_n = M(e^{L(b-a)} - 1) \tag{17.3.12}$$

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. ∎

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:

$$|y(x) - y_n(x)| < \epsilon \ \forall n \ge N \tag{17.3.13}$$

This implies that:

$$|K(s,t)y(s) - K(s,t)y_n(s)| < L\epsilon \ \forall n \ge N \tag{17.3.14}$$

This is equivalent to saying that our iteration converges to the non-homogeneous Fredholm equation:

$$\int_a^t K(s,t)y_n(s) ds \longrightarrow \int_a^t K(s,t)y_{(}s) ds, \ \text{ as } n \to \infty \tag{17.3.15}$$

We have therefore shown the existence of a continuous solution, but what about its uniqueness?

Suppose there is another solution Y so that:

$$|y(t) - Y(t)| \leq P \tag{17.3.16}$$

Let us suppose inductively that:

$$|y(t) - Y(t)| \leq L^{n-1}\frac{(t-a)^{n-1}}{(n-1)!} \tag{17.3.17}$$

Then:

$$|y(t) - Y(t)| = \left| \int_a^t K(s,t)(y(s) - Y(s))ds \right| \tag{17.3.18}$$

$$\leq L^n P\frac{(t-a)^n}{n!} \tag{17.3.19}$$

$$\leq L^n P\frac{(b-a)^n}{n!} \tag{17.3.20}$$

As $n \to \infty$ 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:

$$\begin{cases} y_0 = f(t) \\ y_k(t) = f(t) + \lambda \int_a^t K(s,t)y_{k-1}(s)ds \end{cases} \tag{17.3.21}$$

In this case we find that:

$$|y_n(t) - y_{n-1}(t)| \leq |\lambda|^n L^n M(b-a)^n \tag{17.3.22}$$

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.

17.4 Existence and uniqueness

Consider the Cauchy problem:

$$y' = f(x,y), \ y(a) = c \tag{17.4.1}$$

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:

$$|f(x,y_1) - f(x,y_2)| \leq A|y_1 - y_2|, \ \forall(x,y_1), (x,y_2) \in U \tag{17.4.2}$$

(iii) Defining:

$$M = \sup\{|f(x,y)| : (x,y) \in R\} \tag{17.4.3}$$

then we require:

$$Mh \leq k \tag{17.4.4}$$

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 $|x - a| \leq \hbar$ a solution to the Cauchy problem:

$$y' = f(x, y), \ y(a) = c \tag{17.4.5}$$

and this solution is unique in U.

We will present the proof of a more general result later.

ODE2 First Order ODEs

18.1 Types of first order ODEs

In this chapter we will consider first order differential equations which can be written in **standard form**:

$$\frac{dy}{dx} = f(x, y) \tag{18.1.1}$$

where $f(x, y)$ is a function. Since f can always be written as the ratio of two functions $M(x, y)$ and $-N(x, y)$, a first order ODE in standard form can also be written equivalently in **differential form**:

$$M(x, y)dx + N(x, y)dy = 0 \tag{18.1.2}$$

The main first order ODEs we will be concerned with are:

(i) Separable ODEs: if $M(x, y) = M(x)$ and $N(x, y) = N(x)$, then we can separate the x and y variables and integrate them individually.

(ii) Exact ODEs: if $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$ then the differential form of a first order ODE can be written as the total differential of a function.

(iii) Special inexact ODEs: there are some cases where even though an ODE may not be exact, there are ways to bring it to a separable form by a change of variables.

(iv) Linear ODEs: if $f(x, y) = -p(x)y + q(x)$ then we can find an integrating factor that will separate variables.

(v) Bernoulli equations: linear ODEs but with $q(x)$ containing a x^r term for $r \in \mathbb{R}$.

Before investigating how these first order ODEs may be solved, we state an important theoretical result on the existence and uniqueness of solutions to first order ODEs.

> **Theorem** *(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: $\forall (t_0, y_0) \in \mathcal{D}, \exists Pt$ such that the Cauchy problem:
>
> $$\begin{cases} \dot{y} = f(t, y) \\ y(t_0) = y_0 \end{cases} \tag{18.1.3}$$
>
> 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.

18.2 Separable Differential Equations

Suppose that $M(x, y) = M(x)$ and $N(x, y) = N(y)$:

$$M(x)dx + N(y)dy = 0 \iff \frac{dy}{dx} = -\frac{M(x)}{N(y)} \tag{18.2.1}$$

These are perhaps the simplest to solve, and luckily crop up quite often in physics, especially in classical mechanics when we relate the rate of change of some quantity with another observable. To solve (18.2.1), we simply integrate:

$$\boxed{\int M(x)dx + \int N(y)dy = 0} \tag{18.2.2}$$

For example, consider an electron of charge $-e$, mass m orbiting a proton of charge e at a radius r satisfies the ODE:

$$\frac{dr}{dt} = -\frac{\mu_0 e^4}{12\pi^2 \epsilon_0 m^2 c} \frac{1}{r^2} \tag{18.2.3}$$

This is clearly a separable equation with $M(t) = -\frac{\mu_0 e^4}{12\pi^2 \epsilon_0 m^2 c}$ and $N(r) = \frac{1}{r^2}$. Integrating:

$$\int r^2 dr = -\int \frac{\mu_0 e^4}{12\pi^2 \epsilon_0 m^2 c} dt \implies r(t)^3 = -\frac{\mu_0 e^4}{4\pi^2 \epsilon_0 m^2 c} t + C \tag{18.2.4}$$

where C is a constant we need to determine through suitable initial conditions. Suppose for example that the electron is initially orbiting the proton at a radius r_0 at $t = 0$. Then we find that $r_0^3 = C$. If we want to find the time T it takes for the electron to collapse into the proton, then we must require $r(T) = 0$ and thus:

$$T = \frac{4\pi^2 \epsilon_0 m^2 c}{\mu_0 e^4} r_0^3 \tag{18.2.5}$$

It turns out that for $r_i \approx 5 \times 10^{-11}$ m, it takes the electron just 10^{-11} seconds to collapse! Luckily this paradox is not due to an error in our solution but rather a sign that classical physics cannot provide a coherent description of atomic phenomena.

18.3 Exact Differential Equations

Consider an ODE in full differentials, with solutions $\Phi(x, y)$ such that:

$$\begin{cases} \forall (x, y) \in \mathcal{D}, \ \Phi_x = P(x, y), \ \Phi_y = Q(x, y) \\ d\Phi = 0 \end{cases} \tag{18.3.1}$$

We can then rewrite, using the chain rule, the equation as:

$$P(x, y)dx + Q(x, y)dy = 0 \tag{18.3.2}$$

> **Theorem** (*Exactness condition*)
> 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 first order ODE**.

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 *(Exact)* To solve:

1. Set the equation: $\Phi_x = P(x, y)$, and integrate directly with respect to x:

$$\Phi = \int P(x, y)dx + \phi(y) \tag{18.3.3}$$

2. Substitute this into $\Phi_y = Q(x, y)$:

$$\frac{d}{dy}\left(\int P(x, y)dx + \phi(y) \right) = Q(x, y) \tag{18.3.4}$$

and rearrange to find:

$$\phi(y) = \int \left(Q(x, y) - \frac{d}{dy} \int P(x, y)dx \right) dy \tag{18.3.5}$$

3. Set:

$$\int P(x, y)dx + \int \left(Q(x, y) - \frac{d}{dy} \int P(x, y)dx \right) dy = C \tag{18.3.6}$$

for some constant C.

18.4   Inexact Differential Equations

Inexact differential equations are of the form:

$$\begin{cases} P(x, y)dx + Q(x, y)dy = 0 \\ \frac{\partial P}{\partial y} \neq \frac{\partial Q}{\partial x} \end{cases} \tag{18.4.1}$$

unlike exact differential equations. Firstly, consider an ODE of the following form:

$$\frac{dy}{dx} = f(ax + by). \tag{18.4.2}$$

Strategy *(Inexact)*
To solve:

1. Apply the change of variables $z(x) = ax + by(x)$ to find:

$$\frac{dz}{dx} = a + b\frac{dy}{dx} \tag{18.4.3}$$

2. Substitute the expression for y':

$$\frac{dz}{dx} = a + bf(z) \tag{18.4.4}$$

3. Solve as a separable differential equation:

$$\int \frac{dz}{a + bf(z)} = \int dx \qquad (18.4.5)$$

Next, consider the case where (2.3) is homogeneous, that is:

$$\frac{P(\lambda x, \lambda y)}{P(x, y)} = \frac{Q(\lambda x, \lambda y)}{Q(x, y)}, \ \forall \lambda \neq 0 \qquad (18.4.6)$$

Then:

$$\frac{dy}{dx} = -\frac{P(x, y)}{Q(x, y)} = -\frac{P(\lambda x, \lambda y)}{Q(\lambda x, \lambda y)}|_{\lambda = \frac{1}{x}} = -\frac{P(1, \frac{y}{x})}{Q(1, \frac{y}{x})} = f\left(\frac{y}{x}\right) \qquad (18.4.7)$$

Strategy *(Homogeneous)*
To solve:
   1. Apply the change of variables $u(x)x = y(x)$:

$$x\frac{du}{dx} + u = \frac{dy}{dx} \qquad (18.4.8)$$

2. Substitute into original ODE:

$$f(u) = u + x\frac{du}{dx} \qquad (18.4.9)$$

and separate variables:

$$\int \frac{du}{f(u) - u} = \int \frac{dx}{x}. \qquad (18.4.10)$$

where $f(u) \neq u$.

18.5   Integrating Factor Method

Consider the non homogeneous ODE of the form:

$$\dot{y}(t) = a(t)y(t) + f(t) \qquad (18.5.1)$$

To solve, we wish to multiply the whole equation by a so called **integrating factor** Λ such that $\dot{\Lambda} = \Lambda a(t)$. Then:

$$\Lambda\dot{y}(t) - \overbrace{\Lambda a(t)}^{\dot{\Lambda}} y(t) = \Lambda f(t) \implies \frac{d\Lambda y(t)}{dt} = \Lambda f(t) \qquad (18.5.2)$$

$$\implies \boxed{y(t) = \frac{1}{\Lambda(t)}\left[C + \int_0^t \Lambda(t')f(t')dt'\right]} \qquad (18.5.3)$$

To find the integrating factor, we solve the separable equation:

$$\frac{d\Lambda}{dt} = \Lambda(t)a(t) \implies \int \frac{d\Lambda}{\Lambda} = \int a(t)dt \implies \boxed{\Lambda(t) = \exp\left(\int_0^t a(t')dt'\right)} \qquad (18.5.4)$$

We have proven the following, very useful result:

> **Theorem** *(Integrating factor)*
> Let $\dot{y}(t) = a(t)y(t) + f(t)$ be a first order linear ODE. Its general solution is then given by:
>
> $$y(t) = \frac{1}{\Lambda(t)}\left[C + \int_0^t \Lambda(t')f(t')dt'\right], \ \Lambda(t) = \exp\left(\int_0^t a(t')dt'\right) \tag{18.5.5}$$

18.6 Bernoulli Equations

Finally, let us look at the Bernoulli Equations:

$$y' + a(x)y = b(x)y^n \tag{18.6.1}$$

which due to the y^n term, is non linear.

> **Strategy** *(Bernoulli equation)*
> To solve:
>   1. Divide through by y^n to find:
>
> $$\frac{dy}{dx}y^{-n} + a(x)y^{1-n} = b(x) \tag{18.6.2}$$
>
>   2. Apply a change of variables $u = y^{1-n}$:
>
> $$\frac{du}{dx} + (1-n)a(x)u = (1-n)b(x) \tag{18.6.3}$$
>
>   3. Use integrating factor method.

18.7 Stability and Equilibrium points

> **Definition** *(Equilibrium points)*
> An **equilibrium point** of a differential equation is a constant solution $y' = 0$, $\forall t \in \mathcal{D}$. It is:
>   Stable: if $y \to c$ as $t \to \infty$, in other words the deviation decays.
>   Unstable: if $y \to \infty$ as $t \to \infty$, in other words the deviation grows.

Definition 2.1. We can linearize differential equations by doing a perturbative analysis. Suppose $y = a$ is an equilibrium point of $y' = f(x, y)$. We then induce an arbitrarily small perturbation $y = a + \epsilon(t)$, so that:

$$\frac{d\epsilon}{dt} = \frac{dy}{dt} = f(a,t) + \epsilon\frac{\partial f}{\partial y}(a,t) + O(\epsilon^2) \tag{18.7.1}$$

$$\approx \epsilon\frac{\partial f}{\partial t}(a,t) \tag{18.7.2}$$

If $\dot{\epsilon} > 0$, we have unstable equilibrium, if $\dot{\epsilon}(t) < 0$, we have stable equilibrium.

ODE3 Second Order ODEs

19

19.1 Homogeneous equation

We consider the second order homogeneous differential equation with initial conditions:

$$\begin{cases} a(x)y'' + b(x)y' + c(x)y = 0, \ (x \in [a,b]) \\ y(x_0) = y_0 \\ y'(x_0) = y'_0 \end{cases} \tag{19.1.1}$$

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:
>
> $$\mathsf{Y}(y_1, y_2) = \begin{pmatrix} y_1(x) & y_2(x) \\ y'_1(x) & y'_2(x) \end{pmatrix} \tag{19.1.2}$$
>
> and define their **Wronskian** to be the determinant of the fundamental matrix:
>
> $$W(y_1, y_2)(x) = \begin{vmatrix} y_1(x) & y_2(x) \\ y'_1(x) & y'_2(x) \end{vmatrix} = y_1(x)y'_2(x) - y'_1(x)y_2(x) \tag{19.1.3}$$

> **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:

$$y_1(x_0)y'_2(x_0) - y_2(x_0)y'_1(x_0) \neq 0 \implies y_1(x_0) \neq \frac{y'_1(x_0)}{y'_2(x_0)}y_2(x_0) = cy'_1(x_0) \tag{19.1.4}$$

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: $c_1y_1(x) + c_2y_2(x) = 0$ for some $c_1, c_2 \in \mathbb{R}$ not both zero. Then differentiating with respect to x one finds that:

$$\begin{cases} c_1y_1(x) + c_2y_2(x) = 0 \\ c_1y_1'(x) + c_2y_2'(x) = 0 \end{cases} \tag{19.1.5}$$

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:

$$\begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix} = 0 \implies W(y_1, y_2)(x) = 0, \ \forall x \in [a, b] \tag{19.1.6}$$

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 (19.1.1). Due to the linearity of (19.1.1), we may use the principle of superposition and state that:

$$y(x) = c_1y_1(x) + c_2y_2(x), \ \forall c_1, c_2 \in \mathbb{R} \tag{19.1.7}$$

will also be a solution. We now determine the constants:

$$\begin{pmatrix} y(x_0) \\ y'(x_0) \end{pmatrix} = \begin{pmatrix} y_1(x_0) & y_2(x_0) \\ y_1'(x_0) & y_2'(x_0) \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \mathsf{Y} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{19.1.8}$$

Then, the coefficients are determined as:

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_0' \end{pmatrix} \cdot \mathsf{Y}^{-1} \tag{19.1.9}$$

$$= \frac{1}{W(y_1, y_2)(x_0)} \begin{pmatrix} y_2'(x_0) & -y_2(x_0) \\ -y_1'(x_0) & y_1(x_0) \end{pmatrix} \cdot \begin{pmatrix} y(x_0) \\ y'(x_0) \end{pmatrix} \tag{19.1.10}$$

Note that had we chosen linearly dependent solutions then 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:
>
> $$y'' + p(x)y' + q(x)y = 0 \tag{19.1.11}$$
>
> then:
>
> $$W(y_1, y_2)(x) = W(y_1, y_2)(x_0) \exp\left[-\int_{t_0}^{x} p(s)ds \right] \tag{19.1.12}$$

Proof. The derivative of a determinant is given by Jacobi's formula:

$$\frac{d}{dx} \det\{\mathsf{A}(x)\} = \text{tr}\left(\text{adj } \mathsf{A}(x)\frac{d\mathsf{A}(x)}{dx} \right) \tag{19.1.13}$$

For a 2×2 matrix A:

$$\mathsf{A} = \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \implies \text{adj } \mathsf{A} = \begin{pmatrix} d(x) & -b(x) \\ -c(x) & a(x) \end{pmatrix} \tag{19.1.14}$$

so that:

$$\text{adj } \mathsf{A}(x)\frac{d\mathsf{A}(x)}{dx} = \begin{pmatrix} d(x) & -b(x) \\ -c(x) & a(x) \end{pmatrix} \begin{pmatrix} a'(x) & b'(x) \\ c'(x) & d'(x) \end{pmatrix} \tag{19.1.15}$$

$$= \begin{pmatrix} a'(x)d(x) - b(x)c'(x) & b'(x)d(x) - b(x)d'(x) \\ a(x)c'(x) - a'(x)c(x) & a(x)d'(x) - c(x)b'(x) \end{pmatrix} \tag{19.1.16}$$

and hence:

$$\frac{d}{dx}\det\{\mathsf{A}\} = a'(x)d(x) - b(x)c'(x) + a(x)d'(x) - c(x)b'(x) \tag{19.1.17}$$

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:

$$W'(y_1, y_2)(x) = y_1'(x)y_2'(x) - y_2(x)y_1''(x) + y_1(x)y_2''(x) - y_1'(x)y_2'(x) \tag{19.1.18}$$

$$= y_1(x)y_2''(x) - y_2(x)y_1''(x) \tag{19.1.19}$$

We may now substitute $y_i''(x) + p_i(x)y_i'(x) + q_i(x) = 0$ for $i = 1, 2$ t find that

$$W'(y_1, y_2)(x) = = y_1(x)(-p(x)y_2'(t) - q(x)y_2(x)) - y_2(x)(-p(x)y_1'(x) - q(x)y_1(x)) \tag{19.1.20}$$

$$= -(y_1(x)y_2'(x) - y_2(x)y_1'(x))p(x) \tag{19.1.21}$$

$$= -W(y_1, y_2)(x)p(x) \tag{19.1.22}$$

Separating variables, and integrating from x_0 to x, one finally finds:

$$W(y_1, y_2)(x) = W(y_1, y_2)(x_0) \exp\left[-\int_{x_0}^{x} p(s)ds\right] \tag{19.1.23}$$

as required. ∎   ∎

19.2   Non-homogeneous

Definition 3.4. The second order non-homogeneous differential equation is:

$$\begin{cases} a(t)y'' + b(t)y' + c(t)y = f(t) \\ y(t_0) = y_0 \\ y'(t_0) = y_0' \end{cases} \tag{19.2.1}$$

and its *associated homogeneous equation* is (3.1).

> **Theorem** *(Complementary Function and Particular Integral)*
> The general solution to the non-homogeneous differential equation may be written as the sum:
> $$y(t) = y_{CF}(t) + y_{PI}(t) \tag{19.2.2}$$
> 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. In-

deed:

$$a(t)(y''(t) - y''_{PI}(t)) + b(t)(y'(t) - y'_{PI}(t) + c(t)(y(t) - y_{PI}(t)) \tag{19.2.3}$$
$$= a(t)y''(t) + b(t)y'(t) + c(t)y(t) - (a(t)y''_{PI}(t) + b(t)y'_{PI}(t) + c(t)y_{PI}(t)) \tag{19.2.4}$$
$$= f(t) - f(t) = 0 \tag{19.2.5}$$

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:

$$y(t) - y_{PI}(t) = c_1 y_1(t) + c_2 y_2(t) \implies y(t) = y_{PI}(t) + c_1 y_1(t) + c_2 y_2(t) \tag{19.2.6}$$

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.

19.3 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 $f(t)$:

$g(t)$	$y_{PI}(t)$
$\alpha e^{\beta t}$	$A e^{\beta t}$
$a\cos(\beta t) + b\sin(\beta t)$	$A\cos(\beta t) + B\sin(\beta t)$
$\sum_{i=0}^{n} a_i x^i$	$\sum_{i=0}^{n} A_i x^i$

19.4 Variation of Constants

Assume that we have found the complementary solution to the non-homogeneous equation:

$$y_{CF}(t) = c_1 y_1(y) + c_2 y_2(t) \tag{19.4.1}$$

and look for the particular integral of the form:

$$y_{PI}(t) = \psi_1 y_1 + \psi_2 y_2 \tag{19.4.2}$$

such that:

$$\psi'_1 y_1 + \psi'_2 y_2 = 0 \tag{19.4.3}$$

Differentiating:

$$y'_{PI} = \psi_1 y'_1 + \psi_2 y'_2 \tag{19.4.4}$$
$$y''_{PI}(t) = \psi'_1 y'_1 + \psi'_2 y'_2 + \psi_1 y''_1 + \psi_2 y''_2 \tag{19.4.5}$$

Inserting these into the differential equation and simplifying:

$$\psi_1' y_1' + \psi_2' y_2' = \frac{f(t)}{a(t)} \tag{19.4.6}$$

Let us finally assume that $a(t) = 1$ (which corresponds to rearranging the equation so that the coefficient of y'' is 1), then:

$$\begin{cases} \psi_1' y_1' + \psi_2' y_2' = f(t) \\ \psi_1' y_1 + \psi_2' y_2 = 0 \end{cases} \tag{19.4.7}$$

$$\implies \psi_1' = -\frac{\psi_2' y_2}{y_1} \implies -\frac{\psi_2' y_2}{y_1} y_1' + \psi_2' y_2' = f(t) \tag{19.4.8}$$

$$\implies \psi_2' = \frac{y_1 f(t)}{y_1 y_2' - y_2 y_1'}, \ \psi_1' = -\frac{y_2 f(t)}{y_1 y_2' - y_2 y_1'} \tag{19.4.9}$$

Since y_1 and y_2 are linearly independent, their Wronskian is non-zero, and thus:

$$\psi_1' = -\frac{y_2 f(t)}{W(y_1, y_2)}, \ \psi_2' = \frac{y_1 f(t)}{W(y_1, y_2)}, \tag{19.4.10}$$

which can be integrated directly in (3.4) to get:

$$y_{PI}(t) = -y_1 \int \frac{y_1 f(t)}{y_2 y_2' - y_2 y_1'} dt + y_2 \int \frac{y_1 f(t)}{W(y_1, y_2)} \tag{19.4.11}$$

19.5 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:

$$y(t) = \phi(t) y_1(t) \tag{19.5.1}$$

and evaluate its first and second derivatives.
2. We substitute into the original differential equation.
3. Solve the resulting differential equation by substituting $\psi = \phi'$.

19.6 Euler-Cauchy equations

19.7 Intro to Green's functions

Up until now all our methods have relied on producing a general solution and subsequently fitting it to given boundary conditions. The method of Green's functions (which will prove to be powerful for PDEs too) already takes the boundary conditions from the beginning, building a particular solution directly.

Theorem *()*

ODE 4 Mechanical Vibrations and Resonance Phenomena

Constant coefficient second-order linear ODEs are of particular interest in the area of mechanical vibrations.

20.1   Homogeneous Equation

Consider the homogeneous equation:

$$\ddot{y} + a\dot{y} + by = 0 \tag{20.1.1}$$

and we guess a solution in exponential form:

$$y(t) = Ce^{\lambda t} \tag{20.1.2}$$

Plugging into (4.1) one finds:

$$\lambda^2 + a\lambda + b = 0 \tag{20.1.3}$$

which is called the *auxiliary equation*. The quadratic formula then yields:

$$\lambda_{1,2} = \frac{-a \pm \sqrt{a^2 - 4b}}{2} \tag{20.1.4}$$

which brings us to the following result.

Proposition 1

The solutions to the second order homogeneous ODE with constant coefficients is:

$$y(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} \tag{20.1.5}$$

provided that the solutions λ_1, λ_2 are non-degenerate. In the case where:

$\boxed{a^2 > 4b}$: we have two real, distinct solutions, and by the superposition principle:

$$y(t) = e^{-at/2}\left(C_1 e^{t\sqrt{a^2-4b}/2} + C_2 e^{-t\sqrt{a^2-4b}/2}\right) \tag{20.1.6}$$

$\boxed{a^2 < 4b}$: we have two complex, distinct solutions, and by the superposition principle:

$$y(t) = e^{-at/2}\big(C_1 e^{it\sqrt{4b-a^2}/2} + C_2 e^{-it\sqrt{4b-a^2}/2}\big)$$

Euler's identity then gives:

$$y(t) = e^{-at/2}\big(A\cos\Omega t + B\sin\Omega t\big) = \alpha e^{-at/2}\cos(\Omega t + \phi) \tag{20.1.7}$$

where $\Omega = \frac{\sqrt{4b-a^2}}{2}$.

If instead we have two degenerate solutions, so that $\boxed{a^2 = 4b}$, then we have only found one solution:

$$y_1(t) = Ce^{-at/2}. \tag{20.1.8}$$

We use the method of reduction of variables to find the general solution:

$$y(t) = \psi(t)e^{-at/2} \implies \psi'' + \left(b - \frac{a^2}{4}\right)\psi = 0 \tag{20.1.9}$$

$$\therefore \psi'' = 0 \implies \psi = C_1 t + C_2 \tag{20.1.10}$$

since $b = \frac{a^2}{4}$. Finally:

$$y(t) = (C_1 t + C_2)e^{at/2} \tag{20.1.11}$$

20.2 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:

$$\ddot{y} + \gamma\dot{y} + \omega_0^2 y = 0 \tag{20.2.1}$$

In this case, we may define:

$$\Omega \equiv \frac{\sqrt{\omega_0^2 - \frac{\gamma^2}{4}}}{2} \tag{20.2.2}$$
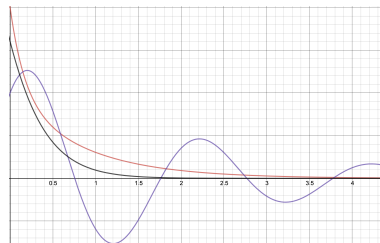
Figure 20.1. Plots of over damped (red), critically damped (black) and under damped (purple) solutions.

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:

$$y(t) = y_0 e^{-\gamma t/2}\left(\cos \Omega t + \frac{\gamma}{2\Omega}\sin \Omega t\right) \tag{20.2.3}$$

Note that this can be rewritten as:

$$y(t) = Ae^{-\gamma t/2}\cos(\Omega t + \phi) \tag{20.2.4}$$

which implies that the period of oscillations is:

$$\tau = \frac{2\pi}{\Omega} \tag{20.2.5}$$

and over a cycle the amplitude is multiplied by:

$$e^{-\gamma \tau/2} = \exp\left(-\frac{2\pi\gamma}{\Omega}\right) \tag{20.2.6}$$

called the *amplitude decay factor*.

$\boxed{\omega_0 < \lambda/2}$, the oscillator is *over damped*. The solution is:

$$y(t) = \frac{y_0}{2\Omega}e^{-\gamma t/2}\left[\left(\Omega + \frac{\gamma}{2}\right)e^{\Omega t} + \left(\Omega - \frac{\gamma}{2}\right)e^{-\Omega t}\right] \tag{20.2.7}$$

$\boxed{\omega_0 = \lambda/2}$, the oscillator is *critically damped*, the solution is:

$$y(t) = y_0 e^{-\gamma t/2}\left(1 + \frac{\gamma}{2}t\right) \tag{20.2.8}$$

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:

$$Q = \frac{2\pi E_0}{PE_\tau} \tag{20.2.9}$$

Proposition 2. *The quality factor for a damped oscillator is:*

$$Q = \frac{\omega_0}{\gamma} \tag{20.2.10}$$

Proof. The energy stored in the oscillator is:

$$E_0 = \frac{1}{2}ky_0^2 = \frac{1}{2}m\omega_0^2 y_0^2 \tag{20.2.11}$$

The under damped solution gives:

$$\begin{cases} y(t) = y_0 e^{-\gamma t/2}\cos\left(\omega_0 t - \phi\right) \\ \dot{y}(t) = -y_0 e^{-\gamma t/2}\left(\frac{\gamma}{2}\cos\left(\omega_0 t - \phi\right) + \omega_0 \sin\left(\omega_0 t - \phi\right)\right) \end{cases} \tag{20.2.12}$$

We then define the energy at t to be the sum of the kinetic and potential energy:

$$E(t) = \frac{1}{2}m\dot{y}^2 + \frac{1}{2}m\omega_0 y^2 \tag{20.2.13}$$

$$\implies \frac{dE(t)}{dt} = m\ddot{y}\dot{y} + m\omega_0 y\dot{y} = m\dot{y}\underbrace{(\ddot{y} + \omega_0 y)}_{-\omega\dot{y}} = -m\gamma\dot{y}^2 \tag{20.2.14}$$

$$\therefore PE_\tau = \int_t^{t+Pt} -m\gamma\dot{y}^2 dt \tag{20.2.15}$$

$$= \int_t^{t+Pt} m\gamma y_0^2 e^{-\gamma t}\left(\frac{\gamma}{2}\cos\left(\omega_0 t - \phi\right) + \omega_0 \sin\left(\omega_0 t - \phi\right)\right)^2 dt \tag{20.2.16}$$

$$\tag{20.2.17}$$

We can now apply the substitution $t' = \omega_0 t - \phi$, so that the limits of integration become 0 and 2π:

$$PE_\tau = m\gamma y_0^2 \int_0^{2\pi} e^{-\gamma\frac{t'+\phi}{\omega_0}}(\frac{\gamma}{2}\cos t' + \omega_0 \sin t')^2 \frac{1}{\omega_0}dt' \tag{20.2.18}$$

$$= \frac{m\gamma y_0^2}{\omega_0}\frac{\pi}{4}(4\omega_0^2 + \gamma^2) \tag{20.2.19}$$

$$= \pi m\gamma y_0^2 \omega_0\left(1 + \frac{\gamma^2}{4\omega^2}\right) \tag{20.2.20}$$

$$= \pi m\gamma y_0^2 \omega_0 \tag{20.2.21}$$

hence:

$$Q = 2\pi\frac{\frac{1}{2}m\omega_0^2 y_0^2}{\pi m\gamma y_0^2 \omega_0} = \frac{\omega_0}{\gamma} \tag{20.2.22}$$

as required. ∎

20.3 Forced Oscillations

Finally, let us consider the damped, forced oscillations equation:

$$\ddot{y} + \gamma\dot{y} + \omega_0^2 y = F\cos\omega t \tag{20.3.1}$$

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):

$$\ddot{z} + \gamma\dot{z} + \omega_0^2 z = Fe^{i\omega t} \tag{20.3.2}$$

and use as a trial solution $z = Ce^{i\omega t}$. Then:

$$C = \frac{F}{\omega_0^2 - \omega^2 + i\omega\gamma} = \frac{F(\omega_0^2 - \omega^2 - i\omega\gamma)}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2} \tag{20.3.3}$$

Next, we define:

$$\cos\phi = \frac{\omega_0^2 - \omega^2}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}}, \sin\phi = \frac{\omega\gamma}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}} \tag{20.3.4}$$

so that:

$$C = \frac{Fe^{-i\phi}}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}} \tag{20.3.5}$$

This finally gives the solution:

$$z(t) = \frac{Fe^{i(\omega_0 t - \phi)}}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}} \tag{20.3.6}$$

Taking the real part yields:

$$y_{PI}(t) = A\cos(\omega t - \phi), \ A = \frac{F}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}} \tag{20.3.7}$$

Finally, the general solution is:

$$y(t) = \underbrace{A\cos(\omega t - \phi)}_{\text{steady state}} + \underbrace{e^{-\gamma t/2}\big[C_1\cos\Omega t + C_2\sin\Omega t\big]}_{\text{transient}} \tag{20.3.8}$$

where the first term is the steady state solution, and the second term is the transient solution, and quickly decays after $t \gg \gamma^{-1}$.
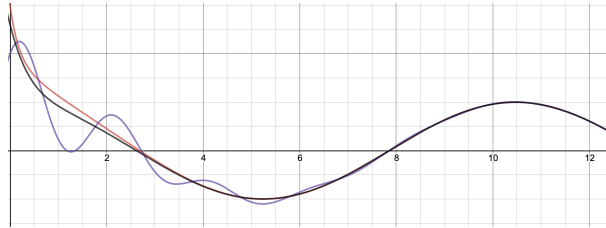
Figure 20.2. Forced solutions for over damped, under damped and critically damped oscillators.

20.4   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 $A'(\omega) = 0$:

$$4\omega_{res}(\omega_0 - \omega_{res}^2) + 2\omega_{res}\gamma^2 = 0 \implies \omega_{res} = \sqrt{\omega_0^2 - \frac{\gamma^2}{2}} \approx \omega_0 \tag{20.4.1}$$

The peak amplitude is then:

$$A_{res} \approx A(\omega_0) = \frac{F}{\omega_0\gamma} \tag{20.4.2}$$

ODE5 General Linear ODEs

21.1 Existence and Uniqueness

We now consider the general theory of linear ODEs.

> **Theorem** *(Existence and uniqueness linear)*
> If all $A_{ij}(t)$ and $f_i(t)$ are continuous on $\mathcal{I} = (t_1, t_2)$, then $\forall t_0 \in \mathcal{I}, \forall \boldsymbol{y}_0 \in \mathbb{R}^n$, the Cauchy problem:
> $$\begin{cases} \dot{\boldsymbol{y}} = \mathsf{A}(t)\boldsymbol{y} + \boldsymbol{f}(t) \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases} \tag{21.1.1}$$
> has a unique solution in $\mathcal{I}$

21.2 Fundamental set and Wronskians

> **Definition** *(Fundamental set of solutions)* The *fundamental system of solutions* of the homogeneous equation:
> $$\dot{\boldsymbol{y}} = \mathsf{A}(t)\boldsymbol{y} + \boldsymbol{f}(t) \tag{21.2.1}$$
> is the set of linearly independent solutions $\{\boldsymbol{y}_1(t), \boldsymbol{y}_2(t), ..., \boldsymbol{y}_n(t)\}$ to the associated homogeneous differential equation:
> $$\dot{\boldsymbol{y}} = \mathsf{A}(t)\boldsymbol{y} \tag{21.2.2}$$

> **Proposition** *(Independent solutions)*
> The following are true for any Cauchy system of the form (3.1):
> a. If $\exists t_0 \in \mathcal{I}$ such that $\{\boldsymbol{y}_i(t_0)\}$ is linearly independent, then the fundamental set of solutions is linearly independent.
> b. If $\exists t_0 \in \mathcal{I}$ such that $\{\boldsymbol{y}_i(t_0)\}$ is linearly dependent, then the fundamental set of solutions is linearly dependent.

Proof.

a. Suppose that $\{\boldsymbol{y}_i(t_0)\}$ is linearly dependent. Then, $\exists\{C_i\}$ not all equal to zero such that, $\forall t \in \mathcal{I}, C_i \boldsymbol{y}_i = \boldsymbol{0}$. However, this is not satisfied for $t = t_0$, thus we have a contradiction.

b. Suppose that $\{\boldsymbol{y}_i(t_0)\}$ is linearly dependent. Then, $\forall t \in \mathcal{I}, \exists\{C_i\}$ such that: $\boldsymbol{y} = C_i\boldsymbol{y}_i = \boldsymbol{0}$. This implies that $\boldsymbol{y}(t_0) = 0$, and by the uniqueness theorem, $\boldsymbol{y}(t) = 0$ identically, and therefore $C_i\boldsymbol{y}_i(t) = (0), \forall t$ as required. ∎

> **Definition** *(Wronskian)*
>
> The fundamental set of solutions can be packed into a matrix, called the **fundamental matrix**:
>
> $$\mathsf{Y}(t) = \begin{pmatrix} \boldsymbol{y}_1(t) \ \boldsymbol{y}_2(t) \ ..., \boldsymbol{y}_n(t) \end{pmatrix} \tag{21.2.3}$$
>
> The **Wronskian** is then defined as the determinant of the fundamental matrix:
>
> $$W(t) = \det \mathsf{Y}(t) \tag{21.2.4}$$

> **Theorem** *(Liouville's formula)*
>
> To compute the Wronskian, one can use Liouville's formula:
>
> $$W(t) = W(t_0) \exp\left[\int_{t_0}^{t} \mathrm{tr}\mathsf{A}(t')dt' \right] \tag{21.2.5}$$

Proof. The derivative of the Wronskian is given by differentiating row by row and then summing:

$$\dot{W}(t) = \sum_{i=1}^{n} \det \mathsf{Y}_i^*(t) \tag{21.2.6}$$

where we define:

$$\mathsf{Y}_i^*(t) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ \vdots & \vdots & & \\ \dot{y}_{i1} & \dot{y}_{i2} & \cdots & \dot{y}_{in} \\ \vdots & \vdots & & \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{pmatrix} \tag{21.2.7}$$

Now note that since $\dot{\mathsf{Y}} = \mathsf{A}\mathsf{Y} \iff \dot{y}_{ik} = \sum_j A_{ij}y_{jk}$, we may write the above as:

$$\mathsf{Y}_i^*(t) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ \vdots & \vdots & & \\ \sum_j A_{ij}y_{j1} & \sum_j A_{ij}y_{j2} & \cdots & \sum_j A_{ij}y_{jn} \\ \vdots & \vdots & & \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{pmatrix} \tag{21.2.8}$$

Now since the determinant is unchanged if we subtract from one row a linear combination of all the others, we can subtract from the ith row the following linear combination of all other rows:

$$\sum_{j\neq i}^{n} A_{ij}\begin{pmatrix} y_{j1} & ... & y_{jn} \end{pmatrix} \tag{21.2.9}$$

which will leave only the a_{ii} coefficients in the matrix:

$$\det(\mathsf{Y}_i^*(t)) = \det \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ \vdots & \vdots & & \\ A_{ii}y_{j1} & A_{ii}y_{j2} & \cdots & A_{ii}y_{jn} \\ \vdots & \vdots & & \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{pmatrix} = A_{ii} \det \mathsf{Y} \tag{21.2.10}$$

Therefore, summing over all i we get:

$$\det \dot{W}(t) = (\text{tr}\mathsf{A}) \det \mathsf{Y}(t) \implies \dot{W}(t) = (\text{tr}\mathsf{A})\mathsf{Y}(t) \tag{21.2.11}$$

which can be solved to yield the required formula. ∎

21.3 Homogeneous ODE

Let us now look at how to solve the homogeneous ODE:

$$\begin{cases} \dot{\boldsymbol{y}}(t) = \mathsf{A}(t)\boldsymbol{y} \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{cases} \tag{21.3.1}$$

If we know a fundamental system of solutions $\boldsymbol{y}_i(t)$, then by the principle of superposition, seeing as the elements of this set form a basis for all solutions, one finds that:

$$\boldsymbol{y}(t) = \mathsf{Y}(t)\boldsymbol{C} \tag{21.3.2}$$

We impose the condition $\boldsymbol{u}(t_0) = \boldsymbol{y}_0$ to get:

$$\boldsymbol{C} = \mathsf{Y}^{-1}(t_0)\boldsymbol{y}_0 \tag{21.3.3}$$

where $\exists \mathsf{Y}^{-1}(t_0)$ since the columns of the Wronskian are all linearly independent (linearly independent solutions). The solution to the Cauchy problem is thus:

$$\boxed{\boldsymbol{y}(t) = \mathsf{Y}(t)\mathsf{Y}^{-1}(t_0)\boldsymbol{y}_0} \tag{21.3.4}$$

21.4 Variation of parameters

Once we have solved the homogeneous equation, we may generalise our results to non-homogeneous differential equations. Consider the following solution

$$\boldsymbol{y}(t) = \mathsf{Y}(t)\boldsymbol{C} + \boldsymbol{y}_{PI}(t) \tag{21.4.1}$$

which subject to the initial condition $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$ yields:

$$\boldsymbol{C} = \mathsf{Y}^{-1}(t_0)[\boldsymbol{y}_0 - \boldsymbol{y}_{PI}(t_0)] \tag{21.4.2}$$

So how can we find the particular integral? It suffices to use the method of variation of constants, $C_i \to \psi_i(t)$:

$$\boldsymbol{y}(t) = \sum_i \boldsymbol{y}_i(t)\psi_i(t) = \mathsf{Y}(t)\boldsymbol{\psi} \tag{21.4.3}$$

We substitute this into the non homogeneous equation:

$$\dot{\boldsymbol{y}}(t) = \dot{\mathsf{Y}}\boldsymbol{\psi} + \mathsf{Y}\dot{\boldsymbol{\psi}} = \mathsf{A}\mathsf{Y}\boldsymbol{\psi} + \boldsymbol{f} \implies \mathsf{Y}\dot{\boldsymbol{\psi}} = \boldsymbol{f} \tag{21.4.4}$$

since $\dot{\mathsf{Y}} = \mathsf{A}\mathsf{Y}$. Noting that $\det \mathsf{Y}(t) = W(t) \neq 0$ then:

$$\dot{\boldsymbol{\psi}} = \mathsf{Y}^{-1}\boldsymbol{f} \implies \boldsymbol{\psi} = \boldsymbol{C} + \int_{t_0}^t Y^{-1}(t)\boldsymbol{f}(t') \tag{21.4.5}$$

so that:

$$\boldsymbol{y}(t) = \underbrace{\mathsf{Y}(t)}_{\text{CF}}\Big[\boldsymbol{C} + \int_{t_0}^t \mathsf{Y}^{-1}(t)\boldsymbol{f}(t')\Big] \tag{21.4.6}$$

We substitute back the expression for $\boldsymbol{C}$ and find:

$$\boxed{\boldsymbol{y}(t) = \mathsf{Y}(t)\left[\mathsf{Y}^{-1}(t_0)\boldsymbol{y}_0 + \int_{t_0}^t \mathsf{Y}^{-1}(t)\boldsymbol{f}(t')dt'\right]} \tag{21.4.7}$$

21.5 Higher Order linear ODEs

Consider the general form of a linear ODE:

$$y^{(n)} + p_{n-1}(t)y^{(n-1)} + ... + p_1(t)y' + p_0(t)y = f(t) \tag{21.5.1}$$

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:

$$y_{CF}(t) = \sum_{i=1}^n c_i y_i(t) \tag{21.5.2}$$

We now perform a variation of parameters and write the particular integral as

$$y_{PI}(t) = \sum_{i=1}^n \psi_i y_i \tag{21.5.3}$$

and assuming that $\sum_{i=1}^n \psi_i' y_i = 0$, differentiating yields:

$$y_{PI}'(t) = \sum_{i=1}^n \psi_i y_i' \tag{21.5.4}$$

In general, we will set:

$$\begin{cases} y_{PI}^{(k)}(t) = \sum_{i=1}^n \psi_i y_i^{(k)}, \ k = 1, ..., n-1 \\ \sum_{i=1}^n \psi_i' y_i^{(k)} = 0, \ k = 0, ..., n-2 \end{cases} \tag{21.5.5}$$

Finally, we evaluate the nth derivative as usual without making any further special assumptions:

$$y_{PI}^{(n)}(t) = \sum_{i=1}^n (\psi_i y_i^{(n)} + \psi_i' y_i^{(n-1)}) \tag{21.5.6}$$

We are now ready to substitute everything into (5.5):

$$\sum_{i=1}^{n} (\psi_i y_i^{(n)} + \psi_i' y_i^{(n-1)}) + p_{n-1}(t) \sum_{i=1}^{n} \psi_i y_i^{(n-1)} + ... + p_1(t) \sum_{i=1}^{n} \psi_i y_i' + p_0(t) \sum_{i=1}^{n} \psi_i y_i = f(t) \quad (21.5.7)$$

which rearranging gives:

$$\sum_{i=1}^{n} \left(\psi_i \left[\sum_{j=0}^{n} p_j y_i^{(j)} \right] \right) + \sum_{i=1}^{n} \psi_i' y_i^{(n-1)} = f(t) \quad (21.5.8)$$

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:

$$\sum_{i=1}^{n} \psi_i' y_i^{(n-1)} = f(t) \quad (21.5.9)$$

We therefore have the following system of equations:

$$\begin{cases} \sum_{i=1}^{n} \psi_i' y_i^{(k)} = 0, \ k = 0, ..., n-2 \\ \sum_{i=1}^{n} \psi_i' y_i^{(n-1)} = f(t) \end{cases} \quad (21.5.10)$$

To solve this system of equations, we will use Cramer's rule. As always, the Wronskian is:

$$W(t) = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \vdots & \vdots & & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix} \quad (21.5.11)$$

To use Cramer's rule, we must successively substitute each column of the Wronskian with $\begin{pmatrix} 0 & 0 & \cdots & f(t) \end{pmatrix}^T$. We will thus denote:

$$W_i = \begin{vmatrix} y_1 & \cdots & y_{i-1} & 0 & \cdots & y_n \\ y_1' & \cdots & y_{i-1}' & 0 & \cdots & y_n' \\ \vdots & & \vdots & \vdots & & \vdots \\ y_1^{(n-1)} & \cdots & y_{i-1}^{(n)} & f(t) & \cdots & y_n^{(n-1)} \end{vmatrix} = f(t) \begin{vmatrix} y_1 & \cdots & y_{i-1} & 0 & \cdots & y_n \\ y_1' & \cdots & y_{i-1}' & 0 & \cdots & y_n' \\ \vdots & & \vdots & \vdots & & \vdots \\ y_1^{(n-1)} & \cdots & y_{i-1}^{(n)} & 1 & \cdots & y_n^{(n-1)} \end{vmatrix} \quad (21.5.12)$$

where the ith column was altered. Cramer's rule finally gives:

$$\psi_i' = \frac{f(t) W_i(t)}{W(t)} \implies u_i = \int \frac{f(t) W_i(t)}{W(t)} dt \quad (21.5.13)$$

which substituting back into (5.6) gives:

$$\boxed{y_{PI}(t) = \sum_{i=1}^{n} \left(y_i(t) \int \frac{f(t) W_i(t)}{W(t)} dt \right)} \quad (21.5.14)$$

ODE6 Sturm-Liouville theory and Green's functions

22.1   Linear differential operators

We will try to study differential equations from a more general point of view. To do so we must introduce the concept of differential operators.

> **Definition** *(Differential operator)* A **differential operator** is a linear operator L on a function space V, such that:
>
> $$L[y] = f(y, y', ..., y^{(n)}, ...), \ \forall y \in V \qquad (22.1.1)$$
>
> where f is a linear function in its arguments, so that:
>   (i) $L[y_1 + y_2] = L[y_1] + L[y_2]$
>   (ii) $L[\alpha y_1] = \alpha L[y_1]$

This notation allow us to write differential equations more succinctly. For example, we may write:

$$\frac{d^2y}{dx^2} + x\frac{dy}{dx} + x^2 y = 0 \qquad (22.1.2)$$

as

$$L[y] = 0, \ \text{ where } L \equiv \frac{d^2}{dx^2} + x\frac{d}{dx} + x^2 \qquad (22.1.3)$$

> **Definition** *(Boundary value problem)* A **boundary value problem** is an equation $L[y] = f$ defined on an interval (a, b) together with a boundary constraints $B[y] = g$, which can be:
>   (i) Dirichlet: if $y(a) = c$, that is we define the value of the solution on a boundary
>   (ii) Neumann: if $y'(a) = c$, that is we define the value of the derivative of the solution on a boundary.
>   (iii) Robin: if $c_1 y(a) + y'(a) = c_2$, that is a combination of Dirichlet and Neumann conditions.
>   (iv) Periodic: if $y(a) = y(b), y'(a) = y'(b)$ if we require the solution and its derivative to match at two boundaries.
> If f is zero the BVP is said to be homogeneous, and if g is zero the BVP is said to have homogeneous boundary constraints.

In most cases solving a BVP is a very difficult, if not impossible task. Often one can exploit the linearity of the differential operator to simplify matters. One way is to express the solution as a superposition of a special set of functions and find the coefficients for this superposition by substitution. This procedure is particularly reminiscent of linear algebra, so it is important to look at the properties of infinite dimensional vector spaces (more detailed discussions can be found in the Functional analysis part).

Definition *(Weight function)*
A **weight function** on $[a, b]$ is a real, non-negative function on the interval with a finite number of zeros.

Definition *(Inner product)*
We define an inner product on the function space $L^2[a, b]$ by:

$$\langle f, g \rangle_w = \int_a^b f^*(x)g(x)\rho(x)dx \tag{22.1.4}$$

where $\rho(x)$ is a weight function. This defines a **Hilbert space**. By convention $\langle f, g \rangle_1 \equiv \langle f, g \rangle$.

Much like on normal vector spaces, we can construct a basis on Hilbert spaces too. Any set of linearly independent functions $\{u_n(x)\}$ that spans the Hilbert space is a basis. One can use the Gram-Schmidt procedure to then produce an orthonormal basis $\{\phi_n(x)\}$. Given any function $f(x)$, we can express it as a superposition of this orthonormal basis:

$$f(x) = \sum_{n=0}^{\infty} c_n \phi_n, \ c_n = \langle \phi_n, f \rangle = \int_a^b \phi^*(x)f(x)\rho(x)dx \tag{22.1.5}$$

Returning to linear operators, we can define the adjoint of an operator as usual

Definition *(Adjoint of a differential operator)*
Let L be a linear operator on a Hilbert space. We define its adjoint $L^\dagger$ so that:

$$\int_a^b (L[f])^*(x)g(x)dx = \int_a^b f^*(x)(L^\dagger[g])(x)dx + \text{ boundary terms} \tag{22.1.6}$$

If $L^\dagger = L$ then the operator is **self-adjoint**, and if in addition the boundary terms vanish then it is **hermitian**.

Example. Consider for example $L[y] = y''(x) + p(x)y'(x)$ on $[0, 1]$. We have that:

$$\int_a^b f^*(x)(L^\dagger[g])(x)dx = \int_0^1 (f'')^* g dx + \int_0^1 (pf')^* g dx \tag{22.1.7}$$

$$= [f'g]_a^b - \int_0^1 (f')^* g' dx + [p^* fg] - \int_0^1 (pf)^* g' dx \tag{22.1.8}$$

$$= [f'g - fg' + p^* fg]_0^1 + \int_0^1 f^*(g'' - p^* g')dx \tag{22.1.9}$$

so we see that:

$$L^{\dagger}[y] = y''(x) - p^*(x)y(x) \tag{22.1.10}$$

If $p(x)$ is purely imaginary then we see that L is self-adjoint. Furthermore, L is hermitian if in addition

$$[f'g - fg' + p^* fg]_0^1 = 0 \tag{22.1.11}$$

◄

If we are also given a set of boundary conditions together with L, then we can find their adjoint by looking at how one can make the boundary terms in (22.1.6) vanish (we need this or else retaking the adjoint of $L^{\dagger}$ would yield several boundary terms $+ L$). Indeed for the boundary terms to vanish for any suitable f, we need both the term at $x = a$ and $x = b$ to be zero.

Example. Suppose we add the BCs: $2y(0) - y'(0) = 0$ and $y'(1) = 0$ to the previous example's linear operator. The boundary condition at $x = 0$ reads:

$$2f(0)g(0) - f(0)g'(0) + p^*(0)f(0)g(0) = 0 \implies g'(0) = (2 + p^*(0))g(0) \tag{22.1.12}$$

Similarly, the boundary condition at $x = 1$ reads:

$$-f(1)g'(1) + p^*(1)f(1)g(1) \implies g'(1) = p^*(1)g(1) \tag{22.1.13}$$

These are the adjoint boundary conditions.

◄

22.2 Eigenfunctions

Continuing with our analogy between finite and infinite dimensional vector spaces, we now seek to find eigenfunctions and eigenvalues of L.

Definition *(Eigenfunctions of differential operators)*

Let L be a linear operator on a Hilbert space, and suppose there is a function ϕ such that:

$$L[\phi](x) = \lambda \rho(x)\phi(x) \tag{22.2.1}$$

where λ is a non-zero constant. Then $\phi(x)$ is a **eigenfunction** of L with eigenvalue λ. We can in addition require the function to satisfy a set of boundary conditions.

Example. Consider for example $L[y] = -y''$ on $[0, \pi]$ and the boundary conditions $y(0) = 0, y'(\pi) = 0$. Let us find the eigenfunctions and eigenvalues of L with unit weight $\rho(x) = 1$. We need to solve

$$-\phi''(x) = \lambda\phi(x) \tag{22.2.2}$$

(i) if $\lambda = 0$ then the general solution is

$$\phi(x) = c_1 x + c_2 \tag{22.2.3}$$

and applying the boundary conditions we see that the only possible eigenfunction is

the trivial one.

(ii) if $\lambda > 0$ then the general solution is

$$\phi(x) = c_1 e^{i\omega x} + c_2 e^{-i\omega x}, \ \omega = \sqrt{\lambda} \tag{22.2.4}$$

and applying the boundary conditions:

$$\begin{cases} c_1 + c_2 = 0 \\ i\omega(c_1 e^{i\omega\pi} - c_2 e^{-i\omega\pi}) = 0 \end{cases} \implies \begin{pmatrix} 1 & 1 \\ e^{i\omega\pi} & -e^{-i\omega\pi} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{22.2.5}$$

we see that the only non-trivial solutions arise when:

$$\det \begin{pmatrix} 1 & 1 \\ e^{i\omega\pi} & -e^{-i\omega\pi} \end{pmatrix} = -e^{-i\omega\pi} - e^{i\omega\pi} = 0 \implies \omega = \left(n + \frac{1}{2}\right) \tag{22.2.6}$$

in which case $c_1 = -c_2$. Consequently, the eigenfunctions are:

$$\phi_n(x) = A \sin\left(n + \frac{1}{2}\right)x, \ n = 0, 1, ... \tag{22.2.7}$$

with eigenvalues:

$$\lambda_n = \left(n + \frac{1}{2}\right)^2, \ n = 0, 1, ... \tag{22.2.8}$$

(iii) if $\lambda < 0$ then the general solution is

$$\phi(x) = c_1 e^{-\omega x} + c_2 e^{\omega x}, \ \omega = \sqrt{\lambda} \tag{22.2.9}$$

and applying the boundary conditions:

$$\begin{cases} c_1 + c_2 = 0 \\ \omega(-c_1 e^{-\omega\pi} + c_2 e^{\omega\pi}) = 0 \end{cases} \implies \begin{pmatrix} 1 & 1 \\ -e^{-\omega\pi} & e^{\omega\pi} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{22.2.10}$$

we see that the only non-trivial solutions arise when:

$$\det \begin{pmatrix} 1 & 1 \\ -e^{-\omega\pi} & e^{\omega\pi} \end{pmatrix} = e^{\omega\pi} - e^{-\omega\pi} = 0 \tag{22.2.11}$$

which has no solutions. Thus there are no non-trivial eigenfunctions with negative eigenvalues.

Finally, we should normalize the eigenfunctions we have found for $\lambda > 0$:

$$\langle \phi_n, \phi_n \rangle = \int_0^\pi \sin^2\left(n + \frac{1}{2}\right)x dx = \frac{\pi}{2} \implies \phi_n(x) \equiv \frac{2}{\pi} \sin\left(n + \frac{1}{2}\right)x \tag{22.2.12}$$

which of course does not affect the corresponding eigenvalues. ◄

There are three properties regarding the eigenfunctions of Hermitian differential operator that are particularly important.

> **Theorem** *(Spectral properties of Hermitian differential operators)*
> Let L be a hermitian differential operator over an interval $[a, b]$ coupled with a set of BCs, and let $L^2_w[a, b]$ be the space of square-integrable functions on $[a, b]$ with weight $w(x)$, satisfying the relevant BCs. Then:
>   (i)   **Real-valued**: L has real eigenvalues,
> (ii)   **Orthonormality**: for a given weight function, eigenfunctions corresponding to distinct eigenvalues are orthogonal.
> (iii)   **Completeness**: the normalized eigenfunctions form an orthonormal basis of the Hilbert space $L^2_w[a, b]$.

Proof. We do not prove (iii) as it is rather involved, but it is a fundamental result that is important to remember.

(i) let ϕ be an eigenfunction of L with eigenvalue λ and weight ρ:

$$L[\phi](x) = \lambda\phi(x)\rho(x) \tag{22.2.13}$$

Then we get that:

$$\int_a^b \phi^*(x)L[\phi](x)dx = \lambda \int_a^b \phi^*(x)\phi(x)\rho(x)dx \tag{22.2.14}$$

$$\implies \int_a^b \phi(x)(L[\phi](x))^*dx = \lambda^* \int_a^b \phi^*(x)\phi(x)\rho(x)dx \tag{22.2.15}$$

$$\tag{22.2.16}$$

Since L is hermitian, we may write that:

$$\int_a^b \phi(x)(L[\phi](x))^*dx = \int -a^b L^\dagger[\phi](x)\phi^*(x)dx = \int -a^b L[\phi](x)\phi^*(x)dx \tag{22.2.17}$$

implying that:

$$(\lambda - \lambda^*)\int_a^b \phi^*(x)\phi(x)\rho(x)dx = 0 \tag{22.2.18}$$

Since the inner product is positive semi-definite, and ϕ is non-trivial, (22.2.18) is only possible if $\lambda = \lambda^*$, that is if λ is real.

(ii) We repeat the calculation above, but with two different eigenfunctions $\phi_i(x), \phi_j(x)$ with distinct eigenvalues λ_i, λ_j

$$L[\phi_i](x) = \lambda_i\phi_i(x)\rho(x) \tag{22.2.19}$$

$$L[\phi_j](x) = \lambda_j\phi_j(x)\rho(x) \tag{22.2.20}$$

Then we get that:

$$\int_a^b \phi_j^*(x) L[\phi_i](x) dx = \lambda_j \int_a^b \phi_j^*(x) \phi_i(x) \rho(x) dx \qquad (22.2.21)$$

$$\int_a^b \phi_i(x) (L[\phi_j](x))^* dx = \lambda_i \int_a^b \phi_j^*(x) \phi_i(x) \rho(x) dx \qquad (22.2.22)$$

$$(22.2.23)$$

Since L is hermitian, we may write that:

$$\int_a^b \phi_i(x) (L[\phi_j](x))^* dx = \int_a^b \phi_j^*(x) L^\dagger[\phi_i](x) dx = \int_a^b \phi_j^*(x) L[\phi_i](x) dx \qquad (22.2.24)$$

implying that:

$$(\lambda_j - \lambda_i) \int_a^b \phi_j^*(x) \phi_i(x) \rho(x) dx = 0 \implies \langle \phi_i, \phi_j \rangle_\rho = 0 \qquad (22.2.25)$$

as desired.

∎

22.3 Sturm-Liouville problems

We now examine a special type of differential operators that props up very frequently in mathematical physics.

Definition *(Sturm-Liouville operator)* A **Sturm-Liouville operator** is a differential operator of the form:

$$L[y] = \frac{d}{dx}\left(p(x) \frac{dy}{dx} \right) - q(x) y \qquad (22.3.1)$$

Theorem *(Sturm-Louville is self-adjoint)*
All Sturm-Liouville operators are self-adjoint, and can be made hermitian on an interval $[a, b]$ by requiring:

$$\left[p(x) \left(\frac{df^*}{dx} g(x) - \frac{dg}{dx} f(x) \right) \right]_a^b = 0 \qquad (22.3.2)$$

Proof. We find that:

$$\int_a^b (L[f])^*(x) g(x) dx = \int_a^b \left[\frac{d}{dx}\left(p(x) \frac{df}{dx} \right) - q(x) f(x) \right]^* g(x) dx \qquad (22.3.3)$$

$$= \int_a^b \left(p(x) \frac{df^*}{dx} \frac{dg}{dx} - q(x) f^*(x) g(x) dx \right) \qquad (22.3.4)$$

$$- \left[p(x) \frac{df^*(x)}{dx} g(x) \right]_a^b \qquad (22.3.5)$$

and similarly

$$\int_a^b f^*(x)(L[g])(x)dx = \int_a^b \left[\frac{d}{dx}\left(p(x)\frac{dg}{dx}\right) - q(x)g(x)\right]f^*(x)dx \tag{22.3.6}$$

$$= \int_a^b \left(p(x)\frac{dg}{dx}\frac{df^*}{dx} - q(x)f^*(x)g(x)dx\right) \tag{22.3.7}$$

$$- \left[p(x)\frac{dg}{dx}f(x)\right]_a^b \tag{22.3.8}$$

Consequently we see that:

$$\int_a^b f^*(x)(L[g])(x)dx = \int_a^b (L[f])^*(x)g(x)dx + \left[p(x)\left(\frac{df^*}{dx}g(x) - \frac{dg}{dx}f(x)\right)\right]_a^b \tag{22.3.9}$$

so L is indeed self-adjoint. Furthermore, if we impose the boundary condition:

$$\left[p(x)\left(\frac{df^*}{dx}g(x) - \frac{dg}{dx}f(x)\right)\right]_a^b = 0 \tag{22.3.10}$$

then we do indeed find that L is hermitian. ∎

It may seem innatural to look at this special form of second order ODEs. However, much like how we can find an integrating factor to write $y' + p(x)y = q$ into $(p(x)\lambda(x))' = q(x)$, we can write any second order linear ODE into a Sturm-Louville ODE. To see why, consider:

$$p(x)y''(x) + r(x)y' + q(x)y = 0 \tag{22.3.11}$$

We multiply by η so that:

$$\eta(x)p(x)y''(x) + \eta(x)r(x)y' + \eta(x)q(x)y = \frac{d}{dx}(\eta(x)p(x)y'(x)) - \eta(x)q(x)y \tag{22.3.12}$$

implying

$$\eta p y'' + \eta r y' = \eta' p y' + \eta[py'' + p'y'] \tag{22.3.13}$$

$$\implies \eta' p y' + \eta(p'y' - ry') = 0 \tag{22.3.14}$$

$$\implies \frac{\eta'}{\eta} = \frac{r - p'}{p} \implies \eta(x) = \exp\left(\int^x \frac{r(s) - p'(s)}{p(s)}ds\right) \tag{22.3.15}$$

is the required integrating factor. We can simplify it to

$$\eta(x) = \frac{1}{p(x)}\exp\left(\int^x \frac{r(s)}{p(s)}ds\right) \tag{22.3.16}$$

Since $\eta(x)L$ is a Sturm-Liouville operator, its eigenfunctions are orthogonal with respect to the unit weight function, implying that the eigenfunctions of L are orthogonal with respect to the weight function $w(x) = \eta(x)$.

Proposition *(Sturm-Liouville integrating factor)*

The second order ODE

$$p(x)y''(x) + r(x)y' + q(x)y = 0 \tag{22.3.17}$$

may be turned into a Sturm-Louville problem by multiplying by the following integrating factor:

$$\eta(x) = \frac{1}{p(x)} \exp\left(\int^x \frac{r(s)}{p(s)} \, ds\right) \tag{22.3.18}$$

This is equivalent to setting the weight function to be equal to the integrating factor $w(x) = \eta(x)$.

Now suppose we have an inhomogeneous Sturm-Liouville problem:

$$L[y](x) = \rho(x)F(x) \equiv f(x) \tag{22.3.19}$$

Since L is hermitian, its eigenfunctions will form an orthonormal basis $\{\phi_k(x)\}$ of the Hilbert space. Thus the solution $y(x)$ to (22.3.19) and $F(x)$ may be expanded as:

$$y(x) = \sum_n y_n \phi_n(x), \ y_n = \int_a^b \phi_n^*(x)y(x)dx \tag{22.3.20}$$

and

$$F(x) = \sum_n F_n \phi_n(x), \ F_n = \int_a^b \phi_n^*(x)F(x)dx \tag{22.3.21}$$

Substituting this into (22.3.19) we get that:

$$L[y](x) = \sum_n y_n L[\phi_n](x) = \sum_n \lambda_n y_n \phi_n(x) = \sum_n F_n \phi_n(x)\rho(x) \tag{22.3.22}$$

$$\implies \boxed{y(x) = \sum_n \frac{F_n}{\lambda_n}\phi_n(x)} \tag{22.3.23}$$

where in the last step we used the linear independence of $\{\phi_n(x)\}$. We can rewrite the above as:

$$y(x) = \sum_n \frac{\langle \phi_n, F\rangle_\rho}{\lambda_n}\phi_n(x) = \sum_n \left(\frac{\phi_n(x)}{\lambda_n}\int_a^b \phi_n^*(x') \underbrace{F(x')\rho(x')}_{f(x')} \, dx'\right) \tag{22.3.24}$$

This motivates the following definition:

Definition *(Green's function)*
Given a Sturm-Liouville operator L, we define its Green's function to be:

$$G(x, x') = \sum_n \frac{1}{\lambda n}\phi_n^*(x)\phi_n(x) \tag{22.3.25}$$

so that the solution to $L[y](x) = f(x)$ becomes an integral:

$$y(x) = \int_a^b G(x, x')f(x')dx' \tag{22.3.26}$$

> **Proposition** *(Parseval's identity)*
> Let $\{\phi_n\}$ be an orthonormal basis for a Hilbert space with weight function ρ, and let f be function with $\langle \phi_n, f \rangle_\rho = f_n$. Then:
> $$\langle f, f \rangle_\rho = \sum_n |f_n|^2 \tag{22.3.27}$$

Proof. ∎

22.4 Green's functions for BVPs

Green's functions for homogeneous BCs

There is another approach to define Green's functions, a distributional approach which makes use of the delta function.

> **Definition** *(Green's function)*
> Let L be a differential operator. Then we define a Green's function $G(x, x')$ to satisfy:
> $$L[G](x, x') = \delta(x - x') \tag{22.4.1}$$

We would like to verify that this definition is equivalent to the linear algebra approach. Indeed let

$$G(x, x') = \sum_n \frac{1}{\lambda_n} Y_n(x) Y_n^*(x') \tag{22.4.2}$$

then operating L (with variable x, not x') on both sides

$$L[G](x, x') = \sum_n Y_n(x) Y_n^*(x') \tag{22.4.3}$$

$$\implies \int_a^b L[G](x, x') Y_m(x') w(x) \, dx = \sum_n Y_n(x) \langle Y_n, Y_m \rangle_w \tag{22.4.4}$$

$$\implies Y_m(x) = \int_a^b L[G](x, x') Y_m(x') w(x') \, dx' \tag{22.4.5}$$

implying that $\delta(x - x')$ as desired.

The utility of Green's functions is best expressed in the following theorem.

> **Theorem** *(Green's functions)*
> Let $G(x, x')$ be a Green's function of the differential operator L satisfying homogeneous boundary conditions $G(a, x') = G(b, x') = 0$. Then the particular solution of $L[y] = f$ satisfying $y(a) = y(b) = 0$ is given by
> $$y(x) = \int_a^b G(x, x') f(x') \, dx' \tag{22.4.6}$$

Proof. Indeed note that

$$L[y] = \int_a^b L[G](x, x')f(x')\, dx' = \int_a^b \delta(x - x')f(x')\, dx = f(x) \tag{22.4.7}$$

and $y(a) = y(b) = 0$ since $G(a, x') = G(b, x') = 0$, as desired. ∎

We now derive a general procedure to calculate Green's functions with homogeneous boundary conditions. We consider the case where L is a second order operator, and begin by noting that $L[G](x, x') = 0$ for $x > x'$ or $x < x'$ so in these two regions, $G(x, x')$ will simply be equal to the complementary function of L. Therefore suppose we have found a fundamental set of solutions $\{y_1, y_2\}$ which also satisfy one of the BCs i.e. $y_1(a) = 0$ and $y_2(b) = 0$.

It follows that the Green's function to the left of x' is proportional to y_1. and to the right of x' it is proportional to y_2, in order to satisfy the boundary conditions:

$$G(x, x') = \begin{cases} A(x')y_1(x), & x \in [a, x') \\ B(x')y_2(x), & x \in (x', b] \end{cases} \tag{22.4.8}$$

but what should happen at x'? Suppose G were discontinuous at $x = x'$, for simplicity we assume a step discontinuity. Then

$$\frac{\partial G}{\partial x} = \delta(x - x') \implies \frac{\partial^2 G}{\partial x^2} = \delta'(x - x') \tag{22.4.9}$$

which can't be the case since $L[G](x, x')$ does not contain derivatives of δ. Therefore the Green's function must be continuous on $[a, b]$, especially at $x = x'$:

$$A(x')y_1(x') = B(x')y_2(x') \tag{22.4.10}$$

To further investigate the behaviour at x' let's integrate $L[G]$ on $(x' - \epsilon, x' + \epsilon)$:

$$\int_{x'-\epsilon}^{x'+\epsilon} \left(\alpha(x)\frac{\partial^2 G}{\partial x} + \beta(x)\frac{\partial G}{\partial x} + \gamma(x)G \right) dx = 1 \tag{22.4.11}$$

Since G is continuous only the second derivative can contribute to the integral as $\epsilon \to 0$ so

$$\lim_{\epsilon \to 0} \int_{x'-\epsilon}^{x'+\epsilon} \alpha(x)\frac{\partial^2 G}{\partial x}\, dx = \alpha(x')\left(\left.\frac{\partial G}{\partial x}\right|_{x'+} - \left.\frac{\partial G}{\partial x}\right|_{x'-} \right) = 1 \tag{22.4.12}$$

implying that

$$A(x')y_1'(x') - B(x')y_2'(x') = \frac{1}{\alpha(x')} \tag{22.4.13}$$

To summarize, we found the gluing conditions for $G(x < x')$ and $G(x > x')$:

$$\begin{cases} G(x, x')|_{x'+} = G(x, x')|_{x'-} \\ \left.\frac{\partial G}{\partial x}\right|_{x'+} - \left.\frac{\partial G}{\partial x}\right|_{x'-} = \frac{1}{\alpha(x')} \end{cases} \tag{22.4.14}$$

which are equivalent to

$$\begin{cases} A(x')y_1(x') = B(x')y_2(x') \\ A(x')y_1'(x') - B(x')y_2'(x') = \dfrac{1}{\alpha(x')} \end{cases} \tag{22.4.15}$$

These can be solved to give

$$A(x') = \frac{y_2(x')}{\alpha(x')W(x')}, \ B(x') = \frac{y_1(x')}{\alpha(x')W(x')} \tag{22.4.16}$$

where $W(x')$ is the Wronskian. Finally we have that

$$G(x, x') = \begin{cases} \dfrac{y_1(x)y_2(x')}{\alpha(x')W(x')}, & x \in [a, x'] \\ \dfrac{y_1(x')y_2(x)}{\alpha(x')W(x')}, & x \in [x', b] \end{cases} \tag{22.4.17}$$

Example: Tension in a string

Suppose we have a string of density μ that is fixed at its ends $x = 0, L$. Letting $y(x, t)$ represent the vertical displacement of the string at (x, t), then one can derive (see Analytical mechanics lecture notes):

$$T\frac{\partial^2 y}{\partial x^2} - \mu g = \mu \frac{\partial^2 y}{\partial t^2}, \ x \in [0, L], \ y(0) = y(L) = 0 \tag{22.4.18}$$

where T is the tension in the string. In the case of a stationary string we have that its profile will satisfy the following ODE:

$$\frac{d^2 y}{dx^2} = \frac{\mu(x)g}{T} \tag{22.4.19}$$

This equation can be solved using normal methods of for well-behaved $\mu(x)$. However, an interesting discussion arises from letting

$$\mu = m\delta(x - x') \tag{22.4.20}$$

equivalent to having a massless string and adding a point mass m at x'. The result is that we get a homogeneous Sturm-Liouville problem

$$\frac{d^2 y}{dx^2} \frac{mg}{T} \delta(x - x'), \ y(0) = y(L) = 0 \tag{22.4.21}$$

The procedure from the preceding section can thus be put to use, the fundamental set of solutions can be chosen to be $\{x, 1 - \frac{x}{L}\}$. Then the Green's function can be decomposed as

$$G(x, x') = \begin{cases} A(x')x, \ 0 \le x < x' \\ B(x')\left(1 - \dfrac{x}{L}\right), \ x' < x \le L \end{cases} \tag{22.4.22}$$

The condition of continuity at x' yields

$$A(x')x' = B(x')\left(1 - \frac{x'}{L}\right) \implies B(x') = \frac{Lx'}{L - x'}A(x') \tag{22.4.23}$$

while the jump condition for the first derivative yields

$$A(x') + 1 = -\frac{1}{L}B(x') \implies A(x') + 1 = -\frac{x'}{L - x'}A(x') \tag{22.4.24}$$

Therefore we find that

$$A(x') = \frac{x'}{L}, \; B(x') = -x' \tag{22.4.25}$$

giving the following Green's function

$$G(x, x') = \begin{cases} -\left(1 - \dfrac{x'}{L}\right)x, \; 0 \leq x \leq x' \\ -\left(1 - \dfrac{x}{L}\right)x', \; x' \leq x \leq L \end{cases} \tag{22.4.26}$$

Physically, this means that the string will form two straight lines, a negative slope on one side and a positive slope to the right.

We can therefore interpret the Green's function as the system's response at x to a unit impulse at x'. Moreover, the beauty of Green's functions is that it gives us the particular integral $y(x)$ for any $\mu(x)$. Indeed the solution to (22.4.19) is

$$y(x) = \frac{g}{T}\left[\frac{x - L}{L}\int_0^x x'\mu(x') \, dx' + x\int_x^L \frac{x' - L}{L}\mu(x') \, dx'\right] \tag{22.4.27}$$

Physically, we are building up our solution by decomposing $\mu(x)$ into several impulses, calculating the system's response, and summing all the contributions to give the displacement at x.

Green's functions for inhomogeneous BCs

Let us now turn to inhomogeneous boundary conditions. Suppose we have found a particular solution y_P of $L[y] = 0$. If we let G be the Green's functions with homogeneous boundary conditions then

$$y(x) = y_P(x) + \int_a^b G(x, x')f(x') \, dx' \tag{22.4.28}$$

is the particular solution with inhomogeneous BCs.

22.5 Green's functions for IVPs

Green's functions can be used to solve IVPs, such as

$$L[y] = f(t), 1y(t_0) = y'(t_0) = 0 \tag{22.5.1}$$

Then the Green's function $G(t, t')$ will solve the following

$$L[y] = \delta(t - t'), \; G(t_0, t') = G'(t_0, t') = 0 \tag{22.5.2}$$

Suppose we have found a fundamental set of solutions $\{y_1, y_2\}$. For $t_0 \leq t < t'$, we find that

$$G^-(t, t') = A(t')y_1(t) + B(t')y_2(t), \; t_0 \leq t < t' \tag{22.5.3}$$

and apply the initial value conditions

$$\begin{pmatrix} y_1(t_0) & y_2(t_0) \\ y_1'(t_0) & y_2'(t_0) \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \mathbf{0} \tag{22.5.4}$$

but since the Wronskian is non-zero, $A = B = 0$ so that the Green's function $G(t, t') = 0$ for $t < t'$. Physically, this means that the system cannot respond in advance at t to an impulse occurring at a later time t', it implies causality.

For $t_0 \leq t' < t$ we construct

$$G^+(t, t') = C(t')y_1(t) + D(t')y_2(t), \; t > t' \tag{22.5.5}$$

and apply the Green's function boundary conditions at $t = t'$:

$$\begin{cases} C(t')y_1(t') + D(t')y_2(t') = 0 \\ C(t')y_1'(t) + D(t')y_2'(t') = \dfrac{1}{\alpha(t')} \end{cases} \tag{22.5.6}$$

Note that we do not have to apply the initial conditions since $t > t_0$. Solving this system will give $G(t, t')$

$$\boxed{G(t, t') = \begin{cases} 0, & t_0 \leq t < t' \\ G^+(t, t'), & t_0 < t' \leq t \end{cases}} \tag{22.5.7}$$

Example. Consider the equation for a driven harmonic oscillator

$$\frac{d^2 y}{dt^2} + y = f(t), \; y(0) = y'(0) = 0 \tag{22.5.8}$$

The fundamental set of solutions is $\{\sin t, \cos t\}$. We need only consider the case where $t > t'$, where we need to solve

$$\begin{cases} C(t')\sin(t') + D(t')\cos(t') = 0 \\ C(t')\cos(t) - D(t')\sin(t') = 1 \end{cases} \tag{22.5.9}$$

We find that

$$C = -D\cot(t') \implies D = -\sin(t'), \; C = \cos(t') \tag{22.5.10}$$

giving the following Green's function

$$G(t, t') = \begin{cases} 0, & t < t' \\ \sin(t - t'), & t > t' \end{cases} \tag{22.5.11}$$

and the following particular integral

$$y(t) = \int_0^t \sin(t - t')f(t') \, dt' \tag{22.5.12}$$

◄

ODE7 Linear systems of ODEs

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.

23.1 Non-degenerate Eigenvalues

Consider the system of ODEs:
$$\dot{\boldsymbol{y}} = \mathsf{A} \cdot \boldsymbol{y} + \boldsymbol{f}(t) \tag{23.1.1}$$

To find the eigenvalues of the matrix A, it suffices to solve the *eigenvalue equation*:

$$\det(\mathsf{A} - \lambda \mathsf{I}) = 0 \tag{23.1.2}$$

and then find the corresponding eigenvectors using the *eigenvector equation*:

$$(\mathsf{A} - \lambda \mathsf{I}) \cdot \boldsymbol{v}_i = 0. \tag{23.1.3}$$

Proposition 3.

If a matrix has non-degenerate eigenvalues λ_i, then the corresponding eigenvectors $\boldsymbol{v}_i$ form a basis.

Proof. Consider (6.3), if $\{\boldsymbol{v}_i\}$ are linearly dependent, then it must be possible to write one as:

$$\boldsymbol{v_n} = \sum_j \alpha_j \boldsymbol{v_j} \tag{23.1.4}$$

Then, on one hand we find:
$$\mathsf{A} \cdot \boldsymbol{v}_n = \lambda_n \boldsymbol{v}_n = \lambda_n \sum_j \alpha_j \boldsymbol{v_j} \tag{23.1.5}$$

and on the other hand:
$$\mathsf{A} \cdot \boldsymbol{v}_n = \sum_j \alpha_j \mathsf{A} \cdot \boldsymbol{v_j} = \sum_j \lambda_j \alpha_j \boldsymbol{v_j}. \tag{23.1.6}$$

Equating the two gives:
$$\sum_j \alpha_j \boldsymbol{v_j}(\lambda_n - \lambda_j) = 0 \implies \alpha_j = 0 \tag{23.1.7}$$

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:

$$\mathsf{R} \equiv (\boldsymbol{v}_1...\boldsymbol{v}_n), \; \mathsf{L} \equiv \operatorname{diag}\{\lambda_i\}$$
$$\Longrightarrow \mathsf{A} \cdot \mathsf{R} = \mathsf{R} \cdot \mathsf{L}$$
$$\Longrightarrow \mathsf{A} = \mathsf{R} \cdot \mathsf{L} \cdot \mathsf{R}^{-1}$$

We may substitute this into the associated homogeneous equation to (6.1) and find:

$$\boldsymbol{y} = \mathsf{R} \cdot \mathsf{L} \cdot \mathsf{R}^{-1} \cdot \boldsymbol{y}$$
$$\Longrightarrow \frac{d}{dt}(\mathsf{R}^{-1} \cdot \boldsymbol{y}) = \mathsf{L} \cdot \mathsf{R}^{-1} \cdot \boldsymbol{y}$$

and hence we get:

$$\dot{\boldsymbol{\zeta}} = \mathsf{L} \cdot \boldsymbol{\zeta} \implies \zeta_i(t) = C_i e^{\lambda_i t} \tag{23.1.8}$$

We transform back into the original coordinates so that:

$$\boldsymbol{y} = \sum_i C_i \boldsymbol{v}_i e^{\lambda_i t} \tag{23.1.9}$$

For simplicity, define the following matrix:

$$E(t) \equiv \operatorname{diag} e^{\lambda_i t} \tag{23.1.10}$$

which recasts (6.4) as:

$$\boldsymbol{y} = \underbrace{\mathsf{R} \cdot \mathsf{E}(t)}_{\mathsf{Y}(t)} \cdot \mathsf{C} \tag{23.1.11}$$

Finally, let us set the Cauchy condition $\boldsymbol{y}(0) = \boldsymbol{y}_0$

$$\boldsymbol{y}_0 = \mathsf{R} \cdot \mathsf{C} \implies \mathsf{C} = \mathsf{R}^{-1} \cdot \boldsymbol{y}_0 \tag{23.1.12}$$

and hence:

$$\boldsymbol{y} = \mathsf{R} \cdot \mathsf{E}(t) \cdot \mathsf{R}^{-1} \cdot \boldsymbol{y}_0 \tag{23.1.13}$$

23.2 Matrix exponentiation

Consider now the more general Cauchy problem (6.1), which can be solved, as always, via the method of variation of parameters:

$$\boldsymbol{y}(t) = \mathsf{Y}(t) \left[\mathsf{Y}^{-1}(t_0) \cdot \boldsymbol{y}_0 + \int_{t_0}^{t} \mathsf{Y}^{-1}(t') \cdot \boldsymbol{f}(t') dt' \right] \tag{23.2.1}$$

Here:

$$\mathsf{Y}^{-1}(t) = \mathsf{E}^{-1}(t) \cdot \mathsf{R}^{-1} \implies \mathsf{Y}(t) \cdot \mathsf{Y}^{-1}(t') = \mathsf{R} \cdot \mathsf{E}(t) \cdot \mathsf{E}^{-1}(t) \cdot \mathsf{R}^{-1} = \mathsf{R} \cdot \mathsf{E}(t - t') \cdot \mathsf{R}^{-1}$$
$$\mathsf{Y}^{-1}(0) = \mathsf{R}^{-1}$$

so that:

$$\boldsymbol{y}(t) = \mathsf{Y}(t_0) \left[\mathsf{Y}^{-1}(t) \cdot \boldsymbol{y}_0 + \int_{t_0}^{t} \mathsf{R} \cdot \mathsf{E}(t - t') \cdot \mathsf{R}^{-1} \cdot \boldsymbol{f}(t') dt' \right] \tag{23.2.2}$$

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:

$$e^{At} = R \cdot E(t) \cdot R^{-1} \tag{23.2.3}$$

23.3 Higher Order Linear Constant Coefficient Equations

Consider the nth order linear ODE with constant coefficients:

$$y^{(n)} + a_{n-1}y^{(n-1)} + ... + a_1\dot{y} + a_0 y = f(t) \tag{23.3.1}$$

which can be reduced into a first order linear system of ODEs:

$$\begin{cases} \dot{p}_{n-1} = -a_{n-1}p_{n-1} - ... - a_1 p_1 - a_0 y + f(t) \\ \dot{p}_{n-2} = p_{n-1} \\ ... \\ \dot{p}_1 = p_2 \dot{y} = p1 \end{cases} \tag{23.3.2}$$

or in matrix form:

$$\frac{d}{dt}\begin{pmatrix} p_{n-1} \\ \vdots \\ p_1 \\ y \end{pmatrix} = \begin{pmatrix} -a_{n-1} & -a_{n-2} & \ldots & -a_1 & -a_0 \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} p_{n-1} \\ \vdots \\ p_1 \\ y \end{pmatrix} + \begin{pmatrix} f \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{23.3.3}$$

It can then be shown that the eigenvalue equation and auxiliary equation are equivalent:

$$\det(A - \lambda I) = 0 \iff \lambda^n + a_{n-1}\lambda^{n-1} + ... + \lambda a_1 + a_0 = 0 \tag{23.3.4}$$

If the matrix A turns out to be diagonalisable, then the solution is:

$$y(t) = \sum_{i=1}^{n} C_i e^{\lambda_i t} \tag{23.3.5}$$

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.

23.4 Triangulation

Theorem 4. (Schur's Triangulation Theorem)

For any matrix A, *there is a unitary transformation that converts it into triangular form:*

$$U^\dagger \cdot A \cdot U = T = \begin{pmatrix} \lambda_1 & & stuff \\ & ... & \\ 0 & & \lambda_n \end{pmatrix} \tag{23.4.1}$$

Proof. Consider one eigenvector of our matrix:

$$A \cdot \boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1 \tag{23.4.2}$$

and consider an orthonormal basis $\{w_i\}$ with $w_1 = v_1$ and $w_i^* \cdot w_j = \delta_{ij}$. As a consequence of this orthonormality, the matrix:

$$R = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{w}_2 & ... & \boldsymbol{w}_n \end{pmatrix} \tag{23.4.3}$$

is unitary, that is, $R^{-1} = R^\dagger$. Therefore, it follows that:

$$R^\dagger \cdot A \cdot \boldsymbol{R} = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_n \end{pmatrix} \cdot R = \begin{pmatrix} \lambda_1 \boldsymbol{v}_1 & A \cdot \boldsymbol{w}_2 & ... & A \cdot \boldsymbol{w}_n \end{pmatrix} = \begin{pmatrix} \lambda_1 & \boldsymbol{v}_1^* \cdot A \cdot \boldsymbol{w}_2 & ... & \boldsymbol{v}_1^* \cdot A \cdot \boldsymbol{w}_2 \\ 0 & & & \\ \vdots & & A_{n-1} & \\ 0 & & & \end{pmatrix} \tag{23.4.4}$$

where A_{n-1} has elements $\boldsymbol{w}_i^* \cdot A \cdot \boldsymbol{w}_j$. We can keep on repeating this to 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:

$$\dot{\boldsymbol{y}} = A \cdot \boldsymbol{y} = U \cdot T \cdot U^\dagger \implies \frac{d}{dt}(U^\dagger \cdot \boldsymbol{y}) = T \cdot \underbrace{U^\dagger \cdot \boldsymbol{y}}_{\zeta} \implies \dot{\zeta} = \boldsymbol{T} \cdot \zeta \tag{23.4.5}$$

where:

$$T = \begin{pmatrix} \lambda_1 & T_{12} & T_{13} & ... & T_{1n} \\ 0 & \lambda_2 & T_{23} & ... & T_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ... & \lambda_n \end{pmatrix} \tag{23.4.6}$$

This system can be solved component by component, since it is made up of first order ODEs. Oence ζ has been found, simply revert back to $\boldsymbol{y}$:

$$\boldsymbol{y} = U \cdot \zeta \tag{23.4.7}$$

23.5 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 $\boldsymbol{v}_1$. Then, $\{\boldsymbol{v}_1...\boldsymbol{v}_k\}$ is called a *Jordan Chain* if they satisfy:

$$\begin{cases} A \cdot \boldsymbol{v}_1 & = \lambda_1 \boldsymbol{v}_1 \\ A \cdot \boldsymbol{v}_2 & = \lambda_1 \boldsymbol{v}_2 + \boldsymbol{v}_1 \\ & ... \\ A \cdot \boldsymbol{v}_k & = \lambda_1 \boldsymbol{v}_k + \boldsymbol{v}_{k-1} \end{cases} \tag{23.5.1}$$

and are therefore linearly independent. We may do so for a set of eigenvalues $\{\lambda_i\}$ and find the Jordan form:

$$R^{-1} \cdot A \cdot R = J = \mathrm{diag}J_i \tag{23.5.2}$$

where for each eigenvalue λ_1 corresponds a Jordan block:

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & \ldots & 0 \\ 0 & \lambda_1 & 0 & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_i \end{pmatrix} \tag{23.5.3}$$

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:

$$\dot{\boldsymbol{y}} = A \cdot \boldsymbol{y} = R \cdot J \cdot R^{-1} \implies \dot{\boldsymbol{\zeta}} = \boldsymbol{J} \cdot \boldsymbol{\zeta} \tag{23.5.4}$$

where $\boldsymbol{\zeta} = R^{-1} \cdot \boldsymbol{y}$.

For example, consider the Jordan block J_1:

$$\begin{cases} \dot{\zeta}_1 = \lambda_1 \zeta_1 + \zeta_2 \\ \ldots \\ \dot{\zeta}_{k-1} = \lambda_1 \zeta_{k-1} + \zeta_k \\ \dot{\zeta}_k = \lambda_1 \zeta_k \end{cases} \implies \begin{cases} \zeta_k = C_k e^{\lambda_1 t} \\ \zeta_{k-1} = (C_{k-1} + C_k t) e^{\lambda_1 t} \\ \ldots \\ \zeta_1 = \left(C_1 + C_2 t + \ldots + C_k \frac{t^{k1}}{(k-1)!} \right) e^{\lambda_1 t} \end{cases} \tag{23.5.5}$$

Finally, reverting back to $\boldsymbol{y}$:

$$\boldsymbol{y}(t) = \zeta_1 \boldsymbol{v}_1 + \ldots + \zeta_k \boldsymbol{v}_k \tag{23.5.6}$$

ODE8 Series solutions methods

24.1 Power Series

We summarise below a set of properties of power series we encountered in the Analysis part:

1. The power series $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ *converges at a point x* if the following limit exists:

$$\lim_{n \to \infty} \sum_{n=0}^{\infty} a_n(x - x_0)^n \qquad (24.1.1)$$

   and *converges absolutely at a point x* if the associated power series:

$$\sum_{n=0}^{\infty} |a_n(x - x_0)^n| \qquad (24.1.2)$$

   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:
$$L = \lim_{n \to \infty} |x - x_0| |\frac{a_{n+1}}{a_n}| \qquad (24.1.3)$$

   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 continous and is infinitely differentiable over the interval of convergence $|x - x_0| < \rho$.

6. The value of a_n is then given by:
$$a_n = \frac{f^{(n)}(x_0)}{n!} \qquad (24.1.4)$$

   and the series is then called the $Taylor series$:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \qquad (24.1.5)$$

   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.

24.2 Series Solutions near ordinary points

Let us consider the second order linear homogeneous ODE:

$$P(x)y'' + Q(x)y' + R(x)y = 0 \tag{24.2.1}$$

> **Definition** *(Ordinary point)*
> Consider the ODE $P(x)y'' + Q(x)y' + R(x)y = 0$, then any point x_0 such that $P(x_0) = 0$ is an **ordinary point**.

Since P is continuous, we may always find an interval containing x_0, and divide the ODE by $P(x)$ obtaining the more approachable equation

$$y'' + p(x)y' + q(x)y = 0 \tag{24.2.2}$$

where $p(x) = \frac{Q(x)}{P(x)}, q(x) = \frac{R(x)}{P(x)}$. We now look for series solutions of the form:

$$y = \sum_{n=0}^{\infty} a_n(x - x_0)^n \tag{24.2.3}$$

with some 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.

To illustrate this method, let us solve **Airy's equation**

$$y'' - xy = 0, \ -\infty < x < \infty. \tag{24.2.4}$$

We note that $P(x) = 1, Q(x) = 0, R(x) = -x$, hence any point is ordinary. We assume that:

$$y = \sum_{n=0}^{\infty} a_n(x)^n \tag{24.2.5}$$

converges in some interval $|x| < \rho$. Then:

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=0}^{\infty} a_n(x)^{n+1} = 0 \tag{24.2.6}$$

and using a shift of index we may rewrite:

$$2a_2 + \sum_{n=1}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=1}^{\infty} a_{n-1}(x)^n = 0. \tag{24.2.7}$$

This is only possible for all x if the coefficients of like powers of x cancel each other out:

$$(n+2)(n+1)a_{n+2} - a_{n-1} = 0, \ \text{ for } n = 1, 2, 3... \tag{24.2.8}$$

This is a second order recurrence relation, so the coefficients will be determined in steps of 3. Note that:

$$a_2 = 0 \implies a_{3n+2} = 0 \tag{24.2.9}$$

Furthermore:

$$a_3 = \frac{a_0}{2 \cdot 3}, \; a_6 = \frac{a_3}{5 \cdot 6} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6}, \; a_9 = \frac{a_6}{8 \cdot 9} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 \cdot 8 \cdot 9} \qquad (24.2.10)$$

which suggests (as can be proven by induction):

$$a_{3n} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 ... (3n-1)(3n)} \qquad (24.2.11)$$

Finally, we also find:

$$a_4 = \frac{a_1}{3 \cdot 4}, \; a_7 = \frac{a_4}{6 \cdot 7} = \frac{a_1}{3 \cdot 4 \cdot 6 \cdot 7}, \; a_{1}0 = \frac{a_7}{9 \cdot 10} = \frac{a_0}{3 \cdot 4 \cdot 6 \cdot 7 \cdot 9 \cdot 10} \qquad (24.2.12)$$

which suggests:

$$a_{3n+1} = \frac{a_1}{3 \cdot 4 \cdot 6 \cdot 7 ... (3n)(3n+1)} \qquad (24.2.13)$$

Finally, we arrive at the general solution to Airy's equation:

$$\boxed{y(x) = a_0 \left(1 + \sum_{n=1}^{\infty} \frac{x^{3n}}{2 \cdot 3 \cdot 5 ... (3n)} \right) + a_1 \left(x + \sum_{n=1}^{\infty} \frac{x^{3n+1}}{3 \cdot 4 \cdot 6 ... (3n+1)} \right)} \qquad (24.2.14)$$

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 we have truly found the general solution.

It remains to be justified that given an ODE of the form:

$$P(x)y''(x) + Q(x)y'(x) + R(x)y(x) = 0 \qquad (24.2.15)$$

we can find a series solution. Consider for example that we have found a solution $y = \phi(x)$ which can be expanded as a Taylor series:

$$\phi(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n \qquad (24.2.16)$$

which converges in the interval $|x - x_0| < \rho, \; \rho > 0$. We can differentiate (24.2.16) m times to find:

$$m! \cdot a_m = \phi^{(m)}(x_0) \qquad (24.2.17)$$

so the coefficients in the series expansion are determined by evaluating derivatives of ϕ at the ordinary point x_0. Can this always be done? Suppose we have W.L.O.G the initial conditions $\phi(x_0) = y_0, \phi'(x_0) = y_0'$ implying that $a_0 = y_0, a_1 = y_0'$. Now note that there exists interval around x_0 for which $P(x) \neq 0$. Consequently we can determine a_2:

$$a_2 = \frac{1}{2}\phi''(x_0) = -\frac{1}{2}(p(x_0)\phi'(x_0) + q(x_0)p(x_0)) \qquad (24.2.18)$$

where $p(x_0) = Q(x_0)/P(x_0)$, $q(x_0) = R(x_0)/P(x_0)$. Similarly we can find a_3 by differentiating (24.2.15) and evaluating at x_0, and continue for all other a_n. Note that since P, Q, R are polynomials they can be differentiated infinitely many times and $P(x_0) \neq 0$ so we can keep repeating this process indefinitely.

It seems like we only need to assume that the functions $p(x) = \frac{Q(x)}{P(x)}$ and $q(x) = \frac{R(x)}{P(x)}$ are smooth near x_0. This condition is unfortunately too weak, the better condition is that they be analytic at x_0. We should therefore re-define our definition of ordinary point to the following:

> **Definition** *(Ordinary point redefined)*
> The point x_0 is an ordinary point of $P(x)y'' + Q(x)y' + R(x)y = 0$ if $p(x) = \frac{Q(x)}{P(x)}$ and $q(x) = \frac{R(x)}{P(x)}$ are analytic at x_0.

Regarding the convergence of our series solution, it is expected that it will be bounded below by the convergence of the Taylor series for p and q, as the following theorem (which we shall not prove) states.

> **Theorem** *(Convergence radius of series solutions)*
> Let x_0 be an ordinary point of:
>
> $$P(x)y'' + Q(x)y' + R(x)y = 0, \tag{24.2.19}$$
>
> Then the general solution may be expressed as a power series about x_0:
>
> $$y = \sum_{n=0}^{\infty} a_n (x - x_0)^n = a_0 y_1 + a_1 y_2 \tag{24.2.20}$$
>
> where 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 ρ_p, ρ_q of the Taylor series for $p = \frac{Q(x)}{P(x)}$ and $q = \frac{R(x)}{P(x)}$ respectively:
>
> $$\rho_y \geq \min\{\rho_p, \rho_q\}. \tag{24.2.21}$$

Note that if $P(x), Q(x), R(x)$ are polynomials with common factors cancelled out, then it follows from a theorem in complex analysis that the Taylor series for $p(x)$ and $q(x)$ have a radius of convergence equal to the distance between x_0 and the closest zero of $P(x)$. This gives us a very powerful tool to determine the convergence properties of series solutions quickly and efficiently. In the case of Airy functions, for example, it is clear that $P(x) = 1$ has no zeros so the radius of convergence is infinite.

24.3 Euler equations

Before developing a full theory for series solutions about singular points, we should consider an illustrative example. The Euler equation reads:

$$x^2 y''(x) + \alpha x y'(x) + \beta y(x) = 0, \ \alpha, \beta \in \mathbb{R} \tag{24.3.1}$$

We begin by assuming $x > 0$ and try the ansatz $y = x^r$ then:

$$r(r-1)x^r + \alpha r x^r + \beta x^r = 0 \implies r^2 + (\alpha - 1)r + \beta = 0 \tag{24.3.2}$$

which has solutions:

$$r_{1,2} = -\frac{1}{2}\left[(\alpha - 1) \pm \sqrt{(\alpha - 1)^2 - 4\beta}\right] \tag{24.3.3}$$

As in the case of second order constant coefficient ODEs, we can distinguish between three different cases:

(i) If $(\alpha - 1)^2 - 4\beta > 0$ then we have a fundamental set of solutions and we find:

$$y(x) = c_1 x^{r_1} + c_2 x^{r_2} \tag{24.3.4}$$

(ii) If $(\alpha - 1)^2 - 4\beta > 0$ then $r_1 = \lambda + i\mu, r_2 = r_1^*$ for some $\lambda, \mu \in \mathbb{R}$. Consequently

$$x^{r_1} = x^\lambda x^{i\mu} = x^\lambda e^{i\mu \ln x} = x^\lambda [\cos(\mu \ln x) + i \sin \sin(\mu \ln x)] \tag{24.3.5}$$

giving the fundamental set of solutions with general form:

$$y(x) = x^\lambda [c_1 \sin(\mu \ln x) + c_2 \cos(\mu \ln x)] \tag{24.3.6}$$

(iii) If $(\alpha - 1)^2 - 4\beta = 0$ then $r_1 = r_2$ so we have only found one solution $y = x^{r_1}$. We could find another one by method of reduction of order, but here we present a different derivation that is much faster. Note that:

$$L[x^r] = (r^2 + (\alpha - 1) + \beta)x^r = \left(r^2 + 2\frac{\alpha - 1}{2}r + \frac{(\alpha - 1)}{4}\right)x^r \tag{24.3.7}$$

$$= (r - \frac{\alpha - 1}{2})^2 x^r = (r - r_1)^2 x^r \tag{24.3.8}$$

so that

$$L\left[\frac{\partial x^r}{\partial r}\right] = L[x^r \ln x] = \frac{\partial}{\partial r}(L[x^r]) = 2(r - r_1)x^r + (r - r_1)^2 x^r \ln x \tag{24.3.9}$$

For the above to vanish we need $r = r_1$ so we have found another solution $y = x^{r_1} \ln x$. Consequently the general solution is given by:

$$y(x) = (c_1 + c_2 \ln x)x^r \tag{24.3.10}$$

Finally, note that if we let $u = -x$ then the Euler equation remains unchanged, so our solutions for $x > 0$ will also hold for $x < 0$ by simply changing the sign of x.

24.4   Frobenius' method

We can rewrite the Euler equations as

$$y''(x) + \frac{\alpha}{x}y'(x) + \frac{\beta}{x^2}q(x) = 0 \tag{24.4.1}$$

which suggests generalizing to ODEs of the form

$$y''(x) + p(x)y'(x) + q(x) = 0 \tag{24.4.2}$$

where we can write the following series expansion

$$p(x) = \frac{p_0}{x} + p_1 + p_2 x + p_3 x^2 ... \tag{24.4.3}$$

$$q(x) = \frac{q_0}{x^2} + \frac{q_1}{x} + q_2 + q_3 x ... \tag{24.4.4}$$

Definition *(Regular singular points)* The equation (24.4.2) has a **regular singular point** at $x = x_0$ if $(x - x_0)p(x)$ and $(x - x_0)^2 q(x)$ are analytic at x_0. A non-regular singular point is **irregular**.

Let us multiply (24.4.2) by x^2 to get:

$$x^2 y''(x) + x(xp(x))y'(x) + x^2 q(x) y(x) = 0 \tag{24.4.5}$$

We saw that when constants α, β got promoted to functions $p(x), q(x)$, it helped to assume a power series solution. Inspired by this we may be able to obtain a solution by multiplying our ansatz x^r by a power series:

$$y(x) = x^r \sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} a_n x^{n+r} \tag{24.4.6}$$

We see that:

$$\sum_{n=0}^{\infty} a_n (n+r)(n+r-1) x^{n+r} + \left(\sum_{m=0}^{\infty} p_m x^m \right) \sum_{n=0}^{\infty} a_n (n+r) x^{n+r} \tag{24.4.7}$$

$$+ \left(\sum_{m=0}^{\infty} q_m x^m \right) \sum_{n=0}^{\infty} a_n x^{n+r} = 0 \tag{24.4.8}$$

We can divide by x^r and collect like-terms. For example the x^0 term reads:

$$a_0[r(r-1) + p_0 r + q_0] = 0 \tag{24.4.9}$$

while more generally the x^k (k>0) coefficient reads

$$a_k(k+r)(k+r-1) + \sum_{n=0}^{k-1}[p_{k-n} a_n (n+r) + q_{k-n} a_n] = 0 \tag{24.4.10}$$

or alternatively:

$$[a_k(k+r)(k+r-1) + p_0 a_k(k+r) + q_0 a_k] + \sum_{n=0}^{k-1}[p_{k-n} a_n (n+r) + q_{k-n} a_n] = 0 \tag{24.4.11}$$

Letting $F(r) = r(r-1) + p_0 r + q_0$ then we get the following recurrence relation

$$\boxed{F(k+r)a_k + \sum_{n=0}^{k-1}[p_{k-n}(n+r) + q_{k-n}]a_n = 0} \tag{24.4.12}$$

Note that for $a_0 \neq 0$ we find that $a_0 F(r)x^r = 0$ so we get the **indicial equation**

$$\boxed{F(r) = r(r-1) + p_0 r + q_0 = 0} \tag{24.4.13}$$

which, as before with the Euler equation has two solutions which may or may not be distinct. However, in our case we have a further problem when r_1 and r_2 differ by some integer. We provide the solution without proof.

Theorem *(Series about regular singular points)*
Suppose we have an ODE of the form given in (24.4.2) with a regular singular point at $x = 0$ so that $xp(x)$ and $x^2 q(x)$ are analytic at $x = 0$:

$$xp(x) = \sum_{n=0}^{\infty} p_n x^n, \ x^2 q(x) = \sum_{n=0}^{\infty} q_n x^n \tag{24.4.14}$$

with radius of convergence ρ. Let r_1 and r_2 be roots of:

$$r(r-1) + p_0 r + q_0 = 0 \tag{24.4.15}$$

with $r_1 \geq r_2$. Then we have two linearly independent solutions on $0 < x < \rho$:
   (i) if $r_1 - r_2$ is not an integer then:

$$y(x) = c_1 x^{r_1} \sum_{n=0}^{\infty} a_n x^n + c_2 x^{r_2} \sum_{n=0}^{\infty} b_n x^n \tag{24.4.16}$$

(ii) if $r_1 = r_2$ then:

$$y(x) = c_1 (1 + \ln x) x^{r_1} \sum_{n=0}^{\infty} a_n x^n + c_2 x^{r_1} \sum_{n=0}^{\infty} b_n x^n \tag{24.4.17}$$

(iii) if $r_1 - r_2 = N$ is an integer then:

$$y(x) = c_1 (1 + a \ln x) x^{r_1} \sum_{n=0}^{\infty} a_n x^n + c_2 x^{r_2} \sum_{n=0}^{\infty} b_n x^n \tag{24.4.18}$$

Proof. The special case of equal roots to the indicial equation can be solved using the same method used for the Euler equation. We know that one solution can be found:

$$y(x) = x^{r_1} \sum_{n=0}^{\infty} a_n x^n \tag{24.4.19}$$

Now recall that:

$$L[y] = a_0 F(r) x^r + \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\{ [a_k F(k+r) \right. \tag{24.4.20}$$

$$\left. + \sum_{n=0}^{k-1} [p_{k-n} a_n (n+r) + q_{k-n} a_n] \right\} x^{k+r} \tag{24.4.21}$$

so treating the coefficients $a_m(r)$ as functions of r then we can set:

$$a_k(r) = -\frac{\sum_{n=0}^{k-1}[p_{k-n}a_n \cdot (n+r) + q_{k-n}a_n]}{F(k+r)} \implies L[y] = a_0 F(r) x^r \tag{24.4.22}$$

giving us the solution to $L[y] = 0$

$$y(x) = x^{r_1} \sum_{n=0}^{\infty} a_n x^n \tag{24.4.23}$$

where r_1 is the root of $F(r)$. Now once again we consider the derivative of $L[y]$ with respect to r, and recall that $F(r) = (r - r_1)^2$ for repeated roots:

$$\frac{\partial}{\partial r} L[y] = L\left[\frac{\partial y}{\partial r}\right] = 2a_0(r - r_1)x^r + a_0(r - r_1)^2 x^r \ln x \tag{24.4.24}$$

which again vanishes when $r = r_1$. Hence we find that:

$$y_2(x) = \left.\frac{\partial y_1}{\partial r}\right|_{r=r_1} = y_1(x) \ln x + \sum_{n=0}^{\infty} b_n x u u^{n+r_1} \tag{24.4.25}$$

where $b_n = a'_n(r)|_{r=r_1}$. We have thus derived the required result. ∎

ODE9 Special functions

ODE10 Distributions

26.1 Introducing the Dirac delta

Suppose we have a random variable x with a gaussian probability distribution:

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/2\sigma^2} \tag{26.1.1}$$

Note that this being a probability density function implies that it is properly normalized, and thus satisfies:

$$\int_\mathbb{R} \rho(x)dx = 1 \tag{26.1.2}$$

We ask what would happen if we let $\sigma \to 0$?. Clearly we must have that for $x \neq 0$, $\lim_{\sigma\to0} \rho(x \neq 0) = 0$ due to the exponential suppression overpowering the $1/\sigma$ factor in the front. However, we must still have that:

$$\lim_{\sigma\to0} \int_\mathbb{R} \rho(x)dx = 1 \tag{26.1.3}$$

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:
> $$\delta(x) \equiv \lim_{\sigma\to0} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/2\sigma^2} \tag{26.1.4}$$
> which satisfies:
> $$\begin{cases} \delta(x) = 0, \ x \neq 0 \\ \int_\mathbb{R} \delta(x)dx = 1 \end{cases} \tag{26.1.5}$$

Statistically, the Dirac delta represents the probability distribution of a random variable with zero variance.

Let's now consider:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} \rho(x')dx' \tag{26.1.6}$$

which is the probability of the variable X having a value smaller than x. As long as $\sigma > 0$ we find that:

$$\frac{d\mathbb{P}(X \leq x)}{dx} = \rho(x') \tag{26.1.7}$$

Should we expect this to hold for $\sigma \to 0$? In this limit we have that:

$$\mathbb{P}(X \leq x) \to \Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{26.1.8}$$

since the probability of X being smaller than one should be zero. Extending (26.1.7) in the $\sigma \to 0$ case gives:

$$\frac{d\Theta(x)}{dx} = \delta(x) \tag{26.1.9}$$

We would like to make this reasoning more rigorous, and will require us to introduce the notion of distributions.

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:

$$\frac{d^2 y}{dx^2} = 0, \ x \neq 0 \tag{26.1.10}$$

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:

$$2T \sin \theta \approx 2T \tan \theta = T \left[\frac{dy}{dx} \right]_{x=-L}^{x=L} = W \tag{26.1.11}$$

which can be solved:

$$y(x) = \begin{cases} -W(L+x)/2T, & -L < x < 0 \\ -W(L-x)/2T, & 0 < x < L \end{cases} \tag{26.1.12}$$

Note that this solution satisfies:

$$T \frac{d^2 y}{dx^2} = W \delta(x) \tag{26.1.13}$$

> **Theorem** *(Delta sifting property)*
> For an object $\delta(x)$ satisfying (26.1.5), letting $f(x) \in L^1(\mathbb{R})$ then:
>
> $$\int_{\mathbb{R}} f(x)\delta(x)dx = f(0) \tag{26.1.14}$$

Proof. Let $\delta > 0$ be small. Then we have that:

$$\int_{\mathbb{R}} f(x)\delta(x)dx = \int_{-\delta}^{\delta} f(x)\delta(x)dx \tag{26.1.15}$$

since $\delta(x) = 0$ for $x \neq 0$. Next:

$$\int_{\mathbb{R}} f(x)\delta(x)dx = \int_{-\delta}^{\delta} (f(x) - f(0) + f(0))\delta(x)dx \tag{26.1.16}$$

$$= \int_{-\delta}^{\delta} (f(x) - f(0))\delta(x)dx + f(0) \tag{26.1.17}$$

Since $f(x)$ is continuous we must have that for any $\epsilon > 0$ there is an $\delta > 0$ such that:

$$|f(x) - f(0)| < \epsilon, \ \forall x \text{ s.t. } |x| < \delta \tag{26.1.18}$$

so we can make the integrand $(f(x) - f(0))$ arbitrarily small:

$$\int_{-\delta}^{\delta} (f(x) - f(0))\delta(x)dx = 0 \tag{26.1.19}$$

and thus

$$\int_{\mathbb{R}} f(x)\delta(x)dx = f(0) \tag{26.1.20}$$

as desired.

It is clear that to rigorously treat the delta function we should not treat it as a normal function as the name would suggest, but rather as a generalized function defined through its action on other, well-behaved functions. ∎

26.2 Rigorous treatment-distributions

We begin by defining exactly how these "well-behaved functions", known as test functions, must behave like. We will not be particularly interested in the exact forms of these test functions, as long as the following conditions are satisfied.

Definition *(Test function)*
A function $\phi : \mathbb{R} \to \mathbb{R}$ is a test function if:
  (i) $\phi(x) \in C^\infty$ (smoothness)
  (ii) $\phi(x) = 0$ has compact support, it vanishes outside some interval.

This definition ensures that $\phi(x) \to 0$ 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. Although the exact form of test functions is not particularly important, we construct an example for sake of clarity. Indee one famous example of a test function can be constructed from:

$$\Phi(x) = \begin{cases} 0, \ x \leq 0 \\ e^{-1/x}, \ x > 0 \end{cases} \tag{26.2.1}$$

Note that this function is infinitely differentiable (with all derivatives vanishing at $x = 0$ since its nth 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 $+\infty$ but we can easily fix this by defining:

$$\phi(x) = \Phi(x)\Phi(1 - x) \tag{26.2.2}$$

which does indeed have a compact support.

Definition *(Action on test functions)*
Let $f(x) \in L^1(\mathbb{R})$ be a measurable function. Then we define its action on a test function $\phi(x)$

---

by:
$$\langle f, \phi \rangle = \int_{\mathbb{R}} f(x)\phi(x)dx \tag{26.2.3}$$

Note that this is not the same as a normal inner product since f and ϕ need not lie in the same vector space. We should ask ourselves if this definition of action is well-defined: can two functions have the same action on a given test function?

Theorem *(Unique definition of distribution)*
Let $f_{1,2} : \mathbb{R} \to \mathbb{R}$ 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:
$$\langle f, \phi \rangle = \int_{\mathbb{R}} f(x)\phi(x) = 0 \tag{26.2.4}$$

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:
$$\langle f, \phi \rangle = \int_{a-\delta}^{a+\delta} f(x)\phi(x)dx > 0 \tag{26.2.5}$$

This however is a contradiction, so we cannot have that $f(a) > 0$ at some $x = a$. Applying this result to $f(x) = f_1(x) - f_2(x)$ immediately reproduces the theorem. ∎

Much like in functional analysis, we will also need to define a property of convergence for sequences of test functions $\{\phi_n(x)\}$.

Definition *(Convergence of test function sequences)* The sequence $\{\phi_n\}$ of test functions converges to zero if:
(i) for all x outside some interval I, $\phi_n(x) = 0$ for all n, so there is a shared compact support for all test functions in the sequence,
(ii) for all k, $\phi_n^{(k)}$ converges uniformly to 0 as $n \to \infty$.

We are now ready to define distributions.

Definition *(Distribution)*
A distribution (or generalised function) $\mathcal{F}$ is a continuous functional mapping from the set of test functions $\mathbb{D}$ to $\mathbb{R}$ defined by the action:
$$\phi \to \langle f, \phi \rangle \in \mathbb{R} \tag{26.2.6}$$

It must satisfy:
(i) Continuous: if $\phi_n \to 0$ then $\langle \mathcal{F}, \phi_n \rangle \to 0$.
(ii) Linear: $\langle \mathcal{F}, \alpha\phi + \beta\psi \rangle = \alpha \langle \mathcal{F}, \phi \rangle + \beta \langle \mathcal{F}, \psi \rangle$ for real constants α, β.

An example of a distribution is the Heaviside distribution $\mathcal{T}$ generated from the Heaviside function $\Theta(x)$:

$$\langle \mathcal{T}, \phi \rangle = \int_0^\infty \phi(x) dx \tag{26.2.7}$$

Clearly this distribution is linear. It is continuous since if $\phi_n(x) \to 0$ then $\langle \mathcal{T}, \phi_n \rangle \to 0$ too.

Proposition *(Delta function as a distribution)*
The delta function $\delta(x)$ defined by the action $\langle \delta, \phi \rangle = \phi(0)$ is a distribution.

Proof. Linearity is trivial. If $\phi_n \to 0$ then by the uniform convergence of test functions $\phi_n(0) \to 0$ so $\langle \delta, \phi_n \rangle \to 0$. ∎

Proposition *(Delta function on $\mathcal{C}^0(\mathbb{R})$)*
If $f(x)$ is continuous at $x = 0$ then

$$\langle \delta_n, f \rangle \to f(0) \text{ as } n \to \infty \tag{26.2.8}$$

where

$$\delta_n = \begin{cases} \frac{n}{2}, \ |x| < \frac{1}{n} \\ 0, \ |x| \geq \frac{1}{n} \end{cases} \tag{26.2.9}$$

so $\delta(x)$ can be applied on continuous functions too:

$$\langle \delta, f \rangle = f(0), \ \forall f \in \mathcal{C}^0(\mathbb{R}) \tag{26.2.10}$$

Proof. For any continuous function $f(x) \in \mathcal{C}^0(\mathbb{R})$:

$$\langle \delta_n, f \rangle = \frac{n}{2} \int_{-1/n}^{1/n} f(x) dx = \frac{n}{2} f(\eta_n) \int_{-1/n}^{1/n} dx = f(\eta_n) \to f(0) \tag{26.2.11}$$

∎

This is a very important result because it means that the action of $\delta(x)$ is not restricted to the space of test functions as is the default case for distributions. It makes sense to apply $\delta(x)$ to normal, continuous functions too.

Definitions *(Operations with distributions)*
Let $F(x)$ be a distribution. Then we define:

$$\langle \mathcal{F}(x-a), \phi(x) \rangle = \langle \mathcal{F}(x), \phi(x+a) \rangle \tag{26.2.12}$$

$$\langle \mathcal{F}(ax), \phi(x) \rangle = \frac{1}{|a|} \langle \mathcal{F}(x), \phi(x/a) \rangle, \ a \neq 0 \tag{26.2.13}$$

Applying this to the delta function we see that:

$$\boxed{\langle \delta(x-a), f(x) \rangle = \int_\mathbb{R} \delta(x-a) f(x) dx = f(a)} \tag{26.2.14}$$

and

$$\boxed{\langle \delta(ax), f(x) \rangle = \int_{\mathbb{R}} \delta(ax) f(x) dx = \frac{1}{|a|} f(0)} \qquad (26.2.15)$$

known as sifting properties.

To see where these definitions come from, let us turn to the integral approach and consider these distributions as normal functions. Then

$$\langle \mathcal{F}(x-a), \phi(x) \rangle = \int_{\mathbb{R}} \mathcal{F}(x-a) \phi(x) dx = \int_{\mathbb{R}} \mathcal{F}(s) \phi(s+a) ds = \langle \mathcal{F}(x), \phi(x+a) \rangle \qquad (26.2.16)$$

and similarly:

$$\langle \mathcal{F}(ax), \phi(x) \rangle = \int_{\mathbb{R}} \mathcal{F}(ax) \phi(x) dx = \frac{1}{|a|} \int_{\mathbb{R}} \mathcal{F}(s) \phi(s/a) ds = \frac{1}{|a|} \langle \mathcal{F}(x), \phi(x/a) \rangle \qquad (26.2.17)$$

where the modulus originates from the fact that if $a < 0$ then the integral bounds are flipped, giving a negative sign.

Similarly, we may use this approach to motivate a definition for the derivative of a distribution. Then we see that:

$$\langle \mathcal{F}', \phi \rangle = \int_{\mathbb{R}} \mathcal{F}'(x) \phi(x) dx = \left[\mathcal{F}(x) \phi(x) \right]_{-\infty}^{\infty} - \int_{\mathbb{R}} \mathcal{F}(x) \phi'(x) dx = - \langle \mathcal{F}, \phi \rangle \qquad (26.2.18)$$

prompting us to define:

> **Definition** *(Differentiating distributions)*
> Let $\mathcal{F}$ be a distribution. Then its derivative $\mathcal{F}'$ is defined so as to satisfy $\langle \mathcal{F}', \phi \rangle = - \langle \mathcal{F}, \phi \rangle$.

We know that the derivative of a test function is a distribution. Are the derivatives of distributions also differentiable then?

> **Theorem** *(Derivatives of distributions are distributions)* Let $\mathcal{F}$ be a distribution, then so is its derivative.

Proof. The action of $\mathcal{F}'$ is linear due to the linearity of $\mathcal{F}$. Also, if $\phi_n \to 0$ then $\phi' \to 0$ by uniform convergence so that:

$$\langle \mathcal{F}', \phi_n \rangle = - \langle \mathcal{F}, \phi'_n \rangle \to 0 \qquad (26.2.19)$$

∎

The above result also tells us that since test functions can be infinitely differentiated, any distribution can also by differentiated infinitely many times. This allows us to prove a wonderful result. Consider the derivative of the Heaviside function:

$$\langle \mathcal{T}', \phi \rangle = - \langle \mathcal{T}, \phi' \rangle = - \int_0^{\infty} \phi'(x) dx = -\phi(0) = - \langle \delta, \phi \rangle \implies \mathcal{T}' = -\delta \qquad (26.2.20)$$

In other words we get the result we previously wanted to prove rigorously, that the derivative of the Heaviside function is the Dirac delta function:

$$\boxed{\frac{d\Theta(x)}{dx} = \delta(x)} \tag{26.2.21}$$

Similarly we see that the derivative of the delta function can be defined via the action:

$$\langle \delta', f \rangle = - \langle \delta, f' \rangle = -f'(0) \tag{26.2.22}$$

It seems like these distributions satisfy typical rules of differential calculus. For example, we can recover the product rule of differentiation using the definition and properties of distributions:

> **Proposition** *(Leibniz rule)*
> If $\mathcal{F}$ is a distribution and $f \in C^\infty(\mathbb{R})$ then $(f\mathcal{F})' = f\mathcal{F}' + f'\mathcal{F}$.

Proof. We have that:

$$\langle (f\mathcal{F})', \phi \rangle = - \langle f\mathcal{F}, \phi' \rangle = - \langle \mathcal{F}, \phi' f \rangle \tag{26.2.23}$$

$$= - \langle \mathcal{F}, (\phi f)' \rangle + \langle \mathcal{F}, f'\phi \rangle = \langle f\mathcal{F}', \phi \rangle + \langle f'\mathcal{F}, \phi \rangle \tag{26.2.24}$$

as desired. ∎

ODE11 Laplace transform methods

27.1 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:
> $$F(s) = \mathcal{L}\{f(t)\} \equiv \int_0^\infty e^{-st} f(t) dt \tag{27.1.1}$$

It is important to note that the Laplace transform is linear due to the linearity of integration so that:
$$\mathcal{L}\{\alpha f(t) + \beta g(t)\} = \alpha \mathcal{L}\{f(t)\} + \beta \mathcal{L}\{g(t)\} \tag{27.1.2}$$

Example. Compute the Laplace transform of the Heaviside function:
$$\Theta(t) = \begin{cases} 1, \ t \geq 0 \\ 0, \ t \leq 0 \end{cases} \tag{27.1.3}$$

We find that:
$$\mathcal{L}\{\Theta(t)\} = \int_0^\infty e^{-st} dt = -\frac{1}{s} \lim_{\tau \to \infty} (e^{-\tau s} - 1) = \frac{1}{s}, \ s > 0 \tag{27.1.4}$$

where we must assume that $s > 0$ for the Laplace transform to be well defined. By the linearity of $\mathcal{L}$ we may write:
$$\mathcal{L}\{c\Theta(t)\} = \frac{c}{s} \tag{27.1.5}$$

Note also that for any function $f(t)$ integrable on $[0, \infty)$:
$$\mathcal{L}\{f(t)\Theta(t)\} = \mathcal{L}\{f(t)\} \tag{27.1.6}$$

so from (27.1.5) we have found that:
$$\mathcal{L}\{c\} = \frac{c}{s} \tag{27.1.7}$$

◀

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:

$$\mathcal{L}\{f(t)\Theta(t)\} = F(s) \implies \mathcal{L}\{F(s)\}^{-1} = f(t)\Theta(t) \tag{27.1.8}$$

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:

$$\mathcal{L}\{t\} = \int_0^\infty te^{-st}dt \tag{27.1.9}$$

which can be solved using Feynman's trick. Let:

$$I(s) = \int_0^\infty te^{-st}dt \tag{27.1.10}$$

Then:

$$I(s) = -\frac{d}{ds}\left(\int_0^\infty e^{-st}\right) = -\frac{d}{ds}\left(\frac{1}{s}\right) = \frac{1}{s^2}, \ s > 0 \tag{27.1.11}$$

so we have that:

$$\mathcal{L}\{t\} = \mathcal{L}\{t\Theta(t)\} = \frac{1}{s^2}, s > 0 \tag{27.1.12}$$

◀

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:

$$\mathcal{L}\{e^{kt}\} = \mathcal{L}\{e^{kt}\Theta(t)\} = \int_0^\infty e^{-(s-k)t}dt = \frac{1}{s-k}, \ s > k \tag{27.1.13}$$

It follows that:

$$\mathcal{L}\{e^{ikt}\} = \mathcal{L}\{\cos kt\} + i\mathcal{L}\{\sin kt\} = \frac{s}{s^2+k^2} + i\frac{k}{s^2+k^2}, \ s > 0 \tag{27.1.14}$$

implying that:

$$\mathcal{L}\{\cos kt\} = \frac{s}{s^2+k^2}, \ \mathcal{L}\{\sin kt\} = \frac{k}{s^2+k^2}, \ s > 0 \tag{27.1.15}$$

◀

Proposition *(Properties of the Laplace transform)* Given a function $f(t) \in L^p(\mathbb{R}^+)$ with Laplace transform:

$$F(s) = \mathcal{L}\{f(t)\} \tag{27.1.16}$$

we have that:

$$\mathcal{L}\{e^{kt}f(t)\} = F(s-k) \tag{27.1.17}$$

and

$$\mathcal{L}\{t^n f(t)\} = (-1)^n \frac{d^n F(s)}{ds^n} \tag{27.1.18}$$

Proof. Firstly:

$$\mathcal{L}\{e^{kt} f(t)\} = \int_0^\infty e^{-(s-k)t} f(t) dt = F(s-k) \tag{27.1.19}$$

Secondly:

$$\mathcal{L}\{t^n f(t)\} = \int_0^\infty t^n f(t) e^{-st} dt = (-1)^n \frac{d^n}{ds^n} \left(\int_0^\infty f(t) e^{-st} dt \right) = (-1)^n \frac{d^n F(s)}{ds^n} \tag{27.1.20}$$

∎

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:

$$\mathcal{L}\{f^{(n)}(t)\} = s^n F(s) - \sum_{i}^{n-1} s^{n-i} f^{(i)}(0) \tag{27.1.21}$$

Proof. We proceed by induction. We have that:

$$\mathcal{L}\{\dot{f}(t)\} = \int_0^\infty \dot{f}(t) e^{-st} dt = \int_0^\infty \frac{d}{dt}(f(t) e^{-st}) dt + \int_0^\infty s f(t) e^{-st} dt \tag{27.1.22}$$

$$= sF(s) + [f(t) e^{-st}]_0^\infty \tag{27.1.23}$$

$$= sF(s) - f(0) \tag{27.1.24}$$

assuming that $f(t)$ is dominated by the exponentially decaying e^{-st} as $t \to \infty$. Let us now suppose that (27.1.21) is true up to n. Then:

$$\mathcal{L}\{f^{(n+1)}(t)\} = \mathcal{L}\{f^{(\dot{n})}(t)\} \qquad\qquad = s\mathcal{L}\{f^{(n)}(t)\} - f^{(n)}(0) \tag{27.1.25}$$

$$= s^{n+1} F(s) + \sum_{i}^{n} s^{n-i} f^{(i)}(0) \tag{27.1.26}$$

∎

Theorem *(Laplace transform of an integral)* Let $f(t')$ be integrable over $[0, t]$. Then:

$$\mathcal{L}\{\int_0^t f(t') dt'\} = \frac{1}{s} F(s) \tag{27.1.27}$$

Proof. Define:

$$g(t) = \int_0^t f(t') dt' \implies \dot{g}(t) = f(t) \tag{27.1.28}$$

Then:

$$\mathcal{L}\{\dot{g}(t)\} = F(s) = s\mathcal{L}\{g(t)\} \implies \mathcal{L}\{g(t)\} = \frac{1}{s}F(s) \tag{27.1.29}$$

as desired. ∎

> **Proposition** *(Second shift theorem)* We have that:
>
> $$\mathcal{L}\{f(t-a)\Theta(t-a)\} = e^{-sa}F(s) \tag{27.1.30}$$

We have that:

$$\mathcal{L}\{f(t-a)\Theta(t-a)\} = \int_0^\infty \Theta(t-a)e^{-st}f(t-a)dt \tag{27.1.31}$$

$$= \int_a^\infty e^{-st}f(t-a)dt = e^{-sa}\int_0^\infty e^{-st'}f(t')dt' = e^{-sa}F(s) \tag{27.1.32}$$

27.2 Solving ODEs with Laplace transforms

Consider the general linear inhomogeneous second order ODE with constant coefficients:

$$a\ddot{x} + b\dot{x} + cx = f(t), \; x(0) = x_0, \dot{x}(0) = 0 \tag{27.2.1}$$

where $f(t)$ has Laplace transform $F(s)$. We can take the Laplace transform of this equation:

$$a(s^2\mathcal{L}\{x(t)\} - sx_0 - 0) + b(s\mathcal{L}\{x(t)\} - x_0) + c\mathcal{L}\{x(t)\} = F(s) \tag{27.2.2}$$

$$\implies (as^2 + bs + c)\mathcal{L}\{x(t)\} - asx_0 - bsx_0 = F(s) \tag{27.2.3}$$

$$\implies \mathcal{L}\{x(t)\} = \frac{F(s) + asx_0 + bx_0}{as^2 + bs + c} \tag{27.2.4}$$

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:

$$\mathcal{L}\{x(t)\} = \frac{asx_0 + bx_0}{(s - \lambda_1)(s - \lambda_2)} = \frac{A}{(s - \lambda_1)} + \frac{B}{(s - \lambda_2)} \tag{27.2.5}$$

where

$$A = \frac{a\lambda_1 - b}{\lambda_1 - \lambda_2}x_0, \; B = \frac{b - a\lambda_2}{\lambda_1 - \lambda_2}x_0 \tag{27.2.6}$$

We can invert the Laplace transform and find that:

$$x(t) = Ae^{\lambda_1 t} + Be^{\lambda_2 t} \tag{27.2.7}$$

as expected. The importance of Laplace transforms is now clear: it is a very useful tool in solving differential equations.

27.3 Convolutions

> **Definition** *(Convolution integral)*
> Given two functions $f(t)$ and $g(t)$, their convolution $(f * g)(t)$ is defined as:
> $$(f * g)(t) = \int_{-\infty}^{\infty} f(s)g(t-s)ds \tag{27.3.1}$$

Intuitively, the convolution integral gives us the overlap between f and g as we shift one relative to the other by t along s.

> **Proposition** *(Causal functions)*
> A function $f(t)$ is causal if $f(t) = 0,\ \forall t < 0$. The convolution of two causal functions f, g is given by:
> $$(f * g)(t) = \int_0^t f(s)g(t-s)ds \tag{27.3.2}$$

Proof. We have that

$$\int_{-\infty}^{\infty} f(s)g(t-s)ds = \int_0^{\infty} f(s)g(t-s)ds = \int_0^t f(s)g(t-s)ds \tag{27.3.3}$$

since $g(t-s) = 0$ for $s > t$. ∎

> **Theorem** *(Convolution theorem for Laplace transforms)*
> The Laplace transform of the convolution of two causal functions is the product of their Laplace transforms:
> $$\mathcal{L}\{(f * g)(t)\} = \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\} \tag{27.3.4}$$

Proof. We have that:

$$\mathcal{L}\{(f * g)(t)\} = \int_0^{\infty} e^{-st} \int_0^t f(x)g(t-x)dx\,dt \tag{27.3.5}$$

$$= \int_0^{\infty} \int_x^{\infty} e^{-st}g(t-x)dt\ f(x)dx \tag{27.3.6}$$

$$= \int_0^{\infty} \int_0^{\infty} e^{-s(v+x)}g(v+x-x)dv\ f(x)dx \tag{27.3.7}$$

$$= \int_0^{\infty} g(v)e^{-sv}dv \int_0^{\infty} f(x)e^{-sx}dx \qquad = \mathcal{L}\{g(t)\}\mathcal{L}\{f(t)\} \tag{27.3.8}$$

as desired. ∎

> **Example.** Let's find the inverse Laplace transform of:
> $$F(s) = \frac{a}{s^4 + a^2 s^2} \tag{27.3.9}$$

We have that:

$$\mathcal{L}^{-1}(F(s)) = \mathcal{L}^{-1}\left(\frac{1}{s^2}\frac{a}{s^2 + a^2}\right) \qquad (27.3.10)$$

and since $\mathcal{L}^{-1}(1/s^2) = t$ and $\mathcal{L}^{-1}(a/(s^2 + a^2)) = \sin(at)$ one finds:

$$\mathcal{L}^{-1}(F(s)) = \mathcal{L}^{-1}(\mathcal{L}\{t * \sin(at)\}) = \int_0^t (t - s)\sin(as)ds = \frac{at - \sin at}{a^2} \qquad (27.3.11)$$

◄

ODE12 Phase plane analysis

28.1 Non-linear systems

> **Definition** *(Autonomous equations and Fixed points)*
> An **autonomous ODE** is a differential equation of the form
>
> $$\dot{x} = f(x), \ x(0) = x_0 \tag{28.1.1}$$
>
> The particular solution, $x(t)$ for $t > 0$ starting with an initial value $x(0) = x_0$ is known as the **orbit of** x_0. Solutions to $f(x) = 0$ are known as **fixed points** or **equilibrium points**.

In general, given an autonomous equation integrating $\frac{1}{f(x)}$ can be a very difficult (if not impossible) task. Nevertheless we can still qualitatively investigate the properties of its solutions by looking at $f(x)$ only.

For example, consider the following logistic differential equation

$$\dot{x} = Ax - Bx^2, \ A, B \neq 0 \tag{28.1.2}$$

We can solve this problem by integrating directly and find that

$$\frac{1}{A} \log \frac{x}{A - Bx} = t + c \implies x(t) = \frac{Ace^{At}}{1 + Bce^{At}} = \frac{Ac}{e^{-At} + Bc} \tag{28.1.3}$$

Therefore, if $x(0) = x_0$ then

$$x_0 = \frac{Ac}{1 + Bc} \implies c = \frac{x_0}{A - Bx_0} \tag{28.1.4}$$

so the particular solution is given by

$$x(t) = \frac{Ax_0}{Ae^{-At} + Bx_0(1 - e^{-At})} \tag{28.1.5}$$

Note that if $x_0 = 0$ then $x(t) = 0$ so the orbit of the origin will just stay at the origin, $x_0 = 0$ is a fixed point. Similarly if $x_0 = \frac{A}{B}$ then $x(t) = \frac{A}{B}$ so we get another fixed point provided A, B have the same sign. If we look at solutions with x_0 close to the origin or $\frac{A}{B}$ (if A, B have the same sign), we see that if $A, B > 0$ then the orbits will flow away from the origin, but towards $x = \frac{A}{B}$. If $A, B < 0$ then the orbits will flow towards the origin but away from $x = \frac{A}{B}$. Finally, if $A < 0, B > 0$ then orbits will flow towards the origin while if $A > 0, B < 0$ they flow away from it.

These results could have also been deduced directly from $f(x) = Ax - Bx^2$. Indeed we have that $f(x) = 0$ is solved when $x = 0$ or $x = \frac{A}{B}$, these are the fixed points we found earlier. If $A, B > 0$ then we get a concave-down parabola which has positive slope near $x = 0$ and negative slope near $x = \frac{A}{B}$. This means that the origin is an unstable point because the orbits of nearby points are flowing away from it, while the other fixed point is stable because the orbits of nearby points are flowing towards it. Similarly if $A > 0, B < 0$ (or $A < 0, B > 0$) then we get an unstable (stable) fixed point at $x = 0$ and no other points since $x > 0$, and finally if $A, B < 0$ then we get an stable fixed point at $x = 0$ and an unstable fixed point at $x = \frac{A}{B}$.

More generally, if x_e is an equilibrium point of $\dot{x} = f(x)$ with $f'(x_e) \neq 0$ then we can linearise the autonomous system near x_e by defining $x(t) = x_e + p(t)$, where p is a small perturbation, and expand $f(x_e + p) \approx f(x_e) + pf'(x_e) = pf'(x_e)$ to get

$$\dot{x} = \dot{p} = pf'(x_e) \implies p(t) = Ae^{tf'(x_e)} \tag{28.1.6}$$

So we see that if $f'(x_e) > 0$ then the perturbation increases exponentially and we get an unstable point at x_e. If instead $f'(x_e) < 0$ then the perturbation decays and we get a stable point at x_e. Therefore the behaviour of $f'(x)$ near x_e gives us information on the error-correction of orbits perturbed from the origin.

> **Theorem** *(Stability of fixed points)*
> Let $x = x_e$ be a fixed point of $\dot{x} = f(x)$. Then x_e is a
> (i) **stable fixed point** provided $f'(x_e) < 0$
> (ii) **unstable fixed point** provided $f'(x_e) > 0$

The principles we have described can be readily applied to higher dimensional systems, most notably 2D systems.

28.2 Equilibrium points in Phase space

We consider a system of two interacting populations, a predator population and a prey population. Let $X(t)$ and $Y(t)$ be the prey and predator population respectively at time t.

For a predator-free environment we should expect an exponential growth in the prey population, governed by $\dot{X} = kX, \ k > 0$, due to unlimited reproduction. However, due to the presence of predators there will be an attenuation term $-AXY$ which should be proportional to the number of encounters between the two populations. Therefore we find that

$$\dot{X} = kX - AXY, \ k, A > 0 \tag{28.2.1}$$

On the other hand, in a prey-free environment we should expect exponential decay of the predator population governed by $\dot{Y} = -hY, \ h > 0$. However, the presence of preys will introduce a term BXY which should be proportional to the number of encounters between the two populations. Therefore we find that

$$\dot{Y} = -hY + BXY, \ h, B > 0 \tag{28.2.2}$$

Since we will be performing Taylor approximations it will prove useful to rescale these equations [1]

---

[1] without this rescaling we can't argue reasonably that X is small, since X could be measured in units, hundreds, thousands... who knows! To say that something is small that something should be dimensionless, and thus usually a ratio

by defining $x = \frac{BX}{h}$ and $y = \frac{AY}{k}$:

$$\dot{x} = kx(1 - y), \; \dot{y} = -hy(1 - x-) \tag{28.2.3}$$

These are known as the Lotka-Volterra equations.

> **Definition** *(Lotka-Volterra equations)*
> The **Lotka-Volterra equations** model two interacting populations X and Y via a system of non-linear differential equations
> $$\dot{X} = kX - AXY, \; \dot{Y} = -hY + BXY \tag{28.2.4}$$
> where k, h, A, B are all positive. They can be rescaled to yield
> $$\dot{x} = kx(1 - y), \; \dot{y} = -hy(1 - x) \tag{28.2.5}$$

The Lotka-Volterra equations are an example of a two-dimensional autonomous system

$$\dot{x} = u(x, y), \; \dot{y} = v(x, y) \tag{28.2.6}$$

> **Definition** *(Phase space terminology)*
> Let
> $$\dot{x} = u(x, y), \; \dot{y} = v(x, y) \tag{28.2.7}$$
> be a two-dimensional autonomous system. We define the point $(x(t), y(t))$ to be a **phase point** in a two-dimensional **phase space**. The position vector $\mathbf{x} = (x(t), \; y(t))^T$ in phase space allows us to write
> $$\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x}), \; \mathbf{u}(\mathbf{x}) = (u(x, y), \; v(x, y))^T \tag{28.2.8}$$
> where $\mathbf{u}(\mathbf{x})$ is a vector field known as the **phase velocity**. A solution $\mathbf{x}(t)$ for $t > 0$ satisfying the initial condition $\mathbf{x}(0) = \mathbf{x}_0$ is called the **orbit of $\mathbf{x}_0$** and the trajectory it traces as it evolves is its **phase path**. The phase velocity determines the flow of the orbits, just like in 1d, since $\mathbf{u}$ tangent to a phase path $\mathbf{x}(t)$ at every $t > 0$, and at a given point it points in the direction in which the orbit passing through that point evolves. A **phase portrait** shows the collection of all phase paths.

According to this definition, the Lotka-Volterra equations become

$$\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x}), \; \mathbf{u}(\mathbf{x}) = \begin{pmatrix} kx(1 - y) \\ -hy(1 - x) \end{pmatrix} \tag{28.2.9}$$

Note that the velocity yield vanishes at $(0, 0)$ and $(1, 1)$, so at these points the phase paths do not evolve: if we set $\mathbf{x}(0) = \mathbf{x}_e$ then $\mathbf{x}(t) = \mathbf{x}_e$ at all $t > 0$, we get an equilibrium point. This makes sense, when both populations vanish we are never going to get any change in the system so $\mathbf{x} = 0$. Instead $\mathbf{x} = (1, 1)$ occurs when the birth and death rates balance out for both populations, so we will again get a balanced, unchanging population.

> **Definition** *(Equilibrium point)*
> The system $\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x})$ has an **equilibrium point** at $\mathbf{x}_e$ if the phase-velocity vanishes at this point $\mathbf{u}(\mathbf{x}_e) = \mathbf{0}$.

We can retrieve a lot of information about a non-linear system's behaviour by looking at its properties near an equilibrium point. Indeed suppose $\mathbf{x}_e$ is an equilibrium point of $\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x})$, and suppose we perturb the system away from this equilibrium by defining

$$\mathbf{x}(t) = \mathbf{x}_e + \mathbf{p}(t) \tag{28.2.10}$$

We are interested in the evolution of $\mathbf{p}(t)$ under this system's dynamics. Assuming $\mathbf{p}$ is small then we can expand $\mathbf{u}$ to first order

$$\mathbf{u}(\mathbf{x}_e + \mathbf{p}) = \mathbf{u}(\mathbf{x}_e) + \mathsf{J}(\mathbf{x}_e)\mathbf{p} \tag{28.2.11}$$

where we used the Jacobian

$$\mathsf{J}(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial u}{\partial x}(\mathbf{x}) & \dfrac{\partial u}{\partial y}(\mathbf{x}) \\ \dfrac{\partial v}{\partial x}(\mathbf{x}) & \dfrac{\partial v}{\partial y}(\mathbf{x}) \end{pmatrix} \tag{28.2.12}$$

Consequently we find that:

> **Proposition** *(Linearisation of non-linear systems)*
> The linearisation of $\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x})$ about an equilibrium point $\mathbf{x}_e$ yields the following evolution for a small perturbation $\mathbf{p}$:
> $$\dot{\mathbf{p}} = \mathsf{J}(\mathbf{x}_e)\mathbf{p} \tag{28.2.13}$$

Note that the solution to $\dot{\mathbf{p}} = \mathsf{J}(\mathbf{x}_e)$ is fully determined by diagonalizing the Jacobian and finding its eigenvalues λ_1, λ_2 and correspondingly its eigenvectors $\mathbf{v}_1, \mathbf{v}_2$:

$$\mathbf{p} = c_1 e^{\lambda_1 t}\mathbf{v}_1 e + c_2 e^{\lambda_2 t}\mathbf{v}_2 \tag{28.2.14}$$

In the case of the Lotka-Volterra equations, we see that

$$\mathsf{J}(\mathbf{x}) = \begin{pmatrix} k(1-y) & -kx \\ hy & -h(1-x) \end{pmatrix} \implies \mathsf{J}(0,0) = \begin{pmatrix} k & 0 \\ 0 & -h \end{pmatrix}, \; \mathsf{J}(1,1) = \begin{pmatrix} 0 & -k \\ h & 0 \end{pmatrix} \tag{28.2.15}$$

so we find that

$$\dot{\mathbf{p}} = \begin{pmatrix} k & 0 \\ 0 & -h \end{pmatrix}\mathbf{p} \implies \mathbf{p} = c_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{kt} + c_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{-ht} \qquad \text{near } (0,0) \quad (28.2.16)$$

and

$$\dot{\mathbf{p}} = \begin{pmatrix} 0 & -k \\ h & 0 \end{pmatrix}\mathbf{p}, \implies \mathbf{p} = c_1 \begin{pmatrix} i\sqrt{\frac{a}{b}} \\ 1 \end{pmatrix} e^{i\sqrt{kh}t} + c_2 \begin{pmatrix} -i\sqrt{\frac{a}{b}} \\ 1 \end{pmatrix} e^{-i\sqrt{kh}t} \qquad \text{near } (1,1) \quad (28.2.17)$$

28.3   Classifying equilibrium points

Suppose we have found the equilibrium points of a non-linear system and have found the eigenvalues and eigenvectors of the Jacobian at each of these points. Let's say for example that the linearised solution near $\mathbf{x}_e$ is

$$\mathbf{p} = c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2 \tag{28.3.1}$$

where λ_1, λ_2 are eigenvalues of $\mathsf{J}(\mathbf{x}_e)$ with corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2$. Note that since $\mathbf{u}(\mathbf{x})$ is well-behaved (more specifically we want it and its inverse $\mathbf{u}^{-1}(\mathbf{x})$ to be differentiable), the Jacobian will be invertible and thus we will get two non-zero eigenvalues. Alternatively, this ensures that the only fixed point of the linearised system occurs at the origin $\mathbf{p} = \mathbf{0}$.

λ both real and $\lambda_1 \gtrless \lambda_2 \gtrless 0$

If both eigenvalues are real and have the same sign, then we will get orbits that either diverge away or converge towards $\mathbf{x}_e$. The former occurs when $\lambda_1 > \lambda_2 > 0$, in which case we get an **unstable node**, while the latter occurs when $\lambda_1 < \lambda_2 < 0$, in which case we get a **stable node**.
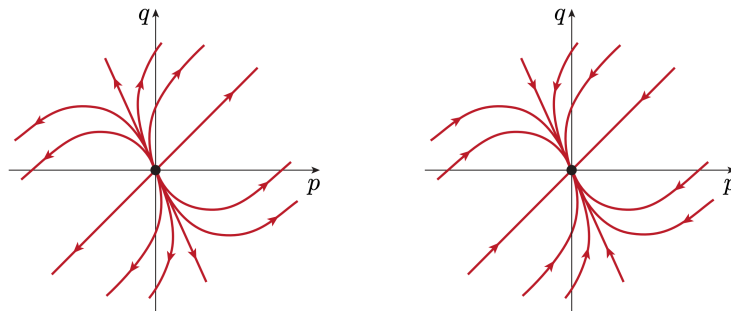
λ both real and $\lambda_1 > 0 > \lambda_2$

If both eigenvalues are real but one is positive and the other negative, then the orbits will flow towards the origin along one eigenvector but away from it along the other eigenvector. We therefore get a **saddle point**, because if we traverse it along one eigenvector it is a stable point, while if we traverse it along the other it is unstable.
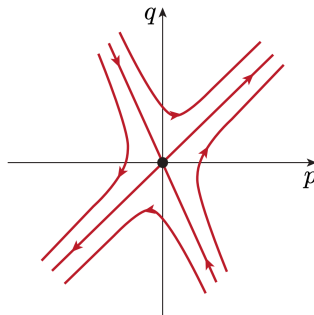
λ both complex and $\mathbf{Re}(\lambda) \neq 0$

Suppose one eigenvalue, $\lambda_1 = \nu + i\omega$, with eigenvector $\mathbf{v}_1 = \mathbf{v}$ is complex. Then since $\mathsf{J}(\mathbf{x}_e)$ is a real matrix, the other eigenvalue $\lambda_2 = \overline{\lambda_1}$ must be the complex conjugate of λ_1, and similarly $\mathbf{v}_2 = \overline{\mathbf{v}}$.

Consequently we find that

$$\mathbf{p}(0) = c_1\mathbf{v} + c_2\overline{\mathbf{v}} \tag{28.3.2}$$

but since the LHS must be real (remember it is a perturbation in phase space), it must hold that $c_1\mathbf{v} + c_2\overline{\mathbf{v}} = \overline{c_2}\mathbf{v} + \overline{c_1}\overline{\mathbf{v}}$ and therefore $c_1 = \overline{c_2} = c$. We finally find that

$$\mathbf{p}(t) = e^{\nu t}(ce^{i\omega t}\mathbf{v} + \overline{c}e^{-i\omega t}\overline{\mathbf{v}}) = 2e^{\nu t}\mathrm{Re}(ce^{i\omega t}\mathbf{v}) \tag{28.3.3}$$

Now it can be seen that if $\nu > 0$ then we will get phase paths that spiral away from the origin, while if $\nu < 0$ the solutions will spiral into the origin. In the first case we get an **unstable spiral** while in the latter we get a **stable spiral**. The sign of ω then determines the direction of the spiral (anti-clockwise for $\omega > 0$ and clockwise for $\omega < 0$).
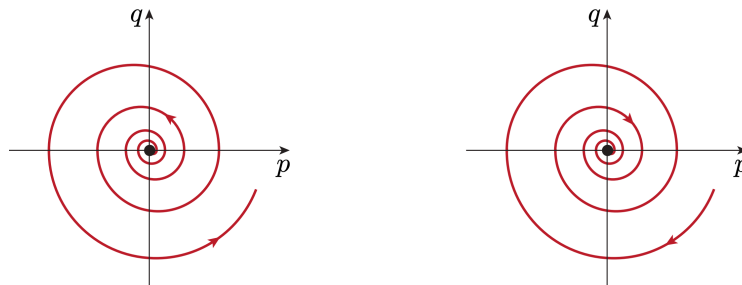
λ **both complex and** $\mathbf{Re}(\lambda) = 0$

If instead $\nu = 0$ then $\mathbf{p}(t) = 2\mathrm{Re}(ce^{i\omega t}\mathbf{v})$ so we get a circle, also known as a **center**.
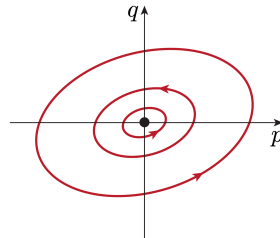
Theorem *(Classifying equilibrium points)*
Consider the linear system $\dot{\mathbf{p}} = \mathsf{J}(\mathbf{x}_{eq})\mathbf{p}$, where J has eigenvalues λ_1, λ_2. The equilibrium point $\mathbf{x}_{eq}$ can be categorised as follows

$\lambda_1, \lambda_2 \in \mathbb{R}$	$\lambda_1 > \lambda_2 > 0$ $\lambda_1 < \lambda_2 < 0$ $\lambda_1 > 0 > \lambda_2$	unstable node stable node saddle point
$\lambda_{1,2} = \nu \pm i\omega$ $\mu, \nu \in \mathbb{R}, \omega \neq 0$	$\nu > 0$ $\nu < 0$ $\nu = 0$	unstable spiral stable spiral center

Example. Let's apply these results to study the phase portrait of the Lotka-Volterra system

$$\dot{x} = kx(1 - y), \ \dot{y} = -hy(1 - x-) \tag{28.3.4}$$

We have already calculated that this system has fixed points at $\mathbf{x} = (0,0)$ and $\mathbf{x} = (1,1)$, with corresponding linearised solutions given by

$$\mathbf{p} = c_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{kt} + c_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{-ht} \qquad \text{near } (0,0) \qquad (28.3.5)$$

$$\mathbf{p} = c_1 \begin{pmatrix} i\sqrt{\frac{a}{b}} \\ 1 \end{pmatrix} e^{i\sqrt{kh}t} + c_2 \begin{pmatrix} -i\sqrt{\frac{a}{b}} \\ 1 \end{pmatrix} e^{-i\sqrt{kh}t} \qquad \text{near } (1,1) \qquad (28.3.6)$$

Since $k, h > 0$ we have that the origin is a saddle, orbits close to it will diverge away along the x-axis and converge in along the y-axis. On the other hand, $(1,1)$ is a center since the Jacobian's eigenvalues $\pm i\sqrt{kh}$ are purely imaginary.
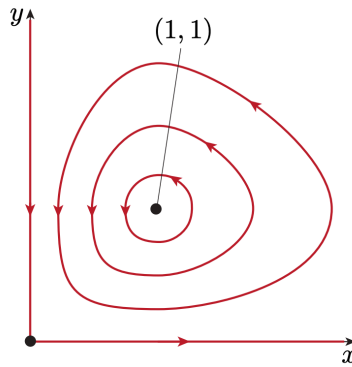
28.4   The Poincare-Bendixson theorem

We saw that in the Lotka-Volterra phase portrait, many (actually all orbits except those through the axes) are closed, and therefore represent oscillating solutions. Centres can often be associated with conserved quantities, constants of motion along the closed orbits, which we shall now investigate.

> **Definition** (*Constant of motion*)
> The quantity $K(x,y)$ is a **constant of motion** *iff* $\dot{K} = 0$ or in other words
>
> $$\dot{x}\frac{\partial K}{\partial x} + \dot{y}\frac{\partial K}{\partial y} \iff \mathbf{u} \cdot \nabla K = 0 \qquad (28.4.1)$$

Since ∇K is normal to the contours of K, this means that the velocity field $\mathbf{u}$ should be tangent to the contours of K. But the velocity field is also tangent to the phase paths! So the contours of K will be given by the phase paths of the system. The phase paths are found by solving

$$\frac{dy}{dx} = \frac{u_y}{u_x} \qquad (28.4.2)$$

which for the Lotka-Volterra system becomes

$$\frac{dy}{dx} = -\frac{hy(1-x)}{kx(1-y)} \implies h(x - \ln x) = k(\ln y - y) + c \tag{28.4.3}$$

It follows that $K(x, y) = h(x - \ln x) - k(\ln y - y)$ must be a constant of motion. Since the phase portrait of the system is given by the contour map of K, it follows that any extrema of a conserved quantity must represent equilibrium points of the phase flow. Not only that, since contours near relative extrema are circles, the corresponding phase flow will be composed of circular orbits. Therefore these equilibrium points are centres. Similarly saddle points of K will give saddle point of the phase portrait.

In the case of the Lotka-Volterra equation we see that

$$\frac{\partial K}{\partial x} = h\left(1 - \frac{1}{x}\right), \ \frac{\partial K}{\partial y} = -k\left(\frac{1}{y} - 1\right) \tag{28.4.4}$$

implying that the equilibrium point is given by $x = y = 1$, as expected. This point is also a global minimum, since $K(x, y)$ is continuous, with one stationary point at $(1, 1)$, and approaches infinity at the origin and at infinity. It follows that all contours of K except for at the stationary point and centre will be closed: the solutions to the Lotka-Volterra system are always periodic in time!

Example. We consider the following nonlinear system

$$\dot{x} = (1 - y^2)x, \ \dot{y} = -(1 - x^2)y \tag{28.4.5}$$

The velocity field is $\mathbf{u} = ((1-y^2)x, \ -(1-x^2)y)^T$, so to find the phase paths we need to solve

$$\frac{dy}{dx} = -\frac{y(1-x^2)}{x(1-y^2)} \tag{28.4.6}$$

which is separable

$$\int \left(\frac{1}{y} - y\right)dy = \int \left(x - \frac{1}{x}\right)dx \implies \ln y - \frac{1}{2}y^2 = \frac{1}{2}x^2 - \ln x + c \tag{28.4.7}$$

Consequently the constant of motion will be $K(x, y) = \ln(xy) - \frac{1}{2}(x^2 + y^2)$. We find its extrema by solving

$$\frac{\partial K}{\partial x} = \frac{1}{x} - x = 0, \ \frac{\partial K}{\partial y} = \frac{1}{y} - y = 0 \tag{28.4.8}$$

so we see that $(1, 1)$ is a stationary point. This is also a fixed point since the $\mathbf{u}(0, 0) = 0$. To test whether we have found a center, let us Taylor expand about $(1, 1)$, with $\log(1 + \epsilon) \approx \epsilon - \frac{1}{2}\epsilon^2$:

$$K(x, y) \approx x + y - x^2 - y^2 \tag{28.4.9}$$

which is indeed maximised at the origin. Therefore $(1, 1)$ is a center of the nonlinear system.

◀

A **limit cycle** is a closed orbit whose neighbouring orbits not open but converge towards it (**stable limit cycle**), diverge away from it (**unstable limit cycle**), or converge in one region

and diverge in the other (**half-stable limit cycle**).

(a)

(b)

(c)

(d)
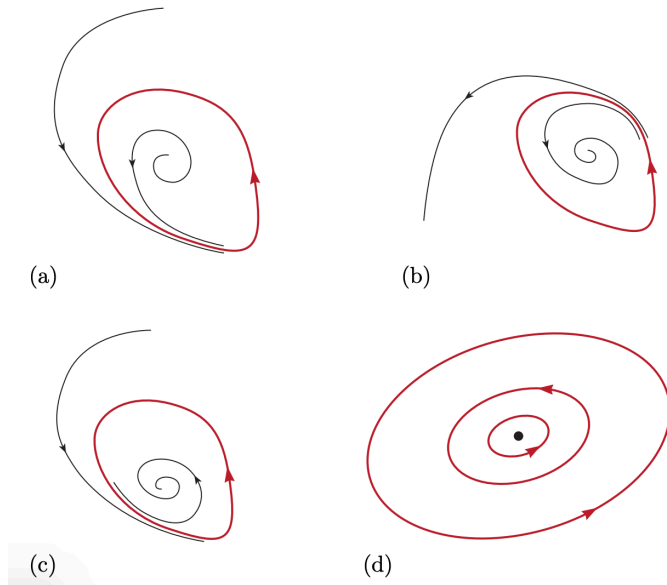
Figure 28.1. (a) stable limit cycle (b) unstable limit cycle (c) half-stable limit cycle (d) center

Example. Consider the following nonlinear system

$$\dot{x} = x(1 - x^2 - y^2)(x^2 + y^2 - 4) - y \tag{28.4.10}$$
$$\dot{y} = y(1 - x^2 - y^2)(x^2 + y^2 - 4) + x \tag{28.4.11}$$

It will be convenient to transform to polar coordinates, where

$$x = r\cos\theta, \ y = r\sin\theta \tag{28.4.12}$$

and therefore

$$\dot{x} = \dot{r}\cos\theta - r\dot{\theta}\sin\theta \tag{28.4.13}$$
$$\dot{y} = \dot{r}\sin\theta + r\dot{\theta}\cos\theta \tag{28.4.14}$$

The nonlinear system then becomes

$$\dot{r}\cos\theta - r\dot{\theta}\sin\theta = r\cos\theta(1 - r^2)(r^2 - 4) - r\sin\theta \tag{28.4.15}$$
$$\dot{r}\sin\theta + r\dot{\theta}\cos\theta = r\sin\theta(1 - r^2)(r^2 - 4) + r\cos\theta \tag{28.4.16}$$

We multiply the first by $\cos\theta$, the second by $\sin\theta$ and add them, this yields

$$\dot{r} = r(1 - r^2)(r^2 - 4) \tag{28.4.17}$$

Similarly we can multiply the first by $\sin\theta$, the second by $\cos\theta$ and subtract them to find

$$\dot{\theta} = 1 \tag{28.4.18}$$

From the first we see that $\theta = t + \theta_0$, the angular flow is always positive (counter-clockwise) and gives no fixed points.
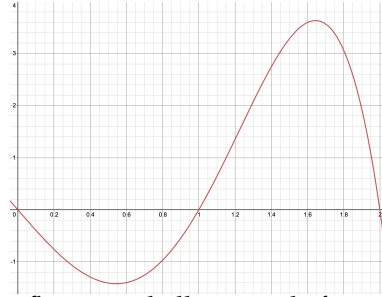
For the radial coordinate, the flow is radially inwards for $0 < r < 1$ and $r > 2$, radially outwards for $1 < r < 2$, with fixed points at $r = 0, 1, 2$. Looking at the slopes of the graph we see that $r = 0$ is a stable node, $r = 2$ is a stable limit cycle (negative slope), and $r = 1$ is an unstable limit cycle (positive slope). ◀

An important result in two-dimensional nonlinear systems theory is known as the Poincaré-Bendixson theorem. It confines the possible behaviour of bounded solutions to three types, ruling out any type of chaotic behaviour. Intuitively, this is a result of the uniqueness of solutions to non-linear systems of ODEs with well-behaved $\mathbf{u}(\mathbf{x})$. Since no two phase paths can ever intersect, any orbit that is contained within a closed orbit can never touch points outside this closed orbit.

Theorem *(Poincaré-Bendixson theorem)*
Consider the two-dimensional nonlinear system $\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x})$ with initial condition $\mathbf{x}(0) = \mathbf{x}_0$, and where $\mathbf{u}$ is continuously differentiable. Suppose a solution to this system is constrained to a closed, bounded region of $\mathbb{R}^2$. Then
   (i) the orbit is a closed path
  (ii) the orbit approaches a closed path as $t \to \infty$
 (iii) the orbit approaches a fixed point as $t \to \infty$

PDE1 First order PDEs

29.1 Introduction

> **Definition** *(Linear first order PDE)* A linear first order partial differential equation is of the form
> $$\sum_{i=1}^{n} a_i(x^1, ..., x^n) \frac{\partial u}{\partial x^i} + c(x, y)u = d(x, y) \tag{29.1.1}$$
>
> where a_i, b, d are continuous functions. In two dimensions this reads
>
> $$a(x, y) \frac{\partial u}{\partial x} + b(x, y) \frac{\partial u}{\partial y} + c(x, y)u = d(x, y) \tag{29.1.2}$$
>
> where a, b, c, d are continuous functions.

As with first order ODEs, there are some special cases of linear first order PDEs that can be considered.

29.2 From ODEs to PDEs

> **Definition** *(Separable linear first order PDE)* A first order PDE is said to be separable if it is of the form
> $$\frac{\partial u}{\partial x^i} = f(x^1, ..., x^{i-1}, x^{i+1}, ...)g(x^i) \tag{29.2.1}$$

Separable PDEs can be solved by integrating directly, this time remembering that all but one variable are not integrating over and should thus be treated as constants. In other words

$$\frac{\partial u}{\partial x^i} = f(x^1, ..., x^{i-1}, x^{i+1}, ...)g(x^i) \implies \int \frac{\partial u}{\partial x^i}\, dx^i = \int f(x^1, ..., x^{i-1}, x^{i+1}, ...)g(x^i)dx^i \tag{29.2.2}$$

Consider for example the following

$$\frac{\partial u}{\partial y} = xy \implies u(x, y) - f(x) = \frac{1}{2}xy^2 \implies u(x, y) = \frac{1}{2}xy^2 + f(x) \tag{29.2.3}$$

Note that the general solution to this first order PDE in y contains an arbitrary function of x, just like the solution to a first order ODE contains an arbitrary constant.

Another case which can be solved in a manner analogous to ODEs are those of the type

$$\frac{\partial u}{\partial x} + a(x,y)u = b(x,y) \tag{29.2.4}$$

These can be solved using the method of integrating factor. Suppose we have found the integrating factor Λ using the usual method, then the above can be written as

$$\frac{\partial}{\partial x}(\Lambda(x,y)u(x,y)) = \Lambda(x,y)b(x,y) \tag{29.2.5}$$

which can be integrated to give

$$\Lambda(x,y)u(x,y) = B(x,y) + f(y) \implies y(x,y) = \frac{B(x,y)}{\Lambda(x,y)} + \frac{f(y)}{\Lambda(x,y)} \tag{29.2.6}$$

Example. For example, consider the following first order PDE

$$\frac{\partial u}{\partial y} + \frac{3}{y}u = y^2, \ y \neq 0 \tag{29.2.7}$$

The integrating factor is

$$\Lambda(x,y) = \exp\left(\int \frac{3}{y}dy\right) = e^{3\ln y} = y^3 \tag{29.2.8}$$

so we find

$$\frac{\partial}{\partial y}(y^3 u(x,y)) = y^5 \implies u(x,y) = \frac{1}{6}y^3 + \frac{f(x)}{y^3} \tag{29.2.9}$$

As another example, consider

$$\frac{\partial u}{\partial y} - xyu = y, \ x \neq 0 \tag{29.2.10}$$

The integrating factor is

$$\Lambda(x,y) = e^{-xy^2/2} \implies \frac{\partial}{\partial y}(e^{-xy^2/2}u(x,y)) = ye^{-xy^2/2} \tag{29.2.11}$$

which can be integrated to give

$$e^{-xy^2/2}u(x,y) = -\frac{e^{-xy^2/2}}{x} + f(x) \implies u(x,y) = -\frac{1}{x} + f(x)e^{xy^2/2} \tag{29.2.12}$$

◀

We can solve certain second order PDEs by transforming it into a first order PDE, as the next example shows.

Example. Consider $\frac{\partial^2 y}{\partial x \partial t} + \frac{1}{t}\frac{\partial u}{\partial x} = x, \ t \neq 0$. We can transform this into a first order PDE by letting $f(x,y) = \frac{\partial u}{\partial x}$:

$$\frac{\partial f}{\partial t} + \frac{1}{t}f = x \tag{29.2.13}$$

This can be solved by the method of integrating factors, it is easy to see that $\Lambda(t) = t$ so that

$$\frac{\partial}{\partial t}(tf) = tx \implies f(x,t) = \frac{1}{2}tx + \frac{1}{t}h(x) \tag{29.2.14}$$

To find $u(x,t)$ we substitute back $f = \frac{\partial u}{\partial x}$ which yields

$$\frac{\partial u}{\partial x} = \frac{1}{2}tx + \frac{1}{t}h(x) \tag{29.2.15}$$

which can be integrated directly to give the general solution:

$$u(x,t) = \frac{1}{4}tx^2 + \frac{1}{t}h(x) + g(t) \tag{29.2.16}$$

◀

29.3 Change of variable and the chain rule

Substitutions of variables are often useful in simplifying PDEs, this requires the use of the chain rule.

For example, consider the first order PDE

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} + u = 2 \tag{29.3.1}$$

We may define $x = \eta$ and $y = \phi - \eta$. Then we see that

$$\frac{\partial u}{\partial \eta} = \frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} \tag{29.3.2}$$

so that the PDE transforms to

$$\frac{\partial u}{\partial \eta} + u = 2 \tag{29.3.3}$$

We can solve this equation for u using the integrating factor method

$$\Lambda(x,y) = e^{\eta} \implies u(\eta) = 2 + e^{-\eta}f(\phi) \tag{29.3.4}$$

which in the old coordinates reads

$$f(x,y) = 2 + e^{-x}f(x+y) \tag{29.3.5}$$

29.4 The method of characteristics

Consider the PDE

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 0 \tag{29.4.1}$$

We know that letting $\chi = x$ and $\eta = x - y$ then it will transform to

$$\frac{\partial u}{\partial \chi} = 0 \implies u(x,y) = f(x-y) \tag{29.4.2}$$

Note that by performing this special change of coordinates we got a PDE depending on only one variable. The corresponding solutions are therefore found by integrating over level curves of the missing variable, in the previous example by keeping $x - y = $ cnst. Such curves are known as characteristic curves of the PDE and are very useful tools.

Consider the general first order PDE

$$a(x, y)\frac{\partial u}{\partial x} + b(x, y)\frac{\partial u}{\partial y} + c(x, y)u = d(x, y) \tag{29.4.3}$$

where $a(x, y) \neq 0$. Thus we may equivalently write

$$\frac{\partial u}{\partial x} + g(x, y)\frac{\partial u}{\partial y} + h(x, y)u = k(x, y) \tag{29.4.4}$$

We define two new variables $\chi = \chi(x, y)$ and $\eta = \eta(x, y)$ so that

$$\frac{\partial u}{\partial \chi} + h(\chi, \eta)u = k(\chi, \eta) \tag{29.4.5}$$

This can be done if we let

$$\frac{\partial u}{\partial \chi} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial \chi} + \frac{\partial u}{\partial y}\frac{\partial x}{\partial \chi} = \frac{\partial u}{\partial x} + g(x, y)\frac{\partial u}{\partial y} \tag{29.4.6}$$

so by solving the following system of PDEs

$$\begin{cases} \frac{\partial x}{\partial \chi} = 1 \\ \frac{\partial y}{\partial \chi} = g(x, y) \end{cases} \tag{29.4.7}$$

Let $\chi = x$, then the first equation is satisfied, and along the characteristic curves $\eta = $ csnt we find that

$$\frac{\partial y}{\partial \chi} = \frac{dy}{dx} = g(x, y) \tag{29.4.8}$$

Thus solving the above ODE will yield the general solutions $\alpha(x, y) = c$ which are the required characteristic curves if we define $\eta = \alpha(x, y)$.

Example. Consider for example

$$x^2\frac{\partial f}{\partial x} - xy\frac{\partial f}{\partial y} + y^2 = 0, \ x \neq 0 \tag{29.4.9}$$

which we can write equivalently as

$$\frac{\partial f}{\partial x} - \frac{y}{x}\frac{\partial f}{\partial y} + \frac{y^2}{x^2} = 0 \tag{29.4.10}$$

To find the characteristic curves we need to solve the ODE

$$\frac{dy}{dx} = -\frac{y}{x} \implies xy = c \tag{29.4.11}$$

so we see that the characteristics are hyperbolae. Hence let $\chi = x$ and $\eta = xy$. Then we have

that

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial \chi}\frac{\partial \chi}{\partial y} + \frac{\partial f}{\partial \eta}\frac{\partial \eta}{\partial y} = x\frac{\partial f}{\partial \eta} \tag{29.4.12}$$

and similarly

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial \chi}\frac{\partial \chi}{\partial x} + \frac{\partial f}{\partial \eta}\frac{\partial \eta}{\partial x} = \frac{\partial f}{\partial \chi} + y\frac{\partial f}{\partial \eta} \tag{29.4.13}$$

The PDE now transforms to

$$\frac{\partial f}{\partial \chi} + \frac{\eta^2}{\chi^4} = 0 \tag{29.4.14}$$

which we can integrate directly to find

$$f(\chi, \eta) = \frac{\eta^2}{3\chi^3} + f(\eta) \implies f(x,y) = \frac{y^2}{3x} + f(xy) \tag{29.4.15}$$

◄

We can apply this reasoning to some second order PDEs by transforming them into first order PDEs. This is best illustrated in the following example.

Example. Consider

$$\frac{\partial^2 u}{\partial x \partial y} + 3x^2\frac{\partial^2 u}{\partial y^2} - 2\frac{\partial u}{\partial y} = 0 \tag{29.4.16}$$

Let $f = \frac{\partial u}{\partial y}$ this becomes a first order PDE

$$\frac{\partial f}{\partial x} + 3x^2\frac{\partial f}{\partial y} - 2f = 0 \tag{29.4.17}$$

To find the characteristic curves we solve $y' = 3x^2 \implies y - x^3 = c$. Thus we define the new coordinates $\chi = x$ and $\eta = y - x^3$. Then by the chain rule

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial \chi} - 3x^2\frac{\partial f}{\partial \eta}, \ \frac{\partial f}{\partial y} = 3x^2\frac{\partial f}{\partial \eta} \tag{29.4.18}$$

We therefore get a PDE in χ only

$$\frac{\partial f}{\partial \chi} - 2f = 0 \tag{29.4.19}$$

which can be integrated directly to give

$$f(\chi, \eta) = g(\eta)e^{2\chi} \implies f(x,y) = g(y - x^3)e^{2x} \tag{29.4.20}$$

This means that

$$\frac{\partial u}{\partial y} = g(y - x^3)e^{2x} \implies u(x,y) = h(y{-}^3)e^{2x} + k(x) \tag{29.4.21}$$

is the general solution. ◄

PDE2 Second order PDEs: an overview

PDE3 Elliptic PDEs: Electrostatics

31.1 Existence and uniqueness theorem

PDE4 Hyperbolic PDEs: Waves

32.1 The wave equation

We are interested in the solutions of the following equation

$$\boxed{\frac{\partial^2 \phi}{\partial t^2} = c^2 \nabla^2 \phi}$$

(32.1.1)

but before we look at the solutions lets try to derive these from a physical perspective.

Newtonian derivation

Consider an elastic, circular rubber membrane Ω of uniform density ρ that is fixed along some boundary $\partial\Omega$ due to a uniform tension force per unit length T. We will also deal with small oscillations only, so $\phi(x, y, t)$ is small. Let's look at the forces acting on an infinitesimal surface element $dA = dx\, dy$. We see that there are four forces acting on four edges of the element, two have magnitude $T dy$ and the other two have magnitude $T dx$. We can model these forces to act only on the midpoint of the edges, as shown below.

The force along x acting on the frontal edge of the membrane is given by

$$F_x^1 = T \cos\theta_2 dy - T \cos\theta_1 dy \approx 0$$

(32.1.2)

$$F_y^1 = T \sin\theta_2 dy - T \sin\theta_1 dy \approx T(\tan\theta_2 - \tan\theta_1)dy$$

(32.1.3)

where we used the small angle approximation. Now note that

$$\tan\theta_2 = \left.\frac{\partial u}{\partial x}\right|_{x+dx,y_1}, \quad \text{and} \quad \tan\theta_1 = \left.\frac{\partial u}{\partial x}\right|_{x,y_2}$$

(32.1.4)

where y_1 and y_2 are midpoints of the left and right edges respectively. Therefore we see that

$$F_y^1 = T\left(\left.\frac{\partial u}{\partial x}\right|_{x+dx,y_1} - \left.\frac{\partial u}{\partial x}\right|_{x,y_2}\right)dy = T\left.\frac{\partial^2 u}{\partial x^2}\right|_{x,y_1} dxdy$$

(32.1.5)

Similarly for the other end we find that

$$F_y^2 = T\left(\left.\frac{\partial u}{\partial x}\right|_{x_1,y+dy} - \left.\frac{\partial u}{\partial x}\right|_{x_2,y}\right)dx = T\left.\frac{\partial^2 u}{\partial y^2}\right|_{x_1,y} dxdy$$

(32.1.6)

Consequently using Newton's second law we find that

$$\frac{\partial^2 u}{\partial t^2}\rho \, dx \, dy = T\nabla^2 u \, dx \, dy \tag{32.1.7}$$

which gives the 2D wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\nabla^2 u, \; c = \sqrt{\frac{T}{\rho}} \tag{32.1.8}$$

We see that the phase velocity of waves propagating on the membrane is given by $\sqrt{\frac{T}{\rho}}$. This derivation however is tricky to generalize if the tension is non-uniform, a simple way to account for this is by taking the lagrangian approach instead.

Variational/Lagrangian derivation

The kinetic energy of the membrane is given by

$$K = \frac{1}{2}\int_\Omega \mu\left(\frac{\partial \phi}{\partial t}\right)^2 dx \, dy \tag{32.1.9}$$

and since we are only working to first order we can approximate

$$dA = \sqrt{1 + \frac{\partial \phi}{\partial x}^2}\sqrt{1 + \frac{\partial \phi}{\partial y}^2} \approx 1 \tag{32.1.10}$$

To work out the potential energy, note that if the displaced membrane has area dS then the associated potential energy is $T(dS - dx \, dy)$ and therefore

$$V = \int_\Omega T \sqrt{1 + \frac{\partial \phi}{\partial x}^2}\sqrt{1 + \frac{\partial \phi}{\partial y}^2} \, dx \, dy - \int_\Omega T \, dx \, dy \tag{32.1.11}$$

$$\approx \frac{1}{2}\int_\Omega T\left[\left(\frac{\partial \phi}{\partial x}\right)^2 + \left(\frac{\partial \phi}{\partial y}\right)^2\right] dx \, dy \tag{32.1.12}$$

Combining these the Lagrangian reads

$$L = \frac{1}{2}\int_\Omega \left\{\mu\left(\frac{\partial \phi}{\partial t}\right)^2 - T\left[\left(\frac{\partial \phi}{\partial x}\right)^2 + \left(\frac{\partial \phi}{\partial y}\right)^2\right]\right\} dx \, dy \tag{32.1.13}$$

and thus the Lagrangian density is

$$\mathcal{L} = \frac{1}{2}\mu\left(\frac{\partial \phi}{\partial t}\right)^2 - \frac{1}{2}T\left[\left(\frac{\partial \phi}{\partial x}\right)^2 + \left(\frac{\partial \phi}{\partial y}\right)^2\right] \tag{32.1.14}$$

The Euler-Lagrange equations are

$$\frac{\partial}{\partial y}\left(\frac{\partial \mathcal{L}}{\partial(\partial_y \phi)}\right) + \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{L}}{\partial(\partial_x \phi)}\right) + \frac{\partial}{\partial t}\left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}}\right) = \frac{\partial \mathcal{L}}{\partial \phi} \tag{32.1.15}$$

and consequently we get

$$\frac{\partial^2 \phi}{\partial t^2} = c^2 \nabla^2 \phi \tag{32.1.16}$$

as desired.

32.2 d'Alembert's solution

Let $\chi = x - ct$ and $\eta = x + ct$ so that

$$\frac{\partial \phi}{\partial x} = \frac{\partial \phi}{\partial \chi}\frac{\partial \chi}{\partial x} + \frac{\partial \phi}{\partial \eta}\frac{\partial \eta}{\partial x} = \frac{\partial \phi}{\partial \chi} + \frac{\partial \phi}{\partial \eta} \tag{32.2.1}$$

and

$$\frac{\partial \phi}{\partial t} = \frac{\partial \phi}{\partial \chi}\frac{\partial \chi}{\partial t} + \frac{\partial \phi}{\partial \eta}\frac{\partial \eta}{\partial t} = c\frac{\partial \phi}{\partial \chi} - c\frac{\partial \phi}{\partial \eta} \tag{32.2.2}$$

Consequently

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{\partial^2 \phi}{\partial \chi^2} + \frac{\partial^2 \phi}{\partial \eta^2} + 2\frac{\partial^2 \phi}{\partial \chi \partial \eta} \tag{32.2.3}$$

$$\frac{\partial^2 \phi}{\partial t^2} = c^2\frac{\partial \phi}{\partial \chi} + c^2\frac{\partial \phi}{\partial \eta} - 2c^2\frac{\partial^2 \phi}{\partial \chi \partial \eta} \tag{32.2.4}$$

from which we get the wave equation in the new coordinates

$$\frac{\partial^2 \phi}{\partial \chi \partial \eta} = 0 \implies \phi(x,t) = f(x - ct) + g(x + ct) \tag{32.2.5}$$

32.3 Energy conservation and Uniqueness

Having derived an expression for the kinetic and potential energy of a vibrating membrane, let us see under what conditions energy is conserved. The Hamiltonian (in N dimensions) is

$$H(t) = \frac{1}{2}\int_\Omega \left[\mu\left(\frac{\partial \phi}{\partial t}\right)^2 + T(\nabla \phi)^2\right] d^n\mathbf{x} \tag{32.3.1}$$

so we see that

$$\frac{dH}{dt} = \int_\Omega \left[\mu\frac{\partial \phi}{\partial t}\frac{\partial^2 \phi}{\partial t^2} + T(\nabla \phi)\cdot\nabla\left(\frac{\partial \phi}{\partial t}\right)\right] d^n\mathbf{x} \tag{32.3.2}$$

$$= \int_\Omega T\left[\frac{\partial \phi}{\partial t}\nabla^2 \phi + (\nabla \phi)\cdot\nabla\left(\frac{\partial \phi}{\partial t}\right)\right] d^N\mathbf{x} \tag{32.3.3}$$

$$= \int_\Omega \nabla\cdot\left(\frac{\partial \phi}{\partial t}\nabla \phi\right) d^N\mathbf{x} \tag{32.3.4}$$

$$= \oint_{\partial\Omega}\left(\frac{\partial \phi}{\partial t}\nabla \phi\right)\cdot\mathbf{n}\, dS \tag{32.3.5}$$

Thus energy is conserved under motion described by the wave equation if $\frac{\partial \phi}{\partial t} = 0$ or if $\mathbf{n}\cdot\nabla\phi = 0$. For example, an electromagnetic wave in vacuum satisfies $\nabla\cdot\mathbf{B} = 0$ and $\nabla\cdot\mathbf{E} = 0$ so energy will be conserved.

This theorem on the conservation of energy can be used to prove another useful result, namely the uniqueness of solutions to the wave equation subject to certain boundary conditions. Indeed suppose we had found two solutions ϕ_1 and ψ_2 such that

$$\frac{\partial^2 \phi_i}{\partial t^2} = c^2 \nabla \phi_i^2, \qquad \begin{cases} \phi_i(\mathbf{x}, 0) = f(x), \ \mathbf{x} \in \Omega \\ \frac{\partial \phi_i}{\partial t}(\mathbf{x}, 0) = g(x), \mathbf{x} \in \Omega \\ \phi_i(\mathbf{x}, t) = h(x), \ \mathbf{x} \in \partial\Omega \end{cases} \tag{32.3.6}$$

We have two initial conditions (as required by the second order time derivative) and a Dirichlet boundary condition (we also have other implicit boundary conditions, such as the well-definedness of the solution in Ω etc...). Then by the linearity of the wave equation $\phi = \phi_1 - \phi_2$ will be another solution, this time satisfying

$$\begin{cases} \phi(\mathbf{x}, 0) = 0, \ \mathbf{x} \in \Omega \\ \frac{\partial \phi}{\partial t}(\mathbf{x}, 0) = 0, \mathbf{x} \in \Omega \\ \phi(\mathbf{x}, t) = 0, \ \mathbf{x} \in \partial\Omega \end{cases} \tag{32.3.7}$$

Therefore we see that $H(t = 0) = 0$, and since $\nabla\phi$ vanishes on $\partial\Omega$, energy conservation implies that $H(t = 0) = H(t) = 0$ at all times t. This implies that

$$\int_\Omega \left[\mu \left(\frac{\partial \phi}{\partial t} \right)^2 + T(\nabla\phi)^2 \right] d^n\mathbf{x} = 0, \ \forall t \tag{32.3.8}$$

and since the integrand is a sum of two positive numbers we find that $\nabla\phi = \frac{\partial\phi}{\partial t} = 0$ on $\Omega \times [0, \infty)$. In other words, $\phi = c$, but using the initial conditions $\phi = 0$ and thus $\phi_1 = \phi_2$, the two original solutions are identical. Note that had we changed the Dirichlet condition with a Neumann condition, then the argument would have been identical until the last step. This time ϕ_1 and ϕ_2 would differ at most by a constant.

Just like for the Laplace equation, this is a very useful result, it shows that if we are able to find a solution to the wave equation satisfying the conditions in (32.3.6), then that will be **the** solution, independent of how we have found it (guessing, summoning a genie, separation of variables etc...)

32.4   The 1D wave equation: strings

32.5   The 2D wave equation: membranes

We move to the two-dimensional case and consider a circular membrane of unit radius made out of a perfectly elastic material of density μ, and that is fixed at its circular boundary by a tension force per unit length T. We are interested in modelling waves on this circular membrane which are describes by the wave equation, it will be useful to use polar coordinates due to the problem's circular geometry:

$$\frac{\partial^2 \phi}{\partial t^2} = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} + \frac{1}{r}\frac{\partial\phi}{\partial r} \right)\phi \tag{32.5.1}$$

Moreover, we also require that

$$\begin{cases} \phi(r, \theta, t) = \phi(r, \theta + 2\pi, t) \\ \phi(1, \theta, t) = 0 \end{cases} \tag{32.5.2}$$

We begin by using separation of variables and let $\phi(r, \theta, t) = R(r)\Theta(\theta)T(t)$, so that

$$\frac{R(r)\Theta(\theta)}{c^2}\frac{d^2T}{dt^2} = \Theta(\theta)T(t)\frac{d^2R}{dr^2} + \frac{\Theta(\theta)T(t)}{r}\frac{dR}{dr} + \frac{R(r)T(t)}{r^2}\frac{d^2\Theta}{d\theta^2} \tag{32.5.3}$$

$$\frac{1}{c^2T(t)}\frac{d^2T}{dt^2} = \frac{1}{R(r)}\frac{d^2R}{dr^2} + \frac{1}{rR(r)}\frac{dR}{dr} + \frac{1}{r^2\Theta(\theta)}\frac{d^2\Theta}{d\theta} \tag{32.5.4}$$

Note that the LHS is a function of t while the RHS is a function of (r, θ), so both must be set to a constant k, yielding

$$\frac{1}{T(t)}\frac{d^2T}{dt^2} = -kc^2, \ \frac{r^2}{R(r)}\frac{d^2R}{dr^2} + \frac{r}{R(r)}\frac{dR}{dr} + kr^2 = -\frac{1}{\Theta(\theta)}\frac{d^2\Theta}{d\theta} \tag{32.5.5}$$

Again we see that for the second equation the LHS depends on r only while the RHS depends on θ, so both sides will also be equal to a constant l

$$\frac{d^2T}{dt^2} = -kc^2T, \ r^2\frac{d^2R}{dr^2} + r\frac{dR}{dr} + (kr^2 - l)R = 0, \ \frac{d^2\Theta}{d\theta} = -l\Theta \tag{32.5.6}$$

The first ODE is solved by $T(t) = Ae^{i\sqrt{k}t} + Be^{-i\sqrt{k}t}$. The other two ODEs are Sturm-Liouiville equations which we have encountered previously! We have that $\Theta(\theta) = Ce^{i\sqrt{l}\theta} + De^{-i\sqrt{l}\theta}$, and similarly $R(r) = EJ_{\sqrt{l}}(\sqrt{k}r) + FY_{\sqrt{l}}(\sqrt{k}r)$. The most general solution we can build up using our separation of variables ansatz is thus given by

$$\phi(r, \theta, t) = \sum_{k,l}(Ae^{\sqrt{k}ct} + Be^{-\sqrt{k}ct})(Ce^{i\sqrt{l}\theta} + De^{-i\sqrt{l}\theta})[EJ_{\sqrt{l}}(\sqrt{k}r) + FY_{\sqrt{l}}(\sqrt{k}r)] \tag{32.5.7}$$

Now we can use our boundary conditions to fix some parameters. Firstly, the Bessel function of the second kind $Y_n(r)$ is ill-defined at the origin, and since we require our solution to be regular over the entire membrane it must be discarded, so $D = 0$. Next, we also require 2π-periodicity in θ, so we need $\sqrt{l} \in \mathbb{N}$ and thus $l = n^2$, $\forall n \in \mathbb{Z}$. Our solution is thus

$$\phi(r, \theta, t) = \sum_{k=1}^{\infty}\sum_{n=0}^{\infty}(Ae^{i\sqrt{k}ct} + Be^{-i\sqrt{k}ct})(C_{n,k}e^{in\theta} + D_{n,k}e^{-in\theta})J_n(\sqrt{k}r) \tag{32.5.8}$$

Finally, we use $\phi(1, \theta, t) = 0$ and the linear independence of our solutions to find that

$$\sum_{k=1}^{\infty}\sum_{n=0}^{\infty}(Ae^{i\sqrt{k}ct} + Be^{-i\sqrt{k}ct})(C_{n,k}e^{in\theta} + D_{n,k}e^{-in\theta})J_n(\sqrt{k}) = 0 \implies J_n(\sqrt{k}) = 0 \tag{32.5.9}$$

Therefore $\sqrt{k} = a_{nk}$ is the kth zero of the nth Bessel function of the first type. We have therefore found that

$$\boxed{\phi(r, \theta, t) = \sum_{k=1}^{\infty}\sum_{n=0}^{\infty}(A_{n,k}e^{ij_{nk}ct} + B_{n,k}e^{-ij_{nk}ct})(C_{n,k}e^{in\theta} + D_{n,k}e^{-in\theta})J_n(j_{nk}r)} \tag{32.5.10}$$

The remaining parameters can be determined using the orthogonality of Sturm-Liouville eigenfunctions and the initial conditions. Usually we are given one condition on $\phi(t = 0)$ and another

on $\frac{\partial \phi}{\partial t}\big|_{t=0}$. For example, suppose that

$$\phi(r, \theta, 0) = 0, \ \frac{\partial \phi}{\partial t}\bigg|_{t=0} = g(r) \tag{32.5.11}$$

so the membrane is initially still but gets a sudden radial impulse at $t = 0$. The first condition implies that $A = -B$ so that

$$\phi(r, \theta, t) = \sum_{k=1}^{\infty} A_{0k} \sin(j_{0k}ct) J_0(j_{0k}r) + \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} (C_{n,k} e^{in\theta} + D_{n,k} e^{-in\theta}) \sin(j_{nk}ct) J_n(j_{nk}r) \tag{32.5.12}$$

For the second condition, we find that

$$\sum_{k=1}^{\infty} cj_{0k} A_{0k} J_0(j_{0k}r) + \sum_{k=1}^{\infty} \sum_{n=0}^{\infty} j_{nk}(C_{n,k} e^{in\theta} + D_{n,k} e^{-in\theta}) J_n(j_{nk}r) = g(r) \tag{32.5.13}$$

The RHS does not depend on θ so only the $n = 0$ term will survive. The solution is

$$\phi(r, \theta, t) = \sum_{k=1}^{\infty} A_{0k} \sin(j_{0k}ct) J_0(j_{0k}r), \ A_{0k} = \frac{1}{cj_{0k}} \frac{2}{[J_1(j_{0k})]^2} \int_0^1 J_0(j_{0k}r) g(r) r \ dr \tag{32.5.14}$$

where we used

$$\int_0^1 J_n(j_{nk}r) J_n(j_{nl}r) r \ dr = \frac{1}{2} [J_{n+1}(j_{nk})]^2 \delta_{kl} \tag{32.5.15}$$

32.6 Fourier transforming to the Hemholtz equation

Let us denote the Fourier transform of $\phi(\mathbf{x}, t)$ as $\tilde{\phi}(\mathbf{k}, t)$ so that

$$\phi(\mathbf{x}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{\phi}(\mathbf{k}, t) e^{i\mathbf{k}\cdot\mathbf{x}} \ d^3\mathbf{k}, \ \tilde{\phi}(\mathbf{k}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(\mathbf{x}, t) e^{-i\mathbf{k}\cdot\mathbf{x}} \ d^3\mathbf{x} \tag{32.6.1}$$

It follows that

$$\partial^2 \phi \partial t^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\partial^2 \tilde{\phi}}{\partial t^2} e^{-i\mathbf{k}\cdot\mathbf{x}} \ d^3\mathbf{k} \tag{32.6.2}$$

and

$$\nabla^2 \phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{\phi} e^{-i\mathbf{k}\cdot\mathbf{x}} (-\mathbf{k}^2) e^{-i\mathbf{k}\cdot\mathbf{x}} \ d^3\mathbf{k} \tag{32.6.3}$$

so the wave-equation takes the form

$$\frac{\partial^2 \tilde{\phi}}{\partial t^2} + c^2 \mathbf{k}^2 \tilde{\phi} = 0 \tag{32.6.4}$$

This is known as the **Hemholtz equation**, it describes the spectral evolution of the solutions to the wave equation as linear (namely each mode evolves linearly).

PDE5 Parabolic PDEs: Heat and Diffusion

33.1 Existence and uniqueness theorem

PDE6 Green's functions for PDEs

34

Part V

Linear Algebra

L1 Vector spaces

35.1 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 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: $(\alpha_1 + \alpha_2)\mathbf{x}_1 = \alpha_1 \mathbf{u} + \alpha_2 \mathbf{u}$
> (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 $W \subseteq W$, 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: $\forall \alpha \in \mathbb{K}, \forall \mathbf{w} \in W, \; \alpha \mathbf{w} \in \mathbb{W}$
>
> (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 $\{\mathbf{v}_1 ... \mathbf{v}_k\}$ is defined as the set of all their linear combinations:
>
> $$\text{Span}(\mathbf{v}_1 ... \mathbf{v}_k) \equiv \left\{ \sum_{i=1}^{k} \alpha_i \mathbf{v}_i : \forall \alpha_i \in \mathbb{K} \right\} \tag{35.1.1}$$

> **Definition 34.5** *(Linear independence)*
> Let V be a vector space over $\mathbb{K}$ and let $\alpha_1 ... \alpha_k \in \mathbb{K}$. Then we say that the set of vectors $\{\mathbf{v}_1 ... \mathbf{v}_k\}$ are linearly independent *iff*:
>
> $$\sum_{i=1}^{k} \alpha_i \mathbf{v}_i = \mathbf{0} \implies \alpha_i = 0, \; \forall 1 \le i \le k \tag{35.1.2}$$
>
> 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 of 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 $\{\mathbf{0}, \mathbf{v}_1, ..., \mathbf{v}_k\} \subseteq V$ is linearly dependent.

Proof. The set $\{\mathbf{0}, \mathbf{v}_1, ..., \mathbf{v}_k\} \subseteq V$ cannot be linearly independent, since for $\alpha \ne 0$ we may write:

$$\alpha \cdot \mathbf{0} + \sum_{i=1}^{k} 0 \cdot \mathbf{v}_i = \mathbf{0} \tag{35.1.3}$$

Therefore, the vectors must be linearly dependent. ∎

35.2 Basis and dimensions

> **Definition 34.6** *(Basis)*
> A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\} \subseteq V$ is said to be a basis of V *iff*:
> (B1) they are linearly independent
> (B2) they generate V : $\text{Span}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k) = V$.
> Then, $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$ are said to be **basis vectors**.

So, given a basis $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$, 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 $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$ be a basis of a vector space V. Then, any vector $\mathbf{v} \in V$ can be expressed as:
> $$\mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i \tag{35.2.1}$$
> where α_i are uniquely determined.

Proof. Suppose that $\mathbf{v}$ can be expressed as two different linear combinations:

$$\mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i = \sum_{i=1}^{k} \alpha_i' \mathbf{v}_i \tag{35.2.2}$$

Then:

$$\sum_{i=1}^{k} (\alpha_i - \alpha_i') \mathbf{v}_i = \mathbf{0} \tag{35.2.3}$$

However, since $\mathbf{v}_i$ are linearly independent by (B1), this implies that $\alpha_i = \alpha_i'$, which is a contradiction. ∎

> **Theorem 34.8** *(Steinitz Exchange theorem)*
> Let $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$ be a basis of V, and let $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_l\} \subsetneq V$. If $l > k$, then $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_l$ are linearly dependent.

Proof. If $\mathbf{w}_1 = \mathbf{0}$, then by Proposition 34.5 $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_l$ are linearly dependent.

Suppose $\mathbf{w}_1 \neq \mathbf{0}$. Then, we may write:

$$\mathbf{w}_1 = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i \iff \mathbf{v}_1 = \frac{1}{\alpha_1}\left(\mathbf{w}_1 - \sum_{i=2}^{k} \alpha_i \mathbf{v}_i\right) \tag{35.2.4}$$

where we assume without loss of generality that $\exists \alpha_1 \neq 0$, since otherwise the sets would be linearly dependent as desired. Hence:

$$\text{Span}\left((\mathbf{w}_1 - \sum_{i=2}^{k} \alpha_i \mathbf{v}_i), \mathbf{v}_2, ..., \mathbf{v}_k\right) = V \tag{35.2.5}$$

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:

$$\text{Span}\left(\mathbf{w}_1, \mathbf{v}_2, ..., \mathbf{v}_k\right) = V \tag{35.2.6}$$

We can repeat this process by replacing $\mathbf{v}_j$ with:

$$\frac{1}{\alpha_j'}\left(\mathbf{w}_j - \sum_{i=1}^{j-1} \alpha_i' \mathbf{w}_i - \sum_{i=j+1}^{k} \alpha_i' \mathbf{v}_i\right), \ \forall \, 1 < j \leq l \tag{35.2.7}$$

so that, by similar logic to before:

$$\text{Span}\left(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{j-1}, \mathbf{w}_j - \sum_{i=1}^{j-1} \alpha_i' \mathbf{w}_i - \sum_{i=j+1}^{k} \alpha_i' \mathbf{v}_i, \mathbf{v}_{j+1}, ..., \mathbf{v}_k\right) = V \tag{35.2.8}$$

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:

$$\text{Span}(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_j, \mathbf{v}_{j+1}, ... \mathbf{v}_l) \tag{35.2.9}$$

Now since $l \geq k$, the end result of reiterating this algorithm will be:

$$\text{Span}(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_k) = V \tag{35.2.10}$$

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 $\mathbf{w}_1 ... \mathbf{w}_k$. For example:

$$\mathbf{w}_{l-k} = \sum_{i}^{k} \beta_i \mathbf{w}_i \implies \mathbf{w}_{l-k} - \sum_{i}^{k} \beta_i \mathbf{w}_i = \mathbf{0} \tag{35.2.11}$$

implying linear dependence. $\blacksquare$

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 $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$ be a basis of V, and let $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_l\} \subsetneq V$. If $\mathbf{w}_1, \mathbf{w}_2, ... \mathbf{w}_l$ 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 :
>
> $$\dim_{\mathbb{K}}(V) \equiv n \tag{35.2.12}$$

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, $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$, 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 $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\} \subseteq V$ 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:
$$\text{Span}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k) = V \tag{35.2.13}$$

If $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}$ 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 $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k$ do not span V (since otherwise we would already have a basis). Then,
$\exists \mathbf{v}_{k+1} \neq \text{Span}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k)$, 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 inde-
pendent set of vectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ does not form a basis. Then, $\exists \mathbf{v}_{n+1} \neq \text{Span}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n)$

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 $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\} \subseteq V$. We can deduce using (D3) that $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\} \subseteq V$ must be a basis of W too, and thus:

$$\text{Span}(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n) = V = W \tag{35.2.14}$$

as desired.

∎

35.3 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:

$$c\mathbf{w}_1 + \mathbf{w}_2 \in W, \ \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \tag{35.3.1}$$

Proof. We wish to prove that (S1)-(S3) are equivalent to 34.3.1. Firstly, note that:

$$(\mathbf{w}_1' + \mathbf{w}_2' \in W) \wedge (c\mathbf{w}_1' \in W, \ \forall \mathbf{w}_1', \mathbf{w}_2' \in W, \forall c \in \mathbb{K}) \tag{35.3.2}$$
$$\implies c\mathbf{w}_1 + \mathbf{w}_2 \in W, \ \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \tag{35.3.3}$$

if we take $\mathbf{w}_1' = c\mathbf{w}_1$. Similarly:

$$c\mathbf{w}_1 + \mathbf{w}_2 \in W, \ \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \tag{35.3.4}$$
$$\implies (\mathbf{w}_1' + \mathbf{w}_2' \in W) \wedge (c\mathbf{w}_1' \in W, \ \forall \mathbf{w}_1', \mathbf{w}_2' \in W, \forall c \in \mathbb{K}) \tag{35.3.5}$$

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:

$$c\mathbf{w}_1 + \mathbf{w}_2 \in W, \ \forall \mathbf{w}_1, \mathbf{w}_2 \in W, \forall c \in \mathbb{K} \implies \mathbf{0} \in W \tag{35.3.6}$$

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 $S = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k\} \subseteq V$. 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:

$$c\mathbf{v}_1 + \mathbf{v}_2 = c\sum_{i=1}^{k} \alpha_i \mathbf{u}_i + \sum_{i=1}^{k} \beta_i \mathbf{u}_i = \sum_{i=1}^{k}(c\alpha_i + \beta_i)\mathbf{u}_i = \sum_{i=1}^{k} \gamma_i \mathbf{u}_i \in \text{Span}(W) \qquad (35.3.7)$$

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 $\mathcal{B}_V = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$ be a basis of V, and $\mathcal{B}_W = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n\}$ be a basis of W, then it follows that $\mathcal{B}_W$ is linearly independent. Hence, using Proposition 34.9, $\dim(W) = m < n = \dim(V)$. ∎

> **Proposition 34.15** *(Union and intersection of subspaces)*
> Let V be a finite dimensional vector space, with subspaces $\{W_1, W_2, ..., W_n\}$. Then $\forall 1 \leq k \leq n$:
> $$W = \bigcap_{i=1}^{k} W_i \text{ is a subspace of } V \ \forall 1 \leq k \leq n \qquad (35.3.8)$$
> and for any two subspaces W_1, W_2:
> $$W_1 \cup W_2 \text{ is a subspace of } V \text{ iff } W_1 \subseteq W_2 \text{ or } W_2 \subseteq W_1 \qquad (35.3.9)$$

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:

$$\alpha \mathbf{u} + \mathbf{v} \in W_i, \forall i \implies \alpha \mathbf{u} + \mathbf{v} \in W \qquad (35.3.10)$$

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:

$$\mathbf{w}_1 + \mathbf{w}_2 \in V \qquad (35.3.11)$$

Now suppose that $\mathbf{w}_1 + \mathbf{w}_2 \in W_1$. Then:

$$(-\mathbf{w}_1) + (\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{w}_2 \in W_1 \qquad (35.3.12)$$

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:

$$(-\mathbf{w}_2) + (\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{w}_1 \in W_1 \tag{35.3.13}$$

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 $W_2 \subseteq W_1$.

($\Longleftarrow$) 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.

$$\blacksquare$$

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:

$$\mathbf{v} + W \equiv \big\{ \mathbf{v} + \mathbf{w} : \forall \mathbf{v} \in \mathbb{V} \big\} \tag{35.3.14}$$

where $\mathbf{w} \in W$. The set of all cosets of V in W is called the quotient space of W modulo V:

$$V/W \equiv \big\{ \mathbf{v} + W : \forall \mathbf{v} \in V \big\} \tag{35.3.15}$$

Finally, the sum of two vector subspaces U, W of V is defined as:

$$U + W \equiv \{ \mathbf{u} + \mathbf{w} : \forall \mathbf{u} \in U, \mathbf{w} \in W \} \tag{35.3.16}$$

Theorem 34.17 *(Dimension of sum of spaces)*

Let U, W be subspaces of a finite dimensional space V. Then:

$$\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W). \tag{35.3.17}$$

Proof. We firstly prove that $\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W)$.

Let $\mathcal{B}_\cap = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r\}$ be a basis for $U \cap W$ so that $\dim(U \cup W) = r$. 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 $\mathcal{B}_\cap$ can be extended to form a basis of U and W.

Hence suppose that $\mathcal{B}_U = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r, \mathbf{u}_1, ..., \mathbf{u}_m\}$ and $\mathcal{B}_W = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r, \mathbf{w}_1, ..., \mathbf{w}_n\}$ are bases for U and W respectively.

Now let $\mathbf{v}' \in U + W$, then it may be expressed as:

$$\mathbf{v}' = \mathbf{u}' + \mathbf{w}' \tag{35.3.18}$$

$$= \underbrace{\left(\sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=1}^{m} \alpha_i' \mathbf{u}_i \right)}_{\mathbf{u}'} + \underbrace{\left(\sum_{i=1}^{r} \beta_i \mathbf{v}_i + \sum_{i=1}^{n} \beta_i' \mathbf{w}_i \right)}_{\mathbf{w}'} \tag{35.3.19}$$

for some $\mathbf{u}' \in U, \mathbf{w}' \in W$. We can rearrange the above equation:

$$\mathbf{v}' = \left(\sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=1}^{m} \alpha_i' \mathbf{u}_i \right) + \left(\sum_{i=1}^{r} \beta_i \mathbf{v}_i + \sum_{i=1}^{n} \beta_i' \mathbf{w}_i \right) \tag{35.3.20}$$

$$= \sum_{i=1}^{r} \gamma_i \mathbf{v}_i + \sum_{i=1}^{m} \alpha_i' \mathbf{u}_i + \sum_{i=1}^{n} \beta_i' \mathbf{w}_i \tag{35.3.21}$$

where $\gamma_i = \alpha_i + \beta_i$. Therefore:

$$\text{Span}(\mathbf{v}_1, ..., \mathbf{v}_r, \mathbf{u}_1, ..., \mathbf{u}_m, \mathbf{w}_1, ..., \mathbf{w}_n) = U \tag{35.3.22}$$

Also, note that $\{\mathbf{v}_1, ..., \mathbf{v}_r, \mathbf{u}_1, ..., \mathbf{u}_m, \mathbf{w}_1, ..., \mathbf{w}_n\}$ is linearly independent. Indeed, suppose:

$$\sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=1}^{m} \beta_i \mathbf{u}_i + \sum_{i=1}^{n} \gamma_i \mathbf{w}_i = \mathbf{0} \implies \sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=1}^{m} \beta_i \mathbf{u}_i = -\sum_{i=1}^{n} \gamma_i \mathbf{w}_i \tag{35.3.23}$$

for some $\alpha_i, \beta_i, \gamma_i \in \mathbb{K}$. The above belongs to U, since $\{\mathbf{v}_1, ..., \mathbf{v}_r, \mathbf{u}_1, ..., \mathbf{u}_m\}$ is a basis of U. Similarly, it also belongs to W. Therefore:

$$\sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=1}^{m} \beta_i \mathbf{u}_i = -\sum_{i=1}^{n} \gamma_i \mathbf{w}_i \in U \cap W \tag{35.3.24}$$

and can therefore be written as:

$$\sum_{i=1}^{r} \alpha_i \mathbf{v}_i + \sum_{i=1}^{m} \beta_i \mathbf{u}_i = -\sum_{i=1}^{n} \gamma_i \mathbf{w}_i = \sum_{i=1}^{r} c_i \mathbf{v}_i \tag{35.3.25}$$

$$\implies \sum_{i=1}^{m} \beta_i \mathbf{u}_i = \sum_{i=1}^{r} d_i \mathbf{v}_i \text{ and } -\sum_{i=1}^{n} \gamma_i \mathbf{w}_i = \sum_{i=1}^{r} c_i \mathbf{v}_i \tag{35.3.26}$$

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 $\{\mathbf{v}_1, ..., \mathbf{v}_r, \mathbf{u}_1, ..., \mathbf{u}_m, \mathbf{w}_1, ..., \mathbf{w}_n\}$ is a basis of $U + W$. It follows that:

$$\dim(U + W) = (r + m) + (r + n) - r = \dim(U) + \dim(W) - \dim(U \cap W) \tag{35.3.27}$$

as was desired. ∎

> **Theorem 34.18** *(Dimension of quotient spaces)*
> Let U, W be subspaces of a finite dimensional space V. Then:
>
> $$\dim(V/W) = \dim(V) - \dim(W) \tag{35.3.28}$$

Proof. Let $\mathcal{B}_W = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n\}$ and $\mathcal{B}_V = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n, \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}$ 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:

$$\mathbf{v} + W = \sum_{i=1}^{n} \alpha_i \mathbf{w}_i + \sum_{i=1}^{m} \beta_i \mathbf{v}_i + V \tag{35.3.29}$$

$$= \sum_{i=1}^{n} \alpha_i(\mathbf{w}_i + W) + \sum_{i=1}^{m} \beta_i(\mathbf{v}_i + W) \tag{35.3.30}$$

$$= \sum_{i=1}^{m} \beta_i(\mathbf{v}_i + W) \tag{35.3.31}$$

$$\in \mathrm{Span}(\mathbf{v}_1 + W, \mathbf{v}_2 + W, ..., \mathbf{v}_m + W\} \tag{35.3.32}$$

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 $\{\mathbf{v}_1 + W, \mathbf{v}_2 + W, ..., \mathbf{v}_m + W\}$ is linearly independent. Indeed:

$$\sum_{i=1}^{m} \alpha_i(\mathbf{v}_i + W) = \mathbf{0} + W \equiv W \tag{35.3.33}$$

implies:

$$\sum_{i=1}^{m} \alpha_i \mathbf{v}_i \in W \implies \sum_{i=1}^{m} \alpha_i \mathbf{v}_i = \sum_{i=1}^{n} \beta_i \mathbf{w}_i \tag{35.3.34}$$

But $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n, \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}$ is linearly independent since it forms a basis for V. Consequently, $\alpha_i = 0$, thus proving linear independence of $\{\mathbf{v}_1 + W, \mathbf{v}_2 + W, ..., \mathbf{v}_m + W\}$.

So we may claim that $\{\mathbf{v}_1 + W, \mathbf{v}_2 + W, ..., \mathbf{v}_m + W\}$ is a basis for V/W. It follows immediately that:

$$\dim(V/W) = (m + n) - n = \dim(V) - \dim(W) \tag{35.3.35}$$

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) $U + W = V$
(DS2) $U \cap W = \{\mathbf{0}\}$
We then say that U and W are complementary spaces, and denote the direct sum as:

$$U \oplus W = V \tag{35.3.36}$$

Instead, given two arbitrary vector spaces V_1, V_2 then their (external) direct sum is defined as:

$$V_1 \oplus V_2 \equiv \{(\mathbf{v}_1, \mathbf{v}_2) : \mathbf{v}_1 \in V_1, \mathbf{v}_2 \in V_2\} \tag{35.3.37}$$

An immediate consequence of this definition is that:

$$\dim(U \oplus W) = \dim(U) + \dim(W) \tag{35.3.38}$$

or more generally that:

$$\dim\left(\bigoplus_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} \dim(W)_i \tag{35.3.39}$$

L2 Matrix algebra

L3 Linear transformations

37.1 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:
>
> $$f : X \to Y \tag{37.1.1}$$
> $$x \mapsto f(x) \tag{37.1.2}$$
>
> Here X is called the **domain** of f, denoted $\mathrm{dom}(f)$. Instead, the set:
>
> $$\mathrm{Im}(f) = \{f(x) : x \in X\} \subseteq Y \tag{37.1.3}$$
>
> is called the **image** of f. If f is a homomorphism, then the set:
>
> $$\mathrm{Ker}(f) = \{x \in X : f(x) = 0\} \tag{37.1.4}$$
>
> is called the **kernel** of f.
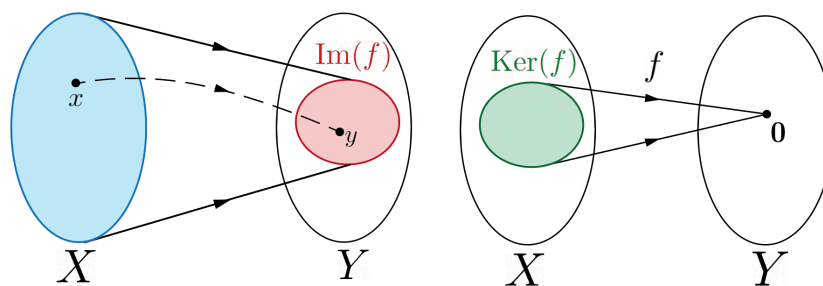
Figure 37.1. Map from X to Y.

> **Definition 46.2** *(Injectivity, Surjectivity, Bijectivity)* Recall that a function $f : X \to Y$ 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 $f : X \to Y$ and $g : Y \to Z$ then their composite map is defined as:

$$g \circ f : X \to Z \tag{37.1.5}$$
$$x \mapsto g(f(x)) \tag{37.1.6}$$

shown in the form of a commutative diagram below:

$$
\begin{array}{ccc}
A & \xrightarrow{\ f\ } & B \\
& \underset{g \circ f}{\searrow} & \downarrow g \\
& & C
\end{array}
\tag{37.1.7}
$$

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:
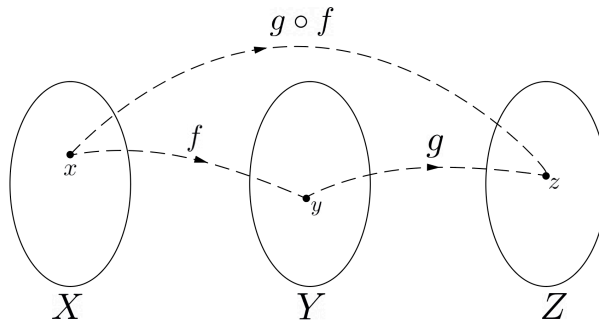
Figure 37.2. Composite map $g \circ f$

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 $\text{id}_X : X \to X$ maps all elements of X to themselves.

Given a map $f : X \to Y$, then $g : Y \to X$ is its inverse is:

$$(g \circ f) = \mathrm{id}_X \quad \text{and} \quad (f \circ g) = \mathrm{id}_Y \tag{37.1.8}$$
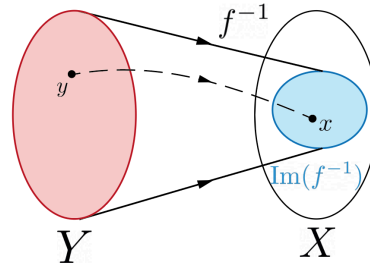
Figure 37.3. Inverse map f^{-1}

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 $f : X \to Y$ has an inverse *iff* f is bijective, and this inverse is unique.

Proof.

($\Longrightarrow$) Suppose f has an inverse $g : Y \to X$. Then:

$$g \circ f = \mathrm{id}_X \quad \text{and} \quad f \circ g = \mathrm{id}_Y \tag{37.1.9}$$

If $f(x) = f(x')$ then $g(f(x)) = \mathrm{id}_X(x) = x$ and $g(f(x')) = \mathrm{id}_X(x') = x'$ so that $x = x'$, thus implying injectivity.

Let $y \in Y$. So $g(y) = x$ for some x, hence $f(g(y)) = \mathrm{id}_Y(y) = y = f(x)$. Hence for any y there exists some x so that $y = f(x)$, giving surjectivity.

($\Longleftarrow$) 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 $y \in Y$:

$$(f \circ g)(y) = f(x) = y \quad \text{and} \quad (g \circ f)(x) = g(y) = x \tag{37.1.10}$$

as required, g is its inverse.

Suppose f is invertible with two inverses g, h. Then:

$$(f \circ g)(x) = (f \circ h)(x) \forall x \in X \tag{37.1.11}$$

then composing with g $g(x) = h(x) \implies g = h$ since this equality holds for any x. $\blacksquare$

One can see this more intuitively. Indeed, if f is surjective, then 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:
> $$(f \circ g)^{-1} = g^{-1} \circ f^{-1} \text{ and } (f^{-1})^{-1} = f \qquad (37.1.12)$$

Proof. The first follows from:
$$(f \circ g)^{-1} \circ (f \circ g) = \mathrm{id} \qquad (37.1.13)$$

and:
$$(g^{-1} \circ f^{-1}) \circ (f \circ g) = g^{-1} \circ \mathrm{id} \circ g = \mathrm{id} \qquad (37.1.14)$$

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:
$$(f^{-1})^{-1} \circ f^{-1} = \mathrm{id} \text{ and } f \circ f^{-1} = \mathrm{id} \qquad (37.1.15)$$

hence by the same logic as before $f = (f^{-1})^{-1}$ as desired. ∎
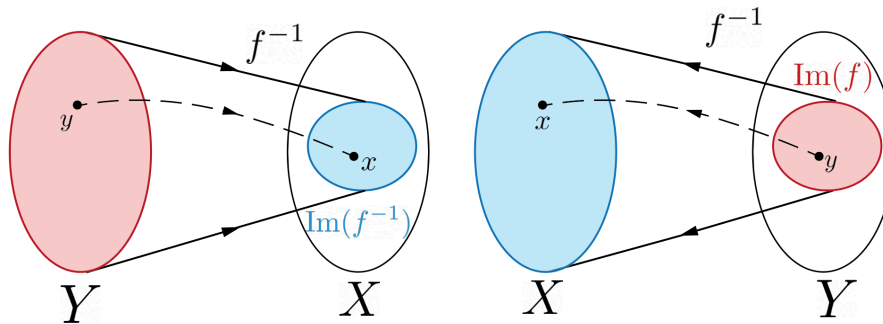
Figure 37.4. Left: diagram of f^{-1} treating it following definition 46.1 (so all elements of Y get mapped). Right: diagram of f^{-1} treating it as the map that undoes f

37.2 What is a linear map?

> **Definition 46.7** *(Linear map)*
> A **linear map** is a map $f : V \to W$ between two vector spaces V and W over a field $\mathbb{F}$ such that it preserves addition and scalar multiplication:
> (L1) $f(\mathbf{v}_1 + \mathbf{v}_2) = f(\mathbf{v}) + f(\mathbf{v}_2)$, $\forall \mathbf{v}_1, \mathbf{v}_2 \in V$
> (L2) $f(\alpha \mathbf{v}_1) = \alpha f(\mathbf{v}_1)$, $\forall \mathbf{v}_1 \in V, \forall \alpha \in \mathbb{F}$
> The set of all linear maps from V to W is denoted $\mathrm{Hom}(V, W)$

> **Proposition 46.8** *(Properties of linear maps)*
> For any linear maps $f : V \to W$:

(i) $f(\mathbf{0}) = \mathbf{0}$, that is, f fixes the zero vector.

(ii) $\mathrm{Ker}(f)$ is a subspace of V and $\mathrm{Im}(f)$ is a subspace of W.

(iii) f is surjective $\iff \mathrm{Im}(f) = W \iff \dim \mathrm{Im}(f) = \dim W$.

(iv) f is injective $\iff \mathrm{Ker}(f) = \{\mathbf{0}\} \iff \dim \mathrm{Ker}(f) = 0$.

(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 $\mathbf{0} \in \mathrm{Ker}(f)$ in point (i). $f(\mathbf{0}) = f(0 \cdot \mathbf{0}) = 0 f(\mathbf{0}) = \mathbf{0}$

(ii) Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathrm{Ker}(f)$. Then $f(\alpha \mathbf{v}_1 + \mathbf{v}_2) = \alpha f(\mathbf{v}_1) + f(\mathbf{v}_2) = \mathbf{0} \implies \alpha \mathbf{v}_1 + \mathbf{v}_2 \in \mathrm{Ker}(f)$. Proposition 34.12 then ensures that $\mathrm{Ker}(f)$ is a subspace of V.

We have already shown that $\mathbf{0} \in \mathrm{Im}(f)$ in point (i). Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathrm{Im}(f)$, so that $\exists \mathbf{v}_1, \mathbf{v}_2 \in V$ such that $\mathbf{w}_i = f(\mathbf{v}_i)$. Then $\alpha \mathbf{w}_1 + \mathbf{w}_2 = \alpha f(\mathbf{v}_1) + f(\mathbf{v}_2) = f(\alpha \mathbf{v}_1 + \mathbf{v}_2) \in \mathrm{Im}(f)$. Proposition 34.12 then ensures that $\mathrm{Im}(f)$ is a subspace of W.

(iii) If f is surjective, then $\forall \mathbf{w} \in W, \exists \mathbf{v} \in V$ such that $\mathbf{w} = f(\mathbf{v}) \in \mathrm{Im}(f)$. So $W \subseteq \mathrm{Im}(f)$, and $\mathrm{Im}(f) \subseteq W$ implying $\mathrm{Im}(f) \subseteq W$.

If instead $\mathrm{Im}(f) = W$, then $W \subseteq \mathrm{Im}(f)$, and $\mathrm{Im}(f) \subseteq W$. Hence $\forall \mathbf{w} \in W, \exists \mathbf{v} \in V$ such that $\mathbf{w} = f(\mathbf{v}) \in \mathrm{Im}(f)$, implying that f is surjective. It follows then that the two spaces have the same dimension.

(iv) Suppose f is injective, and let $\mathbf{v}_1, \mathbf{0} \in \mathrm{Ker}(f)$. Then $f(\mathbf{v}_1 - \mathbf{0} = \mathbf{0}$ since $\mathrm{Ker}(f)$ is a subspace. Consequently $f(\mathbf{v}_1) = f(\mathbf{0})$ and so $\mathbf{v}_1 = \mathbf{0}$. So $\mathrm{Ker}(f) = \{\mathbf{0}\}$.

Suppose $\mathrm{Ker}(f) = \{\mathbf{0}\}$, and let $\mathbf{v}_1, \mathbf{v}_2 \in \mathrm{Ker}(f)$. Then, $f(\mathbf{v}_1) = f(\mathbf{v}_2) \implies f(\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$. Hence $\mathbf{v}_1 - \mathbf{v}_2 \in \mathrm{Ker}(f)$ and so $\mathbf{v}_1 = \mathbf{v}_2$.

It follows that $\dim \mathrm{Ker}(f) = \dim\{\mathbf{0}\} = 0$.

$\blacksquare$

Definition 46.9 *(Rank and nullity)*

The dimension of the image of a linear map f is called its **rank**:

$$\mathrm{rk}(f) = \dim \mathrm{Im}(f) \tag{37.2.1}$$

and the dimension of the kernel of a linear map f is called its **nullity**:

$$\mathrm{null}(f) = \dim \mathrm{Ker}(f) \tag{37.2.2}$$

Theorem 46.10 *(Rank-nullity theorem)*

For any linear map $f : V \to W$:

$$\mathrm{null}(f) + \mathrm{rk}(f) = \dim(V) \tag{37.2.3}$$

Proof. Let $n = \dim V$ and $k = \text{null}(f)$, and let $\{\mathbf{v}_1, ..., \mathbf{v}_k\}$ be a basis for $\text{Ker}(f)$, which we can complete to $\{\mathbf{v}_1, ..., \mathbf{v}_k, \mathbf{v}_{k+1}, ..., \mathbf{v}_n\}$ to form a basis for V. We wish to show that $f(\mathbf{v}_{k+1}...\mathbf{v}_n$ is a basis of $\text{Im}(f)$.

Let is firstly prove that $\text{Im}(f) = \text{Span}(f(\mathbf{v}_{k+1}...\mathbf{v}_n)$. Indeed consider $\mathbf{w} \in \text{Im}(f)$. Then there exists some $\mathbf{v}$ such that:

$$\mathbf{w} = f(\mathbf{v}) = f\left(\sum_{i=1}^{n} \alpha_i \mathbf{v}_i\right) = \sum_{i=1}^{n} \alpha_i f(\mathbf{v}_i) \tag{37.2.4}$$

However, for $i \leq k$ we have that $f(\mathbf{v}_i) = \mathbf{0}$ and so $\text{Im}(f) = \text{Span}(f(\mathbf{v}_{k+1}...\mathbf{v}_n)$.

Now let us prove that $\{\mathbf{v}_{k+1}...\mathbf{v}_n\}$ are linearly independent. Consider:

$$\sum_{i=k+1}^{n} \alpha_i f(\mathbf{v}_i) = \mathbf{0} \implies f\left(\sum_{i=k+1}^{n} \alpha_i \mathbf{v}_i\right) = \mathbf{0} \tag{37.2.5}$$

and consequently $\sum_{i=k+1}^{n} \alpha_i \mathbf{v}_i \in \text{Ker}(f)$. Using the fact that $\{\mathbf{v}_1, ..., \mathbf{v}_k\}$ is a basis of $\text{Ker}(f)$ we find that:

$$\sum_{i=k+1}^{n} \alpha_i \mathbf{v}_i = \sum_{i=1}^{k} \beta_i \mathbf{v}_i \implies \sum_{i=1}^{n} \gamma_i \mathbf{v}_i = \mathbf{0} \tag{37.2.6}$$

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 $\{\mathbf{v}_i, ..., \mathbf{v}_n\}$ forms a basis, and must therefore be linearly independent. Hence $\gamma_i = 0$ and so $\alpha_i = 0$ as well.

Since $\{\mathbf{v}_{k+1}...\mathbf{v}_n\}$ 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:

$$\dim V = n = k + (n - k) = \text{null}(f) + \text{rk}(f) \tag{37.2.7}$$

∎
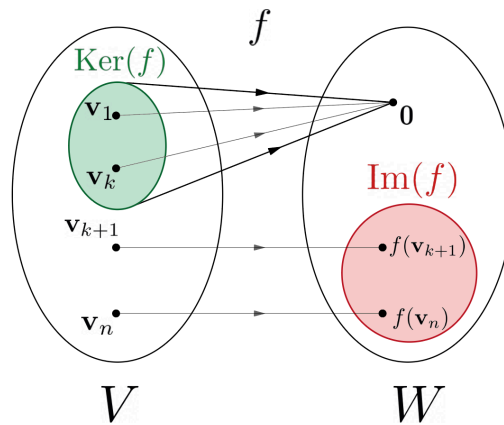
Figure 37.5. Visual depiction of the rank-nullity theorem

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 $\mathrm{Ker}(f)$ to only contain $\mathbf{0}$, and no other vectors. In other words, we need $\mathrm{null}(f) = 0$ and hence $\mathrm{rk}(f) = \dim W$.

Let's justify this more rigorously.

> **Proposition 46.11** *(Consequence of rank-nullity)*
> Let $f : V \to W$ be a linear transformation with $n = \dim V$. Then:
>   (i) f is bijective $\implies$ $\dim V = \dim W$.
>   (ii) f is bijective $\iff$ $\mathrm{null}(f) = 0 \iff \mathrm{rk}(f) = n$.
>   (iii) If it exists, $f^1 : W \to V$ is linear.

Proof.

(i) if f is bijective, then we showed in Proposition 46.8 that $\mathrm{null}(f) = 0$ and $\mathrm{rk}(f) = \dim W$. Using the rank nullity theorem:

$$\dim V = \mathrm{rk}(f) = \dim W \tag{37.2.8}$$

as required.

(ii) Suppose $\dim W = \dim W = n$ (which follows from bijectivity), then from the rank-nullity theorem:

$$n = \dim V = \mathrm{rk}(f) + \mathrm{null}(f) = \dim W \tag{37.2.9}$$

and since f is bijective then:

$$\mathrm{null}(f) = 0 \implies \mathrm{rk}(f) = n \tag{37.2.10}$$

(iii) Set $\mathbf{w}_1 = f(\mathbf{v}_1)$ and $\mathbf{w}_2 = f(\mathbf{v}_2)$. Then:

$$f^{-1}(\alpha \mathbf{w}_1 + \mathbf{w}_2) = f^{-1}(\alpha f(\mathbf{v}_1) + f(\mathbf{v}_2)) \tag{37.2.11}$$
$$= f^{-1}(f(\alpha \mathbf{v}_1 + \mathbf{v}_2)) \tag{37.2.12}$$
$$= \alpha \mathbf{v}_1 + \mathbf{v}_2 \tag{37.2.13}$$
$$= \alpha f^{-1}(\mathbf{w}_1) + f^{-1}(\mathbf{w}_2) \tag{37.2.14}$$

$\blacksquare$

> **Proposition 46.12** *(Linear map given basis of domain)*
> Let V and W be vector spaces, with $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ as a basis of V. For $\mathbf{w}_1, ..., \mathbf{w}_n \in W$ there exists one linear map $f : V \to W$ such that $f(\mathbf{v}_i) = \mathbf{w}_i$.

Proof. One such linear map is:

$$f(\mathbf{v}) = \sum \alpha_i \mathbf{w}_i \tag{37.2.15}$$

where α_i are the coefficients of $\mathbf{v}_i$:

$$\mathbf{v} = \sum \alpha_i \mathbf{v}_i \tag{37.2.16}$$

It is linear since:

$$f(\beta\mathbf{v} + \mathbf{v}') = f\left(\sum \beta\alpha_i\mathbf{v}_i + \sum \alpha_i'\mathbf{v}_i\right) \tag{37.2.17}$$

$$= f\left(\sum(\beta\alpha_i + \alpha_i')\mathbf{v}_i\right) \tag{37.2.18}$$

$$= \sum(\beta\alpha_i + \alpha_i')\mathbf{w}_i \tag{37.2.19}$$

$$= \beta\sum\alpha_i\mathbf{v}_i + \sum\alpha_i'\mathbf{v}_i \tag{37.2.20}$$

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 $\mathbf{v} \in V$:

$$g(\mathbf{v}) = \sum\alpha_i\mathbf{w}_i = f(\mathbf{v}) \implies f = g \tag{37.2.21}$$

as required. ∎

37.3 Isomorphisms

> **Definition 46.13** *(Isomorphism)*
> Two vector spaces V and W are said to be **isomorphic** whenever we can find an invertible linear map $f : V \to W$, 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 $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ and $\{\mathbf{w}_1, ..., \mathbf{w}_n\}$ 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:

$$f(\mathbf{v}) = f(\mathbf{v}') \implies \sum\alpha_i\mathbf{v}_i = \sum\beta_i\mathbf{v}_i \implies \alpha_i = \beta_i \tag{37.3.1}$$

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 $\dim V = \dim W$. ∎

37.4 Linear maps and matrices

> **Definition 46.15** *(Coordinate map)*
> For a vector space V over $\mathbb{K}$, endowed with a basis $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$, then we define the

> **coordinate map** $\varphi : V \mapsto \mathbb{K}^n$ by:
>
> $$\varphi_\beta(\mathbf{v}) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = [\mathbf{v}]_\beta, \ \forall \mathbf{v} \in V \tag{37.4.1}$$
>
> where $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$. Here, $[\mathbf{v}]_\beta$ is known as the **coordinate vector of v relative to** β.
> The inverse of φ_β is ϕ_β which given a basis β associates to a list of scalars $\alpha_1, \alpha_2, ..., \alpha_n$ a
> vector $\mathbf{v}$.

Note that φ and ϕ are clearly linear. Moreover, $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ 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 $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ for V and a basis $\gamma = \{\mathbf{w}_1, ..., \mathbf{w}_m\}$ 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 $1 \leq j \leq n$:
>
> $$f(\mathbf{v}_j) = \sum_{i=1}^{m} a_{ij} \mathbf{w}_i, \tag{37.4.2}$$
>
> 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 $\mathsf{A} = [f]_\beta^\gamma$.

We can draw a **commutative diagram** to demonstrate how matrix representations work:

$$
\begin{array}{ccc}
V & \xrightarrow{\ f\ } & W \\
\phi_\beta \big\uparrow & & \big\uparrow \phi_\gamma \\
\mathbb{K}_n & \xrightarrow[\ [f]_\beta^\gamma\ ]{} & \mathbb{K}_m
\end{array}
$$

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 $f : V \to W$ 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:
>
> $$A = ([f(\mathbf{v}_1)]_\gamma \ [f(\mathbf{v}_2)]_\gamma \ \dots \ [f(\mathbf{v}_n)]_\gamma) \tag{37.4.3}$$
>
> 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 $\mathbf{v} \in V$:

$$f(\mathbf{v}) = f\Big(\sum_{j=1}^n v_j\mathbf{v}_j\Big) = \sum_{j=1}^n v_j f(\mathbf{v}_j) \tag{37.4.4}$$

so that if we let $[f(\mathbf{v}_j)]_\gamma = (a_{1j}\ a_{2j}\ \dots\ a_{mj})^T$ then:

$$[f(\mathbf{v})]_\gamma = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \dots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \dots + a_{2n}v_n \\ \vdots \\ a_{m1}v_1 + a_{m2}v_2 + \dots + a_{mn}v_n \end{pmatrix} \tag{37.4.5}$$

We can express this more compactly as:

$$[f(\mathbf{v})]_\gamma = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \tag{37.4.6}$$

$$= \Big([f(\mathbf{v}_1)]_\gamma \ [f(\mathbf{v}_2)]_\gamma \ \dots \ [f(\mathbf{v}_n)]_\gamma\Big)[\mathbf{v}]_\beta \tag{37.4.7}$$

$$\implies [f(\mathbf{v})]_\gamma = A[\mathbf{v}]_\beta \tag{37.4.8}$$

More abstractly, one may write:

$$f(\mathbf{v}) = f\Big(\sum_{j=1}^n v_j\mathbf{v}_j\Big) = \sum_{j=1}^n v_j \sum_{i=1}^m a_{ij}\mathbf{w}_i = \sum_{i=1}^m \Big(\sum_{j=1}^n v_j a_{ij}\Big)\mathbf{w}_i \tag{37.4.9}$$

so that we end up with:

$$(f(\mathbf{v}))_i = \sum_{j=1}^n v_j a_{ij} = (A\mathbf{v})_i \tag{37.4.10}$$

This however is the expression of matrix multiplication $A\mathbf{v}$, 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 A and B as shown below:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \text{ and } B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix} \tag{37.4.11}$$

Then the image of any basis vector $\mathbf{v}_j \in \beta$ is from Proposition 46.18:

$$[f(\mathbf{v}_j)]_\gamma = (a_{1j}\, a_{2j}\, \dots\, a_{mj})^T = (b_{1j}\, b_{2j}\, \dots\, b_{mj})^T \tag{37.4.12}$$

so jth column of A and B coincide. Since this is true for any $1 \leq j \leq n$ we have that A = 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 $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ and $\gamma = \{\mathbf{w}_1, ..., \mathbf{w}_m\}$ then there exists unique a_{ij} and b_{ij} such that:

$$f(\mathbf{v}_j) = \sum_{i=1}^m a_{ij}\mathbf{w}_i, \text{ and } g(\mathbf{v}_j) = \sum_{i=1}^m b_{ij}\mathbf{w}_i \tag{37.4.13}$$

Then:

$$(\alpha f + g)(\mathbf{v}_j) = \sum_{i=1}^n (\alpha a_{ij} + b_{ij})\mathbf{w}_i \tag{37.4.14}$$

and so:

$$([f + g]_\beta^\gamma)_{ij} = \alpha a_{ij} + b_{ij} = \alpha[f]_\beta^\gamma + [g]_\beta^\gamma \tag{37.4.15}$$

∎

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 $g : V \to W$ and $f : W \to U$ are linear maps:

$$[f \circ g]_\alpha^\gamma = [f]_\beta^\gamma [g]_\alpha^\beta \tag{37.4.16}$$

Proof. Let $C = [f \circ g]_\alpha^\gamma$, $A = [f]_\beta^\gamma$, $B = [g]_\beta^\gamma$. Also, let $\alpha = \{\mathbf{v}_1, ..., \mathbf{v}_n\}, \beta = \{\mathbf{w}_1, ..., \mathbf{w}_m\}, \gamma = \{\mathbf{u}_1, ..., \mathbf{u}_l\}$. 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 $\mathbf{v}_i \in \alpha$:

$$(f \circ g)(\mathbf{v}_i) = f\left(\sum_{k=1}^{m} B_{ki}\mathbf{w}_k\right) = \sum_{k=1}^{m} B_{ki} f(\mathbf{w}_k) \tag{37.4.17}$$

$$= \sum_{k=1}^{m} B_{ki}\left(\sum_{j=1}^{l} A_{jk}\mathbf{u}_k\right) \tag{37.4.18}$$

$$= \sum_{k=1}^{m}\sum_{j=1}^{l} A_{jk}B_{ki}\mathbf{u}_k = \sum_{k=1}^{l} C_{ji}\mathbf{u}_k \tag{37.4.19}$$

where:

$$C_{ji} = \sum_{j=1}^{l} A_{jk}B_{ki} = \mathbf{A}_j \cdot \mathbf{B}^i \tag{37.4.20}$$

However, this is exactly the definition of matrix multiplication we encountered in the previous chapter, hence:

$$[f \circ g]_\alpha^\gamma = \mathsf{C} = \mathsf{AB} = [f]_\beta^\gamma [g]_\alpha^\beta \tag{37.4.21}$$

∎

> **Proposition 46.22** *(Invertibility of linear maps)*
> Let $f : V \to W$ 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:

$$\mathbb{I}_n = [\mathrm{id}_V]_\beta = [f^{-1} \circ f]_\beta = [f^{-1}]_\gamma^\beta [f]_\beta^\gamma \tag{37.4.22}$$

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 $\mathsf{A} = [f]_\beta^\gamma$ is invertible, then there exists B so that $\mathsf{AB} = \mathsf{BA} = \mathbb{I}_n$. Then there exists a map $g \in \mathrm{Hom}(W, V)$ such that:

$$g(\mathbf{w}_j) = \sum_{i=1}^{n} B_{ij}\mathbf{v}_i \tag{37.4.23}$$

where $\gamma = \{\mathbf{w}_1, ..., \mathbf{w}_n\}$ and $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ are bases of W and V respectively. Then:

$$[g \circ f]_\beta = [g]_\gamma^\beta [f]_\beta^\gamma = \mathsf{BA} = \mathbb{I}_n = [\mathrm{id}_V]_\beta \tag{37.4.24}$$

so we conclude that $g \circ f = \mathrm{id}_V$ (and similarly $f \circ g = \mathrm{id}_V$) by the uniqueness of matrix representations.

∎

> **Theorem 46.23** *(All linear maps have a representation)*
> The vector spaces $\mathrm{Mat}_{n,m}(\mathbb{K})$ and $\mathrm{Hom}(V, W)$ are isomorphic.

Proof. Let $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ and $\gamma = \{\mathbf{w}_1, ..., \mathbf{w}_m\}$ be bases for V and W respectively. Then:

$$\Phi : \mathrm{Hom}(V, W) \to \mathrm{Mat}_{n,m}(\mathbb{K}) \tag{37.4.25}$$

$$f \mapsto [f]_\beta^\gamma \tag{37.4.26}$$

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 A then we can always find a linear map such that:

$$f(\mathbf{v}_i) = \sum_{i=1}^{m} A_{ij} \mathbf{w}_i \tag{37.4.27}$$

as warranted by Proposition 46.12. But this implies that $\Phi(f) = [f]_\beta^\gamma = A$ as required. ∎

An immediate consequence is that given two vector spaces V, W of dimensions n and m then $\mathrm{Hom}(V, W)$ 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 A : $\mathbb{F}^n \to \mathbb{F}^m$, its associated linear map is:
>
> $$A\mathbf{v} = \sum_{i=1}^{m} v_i f(\mathbf{v}_i) = \sum_{i=1}^{m} v_i \mathbf{A}^i \tag{37.4.28}$$
>
> then the image of A is $\mathrm{Im}(A) = \mathrm{Span}(\mathbf{A}^1, ..., \mathbf{A}^m)$. Its column rank, rk(A), is the number of linearly independent columns of A. Its row rank is the number of linearly independent rows of A, so rk(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:

$$\mathbf{A}_1 = \sum_{j=1}^{n} \alpha_j \mathbf{A}_j \tag{37.4.29}$$

and let $\boldsymbol{\alpha} = (\alpha_2 \; \alpha_3 \; ... \; \alpha_n)^T$. Let:

$$\mathbf{A}^i = \begin{pmatrix} a_i \\ \mathbf{b}_i \end{pmatrix}, \; \mathbf{b}_i = \begin{pmatrix} b_{2i} \\ b_{3i} \\ \vdots \\ b_{ni} \end{pmatrix} \implies A = \begin{pmatrix} a_1 & a_2 & ... & a_m \\ b_{21} & b_{22} & ... & b_{2m} \\ b_{31} & b_{32} & ... & b_{3m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & ... & b_{nm} \end{pmatrix} \tag{37.4.30}$$

Consequently

$$a_i = (\mathbf{A}^i)_1 = (\mathbf{A}_1)_i = \sum_{j=2}^{n} \alpha_j A_{ji} = \sum_{j=2}^{n} \alpha_j (\mathbf{A}^i)_j \tag{37.4.31}$$

$$= \alpha_2 b_{2i} + \alpha_3 b_{3i} + ... + \alpha_n b_{ni} \tag{37.4.32}$$

$$= \boldsymbol{\alpha} \cdot \mathbf{b}_i \tag{37.4.33}$$

so that

$$\mathbf{A}^i = \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_i \\ \mathbf{b}_i \end{pmatrix} \tag{37.4.34}$$

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 A we get:

$$\mathrm{A} = \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_1 & \boldsymbol{\alpha} \cdot \mathbf{b}_2 & ... & \boldsymbol{\alpha} \cdot \mathbf{b}_m \\ b_{21} & b_{22} & ... & b_{2m} \\ b_{31} & b_{32} & ... & b_{3m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & ... & b_{nm} \end{pmatrix} \longrightarrow \mathrm{A}' = \begin{pmatrix} b_{21} & b_{22} & ... & b_{2m} \\ b_{31} & b_{32} & ... & b_{3m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & ... & b_{nm} \end{pmatrix} \tag{37.4.35}$$

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 A and 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 $\alpha \cdot \mathbf{b}_i$ will not alter the linear dependence of the columns. Suppose that $\mathbf{A}^1$, $\mathbf{A}^2...\mathbf{A}^l$ is the maixmal linearly independent set of columns of A. Then note that

$$c_1 \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_1 \\ \mathbf{b}_1 \end{pmatrix} + c_2 \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_2 \\ \mathbf{b}_2 \end{pmatrix} + ... + c_l \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_l \\ \mathbf{b}_l \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} \cdot (c_1 \mathbf{b}_1 + ... + c_l \mathbf{b}_l) \\ c_1 \mathbf{b}_1 + ... + c_l \mathbf{b}_l \end{pmatrix} \tag{37.4.36}$$

So if $\mathbf{A}^1$, $\mathbf{A}^2...\mathbf{A}^l$ are linearly independent, then $\mathbf{b}_1, \mathbf{b}_2... \mathbf{b}_l$ are also linearly independent. Moreover, adding any other $\mathbf{b}_i$ will result in linear dependence. Indeed, if this were not the case, then $c_1' \mathbf{b}_1 + ... + c_l' \mathbf{b}_l + c_{l+1}' \mathbf{b}_{l+1} \implies c_i' = 0$ and so:

$$\begin{pmatrix} \boldsymbol{\alpha} \cdot (c_1' \mathbf{b}_1 + ... + c_{l+1}' \mathbf{b}_{l+1}) \\ c_1' \mathbf{b}_1 + ... + c_{l+1}' \mathbf{b}_{l+1} \end{pmatrix} = c_1' \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_1 \\ \mathbf{b}_1 \end{pmatrix} + ... + c_{l+1}' \begin{pmatrix} \boldsymbol{\alpha} \cdot \mathbf{b}_{l+1} \\ \mathbf{b}_{l+1} \end{pmatrix} = \mathbf{0} \implies c_i' = 0 \tag{37.4.37}$$

so the columns $\mathbf{A}^1$, $\mathbf{A}^2...\mathbf{A}^{l+1}$ 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 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. ∎

37.5 Change of basis and equivalence

> **Definition 46.26** *(Standard basis)*
> The standard basis of a coordinate vector space $\mathbb{K}^n$ is:
>
> $$\mathbf{e}_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i\text{ith component} \tag{37.5.1}$$
>
> 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**:

$$\mathbf{v} = \sum_{i=1}^{n} v_i \mathbf{e}_i \tag{37.5.2}$$

How do we find its coordinates in some other basis $\beta = \mathbf{v}_1, ..., \mathbf{v}_n$?

Consider the commutative diagram below:

$$\begin{array}{ccc} V & \xleftarrow{\ \mathrm{id}_V\ } & V \\ {\scriptstyle \phi_{\mathrm{id}}}\big\uparrow & & \big\uparrow{\scriptstyle \phi_\beta} \\ \mathbb{K}_n & \xleftarrow{\ \mathsf{P}_\beta\ } & \mathbb{K}_n \end{array}$$

then:

$$\mathsf{P}_\beta = [\mathrm{id}_V]_\beta^{\mathrm{id}} = \left([\mathbf{v}_1]_{\mathrm{id}}\ [\mathbf{v}_2]_{\mathrm{id}}\ \cdots\ [\mathbf{v}_n]_{\mathrm{id}}\right) \tag{37.5.3}$$

whereas:

$$\mathsf{M}_\beta = \mathsf{P}_\beta^{-1} = [\mathrm{id}_V]_{\mathrm{id}}^\beta = \left([\mathbf{e}_1]_\beta\ [\mathbf{e}_2]_\beta\ \cdots\ [\mathbf{e}_n]_\beta\right) \tag{37.5.4}$$

Indeed, notice that:

$$\mathsf{M}_\beta [\mathbf{v}]_{\mathrm{id}} = \left([\mathbf{e}_1]_\beta\ [\mathbf{e}_2]_\beta\ \cdots\ [\mathbf{e}_n]_\beta\right)[\mathbf{v}]_{\mathrm{id}} \tag{37.5.5}$$

$$= v_1[\mathbf{e}_1]_\beta + v_2[\mathbf{e}_2]_\beta + ... + v_n[\mathbf{e}_n]_\beta \tag{37.5.6}$$

$$= [v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + ... + v_n\mathbf{e}_n]_\beta \tag{37.5.7}$$

$$= [\mathbf{v}]_\beta = [\mathrm{id}_V(\mathbf{v})]_\beta \tag{37.5.8}$$

where $v_1, v_2, ..., v_n$ are the components of $[\mathbf{v}]_{\text{id}}$. Similarly:

$$P_\beta[\mathbf{v}]_\beta = ([\mathbf{v}_1]_{\text{id}} \ \ [\mathbf{v}_2]_{\text{id}} \ \ ... \ \ [\mathbf{v}_n]_{\text{id}})[\mathbf{v}]_\beta \tag{37.5.9}$$

$$= v_1'[\mathbf{v}_1]_{\text{id}} + v_2'[\mathbf{v}_2]_{\text{id}} + ... + v_n'[\mathbf{v}_n]_{\text{id}} \tag{37.5.10}$$

$$= [v_1'\mathbf{v}_1 + v_2'\mathbf{v}_2 + ... + v_n'\mathbf{v}_n]_{\text{id}} \tag{37.5.11}$$

$$= [\mathbf{v}]_{\text{id}} = [\text{id}_V(\mathbf{v})]_{\text{id}} \tag{37.5.12}$$

where $v_1', v_2', ..., v_n'$ are the components of $[\mathbf{v}]_\beta$.

> **Definition 46.27** *(Transition matrix)*
> If $\beta = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$ is a basis of a vector space V then:
>
> $$M_\beta = ([\mathbf{e}_1]_\beta \ \ [\mathbf{e}_2]_\beta \ \ ... \ \ [\mathbf{e}_n]_\beta) \tag{37.5.13}$$
>
> called the **transition matrix** maps $\varphi_{\text{id}}(\mathbf{v}) \mapsto \varphi_\beta(\mathbf{v})$ so that:
>
> $$[\mathbf{v}]_\beta = M_\beta[\mathbf{v}]_{\text{id}} \tag{37.5.14}$$

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 $\beta = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$ and γ be two bases of a vector space V. Then:
>
> $$M_{\beta \to \gamma} = ([\mathbf{v}_1]_\gamma \ \ [\mathbf{v}_2]_\gamma \ \ ... \ \ [\mathbf{v}_n]_\gamma) \tag{37.5.15}$$
>
> is the **change of coordinate matrix** mapping $\varphi_\beta(\mathbf{v}) \mapsto \varphi_\gamma(\mathbf{v})$ so that:
>
> $$[\mathbf{v}]_\gamma = M_\beta[\mathbf{v}]_\beta \tag{37.5.16}$$

Proof. Consider the following commutative diagram:

so that it is clear that:

$$M_{\beta \to \gamma} = [\varphi_\gamma \circ \phi_\beta]_{\text{id}} = [\text{id}]_\beta^\gamma \tag{37.5.17}$$

Then

$$M_{\beta \to \gamma} = M_\gamma P_\beta = M_\gamma([\mathbf{v}_1]_{\text{id}} \ \ [\mathbf{v}_2]_{\text{id}} \ \ ... \ \ [\mathbf{v}_n]_{\text{id}}) \tag{37.5.18}$$

$$= (\phi_\gamma([\mathbf{v}_1]_{\text{id}}) \ \ \phi_\gamma([\mathbf{v}_2]_{\text{id}}) \ \ ... \ \ \phi_\gamma([\mathbf{v}_n]_{\text{id}})) \tag{37.5.19}$$

$$= ([\mathbf{v}_1]_\gamma \ \ [\mathbf{v}_2]_\gamma \ \ ... \ \ [\mathbf{v}_n]_\gamma) \tag{37.5.20}$$

as desired. Alternatively, using proposition 46.18 with $f = \text{id}_V$ then:

$$M_{\beta \to \gamma} = [\text{id}]_{\beta}^{\gamma} = ([\mathbf{v}_1]_{\gamma} \; [\mathbf{v}_2]_{\gamma} \; ... \; [\mathbf{v}_n]_{\gamma}) \tag{37.5.21}$$

as found previously. ∎

> **Proposition 46.29** *(Invertibility of change of coordinate matrix)*
> The inverse of $M_{\beta \to \gamma}$ is $M_{\beta \to \gamma}^{-1} = M_{\gamma \to \beta}$.

Proof. Let us firstly prove that $M_{\gamma \to \beta} M_{\beta \to \gamma} = \mathbb{I}_n$. Indeed, let $\beta = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ so that:

$$M_{\gamma \to \beta} M_{\beta \to \gamma} = [\text{id}_V]_{\gamma}^{\beta}[\text{id}_V]_{\beta}^{\gamma} = [\text{id}_V]_{\beta}^{\beta} = ([\mathbf{v}_1]_{\beta} \; [\mathbf{v}_2]_{\beta} \; ... \; [\mathbf{v}_n]_{\beta}) = \mathbb{I}_n \tag{37.5.22}$$

where we used Theorem 46.21. Similarly, one can also find that $M_{\beta \to \gamma} M_{\gamma \to \beta}$ ∎

> **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 $f : V \to W$ is a linear map,
> then:
> $$A' = M_{\gamma \to \gamma'} \, A \, M_{\beta' \to \beta} \tag{37.5.23}$$

Proof. Consider the commutative diagram below:

Then note that:

$$A = \varphi_{\gamma} \circ f \circ \phi_{\beta} \tag{37.5.24}$$
$$A' = \varphi_{\gamma'} \circ f \circ \phi_{\beta'} \tag{37.5.25}$$

hence:

$$A' = \varphi_{\gamma'} \circ f \circ \phi_{\beta'} \tag{37.5.26}$$
$$= \varphi_{\gamma'} \circ (\phi_{\gamma} \circ \varphi_{\gamma}) \circ f \circ (\phi_{\beta} \circ \varphi_{\beta}) \circ \phi_{\beta'} \tag{37.5.27}$$
$$= (\varphi_{\gamma'} \circ \phi_{\gamma}) \circ A \circ (\varphi_{\beta} \circ \phi_{\beta'}) \tag{37.5.28}$$
$$= M_{\gamma' \to \gamma}^{-1} \, A \, M_{\beta' \to \beta} \tag{37.5.29}$$

but we have proven that $M_{\gamma' \to \gamma}^{-1} = M_{\gamma \to \gamma'}$ so that:

$$A' = M_{\gamma \to \gamma'} \, A \, M_{\beta' \to \beta} \tag{37.5.30}$$

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 $M_{\beta'\to\beta}$ converts it to $[\mathbf{v}]_\beta$. Then A maps it to $A[\mathbf{v}]_\beta = [f(\mathbf{v})]_\gamma$ by Proposition 46.18. Finally $M_{\gamma\to\gamma'}$ 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:
> $$B = Q^{-1}AP \tag{37.5.31}$$

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:
> $$\begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \tag{37.5.32}$$
> where $r = \text{rk}(A)$.

Proof. We begin by proving that any linear map $f : V \to W$ has some set of bases β and γ of V and W respectively such that:
$$[f]_\beta^\gamma = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \tag{37.5.33}$$

Set r so that $\text{null}(f) = n-r$, so that $\ker f$ has basis $\mathbf{v}_{r+1}, ..., \mathbf{v}_n$ which we extend to $\mathbf{v}_1, ..., \mathbf{v}_r, \mathbf{v}_{r+1}, ..., \mathbf{v}_n$ a basis for V. We know that $f(\mathbf{v}_1), f(\mathbf{v}_2), ..., f(\mathbf{v}_n)$ is a basis for $\text{Im}(f)$, and can be extended to a basis γ for W. Consequently:
$$[f]_\beta^\gamma = ([f(\mathbf{v}_1)]_\gamma \ [f(\mathbf{v}_2)]_\gamma \ ... \ [f(\mathbf{v}_r)]_\gamma \ [f(\mathbf{v}_{r+1})]_\gamma \ ... \ f(\mathbf{v}_n)]_\gamma) \tag{37.5.34}$$

but $f(\mathbf{v}_i) = \mathbf{0}$ for $r < i \leq n$ so that:
$$[f]_\beta^\gamma = ([f(\mathbf{v}_1)]_\gamma \ [f(\mathbf{v}_2)]_\gamma \ ... \ [f(\mathbf{v}_r)]_\gamma \ \mathbf{0} \ ... \ \mathbf{0}) = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \tag{37.5.35}$$

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:
$$\begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \tag{37.5.36}$$

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 $A \in$

$\mathrm{Mat}_{m,n}(\mathbb{K})$ then there exist $\mathsf{Q}, \mathsf{P} \in \mathrm{Mat}_m(\mathbb{K})$ such that:

$$\mathsf{Q}^{-1}\mathsf{A}\mathsf{P} = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \tag{37.5.37}$$

implying that:

$$(\mathsf{Q}^{-1}\mathsf{A}\mathsf{P})^T == \mathsf{P}^T\mathsf{A}^t(\mathsf{Q}^{-1})^T = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \tag{37.5.38}$$

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.

L4 Solving linear equations

38.1   Structure of solutions

Let's consider a linear map $f : V \to W$. Suppose we wish to find the solutions to

$$f(\mathbf{x}) = \mathbf{b}, \; \mathbf{x} \in V, \mathbf{b} \in W \tag{38.1.1}$$

For $\mathbf{b} \neq 0$, this is known as the inhomogeneous linear equation, whose associated homogeneous equation is:

$$f(\mathbf{x}) = \mathbf{0} \tag{38.1.2}$$

Clearly, the solution of the latter is $\mathrm{Ker}(f)$.

> **Proposition** *(Structure of solutions)*
> Suppose $\mathbf{x}_0 \in V$ is a solution to the inhomogeneous equation (38.1.1). Then the general solution is given by:
> $$\mathbf{x} = \mathbf{x}_0 + \mathrm{Ker}(f) \tag{38.1.3}$$

Proof. Suppose $\mathbf{x} \in V$ is a general solution to (38.1.1) 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 $\mathbf{x} - \mathbf{x}_0 \in \mathrm{Ker}(f)$ or alternatively that $\mathbf{x} = \mathbf{x}_0 + \mathrm{Ker}(f)$. ∎

Suppose we have a linear transformation f represented by a matrix $\mathsf{A} : \mathbb{K}^n \to \mathbb{K}^m$. Then, for $\mathbf{x} \in \mathbb{K}^n$ and $\mathbf{b} \in \mathbb{K}^m$ we consider the linear system of equations $\mathsf{A}\mathbf{x} = \mathbf{b}$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m \end{cases} \tag{38.1.4}$$

We must now consider the following three scenarios:

(i) if $\mathrm{rk}(\mathsf{A}) = m$, then a solution exists for any choice of $\mathbf{b}$. Indeed, since $\mathrm{Im}(\mathsf{A}) = \mathbb{K}^m$ (since $\mathrm{Im}(\mathsf{A}) \subseteq \mathbb{K}^m$ 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 $\dim \mathrm{Ker}(\mathsf{A}) = n - m$. Geometrically, if we let $m = n = 3$ and $\mathbb{K} = \mathbb{R}$, then we see that if $\mathrm{rk}(\mathsf{A}) = 3$ then the linear map f maps the typical Euclidean space to itself.

(ii) if $\mathrm{rk}(A) < m$ and $\mathbf{b} \in \mathrm{Im}(A)$ 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 A. The number of free parameters will be $\dim \mathrm{Ker}(A) = n - \mathrm{rk}(A)$. 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 $\mathrm{Im}(A)$ is a plane the kernel will be a line, any vector on this line may be added, giving a free parameter.

(iii) if $\mathrm{rk}(A) < m$ but $\mathbf{b} \notin \mathrm{Im}(A)$ then solution doesn't exist. Indeed if the vector $\mathbf{b}$ doesn't lie in the space spanned by 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 $A\mathbf{x} = \mathbf{b}$ where $A : \mathbb{K}^n \to \mathbb{K}^m$ is a matrix and $\mathbf{x} \in \mathbb{K}^n$ and $\mathbf{b} \in \mathbb{K}^m$. We define its augmented matrix to be:

$$(A|\mathbf{b}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{array} \right) \tag{38.1.5}$$

Theorem *(Rank of augmented matrix)*
For a matrix $A : \mathbb{K}^n \to \mathbb{K}^m$ and $\mathbf{b} \in \mathbb{K}^n$:

$$\mathbf{b} \in \mathrm{Im}(A) \iff \mathrm{rk}(A) = \mathrm{rk}((A|\mathbf{b})) \tag{38.1.6}$$

Proof. $\implies$ Suppose that $\mathbf{b} \in \mathrm{Im}(A)$. Then adding this vector to A will not alter the space it spans, so that $\mathrm{rk}(A) = \mathrm{rk}((A|\mathbf{b}))$.

$\impliedby$ Suppose that $\mathrm{rk}(A) = \mathrm{rk}((A|\mathbf{b}))$. Then, this means that:

$$\mathrm{Span}(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^n, \mathbf{b}) = \mathrm{Span}(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^n) \tag{38.1.7}$$

showing that $\mathbf{b} \in \mathrm{Im}(A)$. $\blacksquare$

Therefore, if we can find the rank of the augmented matrix and show that it is equal to the rank of the coefficient matrix A, then we have shown that a solution must indeed exist. We can find the rank of matrices using elementary matrix operations.

38.2 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:
>
> $$\begin{pmatrix} \cdots & a_{ij_1} & \cdots & \cdots & \cdots & * \\ \vdots & & a_{2j_2} & & & \vdots \\ \vdots & & & & & \vdots \\ \vdots & & & & a_{rj_r} & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \end{pmatrix}$$
> (38.2.1)
>
> 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:

$$\begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 7 \\ 0 & 0 & 1 & -3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$
(38.2.2)

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:

$$\begin{pmatrix} 1 & 0 & 0 & 3 \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$
(38.2.3)

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 $(A|\mathbf{b})$ 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 $\text{rk}(A) \neq \text{rk}((A|\mathbf{b}))$, and hence $\mathbf{b} \notin \text{Im}(A)$. The system therefore has no solutions.
> (iii) Otherwise, convert to reduced row echelon form, and solve the resulting system of equations.

Example. Consider:

$$\begin{cases} 3x_1 - 11x_2 - 3x_3 = 3 \\ 2x_1 - 6x_2 - 2x_3 = 1 \\ 5x_1 - 17x_2 - 6x_3 = 2 \\ 4x_1 - 8_2 = 7 \end{cases} \tag{38.2.4}$$

We construct the augmented matrix:

$$(A|\mathbf{b}) = \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 2 & -6 & -2 & 1 \\ 5 & -17 & -6 & 2 \\ 4 & -8 & 0 & 7 \end{array} \right) \tag{38.2.5}$$

which we reduce to row echelon form:

$$(A|\mathbf{b}) = \left(\begin{array}{ccc|c} 3 & -11 & - & 3 \\ 2 & -6 & -2 & 1 \\ 5 & -17 & -6 & 2 \\ 4 & -8 & 0 & 7 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & \frac{4}{3} & -1 & -3 \\ 0 & \frac{20}{3} & 4 & 3 \end{array} \right) \tag{38.2.6}$$

$$\rightarrow \left(\begin{array}{ccc|c} 3 & -11 & -3 & 3 \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 4 & 8 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & 0 & -3 & -\frac{21}{4} \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right) \tag{38.2.7}$$

$$\tag{38.2.8}$$

So we see that $\mathrm{rk}(A) = \mathrm{rk}((A|\mathbf{b})) = 3$ implying that $\mathbf{b} \in \mathrm{Im}(A)$, and that a solution to the system exists.

We therefore convert the augmented matrix into reduced row echelon form:

$$(A|\mathbf{b}) \rightarrow \left(\begin{array}{ccc|c} 3 & 0 & -3 & -\frac{21}{4} \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & 0 & 0 & \frac{3}{4} \\ 0 & \frac{4}{3} & 0 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right) \tag{38.2.9}$$

$$\rightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{1}{4} \\ 0 & 1 & 0 & -\frac{3}{4} \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right) \tag{38.2.10}$$

which gives the solution:

$$x_1 = \frac{1}{4}, \ x_2 = -\frac{3}{4}, \ x_3 = 2 \tag{38.2.11}$$

◄

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

$$
\begin{pmatrix}
3 & -11 & -3 & 3 & -8 \\
0 & \frac{4}{3} & 0 & -1 & \frac{1}{3} \\
0 & \frac{4}{3} & -1 & -3 & -\frac{8}{3} \\
0 & \frac{20}{3} & 4 & 3 & \frac{41}{3}
\end{pmatrix}
\tag{38.2.12}
$$

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:

$$
\begin{pmatrix}
3 & -11 & -3 & 3 & -8 \\
0 & \frac{4}{3} & 0 & -1 & \frac{1}{3} \\
0 & 0 & -1 & -2 & -3 \\
0 & 0 & 4 & 8 & 12
\end{pmatrix}
\tag{38.2.13}
$$

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.

38.3 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 $\mathsf{A}' = \mathsf{E}\mathsf{A}$. For example, if E represents the exchange of rows i and j, then:

$$
\mathbb{1} = \begin{pmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_j & \dots & \mathbf{e}_i & \dots & \mathbf{e}_n \end{pmatrix}^T \implies \mathsf{E} = \begin{pmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_j & \dots & \mathbf{e}_i & \dots & \mathbf{e}_n \end{pmatrix}^T
\tag{38.3.1}
$$

so that:

$$
\mathsf{E}_1 \mathsf{A} = \begin{pmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_j & \dots & \mathbf{e}_i & \dots & \mathbf{e}_n \end{pmatrix}^T \begin{pmatrix} \mathbf{A}_1 & \dots & \mathbf{A}_i & \dots & \mathbf{A}_j & \dots & \mathbf{A}_n \end{pmatrix}
\tag{38.3.2}
$$

$$
= \begin{pmatrix}
\mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \dots & \mathbf{e}_1 \cdot \mathbf{A}_n \\
\vdots & \vdots & & \vdots \\
\mathbf{e}_j \cdot \mathbf{A}_1 & \mathbf{e}_j \cdot \mathbf{A}_2 & \dots & \mathbf{e}_j \cdot \mathbf{A}_n \\
\vdots & \vdots & & \vdots \\
\mathbf{e}_i \cdot \mathbf{A}_1 & \mathbf{e}_i \cdot \mathbf{A}_2 & \dots & \mathbf{e}_i \cdot \mathbf{A}_n \\
\vdots & \vdots & & \vdots \\
\mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \dots & \mathbf{e}_n \cdot \mathbf{A}_n
\end{pmatrix}
\tag{38.3.3}
$$

which is indeed the version of A with the ith and jth rows exchanged:

$$A = \mathbb{1}A = (\mathbf{e}_1 \ \cdots \ \mathbf{e}_j \ \cdots \ \mathbf{e}_i \ \cdots \ \mathbf{e}_n)^T (\mathbf{A}_1 \ \cdots \ \mathbf{A}_i \ \cdots \ \mathbf{A}_j \ \cdots \ \mathbf{A}_n) \tag{38.3.4}$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_1 \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_i \cdot \mathbf{A}_1 & \mathbf{e}_i \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_j \cdot \mathbf{A}_1 & \mathbf{e}_j \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_j \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_n \cdot \mathbf{A}_n \end{pmatrix} \tag{38.3.5}$$

Similarly, if E_2 represents multiplication of the ith row by a scalar λ then clearly:

$$E_2 A = (\mathbf{e}_1 \ \cdots \ \lambda\mathbf{e}_i \ \cdots \ \cdots \ \mathbf{e}_n)^T (\mathbf{A}_1 \ \cdots \ \mathbf{A}_i \ \cdots \ \cdots \ \mathbf{A}_n) \tag{38.3.6}$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_1 \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \lambda\mathbf{e}_i \cdot \mathbf{A}_1 & \lambda\mathbf{e}_i \cdot \mathbf{A}_2 & \cdots & \lambda\mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_n \cdot \mathbf{A}_n \end{pmatrix} \tag{38.3.7}$$

which is the version of A with the ith row multiplied by λ.

Finally, suppose that E_3 represents adding a λ-multiple of the jth row to the ith row. Then we find that:

$$E_1 A = (\mathbf{e}_1 \ \cdots \ \mathbf{e}_i \ \cdots \ \mathbf{e}_j + \lambda\mathbf{e}_i \ \cdots \ \mathbf{e}_n)^T (\mathbf{A}_1 \ \cdots \ \mathbf{A}_i \ \cdots \ \mathbf{A}_j \ \cdots \ \mathbf{A}_n) \tag{38.3.8}$$

$$= \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{A}_1 & \mathbf{e}_1 \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_1 \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_i \cdot \mathbf{A}_1 & \mathbf{e}_i \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_j \cdot \mathbf{A}_1 + \lambda\mathbf{e}_i \cdot \mathbf{A}_1 & \mathbf{e}_j \cdot \mathbf{A}_2 + \lambda\mathbf{e}_i \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_j \cdot \mathbf{A}_n + \lambda\mathbf{e}_i \cdot \mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_n \cdot \mathbf{A}_1 & \mathbf{e}_n \cdot \mathbf{A}_2 & \cdots & \mathbf{e}_n \cdot \mathbf{A}_n \end{pmatrix} \tag{38.3.9}$$

which is indeed the version of A with the λ-multiple of the ith row added to the jth 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 $\mathrm{rk}(A) = n$.
>   (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:

$$U = E_k E_{k-1} ... E_1 A = BA \tag{38.3.10}$$

where $E_k, E_{k-1}, ..., E_1$ are elementary matrices. Since they are all invertible, we have that B^{-1} exists.

$\implies$ 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.

$\impliedby$ Suppose $\mathsf{U} = \mathbb{1}$, then:

$$\mathsf{BA} = \mathbb{1} \implies \mathsf{A} = \mathsf{B}^{-1}\mathbb{1} \implies \mathsf{AB} = \mathbb{1} \tag{38.3.11}$$

Note however that $\mathbb{1}$ and B^{-1} are invertible, so A will also be invertible, with $\mathsf{A}^{-1} = \mathsf{B}$.

Therefore, we find that:

$$\mathsf{A}^{-1} = \mathsf{B} = \mathsf{E}_k\mathsf{E}_{k-1}...\mathsf{E}_1\mathbb{1} \implies (\mathsf{A}|\mathbb{1}) \to (\mathbb{1}|\mathsf{A}^{-1}) \tag{38.3.12}$$

so we find A^{-1} by applying the same row operations that row reduce A to $\mathbb{1}$. ∎

We can use the invertibility theorem to find the inverse of matrices.

Example. Let's find the inverse of the following matrix:

$$\mathsf{A} = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 6 & 3 \\ 2 & 3 & 0 \end{pmatrix} \tag{38.3.13}$$

We find that:

$$\left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 1 & 6 & 3 & 0 & 1 & 0 \\ 2 & 3 & 0 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & -1 & 1 & 0 \\ 0 & -5 & -2 & -2 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & -1 & 1 & 0 \\ 0 & 0 & 3 & -\frac{9}{2} & \frac{5}{2} & 1 \end{array}\right) \tag{38.3.14}$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 2 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 4 & 0 & \frac{5}{2} & -\frac{5}{6} & -\frac{1}{3} \\ 0 & 2 & 0 & 2 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array}\right) \tag{38.3.15}$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{3}{2} & \frac{1}{2} & 1 \\ 0 & 2 & 0 & 2 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{3}{2} & \frac{1}{2} & 1 \\ 0 & 1 & 0 & 1 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & 1 & -\frac{3}{2} & \frac{5}{6} & \frac{1}{3} \end{array}\right) \tag{38.3.16}$$

so we see that:

$$\mathsf{A}^{-1} = \frac{1}{6} \begin{pmatrix} -9 & 3 & 6 \\ 6 & -2 & -2 \\ -9 & 5 & 2 \end{pmatrix} \tag{38.3.17}$$

To check:

$$\frac{1}{6} \begin{pmatrix} -9 & 3 & 6 \\ 6 & -2 & -2 \\ -9 & 5 & 2 \end{pmatrix} \begin{pmatrix} 1 & 4 & 1 \\ 1 & 6 & 3 \\ 2 & 3 & 0 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix} = \mathbb{1} \tag{38.3.18}$$

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 $A\mathbf{x} = \mathbf{b}$ has a unique solution for any $\mathbf{b}$
> (c) The system $A\mathbf{x} = \mathbf{0}$ only has a trivial solution.

Proof.

(a) $\implies$ (b) Let A be invertible. Suppose $A\mathbf{x} = \mathbf{b}$, then multiplying by A^{-1} then $A^{-1}A\mathbf{x} = A^{-1}\mathbf{b} \implies \mathbf{x} = A^{-1}\mathbf{b}$. Instead, if $\mathbf{x} = A^{-1}\mathbf{b}$ then multiplying by A we get $A\mathbf{x} = \mathbf{b}$.

(b) $\implies$ (c) Suppose $A\mathbf{x} = \mathbf{b}$ has a unique solution for any $\mathbf{b}$. Then $A\mathbf{x} = \mathbf{0}$ also has a unique solution, which can only be the trivial solution

(c) $\implies$ (a) Suppose that $A\mathbf{x} = \mathbf{0}$ only has a trivial solution. Then this means that when we reduce the augmented matrix:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ a_{31} & a_{32} & a_{33} & 0 \end{pmatrix} \tag{38.3.19}$$

then we must get that:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{38.3.20}$$

since there can only be trivial solutions. So A has a reduced row echelon form of $\mathbb{1}$, proving by the Invertibility theorem that it is invertible.

Alternatively, we could note that if $\dim \mathrm{Ker}(A) = 0$ then $\mathrm{rk}(A) = n$. Therefore A is invertible.

■

> **Example.** Let's consider the system:
>
> $$\begin{cases} x + 4y + z = 4 \\ x + 6y + 3z = 6 \\ 2x + 3y = 9 \end{cases} \tag{38.3.21}$$
>
> which may be written in matrix form as:
>
> $$A\mathbf{x} = \mathbf{b}, \ A = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 6 & 3 \\ 2 & 3 & 0 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 4 \\ 6 \\ 9 \end{pmatrix} \tag{38.3.22}$$

We have already found A^{-1}, so we find that:

$$\mathbf{x} = \mathsf{A}^{-1}\mathbf{b} = \frac{1}{6}\begin{pmatrix} -9 & 3 & 6 \\ 6 & -2 & -2 \\ -9 & 5 & 2 \end{pmatrix}\begin{pmatrix} 4 \\ 6 \\ 9 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 36 \\ -6 \\ 12 \end{pmatrix} = \begin{pmatrix} 6 \\ -1 \\ 2 \end{pmatrix} \tag{38.3.23}$$

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) $(\mathsf{A}^T)^{-1} = (\mathsf{A}^{-1})^T$
(b) $(\mathsf{A}^{-1}))^{-1} = \mathsf{A}$
(c) AB is invertible and $(\mathsf{AB})^{-1} = \mathsf{B}^{-1}\mathsf{A}^{-1}$

Proof. (a) We find that:
$$(\mathsf{A}^T)(\mathsf{A}^{-1})^T = (\mathsf{A}^{-1}\mathsf{A})^T = \mathbb{1}^T = \mathbb{1} \tag{38.3.24}$$

and similarly:
$$(\mathsf{A}^{-1})^T(\mathsf{A}^T) = (\mathsf{AA}^{-1})^T = \mathbb{1}^T = \mathbb{1} \tag{38.3.25}$$

(b) We find that:
$$(\mathsf{A}^{-1})(\mathsf{A}) = \mathbb{1} = (\mathsf{A})(\mathsf{A}^{-1}) \tag{38.3.26}$$

(c) We find that:
$$\mathsf{ABB}^{-1}\mathsf{A}^{-1} = \mathsf{AA}^{-1} = \mathbb{1} \tag{38.3.27}$$

and
$$\mathsf{B}^{-1}\mathsf{A}^{-1}\mathsf{AB} = \mathsf{B}^{-1}\mathsf{B} = \mathbb{1} \tag{38.3.28}$$

as desired.

■

L5 Determinants

39.1 The determinant of a matrix

> **Definition** *(Determinant)*
> The determinant $\det \; : \; \mathbb{K}^n \; \to \; \mathbb{K}$ maps n vectors $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N \; \in \; \mathbb{K}^n$ to a scalar $\det(\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N) \in \mathbb{K}$. It satisfies the following properties:
> (a) $\det(..., \alpha\mathbf{a} + \beta\mathbf{b}, ...) = \alpha \det(..., \mathbf{a}, ...) + \beta \det(..., \mathbf{b}, ...)$.
> (b) $\det(..., \mathbf{a}, ..., \mathbf{b}, ...) = -\det(\mathbf{b}, ..., \mathbf{a}, ...)$
> (c) $\det(\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n) = 1$
> The determinant of a matrix A with column vectors $\mathbf{A}^i$, $1 \leq i \leq n$ is the determinant of these column vectors:
> $$\det \mathsf{A} \equiv \det(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^n) \tag{39.1.1}$$

Note that:
$$\det(..., \mathbf{0}, ...) = \det(..., \mathbf{v} - \mathbf{v}, ...) = \det(..., \mathbf{v}, ...) - \det(..., \mathbf{v}, ...) = 0 \tag{39.1.2}$$

and:
$$\det(..., \mathbf{a}, ..., \mathbf{a}, ...) = -\det(..., \mathbf{a}, ..., \mathbf{a}, ...) \implies \det(..., \mathbf{a}, ..., \mathbf{a}, ...) = 0 \tag{39.1.3}$$

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:
$$\det(..., \mathbf{a}, ..., \alpha\mathbf{a}, ...) = \alpha \det(..., \mathbf{a}, ..., \mathbf{a}, ...) = 0 \tag{39.1.4}$$

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 A is given by the product of its diagonal elements.
> The matrix of an upper or lower triangular matrix A is also given by the product of its diagonal elements.

Proof. Consider a matrix $A = \text{diag}(a_{11}, a_{22}, ..., a_{nn})$. Then:

$$\det A = \det(a_{11}\mathbf{e}_1, a_{22}\mathbf{e}_2, ..., a_{nn}\mathbf{e}_n) = a_{11}a_{22}...a_{nn} \tag{39.1.5}$$

Instead, for an upper triangular matrix:

$$\det A = \det\left(a_{11}\mathbf{e}_1, a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2, ..., \sum_i a_{in}\mathbf{e}_i\right) \tag{39.1.6}$$

$$= \det\left(a_{11}\mathbf{e}_1, a_{12}\mathbf{e}_1, ..., \sum_i a_{in}\mathbf{e}_i\right) + \det\left(a_{11}\mathbf{e}_1, a_{22}\mathbf{e}_2, ..., \sum_i a_{in}\mathbf{e}_i\right) \tag{39.1.7}$$

$$= \det(a_{11}\mathbf{e}_1, a_{22}\mathbf{e}_2, ..., a_{nn}\mathbf{e}_n) \tag{39.1.8}$$

$$= a_{11}a_{22}...a_{nn} \tag{39.1.9}$$

as desired. ∎

> **Theorem** *(Determinant of matrix)*
> For a given $n \times n$ matrix A, we have that:
>
> $$\det A = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^{n} a_{\sigma(i_k)k} \tag{39.1.10}$$

Proof. We start by writing:

$$\det A = \det\left(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^n\right) \tag{39.1.11}$$

$$= \det\left(\sum_{i_1} a_{i_1 1}\mathbf{e}_{i_1}, \sum_{i_2} a_{i_2 2}\mathbf{e}_{i_2}, ..., \sum_k a_{i_n n}\mathbf{e}_{i_n}\right) \tag{39.1.12}$$

$$= \sum_{i_1} a_{i_1 1} \det\left(\mathbf{e}_{i_1}, \sum_{i_2} a_{i_2 2}\mathbf{e}_{i_2}, ..., \sum_{i_n} a_{i_n n}\mathbf{e}_{i_n}\right) \tag{39.1.13}$$

$$= \sum_{i_1 i_2 ... i_n} a_{i_1 1}a_{i_2 2}...a_{i_n n} \det(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, ..., \mathbf{e}_{i_n}) \tag{39.1.14}$$

Now note that the only terms that survive out of this sum are $i_1 \neq i_2 \neq i_3 \neq ... \neq i_n$. In other

words, the only terms surviving are all the permutations of $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n$, so we write:

$$\det A = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{\sigma(i_1)1} a_{\sigma(i_2)2} ... a_{\sigma(i_n)n} \det(\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n) \tag{39.1.15}$$

$$= \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^{n} a_{\sigma(i_k)k} \tag{39.1.16}$$

as desired. ∎

> **Proposition** *(Determinant properties)*
> Let's consider two $n \times n$ matrices A, B. Then:
> (a) $\det(A^T) = \det A$
> (b) $\det(AB) = \det A \cdot \det B$
> (c) A is bijective $\iff \det A \neq 0$ and $\det A^{-1} = \frac{1}{\det A}$

Proof. (a) Let a'_{ij} be the matrix elements of A^T so that $a'_{ij} = a_{ji}$. Then:

$$\det A^T = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^{n} a'_{\sigma(i_k)k} = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{k=1}^{n} a_{k\sigma(i_k)} \tag{39.1.17}$$

$$= \sum_{\sigma \in S_n} \text{sgn}(\sigma^{-1}) \prod_{k=1}^{n} a_{\sigma^{-1}(i_k)k} = \sum_{\rho \in S_n} \text{sgn}(\rho^{-1}) \prod_{k=1}^{n} a_{\rho^{-1}(i_k)k} = \det A \tag{39.1.18}$$

(b) Recall that $(AB)_{ik} = \sum_j a_{ij} b_{jk}$ implying that:

$$(AB)^k = \sum_{ij} a_{ij} b_{jk} \mathbf{e}_i = \sum_j b_{jk} \mathbf{A}^j \tag{39.1.19}$$

Hence, we get that:

$$\det AB = \det\left(\sum_{j_1} b_{j_1 1} \mathbf{A}^{j_1}, \sum_{j_2} b_{j_2 2} \mathbf{A}^{j_2}, ..., \sum_{j_n} b_{j_n n} \mathbf{A}^{j_n} \right) \tag{39.1.20}$$

$$= \sum_{j_1, ... j_n} b_{j_1 1} b_{j_2 2} ... b_{j_n n} \det\left(\mathbf{A}^{j_1}, \mathbf{A}^{j_2}, ..., \mathbf{A}^{j_n} \right) \tag{39.1.21}$$

Again, we see that the only terms that survive are those where $j_1 \neq j_2 \neq ... \neq j_n$, so we will get:

$$\det AB = \sum_{\sigma \in S_n} \text{sgn}(\sigma) b_{\sigma(j_1)1} b_{\sigma(j_2)2} ... b_{\sigma(j_n)n} \det\left(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^n \right) \tag{39.1.22}$$

$$= \det A \cdot \det B \tag{39.1.23}$$

as desired.

(c) ($\implies$) Suppose that A is bijective, and thus invertible. Then:

$$\det\left(AA^{-1} \right) = \det A \cdot \det A^{-1} = \det \mathbb{1} = 1 \tag{39.1.24}$$

implying that $\det A \neq 0$.

($\Longleftarrow$) Suppose that A is not bijective, so that $\mathrm{rk}(A) < n$. 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:

$$\mathbf{A}^i = \sum_j \alpha_j \mathbf{A}^j \tag{39.1.25}$$

It then follows that

$$\det A = \det\left(\mathbf{A}^1, ..., \mathbf{A}^i, ..., \mathbf{A}^n\right) = \det\left(\mathbf{A}^1, ..., \mathbf{A}^i, ..., \mathbf{A}^n\right) \tag{39.1.26}$$

$$= \det\left(\mathbf{A}^1, ..., \sum_j \alpha_j \mathbf{A}^j, ..., \mathbf{A}^n\right) \tag{39.1.27}$$

$$= \sum_j \alpha_j \det\left(\mathbf{A}^1, ..., \mathbf{A}^j, ..., \mathbf{A}^j, ..., \mathbf{A}^n\right) \tag{39.1.28}$$

$$= 0 \tag{39.1.29}$$

as desired.

Using (39.1.24):

$$\det A^{-1} = \frac{1}{\det A} \tag{39.1.30}$$

as desired.

∎

39.2 Laplace expansion

Definition *(Cofactor matrix)*
Consider an $n \times n$ matrix A, then we define the associated (i,j) matrix $A(i,j)$ as the matrix A with the ith row and jth column substituted with $\mathbf{e}_i$ and $\mathbf{e}_j^T$ respectively:

$$A(i,j) = \begin{pmatrix} & & \overset{j\text{th col}}{0} & & \\ & A & \vdots & A & \\ 0 \ldots 0 & & 1 & & 0 \ldots 0 \\ & A & \vdots & A & \\ & & 0 & & \end{pmatrix} \leftarrow i\text{th row} \tag{39.2.1}$$

The (i,j) cofactor coefficient is then defined as the determinant of $A(i,j)$

$$C_{ij} = \det(A(i,j)) \tag{39.2.2}$$

The cofactor matrix is the matrix C whose elements are C_{ij}. The cofactor expansion in the ith row is defined as:

$$\mathrm{cof}_i A = \sum_k a_{ik} C_{ik} = a_{i1} C_{i1} + a_{i2} C_{i2} + ... + a_{in} C_{in} \tag{39.2.3}$$

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 A_{ij} more easily.

> **Proposition** *(Cofactor matrix calculation)*
> Let A be a $n \times n$ matrix and let $\tilde{A}_{ij}$ be the matrix A with the ith row and jth column removed. Then:
> $$C_{ij} = (-1)^{i+j} \det\left(\tilde{A}(i,j)\right) \tag{39.2.4}$$

Proof. Note that the determinant only acquires a sign change when moving columns, and the same goes for rows since $\det\left(A^T\right) = \det A$. Hence, we may move the ith row and jth 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 $\tilde{A}_{ij}$ to be the matrix with the ith and jth rows removed, then if:

$$B(i,j) \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{A}(i,j) & \\ 0 & & & \end{pmatrix} \tag{39.2.5}$$

$$\implies \det(B(i,j)) = (-1)^{i+j+2} \det(A(i,j)) = \det\left(\tilde{A}(i,j)\right) \tag{39.2.6}$$

so that:

$$\det(A(i,j)) = (-1)^{i+j} \det\left(\tilde{A}(i,j)\right) \tag{39.2.7}$$

∎

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 A with cofactor matrix C:
> $$(\det A)\mathbb{1} = C^T A \tag{39.2.8}$$
>
> so that:
> $$\det A = \sum_k (-1)^{k+i} A_{kj} \det\left(\tilde{A}(k,i)\right) \tag{39.2.9}$$

Proof. We find that:

$$(C^T A)_{ij} = \sum_k C_{ki} A_{kj} \tag{39.2.10}$$

$$= \sum_k \det(A(k,i)) A_{kj} \tag{39.2.11}$$

$$= \sum_k A_{kj} \det\left(\mathbf{A}^1 - A_{k1}\mathbf{e}_k, \mathbf{A}^2 - A_{k2}\mathbf{e}_k, ..., \mathbf{e}_k, ..., \mathbf{A}^n - A_{kn}\mathbf{e}_n\right) \tag{39.2.12}$$

---

[1] we can't just exchange the ith row with the first row, since this would alter the order of the rows, and would not give the matrix A with the ith row removed.

Now note that

$$\det\left(\mathbf{A}^1 - A_{k1}\mathbf{e}_k, ..., \mathbf{e}_k, ...\right) = \det\left(\mathbf{A}^1, ..., \mathbf{e}_k, ...\right) - A_{k1}\det(\mathbf{e}_k, ..., \mathbf{e}_k, ...) \tag{39.2.13}$$
$$= \det\left(\mathbf{A}^1, ..., \mathbf{e}_k, ...\right) \tag{39.2.14}$$

Repeating this process we find that:

$$(\mathsf{C}^T\mathsf{A})_{ij} = \sum_k A_{kj}\det\left(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{e}_k, ..., \mathbf{A}^n\right) \tag{39.2.15}$$

$$= \sum_k \det\left(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^j, ..., \mathbf{A}^n\right) \tag{39.2.16}$$

$$= \det\left(\mathbf{A}^1, \mathbf{A}^2, ..., \mathbf{A}^i, ..., \mathbf{A}^n\right)\delta_{ij} \tag{39.2.17}$$

$$= \det\mathsf{A}\,\delta_{ij} \tag{39.2.18}$$

implying that:

$$\mathsf{C}^T\mathsf{A} = \det\mathsf{A}\,\mathbb{1} \tag{39.2.19}$$

as desired. ∎

> **Example.** Consider a 3×3 matrix:
>
> $$\mathsf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{39.2.20}$$
>
> We calculate the cofactor expansion in the first column: Then we find that:
>
> $$C_{11} = a_{22}a_{33} - a_{23}a_{32}, \; C_{21} = a_{13}a_{32} - a_{12}a_{23}, \; C_{31} = a_{12}a_{23}a_{22}a_{13} \tag{39.2.21}$$
>
> so:
>
> $$\det\mathsf{A} = C_{11}a_{11} + C_{21}a_{21} + C_{31}a_{31} \tag{39.2.22}$$
> $$= a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{21}(a_{13}a_{32} - a_{12}a_{23}) + a_{31}(a_{12}a_{23}a_{22}a_{13}) \tag{39.2.23}$$
>
> 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 jth row $\mathbf{A}^j$ of a matrix A is orthogonal to the ith row of its cofactor matrix C, for $i \neq j$. So:
> $$\mathbf{A}^j \cdot \mathbf{C}_i = \sum_k a_{jk}C_{ik} = 0 \tag{39.2.24}$$

Proof. For $i \neq j$, we have that

$$\mathbf{A}^j \cdot \mathbf{C}_i = \sum_k a_{jk}C_{ik} = a_{j1}C_{i1} + a_{j2}C_{i2} + ... + a_{jn}C_{in} \tag{39.2.25}$$

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:

$$a_{j1}C_{i1} + a_{j2}C_{i2} + ... + a_{jn}C_{in} = b_{i1}C_{i1} + b_{i2}C_{i2} + ... + b_{in}C_{in} \tag{39.2.26}$$

Firstly, since C_{ij} takes all the entries of A without the ith row and jth column, it follows that for these cofactors to coincide with those of B, all the elements of B except for the ith row are identical to those of A.

The only change is that for $a_{jk} = b_{ik}$, $k = 1, 2, ..., n$, so the ith row of B is the jth row of A. However, note that the jth row of B must also be the jth row of A as we argued in the previous paragraph, so B has two repeated rows. Therefore, its determinant/cofactor expansion must vanish:

$$a_{j1}C_{i1} + a_{j2}C_{i2} + ... + a_{jn}C_{in} = b_{i1}C_{i1} + b_{i2}C_{i2} + ... + b_{in}C_{in} = 0 \tag{39.2.27}$$

∎

Definition *(Adjoint matrix)*

The adjoint adj A of a matrix A is the transpose of its cofactor matrix C:

$$\text{adj } A = C^T \implies (\det A)\mathbb{1} = (\text{adj } A)A \tag{39.2.28}$$

Theorem *(Inverse of matrix)*

Let A be an invertible $n \times n$ matrix, then:

$$A^{-1} = \frac{1}{\det A}\text{adj } A \tag{39.2.29}$$

Proof. We find that:

$$(\det A)\mathbb{1} = C^T A \implies (\det A)A^{-1} = \text{adj } A \implies A^{-1} = \frac{1}{\det A}\text{adj } A \tag{39.2.30}$$

as desired. ∎

Example. Consider the 2×2 matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \ ad - bc \neq 0 \tag{39.2.31}$$

We find that:

$$\det A = ad - bc \tag{39.2.32}$$

Furthermore:

$$C_{11} = d, \ C_{12} = -c, \ C_{21} = -b, \ C_{22} = d \tag{39.2.33}$$

giving the cofactor matrix:

$$C = \begin{pmatrix} d & -c \\ -b & d \end{pmatrix} \tag{39.2.34}$$

Consequently

$$\text{adj A} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \implies \text{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{39.2.35}$$

◀

39.3 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 $\text{A}\mathbf{x} = \mathbf{b}$. Let B_i be the matrix A with the ith column replaced with $\mathbf{b}$. Then the solution to the system is given by:

$$x_i \det \text{A} = \det \text{B}_i \tag{39.3.1}$$

Proof. We find that:

$$\det \text{B}_i = \det\left(\mathbf{A}^1, ..., \mathbf{b}, ..., \mathbf{A}^n\right) \tag{39.3.2}$$

$$= \sum_i b_i \det\left(\mathbf{A}^1, ..., \mathbf{e}_i, ..., \mathbf{A}^n\right) \tag{39.3.3}$$

$$= \sum_i A_{ij} x_j \det\left(\mathbf{A}^1, ..., \mathbf{e}_i, ..., \mathbf{A}^n\right) \tag{39.3.4}$$

$$= \sum_i x_j \det\left(\mathbf{A}^1, ..., \mathbf{A}^j, ..., \mathbf{A}^n\right) \tag{39.3.5}$$

$$= x_i \det\left(\mathbf{A}^1, ..., \mathbf{A}^i, ..., \mathbf{A}^n\right) = x_i \det \text{A} \tag{39.3.6}$$

as desired. Hence, if A is invertible then the solutions are given by:

$$x_i = \frac{\det \text{B}_i}{\det \text{A}} \tag{39.3.7}$$

■

Example. Consider the following system:

$$\begin{cases} x + 2y + 3z = 0 \\ 2x + 3y + 4z = 1 \\ 3x + 4y + 6z = 2 \end{cases} \tag{39.3.8}$$

Then we see that:

$$\text{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 6 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \tag{39.3.9}$$

so that:

$$B_1 = \begin{pmatrix} 0 & 2 & 3 \\ 1 & 3 & 4 \\ 2 & 4 & 6 \end{pmatrix}, \ B_2 = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 1 & 4 \\ 3 & 2 & 6 \end{pmatrix}, \ B_3 = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 1 \\ 3 & 4 & 2 \end{pmatrix} \tag{39.3.10}$$

Now we evaluate the determinants using the Laplace expansion:

$$\det A = (18 - 16) - 2(12 - 12) + 3(8 - 9) = -1 \tag{39.3.11}$$

$$\det B_1 = -2(6 - 8) + 3(4 - 6) = -2 \tag{39.3.12}$$

$$\det B_2 = (6 - 8) + 3(4 - 3) = 1 \tag{39.3.13}$$

$$\det B_3 = (6 - 4) - 2(4 - 3) = 0 \tag{39.3.14}$$

so that:

$$x = 2, \ y = -1, \ z = 0 \tag{39.3.15}$$

◀

L6 Inner product spaces

40.1   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:
>
> $$\langle \, , \, \rangle : V \times V \to \mathbb{K} \tag{40.1.1}$$
>
> 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:
> $$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \boldsymbol{\eta} \mathbf{v} \tag{40.1.2}$$
>   where $\boldsymbol{\eta} = \mathrm{diag}(1, -1, -1, -1)$.
>   ii. Functional products: for $f, g \in C^0([a, b])$, we define their inner product as:
>
> $$\langle f, g \rangle = \int_a^b f(x)^* g(x) dx \tag{40.1.3}$$
>
> ◀

> **Proposition** *(Inner product is well-defined)*
>
> Let $f, g \in \mathrm{End}(V)$ be two endomorphisms on V, then:
>
> $$\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle \mathbf{v}, g(\mathbf{w}) \rangle , \forall \mathbf{v}, \mathbf{w} \in V \implies f = g \tag{40.1.4}$$

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):

$$\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle \mathbf{v}, g(\mathbf{w}) \rangle \iff \langle \mathbf{v}, f(\mathbf{w}) - g(\mathbf{w}) \rangle = 0 \tag{40.1.5}$$

$$\iff \sum_{ij} v_i^* w_j \langle \mathbf{e}_i, f(\mathbf{e}_j) - g(\mathbf{e}_j) \rangle = 0 \tag{40.1.6}$$

Since this applies for any $\mathbf{v}, \mathbf{w}$, we must therefore have that:

$$\langle \mathbf{e}_i, h(\mathbf{e}_j) \rangle \equiv \langle \mathbf{e}_i, f(\mathbf{e}_j) - g(\mathbf{e}_j) \rangle = 0, \ \forall i, j = 1, 2, \dots \tag{40.1.7}$$

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:
> $$W^\perp = \{\mathbf{v} \in V : \langle \mathbf{v}, \mathbf{w} \rangle = 0, \ \forall \mathbf{w} \in W\} \tag{40.1.8}$$

40.2 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:

$$\mathbf{v} = \underbrace{\langle \mathbf{w}, \mathbf{v} \rangle \, \mathbf{w}}_{\in W} + \underbrace{(\mathbf{v} - \langle \mathbf{w}, \mathbf{v} \rangle \, \mathbf{w})}_{\in W^\perp} \tag{40.2.1}$$

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:
>
> $$\Pi : V \to W \tag{40.2.2}$$
> $$\mathbf{u} + \mathbf{w} \to \mathbf{w} \tag{40.2.3}$$
>
> 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 idompotent maps are projections.

Proof. Let Π be an idempotent map so that $\Pi^2 = \Pi$. Let $\mathbf{v}$, then clearly we have that:

$$\mathbf{v} = \Pi(\mathbf{v}) + (\mathbf{v} - \Pi(\mathbf{v})) = \Pi(\mathbf{v}) + (\mathbb{1} - \Pi)(\mathbf{v}) \tag{40.2.4}$$

Now let us define $W = \{\Pi(\mathbf{v}) \forall \mathbf{v} \in V\} = \mathrm{Im}(\Pi)$ and $U = \{\mathbf{v} - \Pi(\mathbf{v}) : \forall \mathbf{v} \in V\} = \mathrm{Im}(\mathbb{1} - \Pi)$. 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:

$$\Pi(\mathbf{v}) = \Pi(\mathbf{u} + \mathbf{w}) = \Pi(\mathbf{u}) + \Pi(\mathbf{w}) = \mathbf{u} \tag{40.2.5}$$

proving that Π is indeed a projector. $\blacksquare$

40.3 Inner products and matrices

> **Definition** *(Adjoint, Hermitian and Unitary linear map)*
> For a linear map $f \in \mathrm{End}(V)$ on V defined over $\mathbb{K}$, its adjoint map $f^\dagger \in \mathrm{End}(V)$ is define so that it satisfies:
> $$\langle \mathbf{v}, f(\mathbf{w}) \rangle = \langle f^\dagger(\mathbf{v}), \mathbf{w} \rangle, \ \forall \mathbf{v}, \mathbf{w} \in V \tag{40.3.1}$$
> 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 $H^{-1} = H^\dagger$.

> **Proposition** *(Adjoint map properties)*
> For a linear map $f \in \mathrm{End}(V)$, 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) $f \circ f^{-1} = \mathrm{id}_V \implies (f^{-1})^\dagger \circ f^\dagger = \mathrm{id}_V \implies (f^{-1})^\dagger = (f^\dagger)^{-1}$ 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 $f \in \mathrm{End}(V)$, we have that:

$$f(\mathbf{e}_i) = \sum_j A_{ij}\mathbf{e}_j \implies A_{ij} = \langle \mathbf{e}_j, f(\mathbf{e}_i) \rangle \tag{40.3.2}$$

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:

$$B_{ij} = \langle \mathbf{e}_i, f^\dagger(\mathbf{e}_j) \rangle = \langle f(\mathbf{e}_i), \mathbf{e}_j \rangle = \langle \mathbf{e}_j, f(\mathbf{e}_i) \rangle^* = A_{ji}^* \tag{40.3.3}$$

Hence the matrices $\mathsf{A}, \mathsf{A}^\dagger$ representing $f, f^\dagger$ respectively satisfy $\mathsf{A}^\dagger = (\mathsf{A}^*)^T$.

It follows that Hermitian maps have matrix representations $\mathsf{A} = \mathsf{A}^T = (\mathsf{A}^*)^T$ and Unitary maps have matrix representations $\mathsf{A}^{-1} = \mathsf{A}^T = (\mathsf{A}^*)^T$.

> **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 $\langle U(\mathbf{v}), U(\mathbf{w}) \rangle = \langle \mathbf{v}, (U^\dagger \circ U)(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle$. Hence by the well-definedness of inner products, $U^\dagger \circ U = \text{id}_V \iff U^{-1} = U^\dagger$. ∎

40.4   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 $T : V \times V \to \mathbb{R}$ linear in both of its terms:
>
> $$T(\alpha\mathbf{v} + \beta\mathbf{w}, \mathbf{u}) = \alpha T(\mathbf{v}, \mathbf{u}) + \beta T(\mathbf{w}, \mathbf{u}) \tag{40.4.1}$$
> $$T(\mathbf{u}, \alpha\mathbf{v} + \beta\mathbf{w}) = \alpha T(\mathbf{u}, \mathbf{v}) + \beta T(\mathbf{u}, \mathbf{w}) \tag{40.4.2}$$
>
> 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 $T : V \times V \to \mathbb{C}$ linear in both of its terms:
>
> $$T(\alpha\mathbf{v} + \beta\mathbf{w}, \mathbf{u}) = \alpha^* T(\mathbf{v}, \mathbf{u}) + \beta^* T(\mathbf{w}, \mathbf{u}) \tag{40.4.3}$$
> $$T(\mathbf{u}, \alpha\mathbf{v} + \beta\mathbf{w}) = \alpha T(\mathbf{u}, \mathbf{v}) + \beta T(\mathbf{u}, \mathbf{w}) \tag{40.4.4}$$
>
> 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:
>
> $$T(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathsf{A} \mathbf{y} \tag{40.4.5}$$

where $\mathbf{x}, \mathbf{y} \in V$ and $\mathsf{A} \in \mathrm{Mat}_n(V)$. Indeed, we prove linearity in the first argument as follows:

$$T(\alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z}) = (\alpha\mathbf{x} + \beta\mathbf{y})^T \mathsf{A}\mathbf{z} \tag{40.4.6}$$
$$= (\alpha\mathbf{x}^T + \beta\mathbf{y}^T)\mathsf{A}\mathbf{z} \tag{40.4.7}$$
$$= \alpha\mathbf{x}^T \mathsf{A}\mathbf{z} + \beta\mathbf{y}^T \mathsf{A}\mathbf{z} \tag{40.4.8}$$
$$= \alpha T(\mathbf{x}, \mathbf{z}) + \beta T(\mathbf{y}, \mathbf{z}) \tag{40.4.9}$$

The proof is similar for the linearity in second argument. We can extend this example to sesquilinear forms by defining:

$$T(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\dagger \mathsf{A}\mathbf{y} \tag{40.4.10}$$

where $\mathbf{x}, \mathbf{y} \in V$ and $\mathsf{A} \in \mathrm{Mat}_n(V)$. Here $\dagger$ denotes conjugate transposition. We again prove linearity in the first argument as follows:

$$T(\alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z}) = (\alpha\mathbf{x} + \beta\mathbf{y})^\dagger \mathsf{A}\mathbf{z} \tag{40.4.11}$$
$$= (\alpha\mathbf{x}^\dagger + \beta\mathbf{y}^\dagger)\mathsf{A}\mathbf{z} \tag{40.4.12}$$
$$= \alpha\mathbf{x}^\dagger \mathsf{A}\mathbf{z} + \beta\mathbf{y}^\dagger \mathsf{A}\mathbf{z} \tag{40.4.13}$$
$$= \alpha T(\mathbf{x}, \mathbf{z}) + \beta T(\mathbf{y}, \mathbf{z}) \tag{40.4.14}$$

Note also that if A is a symmetric matrix then:

$$T(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathsf{A})\mathbf{y} = (\mathsf{A}^T \mathbf{x})^T \mathbf{y} = \mathbf{y}^T \mathsf{A}\mathbf{x} \tag{40.4.15}$$

so T is a symmetric bilinear form. We can extend this result to sesquilinear forms quite easily by letting A be hermitian. Then

$$T(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\dagger \mathsf{A})\mathbf{y} = (\mathsf{A}^\dagger \mathbf{x})^\dagger \mathbf{y} = \mathbf{y}^\dagger \mathsf{A}\mathbf{x} \tag{40.4.16}$$

◀

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:

$$T(\mathbf{x}, \mathbf{y}) = \sum_{ij} x_i^* y_j T(\mathbf{e}_i, \mathbf{e}_j) \tag{40.4.17}$$

If we let $\mathsf{A}_{ij} = T(\mathbf{e}_i, \mathbf{e}_j)$ then clearly:

$$\mathbf{x}^\dagger \mathsf{A}\mathbf{y} = \sum_{ij} x_i^* A_{ij} y_j = \sum_{ij} x_i^* y_j T(\mathbf{e}_i, \mathbf{e}_j) = T(\mathbf{x}, \mathbf{y}) \tag{40.4.18}$$

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:

$$A = \begin{pmatrix} T(\mathbf{e}_1, \mathbf{e}_1) & T(\mathbf{e}_1, \mathbf{e}_2) & \dots \\ T(\mathbf{e}_2, \mathbf{e}_1) & T(\mathbf{e}_2, \mathbf{e}_2) & \dots \\ \vdots & \ddots & \vdots \end{pmatrix} \tag{40.4.19}$$

Analogously to linear maps, one can also performs changes of basis for sesquilinear/bilinear forms. Suppose that in the basis $\{\mathbf{e}_i\}$ the form T is represented by A so that

$$A_{ij} = T(\mathbf{e}_i, \mathbf{e}_j) \tag{40.4.20}$$

Let us introduce a new basis $\{\mathbf{e}'_i\}$ such that:

$$\mathbf{e}'_i = \sum_m c_{mi} \mathbf{e}_m \tag{40.4.21}$$

then we find

$$A'_{ij} = T(\mathbf{e}'_i, \mathbf{e}'_j) = \sum_{mn} c^*_{mi} c_{nj} T(\mathbf{e}_m, \mathbf{e}_n) = \sum_{mn} c^*_{mi} A_{mn} c_{nj} \tag{40.4.22}$$

Consequently, if we define a change of basis matrix P with components $P_{mn} = c_{mn}$ then we get:

$$A' = P^\dagger A P \tag{40.4.23}$$

We interpret this result as usual. P converts our vector from the original unprimed basis to the new primed basis:

$$\mathbf{x} = \sum_i x'_i \mathbf{e}'_i = \sum_{ij} x'_i c_{ji} \mathbf{e}_j = \sum_j x_j \mathbf{e}_j \tag{40.4.24}$$

where

$$x'_j = \sum_i x'_i c_{ji} \iff [\mathbf{x}]' = P[\mathbf{x}] \tag{40.4.25}$$

Therefore if we want to calculate the form $\mathbf{x}^\dagger A \mathbf{y}$ then we need a $P^\dagger$ to the left of A to convert the components of $\mathbf{x}^\dagger$ to the primed basis, and a P to the right of A to convert the components of $\mathbf{y}$.

L7 Eigen-everything

41.1 Finding eigenvalues and eigenvectors

> **Definition** *(Eigenvalue and eigenvector)*
> Let $f : V \to V$ 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:
>
> $$f(\mathbf{v}) = \lambda \mathbf{v} \tag{41.1.1}$$
>
> The eigenspace of λ is defined as:
>
> $$\mathrm{Eig}_f(\lambda) \equiv \mathrm{Ker}(f - \lambda \mathrm{id}_V) \subseteq V \tag{41.1.2}$$

For a map f to have non-trivial eigenvalues, we require that:

$$\dim \mathrm{Eig}_f(\lambda) = \mathrm{Ker}(f - \lambda \mathrm{id}_V) > 0 \implies \mathrm{rk}(f - \lambda \mathrm{id}_V) < \dim V \tag{41.1.3}$$

This is equivalent, by the proposition on linear systems and invertibility, to setting

$$\det(f - \lambda \mathrm{id}_V) = 0 \tag{41.1.4}$$

We could have also seen this by noting that if $f - \lambda \mathrm{id}_V$ were invertible, then there would only be one $\mathbf{v}$ in $\mathrm{Ker}(f - \lambda \mathrm{id}_V)$, which must be $\mathbf{0}$. Therefore $f - \lambda \mathrm{id}_V$ cannot be invertible, yielding (41.1.4).

> **Definition** *(Characteristic polynomial)*
> The characteristic polynomial of a map $f : V \to V$ is defined as:
>
> $$\chi_f(\lambda) = \det(f - \lambda \mathrm{id}_V) \tag{41.1.5}$$

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:

$$(f - \lambda \mathrm{id}_V)\mathbf{v} = 0 \tag{41.1.6}$$

using one of the methods introduced for solving linear systems.

Example. Let us find the eigenvalues and eigenspaces of

$$A = \begin{pmatrix} 4 & 0 & 4 \\ 0 & 4 & 4 \\ 4 & 4 & 8 \end{pmatrix} \qquad (41.1.7)$$

Its characteristic polynomial is:

$$\chi_A(\lambda) = \begin{vmatrix} 4-\lambda & 0 & 4 \\ 0 & 4-\lambda & 4 \\ 4 & 4 & 8-\lambda \end{vmatrix} \qquad (41.1.8)$$

$$= (4-\lambda)((4-\lambda)(8-\lambda) - 16) - 16(4-\lambda) \qquad (41.1.9)$$

$$= (4-\lambda)(\lambda^2 - 12\lambda) \qquad (41.1.10)$$

$$= \lambda(4-\lambda)(\lambda - 12) \qquad (41.1.11)$$

Clearly, the solutions to $\chi_A(\lambda) = 0$ are $\lambda = 0, 4, 12$.
For $\lambda_1 = 0$, we need:

$$\begin{pmatrix} 4 & 0 & 4 \\ 0 & 4 & 4 \\ 4 & 4 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies x = -z, y = -z \qquad (41.1.12)$$

giving an eigenspace:

$$\text{Eig}_A(\lambda_1) = \left\{ k \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \ \forall k \in \mathbb{R}^* \right\} \qquad (41.1.13)$$

Similarly, for $\lambda_1 = 4$, we need:

$$\begin{pmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies x = -y, z = 0 \qquad (41.1.14)$$

giving an eigenspace:

$$\text{Eig}_A(\lambda_1) = \left\{ k \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \ \forall k \in \mathbb{R}^* \right\} \qquad (41.1.15)$$

Finally, for $\lambda_1 = 12$, we need:

$$\begin{pmatrix} -8 & 0 & 4 \\ 0 & -8 & 4 \\ 4 & 4 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies z = 2x, y = x \qquad (41.1.16)$$

giving an eigenspace:

$$\text{Eig}_A(\lambda_1) = \left\{ k \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \ \forall k \in \mathbb{R}^* \right\} \qquad (41.1.17)$$

> **Proposition** *(Characteristic polynomial)*
> For a matrix $A \in \mathrm{Mat}_n(\mathbb{K})$ 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 $P \in \mathrm{Mat}_n(\mathbb{K})$
>   (ii) $c_n = (-1)^n, c_{n-1} = (-1)^{n-1}\mathrm{tr}A, c_0 = \det A$

Proof.   (i) We have that:

$$\det\big(PAP^{-1} - \lambda\mathbb{1}\big) = \det\big(P(A - \lambda\mathbb{1})P^{-1}\big) = \det(A - \lambda\mathbb{1}) \tag{41.1.18}$$

as desired.

(ii) We have that:

$$c_0 = \chi_A(0) = \det A \tag{41.1.19}$$

Furthermore

$$\chi_A(\lambda) = \prod_{i=1}^{n}(A_{ii} - \lambda) + o(\lambda^{n-2}) \tag{41.1.20}$$

$$= (-1)^n \lambda^n + (-1)^{n-1}\sum_{i=1}^{n} A_{ii} + o(\lambda^{n-1}) \tag{41.1.21}$$

$$= (-1)^n \lambda^n + (-1)^{n-1}\mathrm{tr}A + o(\lambda^{n-1}) \tag{41.1.22}$$

implying that $c_n = (-1)^n$ and $c_{n-1} = (-1)^{n-1}\mathrm{tr}A$.

∎

41.2   Matrix diagonalization

> **Definition** *(Diagonalized map)*
> A linear map $f : V \to V$ 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 $f : V \to V$ can be diagonalised *iff* there exists a basis of V consisting of eigenvectors of f. In this basis, the matrix representation of f is $\mathrm{diag}(\lambda_i)$ where λ_i are the eigenvalues of f.

Proof. ($\implies$) Suppose that f can be diagonalized into the form $\mathrm{diag}(c_i)$ in some basis $\{\mathbf{v}_i\}$. This implies that:

$$f(\mathbf{v}_i) = \sum_{j} A_{ji}\mathbf{v}_j = \sum_{j} \delta_{ji}c_i\mathbf{v}_i \tag{41.2.1}$$

which gives $f(\mathbf{v}_i) = c_i\mathbf{v}_i$, as desired.

($\impliedby$) Let $\mathsf{P} = (\mathbf{v}_1 \, \mathbf{v}_2 \, ... \, \mathbf{v}_n)$ 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 A of f in $\mathcal{B}$:

$$\mathsf{A}' = \mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{P}^{-1}\mathsf{A}(\mathbf{v}_1 \, \mathbf{v}_2 \, ... \, \mathbf{v}_n) \tag{41.2.2}$$

$$= \mathsf{P}^{-1}(\mathsf{A}\mathbf{v}_1 \, \mathsf{A}\mathbf{v}_2 \, ... \, \mathsf{A}\mathbf{v}_n) \tag{41.2.3}$$

$$= \mathsf{P}^{-1}(\lambda_1\mathbf{v}_1 \, \lambda_2\mathbf{v}_2 \, ... \, \lambda_n\mathbf{v}_n) \tag{41.2.4}$$

$$= \mathsf{P}^{-1}\mathsf{P}\,\mathrm{diag}(\lambda_i) \tag{41.2.5}$$

$$= \mathrm{diag}(\lambda_i) \tag{41.2.6}$$

Thus, the new matrix in the basis 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 A is diagonalizable then:

$$\mathrm{tr}\mathsf{A} = \sum_i \lambda_i, \; \det \mathsf{A} = \prod_i \lambda_i \tag{41.2.7}$$

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 ith column of a matrix representation gives the vector that the corresponding ith 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

$$\sigma = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{41.2.8}$$

Its characteristic equation is:

$$\chi_\sigma(\lambda) = \lambda^2 - 1 = 0 \implies \lambda = \pm 1 \tag{41.2.9}$$

For $\lambda_1 = 1$ we find that:

$$\begin{pmatrix} -1 & -i \\ i & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies x = -iy \tag{41.2.10}$$

Its eigenspace is:

$$\mathrm{Eig}_\sigma(\lambda_1) = \left\{ k \begin{pmatrix} -i \\ 1 \end{pmatrix} : k \in \mathbb{R}^* \right\} \tag{41.2.11}$$

Similarly, for $\lambda_2 = -1$ we find that:

$$\begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies x = iy \tag{41.2.12}$$

Its eigenspace is:

$$\mathrm{Eig}_\sigma(\lambda_1) = \left\{ k \begin{pmatrix} i \\ 1 \end{pmatrix} : k \in \mathbb{R}^* \right\} \tag{41.2.13}$$

We therefore choose the eigenbasis with $k = 1$, whose transition matrix is:

$$\mathsf{P} = \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} \tag{41.2.14}$$

whose inverse is (since $\det \mathsf{P} = -i - i = -2i$):

$$\mathsf{P}^{-1} = -\frac{1}{2i} \begin{pmatrix} 1 & -i \\ -1 & -i \end{pmatrix} = \frac{i}{2} \begin{pmatrix} 1 & -i \\ -1 & -i \end{pmatrix} = \frac{1}{2} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \tag{41.2.15}$$

We find that:

$$\sigma' = \frac{1}{2} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{41.2.16}$$

◀

Theorem *(Simultaneous diagonalization)*

Two linear maps f, g are simultaneously diagonalizable, that is, they are diagonalized by the same matrix P, *iff* they commute:

$$[f, g] \equiv f \circ g - g \circ f = 0 \tag{41.2.17}$$

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:

$$[f, g](\mathbf{v}_i) = \lambda_i g(\mathbf{v}_i) - \lambda_i' f(\mathbf{v}_i) = \lambda_i \lambda_i' - \lambda_i \lambda_i' = 0 \tag{41.2.18}$$

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:

$$(f \circ g)(\mathbf{v}_i) = f(g(\mathbf{v}_i)) = (g \circ f)(\mathbf{v}_i) = \lambda_i g(\mathbf{v}_i) \tag{41.2.19}$$

implying that $g(\mathbf{v}_i) \in \mathrm{Eig}_f(\lambda_i)$. 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. ∎

41.3 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

$$\langle \mathbf{v}, A\mathbf{v} \rangle = \lambda = \langle A\mathbf{v}, \mathbf{v} \rangle = \lambda^* \implies \lambda \in \mathbb{R} \tag{41.3.1}$$

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:

$$\langle \mathbf{v}_i, A\mathbf{v}_j \rangle = \lambda_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \lambda_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle \tag{41.3.2}$$

giving:

$$(\lambda_j - \lambda_i) \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \tag{41.3.3}$$

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 $\mathbf{u} \in W^\perp$:

$$\langle \mathbf{w}, f(\mathbf{u}) \rangle = \langle f(\mathbf{w}), \mathbf{u} \rangle = \lambda \langle \mathbf{w}, \mathbf{u} \rangle = 0 \tag{41.3.4}$$

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:
>
> $$A' = P^T A P \tag{41.3.5}$$

where A is the matrix representation of f in some basis $\mathcal{B}$, and

$$P = ([\mathbf{v}_1]_\mathcal{B} \ [\mathbf{v}_2]_\mathcal{B} \ ... \ [\mathbf{v}_n]_\mathcal{B}) \tag{41.3.6}$$

Proof. We have that:

$$P^T P = \begin{pmatrix} [\mathbf{v}_1]_\mathcal{B}^T \\ [\mathbf{v}_2]_\mathcal{B}^T \\ \vdots \\ [\mathbf{v}_n]_\mathcal{B}^T \end{pmatrix} ([\mathbf{v}_1]_\mathcal{B} \ [\mathbf{v}_2]_\mathcal{B} \ ... \ [\mathbf{v}_n]_\mathcal{B}) \tag{41.3.7}$$

$$= \begin{pmatrix} [\mathbf{v}_1]_\mathcal{B}^T[\mathbf{v}_1]_\mathcal{B} & [\mathbf{v}_1]_\mathcal{B}^T[\mathbf{v}_2]_\mathcal{B} & \cdots & [\mathbf{v}_1]_\mathcal{B}^T[\mathbf{v}_n]_\mathcal{B} \\ [\mathbf{v}_2]_\mathcal{B}^T[\mathbf{v}_1]_\mathcal{B} & [\mathbf{v}_2]_\mathcal{B}^T[\mathbf{v}_2]_\mathcal{B} & \cdots & [\mathbf{v}_2]_\mathcal{B}^T[\mathbf{v}_n]_\mathcal{B} \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{v}_n]_\mathcal{B}^T[\mathbf{v}_1]_\mathcal{B} & [\mathbf{v}_n]_\mathcal{B}^T[\mathbf{v}_2]_\mathcal{B} & \cdots & [\mathbf{v}_n]_\mathcal{B}^T[\mathbf{v}_n]_\mathcal{B} \end{pmatrix} \tag{41.3.8}$$

$$= \mathbb{1} \tag{41.3.9}$$

Consequently P is unitary, and hence when diagonalizing A' according to the general procedure:

$$A' = P^{-1}AP = P^T AP \tag{41.3.10}$$

as desired. ∎

Example. Let's diagonalize the following hermitian matrix:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \tag{41.3.11}$$

Its characteristic equation is:

$$\begin{vmatrix} 1-\lambda & 0 & 0 \\ 0 & 2-\lambda & 1 \\ 0 & 1 & 2-\lambda \end{vmatrix} = (1-\lambda)((2-\lambda)^2 - 1) \tag{41.3.12}$$

$$= (1-\lambda)(1-\lambda)(3-\lambda) = 0 \tag{41.3.13}$$

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

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies y = -z \tag{41.3.14}$$

This gives a two-dimensional eigenspace:

$$\text{Eig}_A(\lambda_1) = \left\{ t \begin{pmatrix} k \\ 1 \\ -1 \end{pmatrix}, \ \forall k, t \in \mathbb{R} \right\} \tag{41.3.15}$$

so we can choose as our orthonormal eigenvectors:

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \ \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{41.3.16}$$

Finally, for $\lambda_2 = 3$ then we get that:

$$\begin{pmatrix} -2 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies x = 0, y = z \tag{41.3.17}$$

giving the following eigenspace:

$$\text{Eig}_A(\lambda_2) = \left\{ t \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \ \forall t \in \mathbb{R}^* \right\} \tag{41.3.18}$$

We choose the following eigenvector

$$\mathbf{v}_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \tag{41.3.19}$$

Hence, the orthonormal eigenbasis has a transition matrix:

$$\mathsf{P} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \implies \mathsf{P}^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & -1 \\ \sqrt{2} & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \tag{41.3.20}$$

The diagonalized form of A is then:

$$\frac{1}{2} \begin{pmatrix} 0 & 1 & -1 \\ \sqrt{2} & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{41.3.21}$$

as desired. This concludes our process of orthogonal diagonalization. ◀

Definition *(Normal linear map)*
Let $f : V \to V$ 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 $f : V \to V$ be a normal linear map over V, and let $\mathbf{v} \in \mathrm{Eig}_f(\lambda) \subseteq V$ be an eigenvector of f with eigenvalue λ. Then, we have that $f^\dagger(\mathbf{v}) = \lambda^* \mathbf{v}$.

Proof. Let $g = f - \lambda \mathrm{id}$. Then we find that:

$$g \circ g^\dagger = (f - \lambda \mathrm{id}) \circ (f^\dagger - \lambda^* \mathrm{id}) = f \circ f^\dagger - \lambda^* f - \lambda f^\dagger + |\lambda|^2 \mathrm{id} \tag{41.3.22}$$
$$= f^\dagger \circ f - \lambda^* f - \lambda f^\dagger + |\lambda|^2 \mathrm{id} \tag{41.3.23}$$
$$= g^\dagger \circ g \tag{41.3.24}$$

implying that g is a normal linear map. Consequently:

$$0 = \langle g(\mathbf{v}), g(\mathbf{v}) \rangle = \langle \mathbf{v}, g^\dagger \circ g(\mathbf{v}) \rangle = \langle \mathbf{v}, g \circ g^\dagger(\mathbf{v}) \rangle = \langle g^\dagger(\mathbf{v}), g^\dagger(\mathbf{v}) \rangle \tag{41.3.25}$$

implying that $g^\dagger(\mathbf{v}) = f^\dagger(\mathbf{v}) - \lambda^* \mathbf{v} = 0$ as desired. ∎

> **Theorem** *(Spectral theorem for normal maps)*
> Let $f : V \to V$ 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 $W \equiv \mathrm{Eig}_f(\lambda)$ 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 $\mathbf{u} \in W^\perp$:

$$\langle f(\mathbf{u}), \mathbf{v} \rangle = \langle \mathbf{u}, f^\dagger(\mathbf{v}) \rangle = \lambda^* \langle \mathbf{w}, \mathbf{v} \rangle = 0 \tag{41.3.26}$$
$$\langle f^\dagger(\mathbf{u}), \mathbf{v} \rangle = \langle \mathbf{u}, f(\mathbf{v}) \rangle = \lambda \langle \mathbf{w}, \mathbf{v} \rangle = 0 \tag{41.3.27}$$

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 $n - 1$ orthonormal eigenvectors. Adding $\frac{\mathbf{v}}{|\mathbf{v}|}$ gives the desired list of n linearly independent orthonormal eigenvectors.

($\impliedby$) Suppose that f has an orthonormal eigenvector basis $\{\mathbf{n}_i\}$. Then:

$$f \circ f^\dagger(\mathbf{n}_i) = \lambda_i^* f(\mathbf{n}_i) = |\lambda_i|^2 \mathbf{n}_i \tag{41.3.28}$$

and

$$f^\dagger \circ f(\mathbf{n}_i) = \lambda_i f(\mathbf{n}_i) = |\lambda_i|^2 \mathbf{n}_i \tag{41.3.29}$$

implying that

$$[f, f^\dagger](\mathbf{n}_i) = 0 \tag{41.3.30}$$

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:

$$[f, f^\dagger](\mathbf{v}) = \sum_i \alpha_i [f, f^\dagger](\mathbf{n}_i) = 0, \ \forall \mathbf{v} \in V \implies [f, f^\dagger] = 0 \tag{41.3.31}$$

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.

41.4 Classifying conics

Suppose we have a conic with general equation:

$$Ax^2 + Bxy + Cy^2 + Fx + Gy + H = 0 \tag{41.4.1}$$

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 (41.4.1) as a product of matrices:

$$(x \ y) \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + (F \ G) \begin{pmatrix} x \\ y \end{pmatrix} + H = 0 \tag{41.4.2}$$

Let us define:

$$\mathsf{A} = \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix}, \mathbf{J} = (F \ G), \ \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{41.4.3}$$

then (41.4.2) turns into

$$\mathbf{x}^T \mathsf{A} \mathbf{x} + \mathbf{J}^T \mathbf{x} + H = 0 \tag{41.4.4}$$

It is important to note that A is a symmetric matrix, and can therefore be orthogonally diagonalized. Suppose that $\mathsf{D} = \mathsf{P}^T \mathsf{A} \mathsf{P}$, then we get that:

$$\mathbf{x}^T \mathsf{P} \mathsf{D} \mathsf{P}^T \mathbf{x} + \mathbf{J}^T \mathbf{x} + H = 0 \tag{41.4.5}$$

Let us define the coordinate vectors in the eigenbasis as: $\mathbf{x}' = \mathsf{P}^T \mathbf{x}$ (recall that P represents a change from the eigenbasis, so its inverse/transpose will represent a change to the eigenbasis). Then we find that:

$$(\mathbf{x}')^T \mathsf{D} \mathbf{x}' + \mathbf{J}^T \mathsf{P} \mathbf{x}' + H = 0 \tag{41.4.6}$$

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 $\mathbf{x} \to \mathbf{x}'$ we were really just rotating $\mathbb{R}^2$. Suppose $\mathsf{D} = \mathrm{diag}(\lambda_1, \lambda_2)$ and $\mathbf{J}^T \mathsf{P} = (f \ g)$, then we find:

$$\lambda_1 x'^2 + \lambda_2 y'^2 + fx' + gy' + H = 0 \tag{41.4.7}$$

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:

$$\lambda_1 x'^2 + \lambda_2 y'^2 + fx' + gy' + H = 0 \tag{41.4.8}$$

$$\implies \lambda_1 \left(x' + \frac{f}{2\lambda_1} \right)^2 + \lambda_2 \left(y' + \frac{g}{2\lambda_2} \right)^2 + H - \frac{f^2}{4\lambda_1} - \frac{g^2}{4\lambda_2} = 0 \tag{41.4.9}$$

Letting the translated axes be defined by:

$$\mathbf{x}'' = \mathbf{x}' + \frac{1}{2} \begin{pmatrix} \frac{f}{\lambda_1} \\ \frac{g}{\lambda_2} \end{pmatrix} \tag{41.4.10}$$

then we find that:

$$\lambda_1 x''^2 + \lambda_2 y''^2 + h = 0, \ h = H - \frac{f^2}{4\lambda_1} - \frac{g^2}{4\lambda_2} \tag{41.4.11}$$

which is a conic. It can be rearranged into the more useful form :

$$\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = 1 \tag{41.4.12}$$

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.

Conic	Standard form
Hyperbola	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$
Parabola	$y^2 - 4ax = 0$
Ellipse	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$

If instead we area dealing with a parabola, then we will find that one of the eigenvalues is 0. Suppose WLOG that $\lambda_1 = 0$, then we find:

$$\lambda_2 y'^2 + fx' + gy' + H = 0 \tag{41.4.13}$$

$$\implies \lambda_2 \left(y' + \frac{g}{2\lambda_2} \right)^2 + fx' + H - \frac{g^2}{4\lambda_2} = 0 \tag{41.4.14}$$

Letting the translated axes be defined by:

$$\mathbf{x}'' = \mathbf{x}' + \begin{pmatrix} \frac{H}{f} - \frac{g^2}{4f\lambda_2} \\ \frac{g}{2\lambda_2} \end{pmatrix} \tag{41.4.15}$$

then we find that:

$$y''^2 + \frac{f}{\lambda_2} x'' = 0 \tag{41.4.16}$$

which is a parabola. Had $\lambda_2 = 0$ then we would have found in complete analogy to before:

$$x''^2 + \frac{g}{\lambda_1} y'' = 0 \tag{41.4.17}$$

Example. Let's classify the conic described by:

$$x^2 - 4xy + 4y^2 - 6x - 8y + 5 = 0 \qquad (41.4.18)$$

which may be re-expressed in matrix for as:

$$\mathbf{x}^T \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} \mathbf{x} + \begin{pmatrix} -6 & -8 \end{pmatrix} \mathbf{x} + 5 = 0 \qquad (41.4.19)$$

The eigenvalues of A are easily found to obey

$$(1 - \lambda)(4 - \lambda) - 4 = 0 \implies \lambda(\lambda - 5) = 0 \implies \lambda_1 = 0, \ \lambda_2 = 5 \qquad (41.4.20)$$

For $\lambda_1 = 0$ we get the eigenspace

$$\text{Eig}_A(\lambda_1) = \left\{ k \begin{pmatrix} 2 \\ 1 \end{pmatrix} : k \in \mathbb{R}^* \right\} \qquad (41.4.21)$$

Similarly, for $\lambda_2 = 5$ we get the eigenspace:

$$\text{Eig}_A(\lambda_2) = \left\{ k \begin{pmatrix} 1 \\ -2 \end{pmatrix} : k \in \mathbb{R}^* \right\} \qquad (41.4.22)$$

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:

$$P = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \implies \mathbf{J}^T P = \frac{1}{\sqrt{5}} \begin{pmatrix} -20 & 10 \end{pmatrix} \qquad (41.4.23)$$

Consequently we find:

$$5y'^2 - 4\sqrt{5}x' + 2\sqrt{5}y' + 5 = 0 \implies y'^2 - \frac{4\sqrt{5}}{5}x' + \frac{2\sqrt{5}}{5}y' + 1 = 0 \qquad (41.4.24)$$

We now complete the square:

$$\left(y' + \frac{\sqrt{5}}{5}\right)^2 - \frac{4\sqrt{5}}{5}x' + \frac{4}{5} = 0 \implies y''^2 = \frac{4\sqrt{5}}{5}x'' \qquad (41.4.25)$$

where $y'' = y' + \frac{\sqrt{5}}{5}$ and $x'' = x' - \frac{\sqrt{5}}{5}$. This is therefore a parabola.

◀
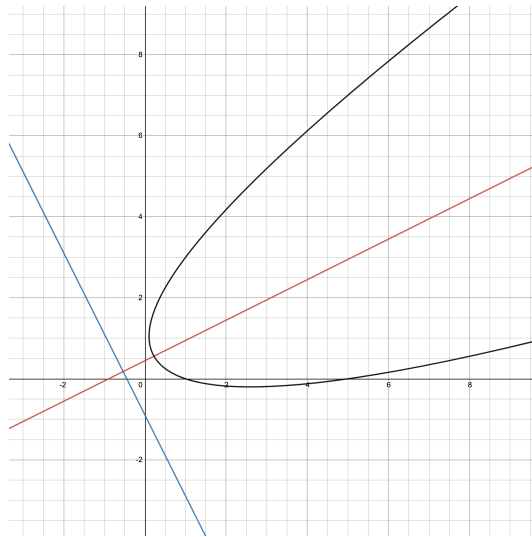
41.5   Matrix exponentials and Lie algebras

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:
>
> $$A^n = PA'^nP^{-1} \qquad (41.5.1)$$

Proof. This follows immediately from:

$$A^n = \underbrace{(PAP^{-1})(PAP^{-1})...(PAP^{-1})}_{n \text{ times}} = PA'^nP^{-1} \qquad (41.5.2)$$

∎

> **Definition** *(Matrix exponential)*
> Suppose that A matrix, then its exponential is defined as:
>
> $$e^A = \sum_{i=0}^{\infty} \frac{1}{n!}A^n = \mathbb{1} + A + \frac{1}{2}A^2 + ... \qquad (41.5.3)$$

41.6 Schur's triangulation theorem

Theorem *(Schur's triangulation theorem)*

Let $A \in \mathrm{Mat}_n(\mathbb{C})$ with eigenvalues $\lambda_1, \lambda_2, \ldots$ which may be degenerate.. Then A is unitarily equivalent to an upper triangular matrix:

$$A = UTU^\dagger \qquad (41.6.1)$$

where:

$$T = \begin{pmatrix} \lambda_1 & \cdots & \\ 0 & \lambda_2 & \\ \vdots & & \ddots \end{pmatrix} \qquad (41.6.2)$$

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 $A \in \mathrm{Mat}_n(\mathbb{C})$ have eigenvalues $\lambda_1, \lambda_2, \ldots$ and eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$, which may be degenerate, and are assumed to have unit norm.

Now $\mathbf{v}_1$ can be used to form an orthonormal basis $\{\mathbf{v}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$. Then, the resulting matrix when we change to this basis will be unitarily equivalent to A:

$$A = V \left(\begin{array}{c|ccc} \lambda_1 & a_{12} & \ldots & a_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{A} & \\ 0 & & & \end{array} \right) V^\dagger \qquad (41.6.3)$$

Clearly, we must have that $\chi_A(\lambda) = (\lambda_1 - \lambda)\chi_{\tilde{A}}(\lambda)$, implying that $\tilde{A}$ has eigenvalues $\lambda_2, \lambda_3, \ldots$ identical to A, and which could be degenerate. We can now use the induction hypothesis, since $\tilde{A} \in \mathrm{Mat}_{n-1}(\mathbb{C})$ we have that it is unitarily equivalent to some upper triangular matrix:

$$\tilde{A} = \widetilde{W} \begin{pmatrix} \lambda_2 & \tilde{a}_{12} & \ldots & \tilde{a}_{1n} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{a}_{n-1n} \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix} \widetilde{W}^\dagger \qquad (41.6.4)$$

Note also that:

$$
\begin{pmatrix} \begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{W} & \\ 0 & & & \end{array} \end{pmatrix}^\dagger
\begin{pmatrix} \begin{array}{c|ccc} \lambda_1 & a_{12} & \dots & a_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{A} & \\ 0 & & & \end{array} \end{pmatrix}
\begin{pmatrix} \begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{W} & \\ 0 & & & \end{array} \end{pmatrix}
=
\begin{pmatrix} \begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{W}^\dagger & \\ 0 & & & \end{array} \end{pmatrix}
\begin{pmatrix} \begin{array}{c|ccc} \lambda_1 & b_{12} & \dots & b_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{A}\tilde{W} & \\ 0 & & & \end{array} \end{pmatrix}
\tag{41.6.5}
$$

$$
=
\begin{pmatrix} \begin{array}{c|ccc} \lambda_1 & c_{12} & \dots & c_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{W}^\dagger\tilde{A}\tilde{W} & \\ 0 & & & \end{array} \end{pmatrix}
=
\begin{pmatrix} \lambda_1 & \tilde{b}_{12} & \dots & \dots & \tilde{b}_{1n} \\ 0 & \lambda_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \tilde{a}_{n-1n} \\ 0 & \dots & \dots & 0 & \lambda_n \end{pmatrix}
\tag{41.6.6}
$$

Defining:

$$
W = \begin{pmatrix} \begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{W} & \\ 0 & & & \end{array} \end{pmatrix}
\tag{41.6.7}
$$

we see that W is unitary and that

$$
A = VW \begin{pmatrix} \begin{array}{c|ccc} \lambda_1 & a_{12} & \dots & a_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{A} & \\ 0 & & & \end{array} \end{pmatrix} W^\dagger V^\dagger
\tag{41.6.8}
$$

Thus A is unitarily triangularizable via $U = VW$. ∎

Theorem *(The Cayley-Hamilton theorem)*
Let $A \in \mathrm{Mat}_n(\mathbb{C})$ have characteristic polynomial $\chi_A(\lambda)$. Then, we have that $\chi_A(A) = 0$ where 0 is the zero element of $\mathrm{Mat}_n(\mathbb{C})$,

---

Proof. We can factorize the characteristic polynomial into the following form due to the Fundamental theorem of algebra:

$$
\chi_A(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)...(\lambda_n - \lambda)
\tag{41.6.9}
$$

implying that

$$
\chi_A(A) = (\lambda_1 \mathbb{1} - A)(\lambda_2 \mathbb{1} - A)...(\lambda_n \mathbb{1} - A)
\tag{41.6.10}
$$

Since $A \in \mathrm{Mat}_n(\mathbb{C})$, Schur's theorem tells us that it can be triangularized unitaril $A = UTU^\dagger$ where T is upper triangular. Therefore:

$$
\chi_A(A) = U(\lambda_1 \mathbb{1} - T)U^\dagger U(\lambda_2 \mathbb{1} - T)U^\dagger...U(\lambda_n \mathbb{1} - T)U^\dagger
\tag{41.6.11}
$$
$$
= U(\lambda_1 \mathbb{1} - T)(\lambda_2 \mathbb{1} - T)...(\lambda_n \mathbb{1} - T)U^\dagger
\tag{41.6.12}
$$

Each of the factors $\lambda_i \mathbb{1} - T$ will be upper triangular with the ith 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 $\mathsf{C} = \mathsf{AB}$. We have that $C_{11} = C_{22} = 0$ since in general for triangular matrices:

$$C_{ii} = \sum_j A_{ij} B_{ji} = A_{ii} B_{ii} \tag{41.6.13}$$

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 $(\lambda_m \mathbb{1} - \mathsf{T})$ so that the first $m - 1$ rows are all zero. Then letting $\mathsf{C} = \prod_{i=1}^{m} (\lambda_i \mathsf{I} - \mathsf{T}) = \mathsf{A}(\lambda_m \mathsf{I} - \mathsf{T})$ we get:

$$C_{lm} = \sum_{l \leq j \leq m} A_{lj} B_{jm} \tag{41.6.14}$$

Now since $A_{lj} = 0$ for all $j < m$, the only term that will survive will be that with $j = m$. Consequently:

$$C_{lm} = A_{lm} B_{mm} = 0 \tag{41.6.15}$$

so the mth column will also be zero. It follows by induction that $(\lambda_1 \mathbb{1} - \mathsf{A})...(\lambda_n \mathbb{1} - \mathsf{A}) = 0$ and thus $\chi_{\mathsf{A}}(\mathsf{A}) = 0$ 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 $\Longleftarrow$ is trivial. Suppose $\mathsf{A} \in \mathrm{Mat}_n(\mathbb{C})$ 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:

$$\mathsf{A} = \left(\begin{array}{c|c} a_{11} & \mathbf{a} \\ \hline 0 & \widetilde{\mathsf{A}} \end{array} \right) \implies \mathsf{A}^\dagger = \left(\begin{array}{c|c} a_{11}^* & 0 \\ \hline \mathbf{a}^* & \widetilde{\mathsf{A}}^\dagger \end{array} \right) \tag{41.6.16}$$

we see that

$$\mathsf{A}^\dagger \mathsf{A} = \left(\begin{array}{c|c} |a_{11}|^2 & \cdots \\ \hline \cdots & \widetilde{\mathsf{A}}^\dagger \widetilde{\mathsf{A}} + ||\mathbf{a}||^2 \end{array} \right) \tag{41.6.17}$$

$$\mathsf{A} \mathsf{A}^\dagger = \left(\begin{array}{c|c} |a_{11}|^2 + ||\mathbf{a}||^2 & \cdots \\ \hline \cdots & \widetilde{\mathsf{A}} \widetilde{\mathsf{A}}^\dagger \end{array} \right) \tag{41.6.18}$$

For these to be equal, we need $a_{11} = 0$, $\mathbf{a} = 0$ and $\widetilde{\mathsf{A}}^\dagger \widetilde{\mathsf{A}} = \widetilde{\mathsf{A}} \widetilde{\mathsf{A}}^\dagger$. Also, since $\mathsf{A} \in \mathrm{Mat}_{n-1}(\mathbb{C})$ is upper triangular and normal, it must be diagonal. Consequently A is diagonal, as required. ∎

> **Theorem** *(The Spectral Theorem)* Let $\mathsf{A} \in \mathrm{Mat}_n(\mathbb{C})$ with eigenvalues $\lambda_1, ..., \lambda_n$, 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

$$A^\dagger A = (UD^\dagger U^\dagger)(UDU^\dagger) = UD^\dagger DU^\dagger \tag{41.6.19}$$
$$AA^\dagger = (UDU^\dagger)(UD^\dagger U^\dagger) = UDD^\dagger U^\dagger \tag{41.6.20}$$

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

$$(UAU^\dagger)^\dagger(UAU^\dagger) = UA^\dagger AU^\dagger = UAA^\dagger U^\dagger = (UAU^\dagger)(UAU^\dagger)^\dagger \tag{41.6.21}$$

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 $D = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ and thus:

$$\sum_{ij} |a_{ij}|^2 = \text{tr}(A^\dagger A) = \text{tr}(UD^\dagger DU^\dagger) = \text{tr}(U^\dagger U)\,\text{tr}(D) = \sum_i |\lambda_i|^2 \tag{41.6.22}$$

$(iii) \implies (ii)$ 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:

$$\sum_{ij} |a_{ij}|^2 = \text{tr}(A^\dagger A) = \text{tr}(T^\dagger T) = \sum_{ij} |T_{ij}|^2 = \sum_i |\lambda_i|^2 \tag{41.6.23}$$

We therefore find that $\sum_i |T_{ii}|^2 = \sum_{ij} |T_{ij}|^2$ which is only possible if T is diagonal.

$\blacksquare$

41.7 Jordan canonical form (make sure to write by October)

L8 Multilinear algebra

42.1 Review of linear algebra

Recall the normal definition of vector spaces.

> **Definition** *(Vector space)*
> A vector space $(V, +, \bullet)$ over a field $\mathbb{F}$ is a set V and two maps:
>
> $$+ : V \times V \to v \tag{42.1.1}$$
>
> $$\bullet : \mathbb{F} \times V \to V \tag{42.1.2}$$
>
> known as addition and scalar multiplication such that for all $u, v, w \in V, \alpha, \beta \in \mathbb{F}$:
>   (i) $v + w = w + v$
> (ii) $(u + v) + w = u + (v + w)$
> (iii) $\exists 0 \in V$ such that $v + 0 = v$
> (iv) $\exists (-v) \in V$ such that $v + (-v) = 0$
>   (v) $\alpha \bullet (\beta \bullet v) = (\alpha\beta) \bullet v$
> (vi) $\alpha \bullet (v + w) = \alpha \bullet v + \alpha \bullet w$
> (vii) $(\alpha + \beta) \bullet v = \alpha \bullet v + \beta \bullet v$
> (viii) given the identity element 1 of $\mathbb{F}$, then $1 \bullet v = v$

The element of a vector space is informally referred to as a **vector**. It is easy to see that the set $\mathcal{P}_n$ of all polynomial functions p up to order $n \in \mathbb{N}$ is indeed a vector space:

$$\mathcal{P}_n = \{p(x) = \sum_{m=0}^{n} p_m x^m : p_m \in \mathbb{R}\} \tag{42.1.3}$$

> **Definition** *(Linear map)*
> Let $(V, +_V, \bullet_V)$ and $(W, +_W, \bullet_W)$ be two vector spaces. Then the map $\phi : V \to W$ is a **linear map** if it is structure preserving:
>   (i) $\phi(u +_V v) = \phi(u) +_W \phi(v)$
> (ii) $\phi(\alpha \bullet_V v) = \alpha \bullet_W \phi(v)$
> If such a map between V and W exists then we write that $V \cong W$.

Definition *(Set of linear maps)*
We define the set of all linear maps between two vector spaces $(V, +_V, \bullet_V)$ and $(W, +_W, \bullet_W)$ (the latter defined over $\mathbb{F}_W$) as $\mathrm{Hom}(V, W)$. Together with the operations $\oplus$ and $\odot$ defined below this set becomes a vector space:

$$\oplus : \mathrm{Hom}(V, W) \times \mathrm{Hom}(V, W) \to \mathrm{Hom}(V, W) \tag{42.1.4}$$
$$(\phi, \varphi) \mapsto \phi \oplus \varphi \tag{42.1.5}$$

and

$$\odot : \mathbb{F}_W \times \mathrm{Hom}(V, W) \to \mathrm{Hom}(V, W) \tag{42.1.6}$$
$$(\alpha, \varphi) \mapsto \alpha \odot \varphi \tag{42.1.7}$$

where

$$(\phi \oplus \varphi)(v) = \phi(v) +_W \varphi(v), \ \forall v \in V \tag{42.1.8}$$
$$(\alpha \odot \varphi)(v) = \alpha \bullet_W \phi(v), \ \forall v \in V, \ \forall \alpha \in \mathbb{F}_W \tag{42.1.9}$$

Definition *(Dual vector space)*
Given a vector space V defined over a field $\mathbb{F}$, then the set V^* is defined as:

$$V^* \equiv \{\phi : V \to \mathbb{F} : \phi \text{ is linear}\} = \mathrm{Hom}(V, \mathbb{F}) \tag{42.1.10}$$

Equipped with $\oplus$ and $\otimes$ defined previously then V^* is defined as the **dual vector space**.

Informally, the element $\phi \in V^*$ is referred to as a **covector**.

Definition *(Dual basis)*
Given a basis $\mathcal{B} = \{e_1, e_2, ..., e_n\} \subset V$ for V then the **dual basis** of the dual space V^* is defined as $\mathcal{B}^* = \{\epsilon^1, \epsilon^2, ..., \epsilon^n\} \subset V^*$ such that:

$$\epsilon^i(\epsilon_j) = \delta_i^j, \ \forall i, j \tag{42.1.11}$$

For the case of $\mathcal{P}_n$, we can choose the basis $\mathcal{B} = \{e_i(x) = x^i : i = 0, 1, ..., n\}$. The dual basis $\mathcal{B}^* = \{\epsilon_i : i = 0, 1, ...n\}$ can be easily identified as:

$$\epsilon^i = \frac{1}{i!} \frac{d^i}{dx^i}\Bigg|_{x=0} \tag{42.1.12}$$

Indeed one finds that

$$\epsilon^i(e_j) = \frac{1}{i!} \frac{d^i}{dx^i}(x^j)\Bigg|_{x=0} = \begin{cases} 0, \ i > j \\ 1, \ i = j \\ 0, \ i < j \end{cases} = \delta_{ij} \tag{42.1.13}$$

42.2 Mltilinear maps

> **Definition** *(Tensor)*
> Let $(V, +, \bullet)$ be a vector space defined over a field $\mathbb{F}$. Then an (r, s)-**tensor** T over V is a map:
>
> $$T : (V^*)^r \times V^s \mapsto \mathbb{F} \tag{42.2.1}$$
>
> that is linear in all its entries, thus a **multi-linear map**.

Consider for example a $(1, 1)$-tensor T. Then we must have that:

$$T(\phi + \varphi, v) = T(\phi, v) + T(\varphi, v) \tag{42.2.2}$$

$$T(\alpha\phi, v) = \alpha T(\phi, v) \tag{42.2.3}$$

$$T(\phi, v + w) = T(\phi, v) + T(\phi, w) \tag{42.2.4}$$

$$T(\phi, \alpha v) = \alpha T(\phi, v) \tag{42.2.5}$$

$$\tag{42.2.6}$$

We can connect this to the typical view of a $(1, 1)$-tensor as a map that takes a vector and maps it to another vector. Indeed let us construct:

$$\phi_T : V \to V \tag{42.2.7}$$

$$v \mapsto T(\cdot, v) \tag{42.2.8}$$

which takes in v and maps it to another map $T(\cdot, v)$ which takes covectors to $\mathbb{F}$. But this map is of the type $V^* \to \mathbb{F}$ and thus $T(\cdot, v) \in (V^*)^* = V$ provided $\dim V < \infty$. It follows that $T(\cdot, v)$ is a vector, so ϕ_T does indeed map a vector to vectors of V.

> **Definition** *(Vectors vs Covectors)*
> Let $\phi \in V^*$ be a covector. Then by definition it must be a $(0, 1)$-tensor since $\phi : V \to \mathbb{F}$ is linear.
> Similarly, let $v \in V = (V^*)^*$ be a vector. Then by definition it must be a $(1, 0)$-tensor since $v : V^* \to \mathbb{F}$.

> **Definition** *(Tensor components)*
> Let T be an (r, s)-tensor over a n-dimensional vector space V with basis $\mathcal{B} = \{e_i : i = 0, 1, ...n\}$ and dual space V^* with dual basis $\mathcal{B}^* = \{\epsilon_i : i = 0, 1, ...n\}$. Then we define the components of T as:
> $$T^{i_1...i_r}_{j_1...j_s} \equiv T(\epsilon^{i_1}, ..., \epsilon^{i_r}, e_{j_1}, ..., e_{j_s}) \tag{42.2.9}$$
> so that:
> $$T(\phi, v) = \sum_{ij} \phi_i v^j T^i_j \tag{42.2.10}$$
> where $\phi = \sum_i \phi_i \epsilon^i \in V^*$ and $v = \sum_j v^j e_j \in V$.

For example, given a $(1, 1)$-tensor T then $T^i_j = T(\epsilon^i, e_j)$. Also, due to the linearity of the tensor map

the tensor components are very useful because they allow us to expand:

$$T(\phi, v) = T\left(\sum_i \phi_i \epsilon^i, \sum_j v^j e_j \right) \tag{42.2.11}$$

$$= \sum_{ij} \phi_i v^j T(\epsilon^i, e_j) \tag{42.2.12}$$

$$= \sum_{ij} \phi_i v^j T^i_j \tag{42.2.13}$$

Note the careful placements of subscripts and superscripts, where basis vectors are given subscripts and basis covectors are given superscripts. This ensures that multiply labelled indices always appear in an up-down arrangement. Moreover, if we assume that an index appearing both up and down is summed over, then we retrieve the **Einstein summation convention**. For example, (45.4.23) becomes

$$T(\phi, v) = \phi_i v^j T^i_j \tag{42.2.14}$$

The study of tensors in specific bases will be the subject of the next two chapters, and will culminate in the mathematical framework of differential geometry.

L9 Tensor algebra

We have studies 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.

43.1   Einstein summation convention

Consider a vector space V with a basis $\mathcal{B} = \{v_1, v_2, ..., v_n\}$, which we modify to $\mathcal{B}' = \{v_1', v_2', ..., v_n'\}$. From our study of linear algebra we know that we can express this change of basis as:

$$\begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix} = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \dots & & \dots \\ A_{n1} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \tag{43.1.1}$$

Alternatively:

$$v_i' = \sum_{j=1}^{n} A_{ij} v_j \tag{43.1.2}$$

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:

$$v_i' = \sum_{j=1}^{n} A_{ij} v_j \longrightarrow A_{ij} v_j \tag{43.1.3}$$

Suppose we wish to calculate the product of three matrices:

$$(ABC)_{il} = \sum_{j=1}^{n} \sum_{k=1}^{n} A_{ij} B_{jk} C_{kl} \tag{43.1.4}$$

In einstein notation, this becomes:

$$(ABC)_{il} = A_{ij} B_{jk} C_{kl} \tag{43.1.5}$$

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.

43.2   Cartesian tensors

A cartesian coordinate system in an n-dimensional space associates a coordinate $(x_1, x_2, ..., x_n)$ to each point in this space, with reference to a set of n basis vectors.

Suppose we have some vector **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:
> $$x_j' = R_{jk} x_k \tag{43.2.1a}$$
> $$x_k = R_{jk} x_j' \tag{43.2.1b}$$
> where $R_{jk} = \mathbf{e}_j' \cdot \mathbf{e}_k$.

Indeed, a general vector **r** may be expressed as:

$$\mathbf{r} = x_k \mathbf{e}_k = x_j' \mathbf{e}_j' \tag{43.2.2}$$

and since $\mathbf{e}_k = (\mathbf{e}_j' \cdot \mathbf{e}_k)\mathbf{e}_j' = R_{jk}\mathbf{e}_j'$ we find that:

$$x_k R_{jk} \mathbf{e}_j' = x_j' \mathbf{e}_j' \implies x_k R_{jk} = x_j' \tag{43.2.3}$$

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:

$$x_j' \mathbf{e}_j' = x_j' R_{jk} \mathbf{e}_k = x_k \mathbf{e}_k \implies x_k R_{jk} = x_j' \tag{43.2.4}$$

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:

$$x'_1 = (\mathbf{e}_1 \cdot \mathbf{e}'_1)x_1 + (\mathbf{e}_2 \cdot \mathbf{e}'_1)x_2 + (\mathbf{e}_3 \cdot \mathbf{e}'_1)x_3 = \cos\phi\, x_1 + \sin\theta\, x_2 \qquad (43.2.5)$$
$$x'_2 = (\mathbf{e}_1 \cdot \mathbf{e}'_2)x_1 + (\mathbf{e}_2 \cdot \mathbf{e}'_2)x_2 + (\mathbf{e}_3 \cdot \mathbf{e}'_2)x_3 = -\sin\phi\, x_1 + \cos\theta\, x_2 \qquad (43.2.6)$$
$$x'_3 = (\mathbf{e}_1 \cdot \mathbf{e}'_3)x_1 + (\mathbf{e}_2 \cdot \mathbf{e}'_3)x_2 + (\mathbf{e}_3 \cdot \mathbf{e}'_3)x_3 = x_3 \qquad (43.2.7)$$

Therefore, we define the **rotation vector**:

$$\mathsf{R} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (43.2.8)$$

so that a vector $\mathbf{x}$:

$$\mathbf{x}' = \mathsf{R}\mathbf{x} \qquad (43.2.9)$$

◀

Theorem *(Rotation operator* R*)*
The rotation operator in three dimensions, defined as:

$$\mathsf{R} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (43.2.10)$$

is unitary/orthogonal, satisfying $\mathsf{R}^T\mathsf{R} = \mathsf{R}\mathsf{R}^T = \mathbb{1}$ and $\det\mathsf{R} = \pm 1$.

Proof. We can write that:

$$\mathbf{x}' = \mathsf{R}\mathbf{x} \iff \mathbf{x} = \mathsf{R}^{-1}\mathbf{x}' \qquad (43.2.11)$$

or in component form:

$$x'_i = R_{ij}x_j \iff x_j = (\mathsf{R}^{-1})_{ji}x_i \qquad (43.2.12)$$

implying that $(\mathsf{R}^{-1})_{ji} = R_{ij}$, or in other words:

$$\boxed{\mathsf{R}^T = \mathsf{R}^{-1}} \qquad (43.2.13)$$

In other words, the rotation matrix 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:

$$1 = \det\mathbb{1}_n = \det(\mathsf{R}\mathsf{R}^T) = \det\mathsf{R}\det\mathsf{R}^T = (\det\mathsf{R})^2 \qquad (43.2.14)$$

so that:

$$\boxed{\det\mathsf{R} = \pm 1} \qquad (43.2.15)$$

Another important property of rigid rotation matrices is that their components are unitary:

$$\boxed{R_{ik}R_{jk} = \delta_{ij},\ R_{ik}R_{il} = \delta_{kl}} \qquad (43.2.16)$$

Indeed:

$$\mathbf{e}'_i \cdot \mathbf{e}'_j = (\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_j \cdot \mathbf{e}_k)\mathbf{e}_k \cdot \mathbf{e}_k = R_{ik}R_{jk} = \delta_{ij} \tag{43.2.17}$$

Similarly:

$$\mathbf{e}_k \cdot \mathbf{e}_l = (\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_i \cdot \mathbf{e}_l)\mathbf{e}'_i \cdot \mathbf{e}'_i = R_{ik}R_{il} = \delta_{kl} \tag{43.2.18}$$

as desired.

We could also see this by using the unitarity of R and write:

$$\mathbb{I} = \mathsf{R}\mathsf{R}^{-1} = \mathsf{R}\mathsf{R}^T \implies \delta_{ij} = R_{ik}R_{jk} \tag{43.2.19}$$

∎

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 :

$$\boxed{\begin{aligned} v'_i &= R_{ij}v_j \\ v_i &= R_{ki}v'_k \end{aligned}}$$

$$\text{(43.2.20a)}$$
$$\text{(43.2.20b)}$$

where R is the rotation matrix as defined (43.2.8).

Example. Consider the quantity $\mathbf{v} = (x_1^2, x_2^2)$, which transforms under rotations as:

$$v'_1 = (x'_1)^2 = (x_1 \cos\theta + x_2 \sin\theta)^2 \tag{43.2.21}$$
$$v'_2 = (x'_2)^2 = (-x_1 \sin\theta + x_2 \cos\theta)^2 \tag{43.2.22}$$

If this were a first order cartesian tensor, we would need:

$$v'_1 = v_1 \cos\theta + v_2 \sin\theta = x_1^2 \cos\theta + x_2^2 \sin\theta \tag{43.2.23}$$
$$v'_2 = -v_1 \sin\theta + v_2 \cos\theta = -x_1^2 \sin\theta + x_2^2 \cos\theta \tag{43.2.24}$$

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:

$$u'_1 = x'_2 = -x_1 \sin\theta + x_2 \cos\theta \tag{43.2.25}$$
$$u'_2 = -x'_1 = x_1 \cos\theta + x_2 \sin\theta \tag{43.2.26}$$

If this were a first order cartesian tensor, we would need:

$$u'_1 = u_1 \cos\theta + u_2 \sin\theta = x_2 \cos\theta - x_1 \sin\theta \tag{43.2.27}$$
$$u'_2 = -u_1 \sin\theta + u_2 \cos\theta = x_2 \sin\theta - x_1 \cos\theta \tag{43.2.28}$$

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:

$$u'_i v'_i = R_{ij} u_j R_{ik} u_k = R_{ij} R_{ik} u_j v_k = \delta_{jk} u_j v_k = u_j v_j \tag{43.2.29}$$

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:

$$T'_{ij} = R_{ik} R_{jl} T_{kl} \tag{43.2.30a}$$
$$T_{kl} = R_{mk} R_{nl} T_{mn} \tag{43.2.30b}$$

or alternatively $\mathsf{T}' = \mathsf{RTR}^T$.

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:

$$(\nabla \mathbf{v})_{ij} = \frac{\partial v_i}{\partial x_j} \tag{43.2.31}$$

We may regard $\left\{ \frac{\partial}{\partial x_j} \right\}$ as a basis, known as the **holonomic basis** or **coordinate basis**. Indeed:

$$\frac{\partial}{\partial x'_i} = \frac{\partial x_j}{\partial x'_i} \frac{\partial}{\partial x_j} = R_{ij} \frac{\partial}{\partial x_j} \implies R_{ij} = \frac{\partial x_j}{\partial x'_i} \tag{43.2.32}$$

from which it follows that ∇ is a first rank cartesian tensor[b].

Hence the components of $\nabla\mathbf{v}$ transform as:

$$(\nabla\mathbf{v})'_{ij} = \frac{\partial v'_i}{\partial x'_j} \tag{43.2.34}$$

$$= \frac{\partial v'_i}{\partial x_k}\frac{\partial x_k}{\partial x'_j} \tag{43.2.35}$$

$$= \frac{\partial}{\partial x_k}(R_{il}v_l)\frac{\partial x_k}{\partial x'_j} \tag{43.2.36}$$

$$= R_{il}\frac{\partial v_l}{\partial x_k}\frac{\partial x_k}{\partial x'_j} \tag{43.2.37}$$

$$= R_{il}R_{jk}(\nabla\mathbf{v})_{lk} \tag{43.2.38}$$

as would be expected from a second order tensor. ◀

---

^{a}this is a first order cartesian tensor
^{b}Similarly:

$$R_{ji} = \frac{\partial x'_i}{\partial x_j} \tag{43.2.33}$$

Definition *(Outer product)*
The outer product of two vectors $\mathbf{v}$, $\mathbf{u}$ is defined as:

$$(\mathbf{u}\otimes\mathbf{v})_{ij} = u_i v_j \tag{43.2.39}$$

and is a second order tensor.

It is easy to see that:

$$(\mathbf{u}\otimes\mathbf{v})'_{ij} = u'_i v'_j = R_{ik}R_{jl}u_k v_l = R_{ik}R_{jl}(\mathbf{u}\otimes\mathbf{v})_{kl} \tag{43.2.40}$$

as desired. Moreover, since $\mathbf{u} = u_i\mathbf{e}_i$ and $\mathbf{v} = v_i\mathbf{e}_i$ then:

$$\mathbf{u}\otimes\mathbf{v} = u_i v_j \mathbf{e}_i\otimes\mathbf{e}_j \tag{43.2.41}$$

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:

$$\mathsf{T} = \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} \tag{43.2.42}$$

Using $s \equiv \sin\theta$ and $c \equiv \cos\theta$ for shorthand we get that the components of T transform as:

$$T'_{11} = (x'_2)^2 = (-x_1 s + x_2 c)^2 = x_1^2 s^2 + x_2^2 c^2 - 2x_1 x_2 cs \tag{43.2.43}$$

$$T'_{12} = -x'_1 x'_2 = -(x_1 c + x_2 s)(-x_1 s + x_2 c) = x_1^2 sc - x_2^2 sc + x_1 x_2(s^2 - c^2) \tag{43.2.44}$$

$$T'_{21} = x_1^2 sc - x_2^2 sc + x_1 x_2(s^2 - c^2) \tag{43.2.45}$$

$$T'_{22} = (x'_1)^2 = (x_1 c + x_2 s)^2 = x_1^2 c^2 + x_2^2 s^2 + 2x_1 x_2 sc \tag{43.2.46}$$

so that:

$$\mathsf{T}' = \begin{pmatrix} x_1^2 s^2 + x_2^2 c^2 - 2x_1 x_2 cs & x_1^2 sc - x_2^2 sc + x_1 x_2 (s^2 - c^2) \\ x_1^2 sc - x_2^2 sc + x_1 x_2 (s^2 - c^2) & x_1^2 c^2 + x_2^2 s^2 + 2x_1 x_2 sc \end{pmatrix} \tag{43.2.47}$$

If T were a tensor then we would find

$$\mathsf{T}' = \mathsf{RTR}^T = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \tag{43.2.48}$$

$$= \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} x_2^2 c - x_1 x_2 s & -x_2^2 s - x_1 x_2 c \\ -x_1 x_2 c + s_x^2 s & x_1 x_2 s + x_1^2 c \end{pmatrix} \tag{43.2.49}$$

$$= \begin{pmatrix} x_1^2 s^2 + x_2^2 c^2 - 2x_1 x_2 cs & x_1^2 sc - x_2^2 sc + x_1 x_2 (s^2 - c^2) \\ x_1^2 sc - x_2^2 sc + x_1 x_2 (s^2 - c^2) & x_1^2 c^2 + x_2^2 s^2 + 2x_1 x_2 sc \end{pmatrix} \tag{43.2.50}$$

so T is indeed a second order cartesian tensor.
More simply, we could have noticed that $\mathsf{T} = (x_2, -x_1) \otimes (x_2, -x_1)$ and since $(x_2, -x_1)$ was proven to be a first order cartesian tensor, T will be a second order cartesian tensor as desired. ◀

Definition *(Cartesian Tensor)* In general, a cartesian tensor is defined as a geometric object 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:

$$T'_{ij...k} = R_{ip} R_{jq} ... R_{kr} T_{pq...r} \tag{43.2.51a}$$
$$T_{ij...k} = R_{pi} R_{qj} ... R_{rk} T'_{pq...r} \tag{43.2.51b}$$

Theorem *(Quotient law)*
Suppose that **B** and **C** are tensors such that its components in any rotated basis satisfy:

$$A_{pq...k...m} B_{ij...k...n} = C_{pq...mij...n} \tag{43.2.52}$$

then **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:

$$A_{pk} B_{ik} = C_{pi}, \ A'_{pk} B'_{ik} = C'_{pi} \tag{43.2.53}$$

and that $B_{jl} = R_{mj} R_{nl} B'_{mn}$, $C'_{pi} = R_{pq} R_{ij} C_{qj}$ then we get that:

$$A'_{pk} B'_{ik} = C'_{pi} \tag{43.2.54}$$
$$= R_{pq} R_{ij} C_{qj} \tag{43.2.55}$$
$$= R_{pq} R_{ij} A_{ql} B_{jl} \tag{43.2.56}$$
$$= R_{pq} R_{ij} A_{ql} R_{mj} R_{nl} B'_{mn} \tag{43.2.57}$$
$$= R_{pq} R_{nl} A_{ql} B'_{in} \tag{43.2.58}$$

so that:

$$(A'_{pk} - R_{pq}R_{nl}A_{ql})B'_{ik} = 0 \qquad (43.2.59)$$

Since this must hold for any B'_{ik} we must have that:

$$A'_{pk} = R_{pq}R_{nl}A_{ql} \qquad (43.2.60)$$

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:

$$\mathsf{T} = \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} \qquad (43.2.61)$$

is a tensor. We have already done this in two ways, but a third way is by using the quotient law:

$$T_{11}x_1^2 = x_1^2 x_2^2 \qquad (43.2.62)$$
$$T_{12}x_1 x_2 = -x_1^2 x_2^2 \qquad (43.2.63)$$
$$T_{21}x_2 x_1 = -x_1^2 x_2^2 \qquad (43.2.64)$$
$$T_{22}x_2^2 = x_1^2 x_2^2 \qquad (43.2.65)$$

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 T is also a tensor. ◀

43.3 The δ_{ij} and ϵ_{ijk} tensors

43.4 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:

$$L_i = m\epsilon_{ijk} r_j \dot{r}_k \qquad (43.4.1)$$

they they will transform under rotations as:

$$L'_i = m\epsilon_{ijk} r'_j \dot{r}'_k \qquad (43.4.2)$$

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.

$$L'_i = m\epsilon_{ijk} R_{jm} R_{kn} r_m \dot{r}_n \qquad (43.4.3)$$

Consider the i, j component of the **inertia tensor**:

$$I_{ij} = \int (\delta_{ij} r^2 - x_i x_j) dm \qquad (43.4.4)$$

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:

$$\langle i|I|j \rangle = \langle i| \int \langle r|r \rangle \, \mathbb{1} - |r \rangle \langle r| \, dm \, |j \rangle \tag{43.4.5}$$

from which it follows that:

$$\boxed{I = \int r^2 \mathbb{1} - dm} \tag{43.4.6}$$

This is a second order cartesian tensor.

43.5 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:

$$\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial u^i}, \quad \mathbf{e}^i = \nabla u_i \tag{43.5.1}$$

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:

$$\mathbf{e}_i \cdot \mathbf{e}^j = \delta_i^j \tag{43.5.2}$$

Proof. We find that:

$$\mathbf{e}_i \cdot \mathbf{e}^j = \frac{\partial \mathbf{r}}{\partial u^i} \nabla u_j \tag{43.5.3}$$

$$= \frac{\partial u_1}{\partial u^i} \frac{\partial u_j}{\partial u_1} + \frac{\partial u_2}{\partial u^i} \frac{\partial u_j}{\partial u_2} + \frac{\partial u_3}{\partial u^i} \frac{\partial u_j}{\partial u_3} \tag{43.5.4}$$

$$= \frac{\partial u_j}{\partial u^i} = \delta_i^j \tag{43.5.5}$$

as desired. $\blacksquare$

Consequently, given some vector $\mathbf{a}$ it may be expanded in both the contravariant and covariant bases. We find that if

$$\mathbf{a} = a^i \mathbf{e}_i = a_i \mathbf{e}^i \tag{43.5.6}$$

where a^i are the contravariant components while a_i are the covariant components, then:

$$\mathbf{a} \cdot \mathbf{e}^j = a^i \delta_i^j = a^j \tag{43.5.7}$$
$$\mathbf{a} \cdot \mathbf{e}_j = a_i \delta_j^i = a_j \tag{43.5.8}$$

as expected.

Let's now consider an infinitesimal vector displacement $d\mathbf{r} = du^i \mathbf{e}_i$. Then, the infinitesimal arc length is:

$$(ds)^2 = du^i du^j \mathbf{e}_i \cdot \mathbf{e}_j = g_{ij} du^i du^j \qquad (43.5.9)$$

where $g_{ij} \equiv \mathbf{e}_i \mathbf{e}_j$ 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:

$$g_{ij} \equiv \mathbf{e}_i \mathbf{e}_j \qquad (43.5.10)$$

Furthermore, the volume element may be expressed as:

$$dV = \sqrt{g} du^i \qquad (43.5.11)$$

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:

$$\mathbf{r} = r \sin\theta \cos\phi \, \mathbf{e}_x + r \sin\theta \sin\phi \, \mathbf{e}_y + r \cos\theta \, \mathbf{e}_z \qquad (43.5.12)$$

The covariant basis is easily found to be:

$$\mathbf{e}_1 = \frac{\partial \mathbf{r}}{\partial r} = \sin\theta \cos\phi \, \mathbf{e}_x + \sin\theta \sin\phi \, \mathbf{e}_y + \cos\theta \, \mathbf{e}_z \qquad (43.5.13)$$

$$\mathbf{e}_2 = \frac{\partial \mathbf{r}}{\partial \theta} = r \cos\theta \cos\phi \, \mathbf{e}_x + r \cos\theta \sin\phi \, \mathbf{e}_y - r \sin\theta \, \mathbf{e}_z \qquad (43.5.14)$$

$$\mathbf{e}_3 = \frac{\partial \mathbf{r}}{\partial \phi} = -r \sin\theta \sin\phi \, \mathbf{e}_x + r \sin\theta \cos\phi \, \mathbf{e}_y \qquad (43.5.15)$$

Consequently:

$$[g_{ij}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} \qquad (43.5.16)$$

As expected from an orthogonal coordinate system, the metric is diagonal. Moreover, we find that the infinitesimal arc length may be expressed as

$$(ds)^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \qquad (43.5.17)$$

while the volume element is found to be:

$$dV = r^2 \sin\theta \, dr d\theta d\phi \qquad (43.5.18)$$

◄

Just like vectors, tensors may also be expressed in both bases:

$$\mathsf{T} = T^{ij}\mathbf{e}_i \otimes \mathbf{e}_j = T_{ij}\mathbf{e}^i \otimes \mathbf{e}^j = T^j_i \mathbf{e}^i \otimes \mathbf{e}_j \tag{43.5.19}$$

Here T^{ij}, T_{ij}, T^j_i 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:

$$\mathbf{a} \cdot \mathbf{b} = a^i \mathbf{e}_i b^j \mathbf{e}_j = g_{ij} a^i b^j = a_i \mathbf{e}^i b_j \mathbf{e}^j = g^{ij} a_i b_j \tag{43.5.20}$$

Finally, we also find that:

$$\mathbf{a} \cdot \mathbf{b} = a^i \mathbf{e}_i b_j \mathbf{e}^j = a^i b_j \delta^j_i = a^i b_i = a_i b^i \tag{43.5.21}$$

Consequently:

$$g_{ij} a^i b^j = a^i b_i, \ g_{ij} a_i b_j = a_i b^i \tag{43.5.22}$$

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:
> $$g_{ij}b^j = b_i, \ g^{ij}b_j = b^i \tag{43.5.23}$$

The special case when $\mathbf{b} = \mathbf{e}_i$ then:

$$\mathbf{e}^i = g^{ij}\mathbf{e}_j, \ \mathbf{e}_i = g_{ij}\mathbf{e}^j \tag{43.5.24}$$

> **Proposition** *(Contravariant and covariant components are inverses)*
> The contravariant g^{ij} and covariant g_{ij} components of a metric tensor obey:
> $$g^{ij}g_{jk} = \delta^i_k \tag{43.5.25}$$

Proof. Consider:

$$a^i = \delta^i_k a^k = g^{ij}a_j = g^{ij}g_{jk}a^k \implies \delta^i_k = g^{ij}g_{jk} \tag{43.5.26}$$

as desired. ∎

43.6 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:

$$\mathbf{e}'_i = \frac{\partial \mathbf{r}}{\partial u'^i}, \ \mathbf{e}'^i = \nabla u'^i \tag{43.6.1}$$

The new covariant basis can be related to the old one by:

$$\mathbf{e}'_i = \frac{\partial \mathbf{r}}{\partial u'^i} = \frac{\partial u^j}{\partial u'^i}\frac{\partial \mathbf{r}}{\partial u^j} = \frac{\partial u^j}{\partial u'^i}\mathbf{e}_j \tag{43.6.2}$$

or alternatively:

$$\mathbf{e}_j = \frac{\partial u'^i}{\partial u^j}\mathbf{e}'_i \tag{43.6.3}$$

Therefore, expanding an arbitrary vector in both covariant bases:

$$\mathbf{a} = a^i\mathbf{e}_i = a^i\frac{\partial u'^j}{\partial u^i}\mathbf{e}'_j = a'^j\mathbf{e}'_j \tag{43.6.4}$$

implying that:

$$a'^j = \frac{\partial u'^j}{\partial u^i}a^i \tag{43.6.5}$$

Similarly, the new contravariant basis can be related to the old one by:

$$\mathbf{e}'^i = \nabla u'^i = \frac{\partial u'^i}{\partial u^j}\nabla u^j = \frac{\partial u'^i}{\partial u^j}\mathbf{e}^j \tag{43.6.6}$$

or alternatively:

$$\mathbf{e}^i = \frac{\partial u^i}{\partial u'^j}\mathbf{e}'^j \tag{43.6.7}$$

Therefore, expanding an arbitrary vector in both contravariant bases we find that:

$$\mathbf{a} = a_i\mathbf{e}^i = a_i\frac{\partial u^i}{\partial u'^j}\mathbf{e}'^j = a'_j\mathbf{e}'^j \tag{43.6.8}$$

implying that covariant components transform as:

$$a'_j = \frac{\partial u^i}{\partial u'^j}a_i \tag{43.6.9}$$

Theorem *(Transformation of co(ntra)variant components)* For a given vector **a**, its covariant and contravariant components transform under a change of basis as:

$$a'^j = \frac{\partial u'^j}{\partial u^i}a^i \tag{43.6.10}$$

$$a'_j = \frac{\partial u^i}{\partial u'^j}a_i \tag{43.6.11}$$

In a completely analogous way, we can show how the co(ntra)variant and mixed components of a second rank tensor transform:

$$T'^{ij} = \frac{\partial u'^i}{\partial u^k}\frac{\partial u'^j}{\partial u^l}T^{kl} \tag{43.6.12}$$

$$T'^i_j = \frac{\partial u'^i}{\partial u^k}\frac{\partial u^l}{\partial u'^j}T^k_l \tag{43.6.13}$$

$$T'_{ij} = \frac{\partial u^k}{\partial u'^i}\frac{\partial u^l}{\partial u'^j}T_{kl} \tag{43.6.14}$$

Proof. We have already proven the result for vectors. For second rank tensors, we find:

$$T'^{ij}\mathbf{e}'_i \otimes \mathbf{e}'_j = T^{ij}\mathbf{e}_i \otimes \mathbf{e}_j \tag{43.6.15}$$

$$T_i^{j'}\mathbf{e}'^i \otimes \mathbf{e}'_j = T_i^{j}\mathbf{e}^i \otimes \mathbf{e}_j \tag{43.6.16}$$

$$T'_{ij}\mathbf{e}'^i \otimes \mathbf{e}'^j = T_{ij}\mathbf{e}^i \otimes \mathbf{e}^j \tag{43.6.17}$$

For example, we would find that:

$$T'^{ij}\mathbf{e}'_i \otimes \mathbf{e}'_j = T^{kl}\mathbf{e}_k \otimes \mathbf{e}_l = T^{kl}\frac{\partial u'^i}{\partial u^k}\frac{\partial u'^i}{\partial u^j}\mathbf{e}'_i \otimes \mathbf{e}'_j \tag{43.6.18}$$

implying

$$T'^{ij} = \frac{\partial u'^i}{\partial u^k}\frac{\partial u'^j}{\partial u^l}T^{kl} \tag{43.6.19}$$

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.
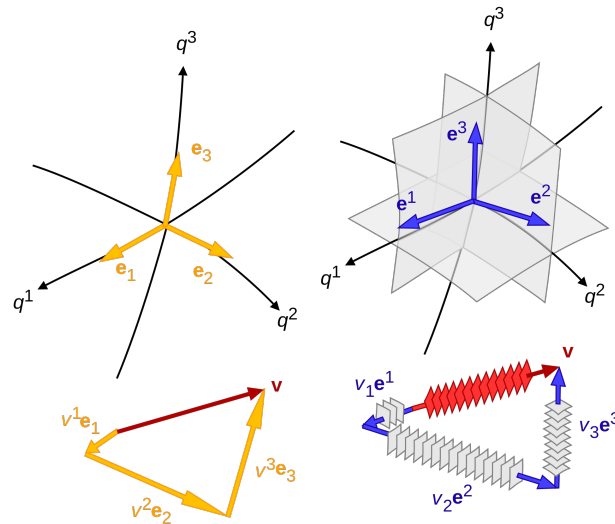
Figure 43.1. By Maschen - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=21043814

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 bass 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 :

$$g_{\mu\nu}dx^\mu dx^\nu = g'_{\alpha\beta}dx'^\alpha dx^\beta \tag{43.6.20}$$

$$= g'_{\alpha\beta}\frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}dx^\mu dx^\nu \tag{43.6.21}$$

$$\implies \left(g_{\mu\nu} - g'_{\alpha\beta}\frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}\right)dx^\mu dx^\nu = 0 \tag{43.6.22}$$

Seeing as this must hold for all dx^μ, dx^ν, it follows that:

$$g_{\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}g'_{\alpha\beta} \tag{43.6.23}$$

proving the required transformation rule. ∎

43.7 Application to special relativity: four-vectors

In special relativity we deal with four vectors whose components transform like:

$$v'^\mu = \frac{\partial x'^\mu}{\partial x^\nu}v^\nu \tag{43.7.1}$$

where $\frac{\partial u'^\mu}{\partial u^\nu}$ are components of the Lorentz transformation. An important four vector is the displacement vector:

$$dx^\mu = (cdt, dx, dy, dz) = (dx^0, dx^1, dx^2, dx^3) = (cdt, d\mathbf{r}) \tag{43.7.2}$$

We also know that the following quantity, known as the proper time, is an invariant under Lorentz transformations:

$$(cd\tau)^2 = c^2 dt^2 - d\mathbf{r}\cdot d\mathbf{r} \tag{43.7.3}$$

This suggests introducing the following metric tensor, known as the Minkowski metric:

$$[\eta_{\mu\nu}] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \implies (cd\tau)^2 = g_{\mu nu}dx^\mu dx^\nu \tag{43.7.4}$$

This allows us to find the covariant components:

$$dx_\mu = (cdt, -d\mathbf{r}) \tag{43.7.5}$$

The importance of the proper time arises when we try to define a velocity four vector. Suppose we naively define it to be:

$$u^\mu = \frac{\partial x^\mu}{\partial t} = \left(c, \frac{d\mathbf{r}}{dt}\right) \tag{43.7.6}$$

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

$$u^\mu = \frac{\partial x^\mu}{\partial \tau} = \frac{\partial x^\mu}{\partial t} \frac{\partial t}{\partial \tau} = \left(c, \frac{d\mathbf{r}}{dt}\right)\frac{dt}{d\tau} \tag{43.7.7}$$

Then, since

$$c^2 d\tau^2 = c^2 dt^2 - d\mathbf{r} \cdot d\mathbf{r} = dt^2 \left(c^2 - \frac{d\mathbf{r}}{dt} \cdot \frac{d\mathbf{r}}{dt}\right) \tag{43.7.8}$$

$$\implies \frac{dt}{d\tau} = \frac{1}{\sqrt{1 - v^2/c^2}} \equiv \gamma \tag{43.7.9}$$

where we defined $\mathbf{v} \equiv \frac{d\mathbf{r}}{dt}$, we find that the velocity four vector reads:

$$u^\mu = \gamma(c, \mathbf{v}) \tag{43.7.10}$$

and thus:

$$u_\mu = \gamma(c, -\mathbf{v}) \tag{43.7.11}$$

L10 Tensor calculus

44.1 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::

$$\boxed{\frac{\partial \mathbf{e}_i}{\partial u^j} = \Gamma_{ij}^k \mathbf{e}_k} \tag{44.1.1}$$

It is easy to see that:

$$\Gamma_{ij}^k = \mathbf{e}^k \cdot \frac{\partial \mathbf{e}_i}{\partial u^j} \tag{44.1.2}$$

Also, we can differentiate the relation $\mathbf{e}^i \cdot \mathbf{e}_j$ to find:

$$\frac{\partial}{\partial u^k}(\mathbf{e}^i \cdot \mathbf{e}_j) = \frac{\partial \mathbf{e}^i}{\partial u^k} \cdot \mathbf{e}_j + \mathbf{e}^i \cdot \frac{\partial \mathbf{e}_j}{\partial u^k} = 0 \tag{44.1.3}$$

$$\implies \frac{\partial \mathbf{e}^i}{\partial u^k} \cdot \mathbf{e}_j = -\Gamma_{jk}^i \tag{44.1.4}$$

$$\boxed{\implies \frac{\partial \mathbf{e}^i}{\partial u^j} = -\Gamma_{kj}^i \mathbf{e}^k} \tag{44.1.5}$$

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:

$$\Gamma_{ij}^{\prime k} = \mathbf{e}^{\prime k} \cdot \frac{\partial \mathbf{e}_i'}{\partial u^{\prime j}} \tag{44.1.6}$$

$$= \frac{\partial u^{\prime k}}{\partial u^n} \mathbf{e}^n \cdot \frac{\partial}{\partial u^{\prime k}}\left(\frac{\partial u^l}{\partial u^{\prime i} \mathbf{e}_l}\right) \tag{44.1.7}$$

$$= \frac{\partial u^{\prime k}}{\partial u^n} \mathbf{e}^n \cdot \left(\frac{\partial^2 u^l}{\partial u^{\prime j} \partial u^{\prime i}} \mathbf{e}_l + \frac{\partial u^l}{\partial u^{\prime i}} \frac{\partial \mathbf{e}_l}{\partial u^m} \frac{\partial u^m}{\partial u^{\prime j}}\right) \tag{44.1.8}$$

$$= \frac{\partial u^{\prime k}}{\partial u^n} \frac{\partial^2 u^l}{\partial u^{\prime j} \partial u^{\prime i}} \delta_l^n + \frac{\partial u^{\prime k}}{\partial u^n} \frac{\partial u^l}{\partial u^{\prime i}} \frac{\partial u^m}{\partial u^{\prime j}} \mathbf{e}_n \cdot \frac{\partial \mathbf{e}_l}{\partial u^m} \tag{44.1.9}$$

$$= \frac{\partial u^{\prime k}}{\partial u^l} \frac{\partial^2 u^l}{\partial u^{\prime j} \partial u^{\prime i}} + \Gamma_{lm}^n \frac{\partial u^{\prime k}}{\partial u^n} \frac{\partial u^l}{\partial u^{\prime i}} \frac{\partial u^m}{\partial u^{\prime j}} \tag{44.1.10}$$

We can find another expression for the Christoffel symbols that allow for faster computation. Consider the derivative of the metric tensor:

$$\frac{dg_{ij}}{du^k} = \frac{\partial \mathbf{e}_i}{\partial u^k} \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \frac{\partial \mathbf{e}_j}{\partial u^k} \tag{44.1.11}$$

$$= \Gamma^l_{ik}\mathbf{e}_l \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \mathbf{e}_l\Gamma^l_{jk} \tag{44.1.12}$$

$$\implies \frac{dg_{ij}}{du^k} = \Gamma^l_{ik}g_{lj} + \Gamma^l_{jk}g_{il} \tag{44.1.13}$$

Now note that Γ^l_{ik} is symmetric, since:

$$\frac{\partial \mathbf{e}_i}{\partial u^j} = \frac{\partial^2 \mathbf{r}}{\partial u^i u^j} = \frac{\partial \mathbf{e}_j}{\partial u^i} \tag{44.1.14}$$

Consequently, we can simply permute the indices in (44.1.13) and find the following:

$$\frac{dg_{ik}}{du^j} = \Gamma^l_{ij}g_{lk} + \Gamma^l_{jk}g_{il} \tag{44.1.15}$$

$$\frac{dg_{kj}}{du^i} = \Gamma^l_{ik}g_{lj} + \Gamma^l_{ji}g_{kl} \tag{44.1.16}$$

implying that:

$$\frac{dg_{ij}}{du^k} + \frac{dg_{ik}}{du^j} - \frac{dg_{kj}}{du^i} = 2\Gamma^l_{jk}g_{il} \tag{44.1.17}$$

$$\iff \left(\frac{dg_{ij}}{du^k} + \frac{dg_{ik}}{du^j} - \frac{dg_{kj}}{du^i}\right)g^{im} = 2\Gamma^l_{jk}g_{il}g^{im} \tag{44.1.18}$$

$$\boxed{\iff \Gamma^m_{jk} = \frac{1}{2}g^{im}\left(\frac{dg_{ij}}{du^k} + \frac{dg_{ik}}{du^j} - \frac{dg_{kj}}{du^i}\right)} \tag{44.1.19}$$

44.2   Differentiating tensors

Let's consider a contravariant vector:

$$x'^i = \frac{\partial x'^i}{\partial x^j}x^j \tag{44.2.1}$$

Its time derivative is transforms following:

$$\frac{\partial x'^i}{\partial u'^j} = \frac{\partial u^k}{\partial u'^j}\frac{\partial x'^i}{\partial u^k} \tag{44.2.2}$$

$$= \frac{\partial u^k}{\partial u'^j}\frac{\partial}{\partial u^k}\left(\frac{\partial u'^i}{\partial u^l}x^l\right) \tag{44.2.3}$$

$$= \frac{\partial u^k}{\partial u'^j}\left(\frac{\partial^2 u'^i}{\partial u^l \partial u^k}x^l + \frac{\partial u'^i}{\partial u^l}\frac{\partial x^l}{\partial u^k}\right) \tag{44.2.4}$$

$$= \frac{\partial u^k}{\partial u'^j}\frac{\partial^2 u'^i}{\partial u^j \partial u^k}x^j + \frac{\partial u^k}{\partial u'^j}\frac{\partial u'^i}{\partial u^l}\frac{\partial x^l}{\partial u^k} \tag{44.2.5}$$

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:

$$\frac{\partial \mathbf{x}}{\partial u^j} = \frac{\partial x^i}{\partial u^j}\mathbf{e}_i + x^i\Gamma^k_{ij}\mathbf{e}_k \tag{44.2.6}$$

$$= \left(\frac{\partial x^i}{\partial u^j} + x^k\Gamma^i_{kj}\right)\mathbf{e}_i \tag{44.2.7}$$

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:
>
> $$\boxed{v^i_{:j} = \frac{\partial x^i}{\partial u^j} + x^k\Gamma^i_{kj}} \tag{44.2.8}$$
>
> and similarly for the covariant representation:
>
> $$\boxed{v_{i:j} = \frac{\partial x_i}{\partial u^j} - x_k\Gamma^k_{ij}} \tag{44.2.9}$$

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:
>
> $$\mathbb{T} = T^{ij}(\mathbf{e}_i \otimes \mathbf{e}_j) \tag{44.2.10}$$
>
> Consequently:
>
> $$\frac{\partial \mathbb{T}}{\partial u^k} = \frac{\partial T^{ij}}{\partial u^k}(\mathbf{e}_i \otimes \mathbf{e}_j) + T^{ij}\frac{\partial}{\partial u^k}(\mathbf{e}_i \otimes \mathbf{e}_j) \tag{44.2.11}$$
>
> We can simplify the second term on the RHS:
>
> $$\frac{\partial}{\partial u^k}(\mathbf{e}_i \otimes \mathbf{e}_j) = \frac{\partial \mathbf{e}_i}{\partial u^k} \otimes \mathbf{e}_j + \mathbf{e}_i \otimes \frac{\partial \mathbf{e}_j}{\partial u^k} \tag{44.2.12}$$
>
> $$= \Gamma^l_{ik}\mathbf{e}_l \otimes \mathbf{e}_j + \mathbf{e}_i \otimes \mathbf{e}_l\Gamma^l_{jk} \tag{44.2.13}$$
>
> giving
>
> $$\frac{\partial \mathbb{T}}{\partial u^k} = \frac{\partial T^{ij}}{\partial u^k}(\mathbf{e}_i \otimes \mathbf{e}_j) + T^{ij}(\Gamma^l_{ik}\mathbf{e}_l \otimes \mathbf{e}_j + \mathbf{e}_i \otimes \mathbf{e}_l\Gamma^l_{jk}) \tag{44.2.14}$$
>
> $$= \left(\frac{\partial T^{ij}}{\partial u^k} + T^{lj}\Gamma^i_{lk} + T^{il}\Gamma^j_{lk}\right)(\mathbf{e}_i \otimes \mathbf{e}_j) \tag{44.2.15}$$
>
> $$= T^{ij}_{:k}(\mathbf{e}_i \otimes \mathbf{e}_j) \tag{44.2.16}$$

where we defined a covariant derivative:

$$\boxed{T^{ij}_{:k} \equiv \frac{\partial T^{ij}}{\partial u^k} + T^{lj}\Gamma^i_{lk} + T^{il}\Gamma^j_{lk}}$$

(44.2.17)

◄

44.3 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:

$$\nabla\phi \equiv \phi_{:i}\mathbf{e}^i = \frac{\partial\phi}{\partial u^i}\mathbf{e}^i$$

(44.3.1)

Divergence

The divergence of a vector field $\mathbf{v}$ is:

$$\nabla \cdot \mathbf{x} \equiv v^i_{:i} = \frac{\partial v^i}{\partial u^i} + x^k\Gamma^i_{ki}$$

(44.3.2)

Note that the Christoffel symbol simplifies significantly:

$$\Gamma^i_{ki} = \frac{1}{2}g^{il}\left(\frac{\partial g_{il}}{\partial u^k} + \frac{\partial g_{kl}}{\partial u^i} - \frac{\partial g_{ki}}{\partial u^l}\right)$$

(44.3.3)

$$= \frac{1}{2}g^{il}\left(\frac{\partial g_{il}}{\partial u^k} + \frac{\partial g_{kl}}{\partial u^i} - \frac{\partial g_{kl}}{\partial u^i}\right)$$

(44.3.4)

$$= \frac{1}{2}g^{il}\frac{\partial g_{il}}{\partial u^k}$$

(44.3.5)

giving:

$$\nabla \cdot \mathbf{x} = \frac{\partial v^i}{\partial u^i} + \frac{1}{2}x^k g^{il}\frac{\partial g_{il}}{\partial u^k}$$

(44.3.6)

> **Proposition** *(Important determinant identity)*
> Let $\mathsf{A} = (a_{ij})$, $\mathsf{B} = (b^{ij})$ and $\mathsf{A} = \mathsf{B}^{-1}$. Then:
>
> $$\frac{\partial|\mathsf{A}|}{\partial u^k} = |\mathsf{A}|b^{ji}\frac{\partial a_{ij}}{\partial u^k}$$
>
> (44.3.7)

Proof. We have that if the cofactor of the element a_{ij} is c^{ij} then:

$$b^{ij} = \frac{1}{|\mathsf{A}|}c^{ji}$$

(44.3.8)

Consequently, since $|\mathsf{A}| = a_{ij}c^{ij}$ with fixed i:

$$\frac{\partial|\mathsf{A}|}{\partial a_{ij}} = c^{ij} = |\mathsf{A}|b^{ij} \tag{44.3.9}$$

implying:

$$\frac{\partial|\mathsf{A}|}{\partial u^k} = \frac{\partial|\mathsf{A}|}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial u^k} = |\mathsf{A}|b^{ij}\frac{\partial a_{ij}}{\partial u^k} \tag{44.3.10}$$

as desired. ∎

We can apply this proposition to the determinant of the metric tensor:

$$\frac{\partial|g|}{\partial u^k} = |g|g^{ij}\frac{\partial g_{ij}}{\partial u^k} \implies g^{ij}\frac{\partial g_{ij}}{\partial u^k} = \frac{1}{|g|}\frac{\partial|g|}{\partial u^k} = \frac{1}{\sqrt{|g|}}\frac{\partial\sqrt{|g|}}{\partial u^k} \tag{44.3.11}$$

and hence when substituted into (44.3.6) we get:

$$\nabla \cdot \mathbf{x} = \frac{\partial v^i}{\partial u^i} + \frac{1}{\sqrt{|g|}}\frac{\partial\sqrt{|g|}}{\partial u^i}x^i \tag{44.3.12}$$

or alternatively:

$$\boxed{\nabla \cdot \mathbf{v} = \frac{1}{\sqrt{|g|}}\frac{\partial(\sqrt{|g|}x^i)}{\partial u^i}} \tag{44.3.13}$$

Laplacian

Curl

44.4   Geodesics

Part VI

Differential Geometry

DG1 Topology

45.1 Why differential geometry?

Broadly speaking, a manifold is a space that looks locally like $\mathbb{R}^n$ even though it may not be so globally. For example, the surface of the Earth may be viewed as a locally flat surface embedded in $\mathbb{R}^3$. Thus most tools from multivariable calculus may be employed. In general however, we will discover that space-time does not seem to be naturally embedded in some higher-dimensional Euclidean space. Traditional calculus is unequipped to deal with such objects, we must develop a new formalism which is known nowadays as modern differential geometry.

45.2 Topology

We begin by introducing some important basic concepts from topology. Indeed, at its most fundamental (i.e. with least structure) level space-time may be viewed as a simple collection of points. This alone is not enough to talk about things such as derivatives etc..., extra conditions must be satisfied. However, we do not want to endow space-time with too much structure, just enough so that good old calculus may be used. For example, the way we may classify these points into sub-collections will be useful in giving structure to space-time. The first step is thus to define a topological space.

> **Definition** *(Topological space)*
> A **topological space** $(X, \mathcal{O})$ is a set X equipped with a collection $\mathcal{O} \subseteq \mathcal{P}(X)$ of its subsets, which we refer to as **open sets** forming the **topology** $\mathcal{O}$ of X, such that the following conditions are satisfied:
>   (i) the empty set $\emptyset$ and X are open sets
> (ii) if $A, B \in \mathcal{O}$ then $A \cap B$ is also open
> (iii) if $A_i \in \mathcal{O}$ for some $\{i\}$ then $\bigcup_i A_i$ is open.

> **Definition** *(Closed set)*
> Let $(X, \mathcal{O})$ be a topological space. Then $A \subseteq X$ is closed (relative to the topology $\mathcal{O}$) if $X \setminus A \in \mathcal{O}$.

There are trivial examples of topologies. Let X be any set, then $\mathcal{O} = \{\emptyset, X\}$ is a topology known as the **chaotic topology**. Also, $\mathcal{O} = \mathcal{P}(X)$ is a topology known as the **discrete topology**.

A very relevant example is the standard topology of the d-dimensional Eulidean space $\mathbb{R}^d$. Before defining it let us discuss what it means to be in the neighborhood of a point x.

> **Definition** *(Neighborhood)*
> Let us define for all $x \in \mathbb{R}^d$ and $d \in \mathbb{R}^+$:
>
> $$N_d(x) = \{y \in \mathbb{R}^d : ||y - x|| < d\} \tag{45.2.1}$$
>
> which is an open (since the inequality is strict) ball centered at x with radius d. This is the **neighborhood** of x of radius d.
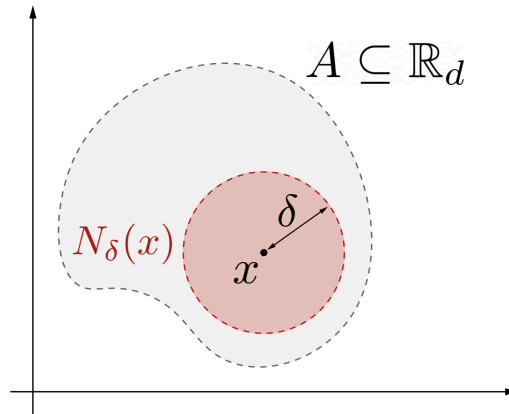
> **Theorem** *(Standard topology)*
> The following is a topology on $\mathbb{R}^d$ known as the **standard topology** $\mathcal{O}_{\mathbb{R}^d}$:
>
> $$A \in \mathcal{O}_{\mathbb{R}^d} \iff \forall x \in A, \exists \delta \in \mathbb{R}^+ \text{ s.t. } N_\delta(x) \subseteq A \tag{45.2.2}$$
>
> that is, an open set is a member of this topology if we can always form a neighborhood of x contained within A.

Proof. We check that the three topological axioms are satisfied:

(i) Firstly $\emptyset \in \mathcal{O}_{\mathbb{R}^d}$ since the neighborhood of the empty set is empty for any $r \in \mathbb{R}^+$ and $\emptyset \subseteq \emptyset$. Similarly $\mathbb{R}^d \in \mathcal{O}_{\mathbb{R}^d}$ since the neighborhood of any arbitrary point for any $r \in \mathbb{R}^+$ must by definition lie in $\mathbb{R}^d$

(iii) Let $A, B \in \mathcal{O}_{\mathbb{R}^d}$ be open. This implies that $\exists r_A, r_B \in \mathbb{R}^+$ such that $N_{r_A}(x) \in A$ and $N_{r_B}(x)$. Letting $r = \min\{r_A, r_B\}$ then $N_r(x) \subseteq U$ an $N_r(x) \subseteq V$ implying that $N_r(x) \subseteq U \cup V$ as desired.

(iv) Let $A_i \in \mathcal{O}$ for given $\{i\}$. Then for any j there exists $r_j \in \mathbb{R}^+$ such that $N_{r_j}(\mathcal{O}) \in A_j$ implying that $N_{r_j}(\mathcal{O}) \in \cup_i A_i$ as desired.

∎

> **Theorem** *(Induced topology)*
> Let $(X, \mathcal{O})$ be a topological space, and let $B \subseteq X$. Then:
>
> $$\mathcal{O}|_B = \{A \cap B | A \in \mathcal{O}\} \tag{45.2.3}$$

is a topology on b known as the **induced topology**.

Proof. We verify the topology axioms:

(i) since $\emptyset \in \mathcal{O}$ we have that $\emptyset \cap B = \emptyset$ is also an open set. Similarly since $X \in \mathcal{O}$ we have that $X \cap B = B$ is open as desired.

(ii) let $U, V \in \mathcal{O}|_N$. Then $\exists S, T \in \mathcal{O}$ such that $U = S \cap B$ and $V = T \cap B$, implying that:

$$U \cap V = (S \cap B) \cap (T \cap B) = \underbrace{(S \cap T)}_{\in \mathcal{O}} \cap B \tag{45.2.4}$$

(ii) let $U_i \in \mathcal{O}|_N$ for some $\{i\}$. Then $\exists S_i \in \mathcal{O}$ such that $U_i = S_i \cap B$ so that

$$\bigcup_i U_i = \bigcup_i (S_i \cap B) = \overbrace{\left(\bigcup_i S_i\right)}^{\in \mathcal{O}} \cup B \tag{45.2.5}$$

$\blacksquare$

An interesting example is the set $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| = 1\}$. We can establish a topology on $\mathcal{C}$ by inducing the standard topology on $\mathcal{S}$:

$$\mathcal{O} = \mathcal{O}_{\mathbb{R}^d}|_\mathcal{S} \tag{45.2.6}$$

This is because the empty set is the intersection of the $\emptyset \in \mathcal{O}_{\mathbb{R}_d}$ and $\mathcal{S}$, similarly $\mathcal{S}$ is the intersection of $\mathbb{R}^d \in \mathcal{O}_{\mathbb{R}^d}$. Also, given any subset of $\mathcal{S}$, it can be written as the intersection of another circle and $\mathcal{S}$ as shown below.

> **Definition** *(Product topology)*
> Let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$. Then $\mathcal{O}_{A,B}$ is a topology, the **product topology** on $A \times B$ defined by:
>
> $$U \in \mathcal{O}_{A \times B} \iff \forall x = (a, b) \in U, \ \exists S \in \mathcal{O}_A, T \in \mathcal{O}_B \text{ s.t. } S \times T \subseteq U \text{ and } a \in S, b \in T \tag{45.2.7}$$

Proof. We verify the topology axioms:

(i) of course $\emptyset \in \mathcal{O}_{A \times B}$ since $\nexists x \in \emptyset$, so any statement following this assumption is trivially satisfied. Similarly $U = A \times B \in \mathcal{O}_{A \times B}$ since given $x = (a, b) \in U$ then $a \in A \in \mathcal{O}_A, b \in B \in \mathcal{O}_B$ with $A \times B \subseteq U$ as desired.

(ii) Let $U, V \in \mathcal{O}_{A \times B}$, and let $S_{U,V} \in \mathcal{O}_A$ and $T_{U,V} \in \mathcal{O}_B$ be the open sets satisfying the topological conditions:

$$\forall x = (a, b) \in U, \ \exists S_U \in \mathcal{O}_A, T_U \in \mathcal{O}_B \text{ s.t. } S_U \times T_U \subseteq U \text{ and } a \in S_U, b \in T_U \tag{45.2.8}$$
$$\forall x = (a, b) \in V, \ \exists S_V \in \mathcal{O}_A, T_V \in \mathcal{O}_B \text{ s.t. } S_V \times T_V \subseteq U \text{ and } a \in S_V, b \in T_V \tag{45.2.9}$$

Let $x = (a, b) \in U \cap V$, and consider $S_U \cap S_V \in \mathcal{O}_A$ and $T_U \cap T_V \in \mathcal{O}_B$. Then clearly $a \in S_U \cap S_V$ and $b \in T_U \cap T_V$, and moreover:

$$(S_U \cap S_V) \times (S_U \cap S_V) = (S_U \times T_U) \cap (S_V \times T_V) \subseteq U \times V \tag{45.2.10}$$

as desired.

(iii) Let $U_i \in \mathcal{O}_{A \times B}$ for some set $\{i\}$. Then, we must have that:

$$\forall i : \forall x = (a, b) \in U_i, \ \exists S_i \in \mathcal{O}_A, T_i \in \mathcal{O}_B \text{ s.t. } S_i \times T_i \subseteq U_i \text{ and } a \in S_i, b \in T_i \tag{45.2.11}$$

Let $x = (a, b) \in \bigcup_i U_i$, and let us consider the unions $\bigcup_i S_i$ and $\bigcup_i T_i$. Clearly $a \in \bigcup_i S_i$ and $b \in \bigcup_i T_i$. Moreover:

$$\left(\bigcup_i S_i \right) \times \left(\bigcup_i T_i \right) = \bigcup_i (S_i \times T_i) \subseteq \bigcup_i U_i \tag{45.2.12}$$

as desired.

∎

This definition can be easily extended to any number of Cartesian products. Also, you can always fit a circle in a square and vice-versa, $S \times T$ and $N_d(x)$ can be fit into each other. Therefore, $\mathcal{O}_{\mathcal{R}^d} = \mathcal{O}_{\mathcal{R} \times \ldots \times \mathcal{R}}$.

> **Definition** *(Convergence)*
> A sequence $q : \mathbb{N} \to X$ on a topological space $(X, \mathcal{O})$ converges to a point $a \in X$ if:
>
> $$\forall U \in \mathcal{O} \text{ with } a \in U, \exists N \in \mathbb{N} \text{ such that } \forall n > N, \ q(n) \in U \tag{45.2.13}$$

> **Theorem** *(Sequence convergence)*
> The sequence $q : \mathbb{N} \to \mathbb{R}$ converges to $a \in \mathbb{R}^d$ if:
>
> $$\forall \epsilon > 0, \ \exists N \in \mathbb{N} \text{ s.t. } ||q(n) - a|| < \epsilon, \ \forall n > N \tag{45.2.14}$$

Proof. Suppose the sequence $q : \mathbb{N} \to \mathbb{R}^d$ with standard topology converges to a. Let $\epsilon > 0$, and define:

$$U_\epsilon = \{ x \in \mathbb{R}^d : ||x - a|| < \epsilon \} \tag{45.2.15}$$

This, as we shall prove soon, is an open set. Also, we have that $a \in U_\epsilon$, so the definition of sequence convergence implies that:

$$\exists N \in \mathbb{N} \text{ such that } ||q(n) - a|| < \epsilon, \ \forall n > N \tag{45.2.16}$$

as desired. ∎

> **Definition** *(Continuity)*
> Let $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ be topological spaces. Then the map $\phi : X \to Y$ is **continuous** if for all open sets $U \in \mathcal{O}_Y$ the set $\text{preIm}_\phi U \in \mathcal{O}_X$.

In other words, a continuous function maps open sets to open sets without disrupting the topologies of the domain and image. Consider for example $\phi : X \to Y$ where X is equipped with the topology $\mathcal{P}(X)$ while Y is equipped with another topology $\mathcal{O}_Y$. Then ϕ must be continuous since the pre-image of any $V \in \mathcal{O}_Y$ is necessarily a subset of the domain X, and must therefore be included in the power set $\mathcal{P}(X)$. Similarly $\phi : X \to Y$ where X is equipped with any topology and Y is equipped with the chaotic topology $\{\emptyset, Y\}$ is also continuous.

We can visualize the definition of a continuous function asa map that preserved the "open-ness" of subsets. It is good sanity check to show that this topological definition of continuity yields the
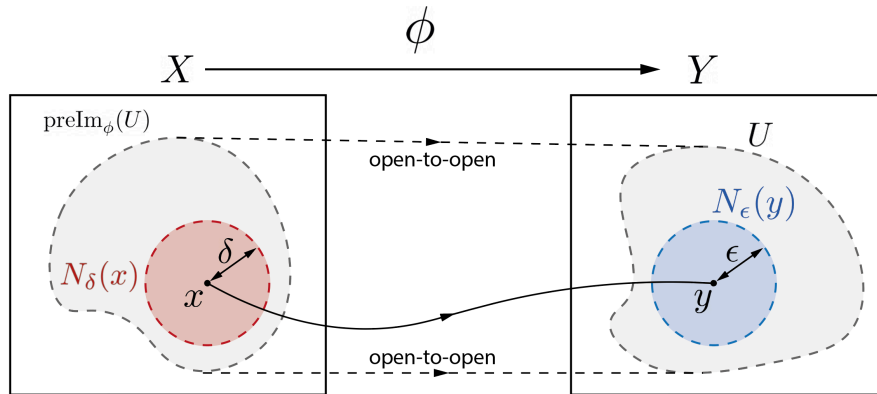
epsilon-delta definition of continuity we are familiar with from Real analysis.

> **Theorem** *(Continuity of $\mathbb{R}^n$ maps)*
> Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a map, where $\mathbb{R}^m$ and $\mathbb{R}^n$ are endowed with their standard topologies. The map ϕ is continuous according according to the topological definition *iff* it is continuous according to the epsilon-delta definition.

Proof. $\implies$ Suppose that the map f is topologically continuous. First, we claim that, given $\epsilon > 0$, the neighbourhood $N_\epsilon(f(x))$ is open

$$N_\epsilon(f(x)) = \{z \in \mathbb{R}^n : ||f(x) - z|| < \epsilon\} \tag{45.2.17}$$

Indeed given $z \in N_\epsilon(f(x))$ then we see that choosing $0 < d + ||z - f(x)|| < \epsilon$
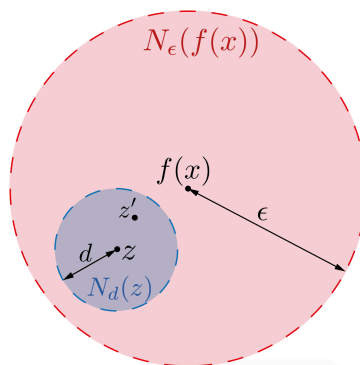
and letting $z' \in N_d(z)$ so that $||z' - z|| < d$ then:

$$||z' - f(x)|| < ||z' - z|| + ||z - f(x)|| < d + ||z - f(x)|| < \epsilon \tag{45.2.18}$$

implying that

$$N_d(z) = \{z' \in \mathbb{R}^n : ||z' - z|| < d\} \subseteq N_\epsilon(f(x)) \tag{45.2.19}$$

as desired. Now since $N_\epsilon(f(x))$ is an open set, its preimage $\text{preIm}_f(N_\epsilon(f(x)))$ must also be open, and since x is in this preimage (because $f(x) \in N_\epsilon(f(x))$) we have that $\exists \delta > 0$ such that

$$N_\delta(x) = \{y \in \mathbb{R}^m : ||x - y|| < \delta\} \subseteq \text{preIm}_f(N_\epsilon(f(x))) \implies f(N_\delta(x)) \subseteq N_\epsilon(f(x)) \tag{45.2.20}$$

that is, if $||x - y|| < \delta$ then $||f(y) - f(x)|| < \epsilon$ as desired.

$\impliedby$ Suppose that the map f is continuous according to the epsilon-delta definition, so that if $x \in \mathbb{R}^m$, then for all $\epsilon > 0$ there exists $\delta > 0$ such that $||x - y|| < \delta \implies ||f(x) - f(y)|| < \epsilon$. Consequently:

$$N_\delta(x) = \{y \in \mathbb{R}^m : ||x - y|| < \delta\} \subseteq \{y \in \mathbb{R}^m : ||f(x) - f(y)|| < \epsilon\} \tag{45.2.21}$$

Let $A \in \mathbb{R}^n$ be an open set in the standard topology, and let $z \in A$. Then there exists $\epsilon > 0$ such that $N_\epsilon(z) \subseteq A$. Consider $\text{preIm}_f(A)$, we claim this is open. Indeed let $x \in \text{preIm}_f(A)$. Then by the definition of $\epsilon - \delta$ continuity:

$$f(N_\delta(x)) = \{f(y) : ||x - y|| < \delta\} \subseteq \{y : ||f(x) - f(y)|| < \epsilon\} \subseteq \{z : ||f(x) - z|| < \epsilon\} = N_\epsilon(z) \subseteq A \tag{45.2.22}$$

implying that $N_\delta(x) \subseteq \text{preIm}(A)$, as desired. So if A is open then $\text{preIm}_f(A)$ is also open, as desired. ∎

> **Definition** *(Homeomorphism)*
> Let $\phi : X \to Y$ be a bijective map, with $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ being topological spaces. Then ϕ is a **homeomorphism** if
>   (a) $\phi : X \to Y$ is continuous
>   (b) $\phi^{-1} : Y \to X$ is continuous If such a ϕ exists then $(X, \mathcal{O}_X) \cong (Y, \mathcal{O}_Y)$, they are homeomorphic.

45.3 Topological manifolds

Now that we have defined a topological space we can add additional structure, namely that of a manifold which, as hinted earlier, is a topological space that locally looks Euclidean.

> **Definition** *(Topological manifold)*
> We define a d-dimensional **topological manifold** if it is a topological space $(X, \mathcal{O})$ if for all $x \in X$, there exists an open set $A \subseteq X$ with $x \in A$ such that there exists a map $\exists \phi : A \to \phi(A) \in \mathcal{O}_{\mathbb{R}^d}$ satisfying:
>   (i) ϕ is invertible
>   (ii) ϕ is continuous
>   (iii) ϕ^{-1} is continuous

where we choose the topology $\mathcal{O}$ on A and $\mathcal{O}_{\mathbb{R}^d}$ on $\mathbb{R}^d$ (i.e. A is homeomorphic to $\mathbb{R}^d$). The choice (A, ϕ) is defined as a **chart**.

What this means is that all points x in the topological manifold $(X, \mathcal{O})$ lie in some open set A that is homeomorphic to $\mathbb{R}^d$. Consider as an example a wire $M \subset \mathbb{R}^2$ that bifurcates into two branches. We
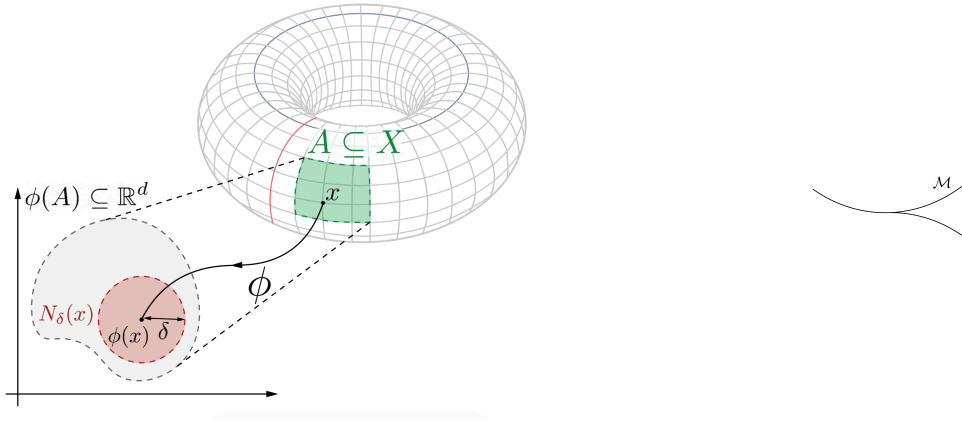
can embed the topology $\mathcal{O}_{\mathbb{R}^2}|_M$ on M creating a topological space. This however is not a topological manifold, since for the bifurcation point we cannot find an open set it lies in that can be mapped continuously bijectively to $\mathbb{R}^2$.

Definition *(Atlas)*

The set of charts $\mathcal{A} = \{(A_i, \phi_i : i \in S\}$ is an **atlas** of the topological space $(X, \mathcal{O})$ provided that:

$$X = \bigcup_{i \in S} A_i \tag{45.3.1}$$

Definition *(Chart components)*

For a given chart (A, ϕ) of an n-manifold we define its **chart components** ϕ^μ as the maps:

$$\phi^\mu : A \to \mathbb{R}^n \tag{45.3.2}$$
$$p \mapsto [\phi(p)]^\mu \tag{45.3.3}$$

which map $p \in A$ to the μth component of $\phi(p) \in \mathbb{R}^n$.
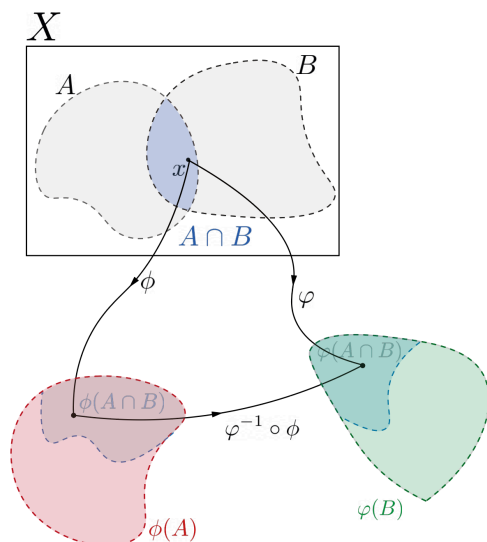
Consider two charts (A, ϕ) and (B, φ) with $A \cap B \neq \emptyset$. Then this implies that given some $x \in A \cap B$, it can be mapped to $\mathbb{R}^d$ using two charts, ϕ or φ. Can we find a way to map between $\phi(x)$ and $\varphi(x)$? Well it is easy to see from the commutative diagram that $\psi \equiv \varphi \circ \phi^{-1}$ does precisely this. We call ψ the chart transition map, it is akin to a change of basis map.

The reason topological manifold are so useful in physics is that it allows us to describe objects (for ex-

ample a curve $\gamma : \mathbb{R} \to X$) in the real world by looking at how they may be charted (e.g. consider the charts (ϕ, X) and (φ, X)). The real physical object is γ but physically we describe them using $\phi \circ \gamma$ or $\varphi \circ \gamma$. We must however make sure that whatever properties we are looking, it does not matter what charts we are looking at. The advantage of this approach is that we have all the tools of multivariable analysis in $\mathbb{R}^d$ at our disposal. The disadvantage is that we are blurring the lines between the real world and the charts we are studying them in by replacing γ, the real world object, with its mathematical description $\phi \circ \gamma$ within the chart (X, ϕ).

For example suppose $\phi \circ \gamma$ is continuous, how do we know that $\varphi \circ \gamma$ is also continuous? From the commutative diagram we see that $\varphi \circ \gamma = (\varphi \circ \phi^{-1}) \circ (\phi \circ \gamma)$ so $\varphi \circ \gamma$ is indeed continuous.

$$\varphi(X) \subseteq \mathbb{R}^d$$
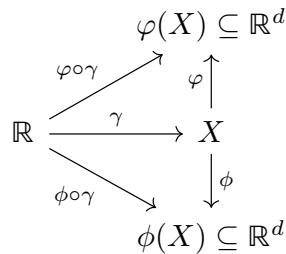
Can we say the same about differentiability? If we have that $\phi \circ \gamma$ is differentiable, we still do not know if this differentiability is shared in all charts since $\varphi \circ \phi^{-1}$ is not necessarily differentiable (they are continuous but can have kinks). We need to find some way to examine differentiability using charts.

45.4 Multilinear algebra

Recall the normal definition of vector spaces.

> **Definition** *(Vector space)*
> A vector space $(V, +, \bullet)$ over a field $\mathbb{F}$ is a set V and two maps:
>
> $$+ : V \times V \to V \tag{45.4.1}$$
> $$\bullet : \mathbb{F} \times V \to V \tag{45.4.2}$$
>
> known as addition and scalar multiplication such that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V, \alpha, \beta \in \mathbb{F}$:
>   (i) $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$
> (ii) $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$
> (iii) $\exists 0 \in V$ such that $\mathbf{v} + 0 = \mathbf{v}$
> (iv) $\exists (-\mathbf{v}) \in V$ such that $\mathbf{v} + (-\mathbf{v}) = 0$
>   (v) $\alpha \bullet (\beta \bullet \mathbf{v}) = (\alpha\beta) \bullet \mathbf{v}$
> (vi) $\alpha \bullet (\mathbf{v} + \mathbf{w}) = \alpha \bullet \mathbf{v} + \alpha \bullet \mathbf{w}$
> (vii) $(\alpha + \beta) \bullet \mathbf{v} = \alpha \bullet \mathbf{v} + \beta \bullet \mathbf{v}$
> (viii) given the identity element 1 of $\mathbb{F}$, then $1 \bullet \mathbf{v} = \mathbf{v}$

The element of a vector space is informally referred to as a **vector**. It is easy to see that the set $\mathcal{P}_n$ of all polynomial functions p up to order $n \in \mathbb{N}$ is indeed a vector space:

$$\mathcal{P}_n = \{p(x) = \sum_{m=0}^{n} p_m x^m : p_m \in \mathbb{R}\} \tag{45.4.3}$$

Definition *(Linear map)*
Let $(V, +_V, \bullet_V)$ and $(W, +_W, \bullet_W)$ be two vector spaces. Then the map $\phi : V \to W$ is a **linear map** if it is structure preserving:
  (i) $\phi(\mathbf{u} +_V \mathbf{v}) = \phi(\mathbf{u}) +_W \phi(\mathbf{v})$
  (ii) $\phi(\alpha \bullet_V \mathbf{v}) = \alpha \bullet_W \phi(\mathbf{v})$
If such a map between V and W exists then we write that $V \cong W$.

Definition *(Set of linear maps)*
We define the set of all linear maps between two vector spaces $(V, +_V, \bullet_V)$ and $(W, +_W, \bullet_W)$ (the latter defined over $\mathbb{F}_W$) as $\mathrm{Hom}(V, W)$. Together with the operations $\oplus$ and $\odot$ defined below this set becomes a vector space:

$$\oplus : \mathrm{Hom}(V, W) \times \mathrm{Hom}(V, W) \to \mathrm{Hom}(V, W) \tag{45.4.4}$$
$$(\phi, \varphi) \mapsto \phi \oplus \varphi \tag{45.4.5}$$

and

$$\odot : \mathbb{F}_W \times \mathrm{Hom}(V, W) \to \mathrm{Hom}(V, W) \tag{45.4.6}$$
$$(\alpha, \varphi) \mapsto \alpha \odot \varphi \tag{45.4.7}$$

where

$$(\phi \oplus \varphi)(\mathbf{v}) = \phi(\mathbf{v}) +_W \varphi(\mathbf{v}), \ \forall \mathbf{v} \in V \tag{45.4.8}$$
$$(\alpha \odot \varphi)(\mathbf{v}) = \alpha \bullet_W \phi(\mathbf{v}), \ \forall \mathbf{v} \in V, \ \forall \alpha \in \mathbb{F}_W \tag{45.4.9}$$

Definition *(Dual vector space)*
Given a vector space V defined over a field $\mathbb{F}$, then the set V^* is defined as:

$$V^* \equiv \{\phi : V \to \mathbb{F} : \phi \text{ is linear}\} = \mathrm{Hom}(V, \mathbb{F}) \tag{45.4.10}$$

Equipped with $\oplus$ and $\otimes$ defined previously then V^* is defined as the **dual vector space**.

Informally, the element $\phi \in V^*$ is referred to as a **covector**, for reasons that shall be clearer soon.

Definition *(Dual basis)*
Given a basis $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\} \subset V$ for V then the **dual basis** of the dual space V^* is defined as $\mathcal{B}^* = \{\epsilon^1, \epsilon^2, ..., \epsilon^n\} \subset V^*$ such that:

$$\epsilon^i(\mathbf{e}_j) = \delta_i^j, \ \forall i, j \tag{45.4.11}$$

For the case of $\mathcal{P}_n$, we can choose the basis $\mathcal{B} = \{e_i(x) = x^i : i = 0, 1, ..., n\}$. The dual basis $\mathcal{B}^* = \{\epsilon_i : i = 0, 1, ...n\}$ can be easily identified as:

$$\epsilon^i = \frac{1}{i!}\frac{d^i}{dx^i}\bigg|_{x=0} \tag{45.4.12}$$

Indeed one finds that

$$\epsilon^i(e_j) = \frac{1}{i!}\frac{d^i}{dx^i}(x^j)\bigg|_{x=0} = \begin{cases} 0, & i > j \\ 1, & i = j \\ 0, & i < j \end{cases} = \delta_{ij} \tag{45.4.13}$$

so the derivative operator may be regarded as a covector in the dual space of P_n. This will turn out to be a more general property of differential geometry.

> **Definition** *(Tensor)*
> Let $(V, +, \bullet)$ be a vector space defined over a field $\mathbb{F}$. Then an (r, s)-**tensor** T over V is a map:
>
> $$T : (V^*)^r \times V^s \mapsto \mathbb{F} \tag{45.4.14}$$
>
> that is linear in all its entries, thus a **multi-linear map**.

Consider for example a $(1, 1)$-tensor T. Then we must have that:

$$T(\phi + \varphi, \mathbf{v}) = T(\phi, \mathbf{v}) + T(\varphi, \mathbf{v}) \tag{45.4.15}$$
$$T(\alpha\phi, \mathbf{v}) = \alpha T(\phi, \mathbf{v}) \tag{45.4.16}$$
$$T(\phi, \mathbf{v} + \mathbf{w}) = T(\phi, \mathbf{v}) + T(\phi, \mathbf{w}) \tag{45.4.17}$$
$$T(\phi, \alpha\mathbf{v}) = \alpha T(\phi, \mathbf{v}) \tag{45.4.18}$$
$$\tag{45.4.19}$$

We can connect this to the typical view of a $(1, 1)$-tensor as a map that takes a vector and maps it to another vector. Indeed let us construct:

$$\phi_T : V \to V \tag{45.4.20}$$
$$\mathbf{v} \mapsto T(\cdot, \mathbf{v}) \tag{45.4.21}$$

which takes in $\mathbf{v}$ and maps it to another map $T(\cdot, \mathbf{v})$ which takes covectors to $\mathbb{F}$. But this map is of the type $V^* \to \mathbb{F}$ and thus $T(\cdot, \mathbf{v}) \in (V^*)^* = V$ provided $\dim V < \infty$. It follows that $T(\cdot, \mathbf{v})$ is a vector, so ϕ_T does indeed map a vector to vectors of V.

> **Theorem** *(Vectors vs Covectors)*
> Let $\phi \in V^*$ be a covector. Then by definition it must be a $(0, 1)$-tensor since $\phi : V \to \mathbb{F}$ is linear.
> Similarly, let $v \in V = (V^*)^*$ be a vector. Then by definition it must be a $(1, 0)$-tensor since $\mathbf{v} : V^* \to \mathbb{F}$.

> **Definition** *(Tensor components)*
> Let T be an (r, s)-tensor over a n-dimensional vector space V with basis $\mathcal{B} = \{\mathbf{e}_i : i =$

$0, 1, ...n\}$ and dual space V^* with dual basis $\mathcal{B}^* = \{\epsilon_i : i = 0, 1, ...n\}$. Then we define the components of T as:

$$T^{i_1...i_r}_{j_1...j_s} \equiv T(\epsilon^{i_1}, ..., \epsilon^{i_r}, \mathbf{e}_{j_1}, ..., \mathbf{e}_{j_s}) \tag{45.4.22}$$

so that:

$$T(\phi, \mathbf{v}) = \sum_{ij} \phi_i v^j T^i_j \tag{45.4.23}$$

where $\phi = \sum_i \phi_i \epsilon^i \in V^*$ and $\mathbf{v} = \sum_j v^j \mathbf{e}_j \in V$.

For example, given a $(1, 1)$-tensor T then $T^i_j = T(\epsilon^i, \mathbf{e}_j)$. Also, due to the linearity of the tensor map the tensor components are very useful because they allow us to expand:

$$T(\phi, \mathbf{v}) = T\left(\sum_i \phi_i \epsilon^i, \sum_j v^j \mathbf{e}_j \right) \tag{45.4.24}$$

$$= \sum_{ij} \phi_i v^j T(\epsilon^i, \mathbf{e}_j) \tag{45.4.25}$$

$$= \sum_{ij} \phi_i v^j T^i_j \tag{45.4.26}$$

Note the careful placements of subscripts and superscripts, where basis vectors are given subscripts and basis covectors are given superscripts. This ensures that multiply labelled indices always appear in an up-down arrangement. Moreover, if we assume that an index appearing both up and down is summed over, then we retrieve the **Einstein summation convention**. For example, (45.4.23) becomes

$$T(\phi, \mathbf{v}) = \phi_i v^j T^i_j \tag{45.4.27}$$

DG2 Differentiable manifolds

46.1 Differenti

We are now ready to tackle the problem mentioned in section 45.3 about the consistency in studying the differentiability of topological spaces. The key lies in choosing the charts on the topological space from a restricted set (which is still an atlas) so that all transition functions are differentiable. Manifolds on which this can be done are known as **differentiable manifolds**.

> **Definition** *(Compatibility)*
> Let $\square$ be some "property" (e.g. differentiability) of a topological space. Two charts (A, ϕ_A) and (B, ϕ_B) are $\square$-**compatible** if:
> (i) $A \cap B = \emptyset$ **or**
> (ii) $\phi_B \circ \phi_A^{-1}$ and $\phi_A \circ \phi_B^{-1}$ restricted to the domain $A \cap B$ are $\square$.
> Moreover, an atlas $\mathcal{A}_\square$ is an $\square$-compatible atlas if given any two charts in $\mathcal{A}$, they are $\square$-compatible.

If this happens then we can define a corresponding type of $\square$-manifold $(X, \mathcal{O}, \mathcal{A}_\square)$.

It can be shown that any C^k-atlas $\mathcal{A}_{C^k}$ contains a C^∞-atlas. This is a substantial result because for physical applications we can always consider C^∞-manifolds (smooth manifolds) without loss of generality.

> **Definition** *(Topological isomorphism)*
> Two topological spaces $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ are **topologically isomorphic** if there exists a bijective homomorphism between them $\phi : A \to B$.

> **Definition** *(Smooth map)*
> A map $\varphi : M \to N$ for two smooth manifolds $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$ is said to be **smooth** if for any two charts (X, ϕ_X) of $(M, \mathcal{O}_M)$ and (Y, ϕ_Y) of $(N, \mathcal{O}_N)$, the map $\phi_Y \circ \varphi \circ \phi_X^{-1}$ is smooth.

> **Definition** *(Diffeomorphism)*
> Two smooth manifolds $(M, \mathcal{O}_M, \mathcal{A}_M)$ and $(N, \mathcal{O}_N, \mathcal{A}_N)$ are **diffeomorphic** if there exists a bijection $\phi : M \to N$ such that both ϕ and ϕ^{-1} are smooth.

> **Definition** *(Smooth structure of topological manifolds)*
> In $d = 4$ dimensions, the number of smooth manifolds that can be made out of a given topological manifold (up to a diffeomorphism) is uncountably infinite.

We can now start talking about tangent vectors and spaces. Consider for example a sphere $X = S^2$ embedded in $\mathbb{R}^3$. We know that given a point p on this sphere, we can define a plane tangent to it at p. Any vector on this plane would then be a tangent vector. This visualization however relies on the fact that the sphere can be embedded in a larger Euclidean space, something that cannot be generally done. Thus we must develop a more abstract concept of tangent vector that does not live outside of the manifold it is tangent to.
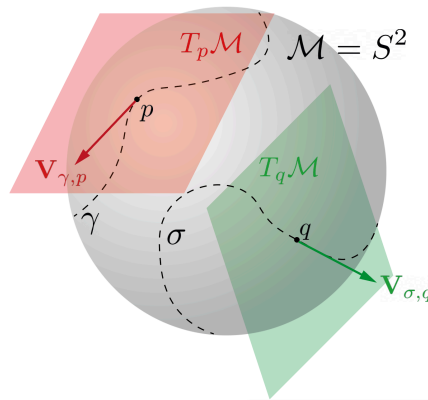
It turns out that the correct definition of a tangent vector is as a directional derivative.

> **Definition** *(Tangent vector)*
> Let $(X, \mathcal{O}, \mathcal{A})$ be a smooth manifold, and let $\gamma : \mathbb{R} \to X$ be a smooth parametrized curve, with $\gamma(t_p) = p \in X$. Then the **tangent vector** $\mathbf{V}_{\gamma,p}$ of γ at the point p is the linear map:
> $$\mathbf{V}_{\gamma,p} : C^\infty(X) \to \mathbb{R} \tag{46.1.1}$$
> $$f \mapsto (f \circ \gamma)'(t_p) \tag{46.1.2}$$
> where $C^\infty(X) = \{f : X \to \mathbb{R} | f \text{ is smooth}\}$ and $'$ represents differentiation with respect to the parametrization parameter t.

In other words, suppose f is some scalar field, for example temperature. $f \circ \gamma$ gives the temperature as we go along the path γ, and thus the velocity is defined as the rate of change of this temperature field as we run along γ. This is shown well in the following commutative diagram:

$$\mathbb{R} \xrightarrow{\ \gamma\ } X \xrightarrow{\ \phi\ } \mathbb{R}^d$$

$$f \circ \gamma \searrow \quad \downarrow f \quad \swarrow f \circ \phi^{-1}$$

$$\mathbb{R}$$
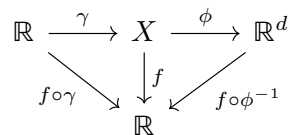
Differentiating f itself doesn't make a lot of sense since it is not a typical type of map that is encountered in analysis. We resolved this issue by looking at the parametrization of f, which is a "normal" map (i.e. from $\mathbb{R}$ to $\mathbb{R}$) and can thus be differentiated in the usual way.

We can write this result using the less rigorous, but more familiar differential notation

$$\mathbf{V}_{\gamma,p}(f) = \frac{d}{dt}f(\gamma(t))\bigg|_{t=t_p} \tag{46.1.3}$$

> **Definition** *(Tangent space)*
> For each point $p \in X$ we define the **tangent space to X at p** as the set of tangent vectors at p for all smooth curves γ on X.
>
> $$T_pX = \{\mathbf{V}_{\gamma,p} : \gamma \text{ is a smooth curve}\} \tag{46.1.4}$$

Note that the view that the tangent space is a plane works fine for classical geometry, but it requires the concept of an ambient space surrounding the topological space X. If for example we're considering the universe, then it doesn't make sense to refer to a tangent plane outside of it. The strength of this new definition of tangent space is that it makes no reference to things outside of X.

It is interesting to look at the structure of this tangent space, is it a vector space for example? We can equip the tangent space T_pX with the following operators:

$$\oplus : T_pX \times T_pX \to \mathrm{Hom}(C^\infty(X), \mathbb{R}) \tag{46.1.5}$$

$$(\mathbf{V}_{\gamma_1,p} \oplus \mathbf{V}_{\gamma_2,p})(f) \mapsto \mathbf{V}_{\gamma_1,p}(f) +_{\mathbb{R}} \mathbf{V}_{\gamma:2,p}(f) \tag{46.1.6}$$

and

$$\odot : \mathbb{R} \times T_pX \to \mathrm{Hom}(C^\infty(X), \mathbb{R}) \tag{46.1.7}$$

$$(\alpha \odot \mathbf{V}_{\gamma,p})(f) \mapsto \alpha \cdot_{\mathbb{R}} \mathbf{V}_{\gamma,p}(f), \; \forall \alpha \in \mathbb{r} \tag{46.1.8}$$

with which the vector space axioms are satisfied. It is yet not understood however if the closure relation is satisfied, that is if $\alpha \cdot_{\mathbb{R}} \mathbf{V}_{\gamma,p}(f)$ and $\mathbf{V}_{\gamma_1,p}(f) +_{\mathbb{R}} \mathbf{V}_{\gamma_2,p}(f)$ belong to the tangent space themselves, and are thus tangent vectors to some smooth curve on X.

Thus, suppose there is some curve σ so that $\mathbf{V}_{\sigma,p} = \alpha \cdot_{\mathbb{R}} \mathbf{V}_{\gamma,p}(f)$. Intuitively we should expect (from now on any $+, \cdot$ operation is implicitly defined on the field $\mathbb{R}$):

$$\sigma : \mathbb{R} \to X \tag{46.1.9}$$

$$t \mapsto \gamma(\alpha t + t_p) \tag{46.1.10}$$

to do the job. Indeed $\sigma(0) = \gamma(t_p) = p$ so this curve passes through the required point. Also, defining $\mu_\alpha : \lambda \to \alpha\lambda + \lambda_0$ then (we do the calculation two ways, first using the rigorous map notation and then using the familiar calculus notation):

$$\mathbf{V}_{\sigma,p} = (f \circ \sigma)'(0) = (f \circ \gamma \circ \mu_\alpha)'(0) = \alpha \cdot (f \circ \gamma)'(\lambda_0) = \alpha \cdot v_{\gamma,p} \tag{46.1.11}$$

$$\mathbf{V}_{\sigma,p} = \frac{d}{dt}f(\sigma(t))\bigg|_{t=0} = \frac{d}{dt}f(\gamma(at + t_p))\bigg|_{t=0} = \alpha \cdot \frac{d}{dt}f(\gamma(t))\bigg|_{t=t_p} = \alpha \cdot \mathbf{V}_{\gamma,p} \tag{46.1.12}$$

The proof for $\oplus$ is slightly more involved, but the calculation will be used heavily later. We choose a chart (A, ϕ) where $p \in A$, keeping in mind that if we make reference to a specific property of this chart that is not compatible with the chosen smooth atlas then the proof will be faulty. We define

the curve:

$$\sigma : \mathbb{R} \to X \tag{46.1.13}$$
$$t \mapsto \phi^{-1}((\phi \circ \gamma_1)(t + t_p) + (\phi \circ \gamma_2)(t + t'_p) - (\phi \circ \gamma)(t_p)) \tag{46.1.14}$$

where $\gamma_1(t_p) = \gamma_2(t'_p) = p$. We summarize the situation in the situation in the commutative diagrams below, :

where we defined $\mathbf{x}(t) = (\phi \circ \gamma)(t) \in \mathbb{R}^d$. Firstly, note that $\sigma(0) = \phi^{-1}(\phi(p) + \phi(p) - \phi(p)) = p$ so this curve does indeed contain p. Moreover:

$$\mathbf{V}_{\sigma,p}(f) = (f \circ \sigma)'(0) = ((f \circ \phi^{-1}) \circ (\phi \circ \sigma))'(0) \tag{46.1.15}$$
$$\mathbf{V}_{\sigma,p}(f) = \frac{d}{dt} f(\sigma(t))\bigg|_{t=0} = \frac{d}{dt}(f \circ \phi^{-1})(\phi(\sigma(t)))\bigg|_{t=0} \tag{46.1.16}$$

We recognize $\mathbf{x}(t) = (\phi \circ \sigma)(t)$ Note that $(f \circ \phi^{-1})$ is a map from $\mathbb{R}^d$ to $\mathbb{R}$ so we must use the multi-dimensional chain rule:

$$\mathbf{V}_{\sigma,p}(f) = (\phi \circ \sigma)^{\mu\prime}(0) \cdot (\partial_\mu (f \circ \phi^{-1}))(\mathbf{x}_p) \tag{46.1.17}$$
$$\mathbf{V}_{\sigma,p}(f) = \frac{d}{dt} \phi^\mu(\sigma(t))\bigg|_{t=0} \cdot \frac{\partial}{\partial x^\mu}(f \circ \phi^{-1})(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}_p} \tag{46.1.18}$$

We substitute (46.1.13) into the first term in the above:

$$(\phi \circ \sigma)^{\mu\prime}(0) = (\phi \circ \gamma_1)^{\mu\prime}(t_p) + (\phi \circ \gamma_2)^{\mu\prime}(t'_p) \tag{46.1.19}$$
$$\frac{d}{dt} \phi^\mu(\sigma(t))\bigg|_{t=0} = \frac{d}{dt} \phi^\mu(\gamma_1(t))\bigg|_{t=t_p} + \frac{d}{dt} \phi^\mu(\gamma_2(t))\bigg|_{t=t'_p} \tag{46.1.20}$$

so that:

$$\mathbf{V}_{\sigma,p}(f) = (\phi \circ \gamma_1)^{\mu\prime}(t_p) \cdot (\partial_\mu (f \circ \phi^{-1}))(\mathbf{x}_p) + (\phi \circ \gamma_2)^{\mu\prime}(t'_p) \cdot (\partial_\mu (f \circ \phi^{-1}))(\mathbf{x}_p) \tag{46.1.21}$$
$$\mathbf{V}_{\sigma,p}(f) = \frac{d}{dt} \phi^\mu(\gamma_1(t))\bigg|_{t=t_p} \cdot \frac{\partial}{\partial x^\mu}(f \circ \phi^{-1})(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}_p} + \frac{d}{dt} \phi^\mu(\gamma_2(t))\bigg|_{t=t'_p} \cdot \frac{\partial}{\partial x^\mu}(f \circ \phi^{-1})(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}_p} \tag{46.1.22}$$

Using the inverse chain rule this becomes:

$$\mathbf{V}_{\sigma,p}(f) = (f \circ \gamma)'(\lambda_0) + (f \circ \gamma')'(\lambda_1) = v_{\gamma,p}(f) + v_{\gamma',p}(f) \tag{46.1.23}$$

as desired. Note that we made no restriction on what chart to use, any chart would have worked as well as long as it was in a smooth atlas. To conclude, we have proven that:

Theorem *(The tangent space is a vector space)* The tangent space $T_p X$ is a vector space when equipped with the operations:

$$\oplus : T_p X \times T_p X \to T_p X \tag{46.1.24}$$

$$(v_{\gamma,p} \oplus v_{\gamma',p})(f) \mapsto v_{\gamma,p}(f) +_{\mathbb{R}} v_{\gamma',p}(f) \tag{46.1.25}$$

and

$$\odot : \mathbb{R} \times T_p X \to T_p X \tag{46.1.26}$$

$$(\alpha \odot v_{\gamma,p})(f) \mapsto \alpha \cdot_{\mathbb{R}} v_{\gamma,p}(f), \ \forall \alpha \in \mathbb{r} \tag{46.1.27}$$

Let $(A, \phi) \in \mathcal{A}_{C^\infty}$ be a chart in a smooth atlas. Let us define the curve:

$$\gamma : \mathbb{R} \to A \tag{46.1.28}$$

$$t \mapsto \gamma(t) \tag{46.1.29}$$

such that $\gamma(t_p) = p \in A$ and let $\mathbf{x}(t_p) = \phi(\gamma(t_p)) = \mathbf{x}_p$. Then we have that by the same calculation as before:

$$\mathbf{V}_{\gamma,p}(f) = (f \circ \gamma)'(t_p) = ((f \circ \phi^{-1}) \circ (\phi \circ \gamma))'(t_p) \tag{46.1.30}$$

$$= (\phi \circ \gamma)^{\mu'}(t_p) \cdot (\partial_\mu (f \circ \phi^{-1}))(\mathbf{x}_p) \tag{46.1.31}$$

$$= \frac{d}{dt} \phi^\mu(\gamma(t)) \bigg|_{t=t_p} \cdot \frac{\partial}{\partial x^\mu}(f \circ \phi^{-1})(\mathbf{x}) \bigg|_{\mathbf{x}=\mathbf{x}_p} \tag{46.1.32}$$

We now define the partial derivative for f as the usual partial derivative of the chart representation of f:

$$\frac{\partial}{\partial x^\mu}\bigg|_{t=t_p} f = \frac{\partial}{\partial x^\mu}(f \circ \gamma)(t)\bigg|_{t=t_p} = \frac{\partial}{\partial x^\mu}(f \circ \phi^{-1})(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}_p} \tag{46.1.33}$$

giving:

$$\mathbf{V}_{\gamma,p}(f) = \frac{dx^\mu(t)}{dt}\bigg|_{t=t_p} \cdot \frac{\partial}{\partial x^\mu}\bigg|_{t=t_p} f \tag{46.1.34}$$

We may thus define the components of the velocity vector:

Definition *(Velocity components)*
Let $(A, x) \in \mathcal{A}$ be a chart a smooth atlas. Let us define the curve:

$$\gamma : \mathbb{R} \to A \tag{46.1.35}$$

$$t \mapsto \gamma(t) \tag{46.1.36}$$

Then the velocity vector of γ at p is:

$$\mathbf{V}_{\gamma,p} = \frac{dx^\mu(t)}{dt}\bigg|_{t=t_p} \cdot \frac{\partial}{\partial x^\mu}\bigg|_{t=t_p} \tag{46.1.37}$$

where $\frac{dx^\mu}{dt}\big|_{t=t_p}$ are the **velocity components** and $\frac{\partial}{\partial x^\mu}\big|_{t=t_p}$ are the **basis elements of $T_p X$ induced by the chart** (A, x).

> **Definition** *(Chart induced basis for tangent space)*
> Given a chart (A, ϕ) with $p \in A$ in a smooth atlas then $\left\{\frac{\partial f}{\partial x^\mu}\big|_p\right\} \subset T_p X$ is a basis of $T_p A$.

Proof. We have seen that all velocity vectors $\mathbf{V}_{\gamma,p}$ may be written in this basis so $\mathrm{Span}\left\{\frac{\partial f}{\partial x^\mu}\big|_p\right\} = T_p A$. To check linear independence we let $f = \phi^\nu$:

$$V^\mu \frac{\partial}{\partial x^\mu}\bigg|_{t=t_p} \phi^\nu = 0 \implies V^\mu \frac{\partial}{\partial x^\mu}(\phi^\nu \circ \gamma)(t)\bigg|_{t=t_p} = 0 \implies V^\mu \frac{\partial x^\nu}{\partial x^\mu} = 0 \implies V^\mu \delta^\nu_\mu = V^\nu = 0$$

$$(46.1.38)$$

as desired. ∎

46.2 Vectors and 1-forms

It is interesting to see how the tangent vector components transform under a change of chart. Consider two overlapping charts (A, ϕ) and $(\tilde{A}, \varphi)$ with $p \in U \cap \tilde{U}$. Let $\mathbf{V} \in T_p M$ be a tangent vector at p along some smooth curve γ, and which can therefore be expanded in the two chart induced bases $\left\{\frac{\partial}{\partial x^\mu}\right\}$ and $\left\{\frac{\partial}{\partial \tilde{x}^\mu}\right\}$ [1]:

$$\mathbf{V} = V^\mu \frac{\partial}{\partial x^\mu} = \tilde{V}^\nu \frac{\partial}{\partial \tilde{x}^\nu} \tag{46.2.1}$$

We see that for $f \in C^\infty(X)$:

$$\frac{\partial}{\partial x^\mu} f = \partial_\mu (f \circ \phi^{-1})(\mathbf{x}) \tag{46.2.2}$$

$$= \partial_\mu ((f \circ \tilde{\phi}^{-1}) \circ (\tilde{\phi} \circ \phi^{-1}))(\mathbf{x}) \tag{46.2.3}$$

$$= (\partial_\mu (\tilde{\phi} \circ \phi^{-1}))^\nu(\mathbf{x}) \cdot \partial_\nu (f \circ \tilde{\phi}^{-1}))(\tilde{\mathbf{x}}) \tag{46.2.4}$$

$$= (\partial_\mu (\tilde{\phi}^\nu \circ \phi^{-1}))(\mathbf{x}) \cdot \partial_\nu (f \circ \tilde{\phi}^{-1}))(\tilde{\mathbf{x}}) \tag{46.2.5}$$

$$= \frac{\partial \tilde{x}^\nu}{\partial x^\mu} \frac{\partial}{\partial \tilde{x}^\nu} f \tag{46.2.6}$$

Consequently:

$$\tilde{V}^\nu = \frac{\partial \tilde{x}^\nu}{\partial x^\mu} V^\mu \tag{46.2.7}$$

so the tangent vector components V^ν transforms contravariantly, they are **contravariant vector components**.

Since the tangent space $T_p X$ is a vector space, we can consider its dual vector space.

> **Definition** *(Cotangent space)*
> Given the tangent space $T_p X$ for $p \in x$ then its **cotangent space** is defined as $T_p^* X \equiv \mathrm{Hom}(T_p X, \mathbb{R})$.

For example, let $f \in C^\infty(X)$ be a smooth function. We introduce the following map in the cotangent

---

[1] where $\tilde{x}^\mu = (\tilde{\phi} \circ \gamma)(t)$ and $x^\mu = (\phi \circ \gamma)(t)$

space:

$$df|_p : T_pX \to \mathbb{R} \tag{46.2.8}$$

$$\mathbf{V} \to \mathbf{V}(f) \tag{46.2.9}$$

known as the **gradient** of f at p. This is a $(0,1)$-tensor over T_pX, a covector as defined earlier, also known as a **1-form**. Its components in the chart induced basis are:

$$(df|_p)_\mu = df|_p(\mathbf{e}_\mu) = df|_p\left(\left.\frac{\partial}{\partial x^\mu}\right|_p\right) = \left.\frac{\partial}{\partial x^\mu}\right|_p f \tag{46.2.10}$$

giving the standard definition of gradient from vector calculus.

> **Theorem** *(Chart induced basis for cotangent space)*
> Let (A, φ) be a chart in the smooth atlas. Then $\{d\varphi^1|_p, d\varphi^2|_p, ..., d\varphi^d|_p\}$ is the dual basis of T_p^*X.

Proof. We have that:

$$d\varphi^\mu|_p(\mathbf{e}_\nu) = d\varphi^\mu|_p\left(\left.\frac{\partial}{\partial x^\nu}\right|_p\right) = \left.\frac{\partial}{\partial x^\nu}\right|_p \varphi^\mu = \left.\frac{\partial x^\mu}{\partial x^\nu}\right|_p = \delta^\mu_\nu \tag{46.2.11}$$

so $\{d\varphi^1|_p, d\varphi^2|_p, ..., d\varphi^d|_p\}$ is indeed the dual basis. $\blacksquare$

Again, much like with tangent vectors we are interested in how the components of a 1-form transform. We have that given $\boldsymbol{\omega} \in T_p^*X$ it can be written as:

$$\boldsymbol{\omega} = \omega_\mu dx^\mu|_p = \tilde{\omega}_\nu d\tilde{x}^\nu|_p \tag{46.2.12}$$

Therefore:

$$d\varphi^\mu|_p(\mathbf{V}) = \mathbf{V}(\varphi^\mu) = (\varphi^\mu \circ \gamma)'(t_p) \tag{46.2.13}$$

$$= ((\varphi^\mu \circ (\tilde{\varphi}^{-1})^\nu) \circ (\tilde{\varphi}^\nu \circ \gamma))'(t_p) \tag{46.2.14}$$

$$= (\tilde{\varphi}^\nu \circ \gamma)'(t_p) \cdot (\varphi^\mu \circ (\tilde{\varphi}^{-1})^\nu)'(\tilde{x}^\nu_p) \tag{46.2.15}$$

$$= \mathbf{V}(\tilde{\varphi}^\nu) \cdot \tilde{\partial}_\nu(\varphi^\mu \circ (\tilde{\varphi}^{-1})^\nu)(\tilde{x}^\nu_p) \tag{46.2.16}$$

and since:

$$\tilde{\partial}_\nu(\varphi^\mu \circ (\tilde{\varphi}^{-1})^\nu)(x^\nu_p) = \left.\frac{\partial}{\partial \tilde{x}^\nu}\varphi^\mu \circ (\tilde{\varphi}^{-1})^\nu\right|_{\tilde{x}^\nu_p} = \frac{\partial\varphi^\mu(p)}{\partial \tilde{x}^\nu} = \frac{\partial x^\mu}{\partial \tilde{x}^\nu} \tag{46.2.17}$$

we get that:

$$d\varphi^\mu|_p(\mathbf{V}) = \frac{\partial x^\mu}{\partial \tilde{x}^\nu}d\tilde{x}^\nu|_p(\mathbf{V}) \tag{46.2.18}$$

and hence the covector components of $\boldsymbol{\omega}$ transform as:

$$\tilde{\omega}_\nu = \frac{\partial x^\mu}{\partial \tilde{x}^\nu}\omega_\mu \tag{46.2.19}$$

This transformation law is in sharp contrast to (46.2.7), the transforming matrices are inverses of

each other:

$$\tilde{\mathbf{V}} = \mathsf{J}\mathbf{V}, \; \tilde{\boldsymbol{\omega}} = \mathsf{J}^{-1}\boldsymbol{\omega} \tag{46.2.20}$$

where J is known as the **Jacobian matrix**.

Definition *(Push-forward)*

Let $\phi : X \to Y$ be a smooth map between two smooth manifolds. Then we define the **push-forward** ϕ_* of ϕ at the point $p \in X$ is the linear map:

$$\phi_* : T_p X \to T_{\phi(p)} Y \tag{46.2.21}$$
$$\mathbf{V} \mapsto \phi_*(\mathbf{V}) \tag{46.2.22}$$

where we define for any $f \in C^\infty(X)$

$$\phi_*(\mathbf{V})(f) = \mathbf{V}(f \circ \phi) \tag{46.2.23}$$

We call ϕ_* the **derivative** of ϕ.

Note that $f \circ \phi \in C^\infty(X)$. Also note that the definition of $\phi^*(\mathbf{V})f$ is forced upon us since there are no other linear maps from the two tangent spaces. Let's apply this definition to the tangent vector
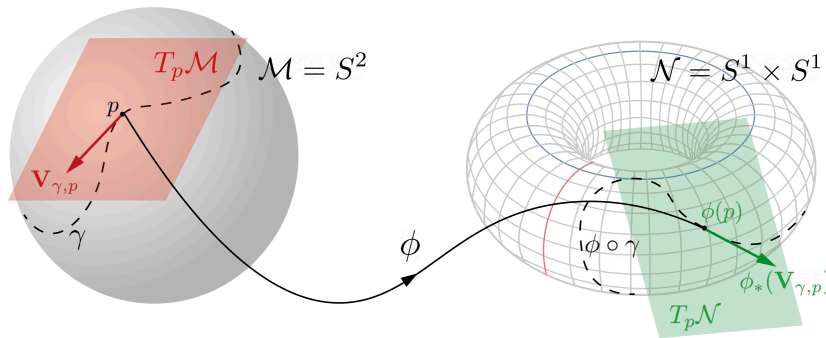
$\mathbf{V}_{\gamma,p} \in T_p X$ of some curve γ at $p \in X$. Then:

$$\phi_*(\mathbf{V}_{\gamma,p})(f) = \mathbf{V}_{\gamma,p}(f \circ \phi) = ((f \circ \phi) \circ \gamma)'(t_p) = (f \circ (\phi \circ \gamma))(t_p) \tag{46.2.24}$$

but the tangent vector to $\phi \circ \gamma$ at $\phi(p)$ is:

$$\mathbf{V}_{\phi \circ \gamma, \phi(p)}(f) = (f \circ (\phi \circ \gamma))(t_p) \tag{46.2.25}$$

since $(\phi \circ \gamma)(t_p) = \phi(p)$.

Therefore, the push forward map maps a tangent vector at a point along some curve to the tangent vector at the image of that point along the image of that curve:

$$\phi_*(\mathbf{V}_{\gamma,p}) = \mathbf{V}_{\phi \circ \gamma, \phi(p)} \tag{46.2.26}$$

Definition *(Pull-back)*

Let $\phi : X \rightarrow Y$ be a smooth map between two smooth manifolds. Then we define the **pull-back** ϕ^* of ϕ at the point $\phi(p) \in Y$ is the linear map:

$$\phi^* : T_p^* X \leftarrow T_{\phi(p)}^* Y \tag{46.2.27}$$

$$\phi^*(\boldsymbol{\omega}) \hookleftarrow \boldsymbol{\omega} \tag{46.2.28}$$

where we define for any $\boldsymbol{\omega} \in T_{\phi(p)}^* X, \mathbf{V} \in T_p X$:

$$\phi^*(\boldsymbol{\omega})(\mathbf{V}) = \boldsymbol{\omega}(\phi_*(\mathbf{V})) \tag{46.2.29}$$

Working in a local basis y^μ for Y and ∂_ν for X we find that

$$(\phi^*)^\mu{}_\nu = \phi^*(dy^\mu)(\partial_\nu) = dy^\mu(\phi_*(\partial_\nu)) = (\phi_*)^\mu{}_\nu \tag{46.2.30}$$

so the components of the push-forwards and pull-backs are the same! These can be expressed more simply as:

$$dy^\mu(\phi_*(\partial_\nu)) = \phi_*(\partial_\nu)y^\mu = \frac{\partial}{\partial x^\nu}(y^\mu \circ \phi) = \frac{\partial X^\mu}{\partial x^\nu} \tag{46.2.31}$$

where we defined $X^\mu = y^\mu \circ \phi$.

46.3 Interlude: Embeddings and immersions

The concept of an embedding/immersion is quite intuitive: we want some smooth manifold to "sit" within some $\mathbb{R}^n$. As we shall learn, there are two known ways a manifold can sit, one way gives an immersion while the other gives an embedding.

Definition *(Immersion)*

Let $\phi : M \rightarrow \mathbb{R}^n$ be a smooth map. Then ϕ is an **immersion** of M into $\mathbb{R}^n$ if ϕ_* is injective at any $p \in M$.

Consider the map $\phi : S^1 \rightarrow \mathbb{R}^2$ as shown below.

Even though ϕ is not injective, it is easy to see that ϕ_* is injective so this is indeed an immersion.

Definition *(Embedding)*

Let $\phi : M \rightarrow \mathbb{R}^n$ be an immersion. Then it is an **embedding** if $\phi(M)$ is homeomorphic to N.

We see that ϕ defined previously is not an embedding since $\phi(S^1)$ is not a manifold and cannot thus be homeomorphic.

> **Theorem** *(Whitney embedding theorems)*
> Any smooth manifold M can be:
> (i) embedded in $\mathbb{R}^{2\dim M}$.
> (ii) immersed in $\mathbb{R}^{2\dim M - 1}$

There is a weaker form of Whitney's theorem that is quite amusing:

> **Theorem** *(Immersion theorem)*
> Any smooth manifold M can be immersed in $\mathbb{R}^{2\dim M - a(\dim M)}$ where $a(\dim M)$ is the number of 1s appearing in the binary form of $\dim M$.

46.4 The tangent bundle

We begin by introducing the concept of bundles, fibres and sections that will be fundamental in defining tensor fields.

> **Definition** *(Bundle)*
> A **bundle** is a triple (X, Y, π) where X, Y are topological manifolds and π is a continuous surjective map from X to Y.

> **Definition** *(Fibre bundle)*
> Let $p \in Y$ and let (X, Y, π) be a bundle. Then $\mathrm{preIm}_\pi(p)$ is the **fibre** at point p. Moreover, let Z be another manifold. If for all $p \in Y$, $\mathrm{preIm}_\pi(p)$ is homeomorphic to some manifold M, then (X, Y, π) is a **fibre bundle**.

> **Definition** *(Section)*
> Let (X, Y, π) be a bundle. Then $\sigma : Y \to X$ is a **section** of (X, Y, π) if $\pi \circ \sigma = \mathrm{id}_M$.

We see that the section σ maps $p \in Y$ to some $\sigma(p) \in \mathrm{preIm}_\pi(p)$ so that π maps it back to itself.
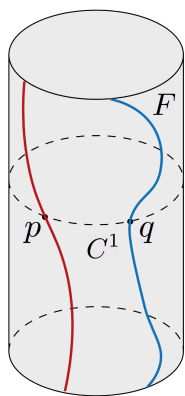
We can visualize bundles, fibres and sections as in Figure **??**. Here we consider the surjective map $\pi : F \to \mathcal{C}^1$ mapping the cylindrical manifold F to the ring manifold $\mathcal{C}^1$. The green line crossing p can be viewed as the fibre at p since it is π's preimage at that point. For the same reason the blue line is the fibre at q. A section σ would then map p to a point on the green line.

A special type of bundle is the tangent bundle, defined as the disjoint union of the tangent spaces at all points on a manifold.

> **Definition** *(Tangent bundle)*
> Let X be a smooth manifold. Then its tangent bundle TX is defined as:
> $$TX = \coprod_{p \in X} T_p X \tag{46.4.1}$$

Similarly we may define the projective map of the tangent bundle as follows.

> **Definition** *(Bundle projection)*
> Let TM be the tangent bundle of some smooth manifold X. Then the bundle projection of the tangent bundle is
>
> $$\pi : TX \to X \qquad\qquad (46.4.2)$$
> $$\mathbf{V} \mapsto p \qquad\qquad (46.4.3)$$
>
> where p is the point in X such that $\mathbf{V} \in T_p X$.

Note that since the tangent spaces $T_p X$ are disjoint, there exists only one such p. Note however that to claim that this indeed a bundle would require TX to be itself a manifold. We now prove this.

We can make TX a smooth manifold by using the smooth atlas $\mathcal{A}_X$ on X. Take some chart $(U, \phi) \in \mathcal{A}_X$ and consider $(\mathrm{preIm}_\pi(U), \eta)$ where η is some chart map to $\mathbb{R}^{2\dim X}$.

Let $\mathbf{V} \in TX$. Then we can define:

$$\eta(\mathbf{V}) = (\phi^1 \pi(\mathbf{V}), ..., \phi^{\dim X} \pi(\mathbf{V}), ...) \qquad\qquad (46.4.4)$$

where we are missing $\dim X$ coordinates that specify how "far" $\mathbf{V}$ is from U. The basis of $T_{\pi(\phi)} X$ (this tangent space contains $\mathbf{V}$) is

$$\left\{ \left. \frac{\partial}{\partial x^\mu} \right|_{\pi(x)} \right\} \qquad\qquad (46.4.5)$$

which can be used to expand $\mathbf{V}$ as:

$$\mathbf{V} = V^\mu \left. \frac{\partial}{\partial x^\mu} \right|_{\pi(x)} \qquad\qquad (46.4.6)$$

We define the missing components as V^μ:

$$\eta(\mathbf{V}) = (\phi^1 \pi(\mathbf{V}), ..., \phi^{\dim X} \pi(\mathbf{V}), V^1, ..., V^{\dim X}) \qquad\qquad (46.4.7)$$

It is easy to verify that any two such charts are C^∞-compatible making the tangent bundle a smooth manifold. Now that we know that the tangent bundle is a smooth manifold, we can provide a rigorous definition for tensor fields (that is not simply "something that assigns a tensor at every point on a manifold")..
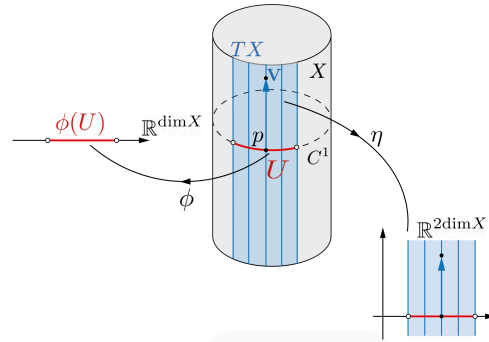
46.5 Tensor fields

> **Definition** *(Vector field)*
> Let X be a smooth manifold and let TX be its tangent bundle with bundle projection π. Then a **vector field** is a smooth section of TX. We further define the set of all vector fields $\Gamma(TX)$.

Intuitively, this means that a vector field smoothly maps to each point p in the manifold X a tangent vector in its fibre $\mathbf{V} \in T_p X$. Note that had TX not been a smooth manifold we would not have been able to define the vector field as a smooth section.

We can equip $\Gamma(TX)$ with two operations:

$$\oplus : \Gamma(TX) \times \Gamma(TX) \to \Gamma(TX) \tag{46.5.1}$$

$$(\sigma, \tau) \mapsto \sigma \oplus \tau \tag{46.5.2}$$

where

$$(\sigma \oplus \tau)(p) = \sigma(p) + \tau(p) \tag{46.5.3}$$

and

$$\odot : C^\infty(X) \times \Gamma(TX) \to \Gamma(TX) \tag{46.5.4}$$

$$(G, \sigma) \mapsto f \odot \sigma \tag{46.5.5}$$

where

$$(f \odot \sigma)(p) = f(p)\dot{\sigma}(p) \tag{46.5.6}$$

It is easy to check that these two operations satisfy the vector space axioms.

Note that $(C^\infty(X), +, \cdot)$ where $\cdot$ is the familiar scalar multiplication is a vector space over $\mathbb{R}$. In contrast, $(C^\infty(X), +, \dot{)}$ is not a vector space but has the structure of a ring since if a function in $C^\infty(X)$ has a zero somewhere then it will not have a multiplicative inverse function. Therefore, $\Gamma(TX)$ satisfies the typical vector space axioms when equipped with $\oplus$ and $\odot$ only that it is defined over a ring rather than a field. Such spaces are known as ring modules, as we shall now see.

> **Definition** *(Ring)*
> A **ring** is a triplet $(R, +, \bullet)$ satisfying $C^+ A^+ N^+ I^+ (C^\bullet) A^\bullet (N^\bullet)(I^\bullet)D$ where $\bullet$-commutativity only applies to commutative rings, $\bullet$-neutrality only applies to unital element, and $\bullet$-invertibility only applies to division rings. Commutative, unital division rings are fields.

Consider once again $(C^\infty(X), +, \bullet)$. It is easy to see that it is a commutative, unital ring.

> **Definition** *(Ring module)*
> Let $(M, \oplus, \odot)$ defined over a ring R equipped with two operations:
>
> $$\oplus : M \times M \to M \tag{46.5.7}$$
>
> $$\odot : R \times M \to M \tag{46.5.8}$$
>
> satisfying the vector space axioms is an R-module (ring-module).

Therefore, $(\Gamma(TX), \oplus, \odot)$ is a $C^\infty(X)$-module.

Note that this definition implies that modules don't necessarily have a basis (but division modules do).

Consider for example the sphere S^2 and let $\mathbf{v} \in \Gamma(TS^2)$ be a vector field on it. There is an important result in algebraic topology, known as the **hairy ball theorem**, that states that one cannot have a

vector field on a sphere that is smooth and non-zero everywhere. This is a problem because it means that one cannot produce a set of linearly independent vector fields on S^2.

Generally the following result holds:

> **Theorem** *(Division ring basis)*
> Let D be a division ring and let V be a D-module. Then V has a basis.

> **Definition** *(R-module terminology)*
> A **free module** is a module over a ring that possesses a basis. A **projective module** is a module Γ over a ring R such that there exists another R-module Q whose direct sum gives a free module:
> $$\exists Q \text{ s.t. } \Gamma \oplus Q = F \tag{46.5.9}$$

> **Theorem** *(Serre, Swan et al)*
> The set of all smooth functions of a vector fibre bundle over a smooth manifold X is finitely generated projective $C^\infty(X)$-module $\Gamma(E)$.

> **Theorem** *(Homomorphism space is finitely generated)*
> Let P, Q be finitely generated (projective) modules over a commutative ring R. Then:
> $$\operatorname{Hom}_R(P, Q) = \{\varphi : P \to Q, \ \varphi \text{ is linear }\} \tag{46.5.10}$$
> equipped with $\oplus$ and $\odot$ is a finitely generated (projective) module.

This result in particular shows that:

$$\operatorname{Hom}_{C^\infty(X)}(\Gamma(TX), C^\infty(X)) = \Gamma(TX)^* = \Gamma(T^*X) \tag{46.5.11}$$

is a module.

> **Definition** *(Tensor field)*
> An (r, s)-tensor field T on a smooth manifold X is a $C^\infty(X)$-multilinear map:
> $$T : \underbrace{\Gamma(T^*X) \times ... \times \Gamma(T^*X)}_{r} \times \underbrace{\Gamma(TX) \times ... \times \Gamma(TX)}_{s} \to C^\infty(X) \tag{46.5.12}$$
> and the space of all (r, s)-tensor fields is denoted by $\mathcal{T}_s^r$.

We can define the differential operator the same way for tensor fields:

$$df : \mathcal{T}_s^r \to C^\infty(M) \tag{46.5.13}$$
$$T \mapsto T(\boldsymbol{\omega}_1, ..., \mathbf{V}_1, ...) \tag{46.5.14}$$

DG3 Differentiable forms

47.1 Differentiable forms

> **Definition** *(Differentiable form)*
> Let X be a smooth manifold, then a **differentiable n-form** is a $(0, n)$-tensor field ω:
>
> $$\omega : \underbrace{\Gamma(TX) \times ... \times \Gamma(TX)}_{n} \to C^\infty(X) \tag{47.1.1}$$
>
> that is totally anti-symmetric:
>
> $$\omega(\mathbf{V}_1, ..., \mathbf{V}_n) = \text{sgn}(\pi)\omega(\mathbf{V}_{\pi(1)}, ..., \mathbf{V}_{\pi(n)}), \ \forall \pi \in S_n, \mathbf{V}_i \in \Gamma(TM) \tag{47.1.2}$$
>
> where S_n is the symmetric group of order n. We denote the set of all n-forms on X by $\Omega^n(X)$.

It is important to note that $\Omega^0(X) = C^\infty(X)$ and $\Omega^1(X) = \Gamma(T^*X)$.

> **Definition** *(Wedge product)*
> The **wedge product** is defined as:
>
> $$\wedge : \Omega^n(X) \times \Omega^m(X) \to \Omega^{n+m}(X) \tag{47.1.3}$$
>
> $$(\omega_1, \omega_2) \mapsto \omega_1 \tau \omega_2 \tag{47.1.4}$$
>
> where:
>
> $$(\omega_1 \wedge \omega_2)(\mathbf{V}_1, ..., \mathbf{V}_{n+m}) = \frac{1}{n!} \frac{1}{m!} \sum_{\pi \in S_{n+m}} \text{sgn}(\pi)(\omega_1 \otimes \omega_2)(\mathbf{V}_{\pi(1)}, ..., \mathbf{V}_{\pi(n+m)}) \tag{47.1.5}$$

For example, if $\omega_1, \omega_2 \in \Omega^1(X)$ are one forms then:

$$\omega_1 \wedge \omega_2 = \omega_1 \otimes \omega_2 - \omega_2 \otimes \omega_1 \tag{47.1.6}$$

> **Theorem** *(Basis for $\Gamma^n(X)$)*
> Let X be a smooth n-manifold with a chart (U, φ). Taking the wedge product of the chart

induced covariant basis $\{d\varphi^\mu\}$:

$$\{d\varphi^{\mu_1} \wedge d\varphi^{\mu_2}... \wedge d\varphi^{\mu_k} : 1 \le \mu_1 < ... < \mu_k \le n\} \tag{47.1.7}$$

gives an ordered basis for $\Omega^k(X)$.

Proof. We begin by proving that this basis spans $\Omega^n(X)$. Let $\omega \in \Omega^k(X)$ and define:

$$\omega_{\mu_1...\mu_n} = \omega(\mathbf{e}_{\mu_1}, ..., \mathbf{e}_{\mu_k}) \tag{47.1.8}$$

where $\{\mathbf{e}_\mu\}$ is the chart-induced contravariant basis. We claim that

$$\omega = \omega_{\mu_1...\mu_k} d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k} \tag{47.1.9}$$

To see why this holds, note that since ω is multi-linear we just need to look at the action of the candidate basis on $(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k})$:

$$\omega_{\mu_1...\mu_k}(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k})(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k}) \tag{47.1.10}$$

Due to the duality of $\mathbf{e}_{\nu_j}$ and $d\varphi^{\mu_i}$, the contributing terms will be delta functions $\delta^{\mu_i}_{\nu_j}$. It is helpful to consider the $k = 2$ case:

$$\omega_{\alpha\beta}(d\varphi^\alpha \wedge d\varphi^\beta)(\mathbf{e}_\mu, \mathbf{e}_\nu) = \omega_{\alpha\beta}(\delta^\alpha_\mu \delta^\beta_\nu - \delta^\alpha_\nu \delta^\beta_\mu) = \frac{1}{2}(\omega_{\mu\nu} - \omega_{\nu\mu}) = \omega_{\mu\nu} \tag{47.1.11}$$

where the $\frac{1}{2}$ comes because the sum is restricted to $\alpha < \beta$. Note that these Kronecker-deltas act independently of each other, that is given a (μ, ν) then only one of these products of δ will be non-zero, yielding:

$$\omega_{12}(d\varphi^1 \wedge d\varphi^2)(\mathbf{e}_1, \mathbf{e}_2) = \omega_{12} \tag{47.1.12}$$
$$\omega_{12}(d\varphi^1 \wedge d\varphi^2)(\mathbf{e}_2, \mathbf{e}_1) = -\omega_{12} = \omega_{21} \tag{47.1.13}$$

as desired. Hence more generally we have that:

$$(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k})(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k}) = \sum_{\pi \in S_k} \text{sgn}(\pi)(d\varphi^{\mu_1} \otimes ... \otimes d\varphi^{\mu_k})(\mathbf{e}_{\pi(\nu_1)}, ..., \mathbf{e}_{\pi(\nu_k)}) \tag{47.1.14}$$

$$= \sum_{\pi \in S_k} \text{sgn}(\pi)\delta^{\mu_1}_{\pi(\nu_1)}...\delta^{\mu_k}_{\pi(\nu_k)} \tag{47.1.15}$$

When we contract this with $\omega_{\mu_1...\mu_k}$, the sum over $\mu_1...\mu_k$ is restricted to $1 \le \mu_1 < \mu_2 < ... < \mu_{k-1} < \mu_k \le n$, giving $\frac{1}{k!}$ many terms as the unrestricted sum[1]. Consequently

$$\omega_{\mu_1...\mu_k}(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k})(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k}) = \frac{1}{k!} \sum_{\pi \in S_k} \text{sgn}(\pi)\omega_{\mu_1...\mu_k}\delta^{\mu_1}_{\pi(\nu_1)}...\delta^{\mu_k}_{\pi(\nu_k)} \tag{47.1.16}$$

$$= \frac{1}{k!} \sum_{\pi \in S_k} \text{sgn}(\pi)\omega_{\pi(\nu_1)...\pi(\nu_k)} = \omega(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k}) \tag{47.1.17}$$

---

[1]indeed if $\omega_{\mu_1...\mu_k}(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k})$ appears in the restricted sum, then the unrestricted sum will contain $\omega_{\mu_1...\mu_k}(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k})$ plus all $k! - 1$ permutations of the indices. Since ω is anti-symmetric and so is the wedge product, permuting indices gives no sign change so the unrestricted sum will indeed be $k!$ times larger.

where by anti-symmetry $\text{sgn}(\pi)\omega_{\pi(\nu_1)...\pi(\nu_k)} = \omega(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k})$. Thus every form $\omega \in \Omega^k(X)$ can be expanded locally in the basis $\{d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k}\}$.

Linear independence immediately follows from the fact that if:

$$\omega_{\mu_1...\mu_k}(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k}) = 0 \qquad (47.1.18)$$

then $\omega_{\mu_1...\mu_k}(d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_k})(\mathbf{e}_{\nu_1}, ..., \mathbf{e}_{\nu_k}) = \omega_{\nu_1...\nu_k} = 0$ as desired. ∎

Since we can define bases for differential forms we should be interested in their transformation properties.

Theorem *(Transformation of n-form)*
Let M be a smooth n-manifold and let ω, τ be two differentiable 1-forms on M related by:

$$\omega^\mu = A^\mu{}_\nu \tau^\nu \qquad (47.1.19)$$

Then:

$$\omega^{\mu_1} \wedge ... \wedge \omega^{\mu_n} = (\det A)\tau^{\nu_1} \wedge ... \wedge \tau^{\nu_n} \qquad (47.1.20)$$

Proof. We have that:

$$\omega^{\mu_1} \wedge ... \wedge \omega^{\mu_n} = (A^{\mu_1}_{\nu_1}\tau^{\nu_1}) \wedge ... \wedge (A^{\mu_n}_{\nu_n}\tau^{\nu_n}) \qquad (47.1.21)$$

$$= \sum_{\sigma \in S_n} A^{\mu_1}_{\sigma(\nu_1)}...A^{\mu_n}_{\sigma(\nu_n)}\tau^{\sigma(\nu_1)} \wedge ... \wedge \tau^{\sigma(\nu_n)} \qquad (47.1.22)$$

We can now rearrange the $\tau^{\sigma n u_1}$ so that the run from τ^{ν_1} to τ^{ν_n} which of course comes with a $\text{sgn}(\sigma)$ factor:

$$\omega^{\mu_1} \wedge ... \wedge \omega^{\mu_n} = \sum_{\sigma \in S_n} \text{sgn}(\sigma)A^{\mu_1}_{\sigma(\nu_1)}...A^{\mu_n}_{\sigma(\nu_n)}\tau^{\nu_1} \wedge ... \wedge \tau^{\nu_n} \qquad (47.1.23)$$

$$= (\det A)\tau^{\nu_1} \wedge ... \wedge \tau^{\nu_n} \qquad (47.1.24)$$

as desired. ∎

An immediate consequence of this theorem is that:

$$\boxed{d\varphi^{\mu_1} \wedge ... \wedge d\varphi^{\mu_n} = \det\left(\frac{\partial x^\mu}{\partial \tilde{x}^\nu}\right)d\tilde{\varphi}^{\nu_1} \wedge ... \wedge d\tilde{\varphi}^{\nu_n}} \qquad (47.1.25)$$

which will be a fundamental result in generalizing our notion of integration on manifolds. Of course, this also implies that when a n-form ω is expanded in these two bases as [2]:

$$\omega = a\,d\varphi^1 \wedge ... \wedge d\varphi^n = \tilde{a}\,d\tilde{\varphi}^1 \wedge ... \wedge d\tilde{\varphi}^n \qquad (47.1.26)$$

---

[2]note that there is only one way to order the differentials so that the superscripts are increasing

the components transform according to

$$a' = a \det\left(\frac{\partial x^\mu}{\partial \widetilde{x}^\nu}\right)$$
(47.1.27)

It will also be useful to extend our notion of push-forward and pull-back to n-forms.

Definition *(Pull-back)*
Let $\omega \in \Omega^n(Y)$ and let $\phi : X \to Y$ be smooth. Then we define the pull-back of ω as:

$$\Phi^*(\omega)(\mathbf{V}_1, ..., \mathbf{V}_n) = \omega(\Phi_*(\mathbf{V}_1), ..., \Phi_*(\mathbf{V}_n))$$
(47.1.28)

Theorem *(Pull-back of wedge product)*
The pull-back distributes over the wedge product:

$$\Phi^*(\omega_1 \wedge \omega_2) = \Phi^*(\omega_1) \wedge \Phi^*(\omega_2)$$
(47.1.29)

Proof. It is easy to verify that if ω_1 is a p-form and ω_2 is a q-form and $n = p + q$ then:

$$\Phi^*(\omega_1 \wedge \omega_2)(\mathbf{V}_1, ..., \mathbf{V}_n) = (\omega_1 \wedge \omega_2)(\Phi_*(\mathbf{V}_1), ..., \Phi_*(\mathbf{V}_n))$$
(47.1.30)

$$= \frac{1}{n!}\frac{1}{m!} \sum_{\pi \in S_{n+m}} \text{sgn}(\pi)(\omega_1 \otimes \omega_2)(\Phi_*(\mathbf{V}_{\pi(1)}), ..., \Phi_*(\mathbf{V}_{\pi(n)}))$$
(47.1.31)

$$= \frac{1}{n!}\frac{1}{m!} \sum_{\pi \in S_{n+m}} \text{sgn}(\pi)(\Phi^*(\omega_1) \otimes \Phi^*(\omega_2))(\mathbf{V}_{\pi(1)}), ..., \mathbf{V}_{\pi(n)})$$
(47.1.32)

$$= \Phi^*(\omega_1) \wedge \Phi^*(\omega_2)$$
(47.1.33)

as desired. ∎

It would be nice to have a space that is closed under the wedge product.

Define the following space:

$$\Omega(X) = \bigoplus_{i=0}^{\dim X} \Omega^i(X)$$
(47.1.34)

Then $(\Omega(X), +, \cdot, \wedge)$ is known as the **Grassman algebra** on X, and:

$$\wedge : \Omega(X) \times \Omega(X) \to \Omega(X)$$
(47.1.35)

For example, let $\sigma = \omega_1 + \omega_2$ where $\omega_1 \in \Omega^1(X)$ and $\omega_2 \in \Omega^2(X)$, and let $\tau \in \Omega^n(X)$. Then:

$$\tau \wedge \sigma = \tau \wedge (\omega_1 + \omega_2) := \tau \wedge \omega_1 + \tau \wedge \omega_2$$
(47.1.36)

where since $\tau \wedge \omega_1 \in \Omega^{n+1}(X)$ and $\tau \wedge \omega_2 \in \Omega^{n+3}(X)$, the addition $+$ of these forms must be in the Grassman algebra $\Omega(X)$.

Let's look at the commutativity of the wedge product:

> **Proposition** *(Wedge commutativity)*
>
> Let $\omega_1 \in \Omega^n(X)$ and $\omega_2 \in \Omega^m(X)$ be two forms. Then:
>
> $$\omega_1 \wedge \omega_2 = (-1)^{nm} \omega_2 \wedge \omega_1 \tag{47.1.37}$$

Proof. We have that:

$$(\omega_1 \wedge \omega_2)(\mathbf{V}_1, ..., \mathbf{V}_{n+m}) = \frac{1}{n!m!} \sum_{\pi \in S_{n+m}} \mathrm{sgn}(\pi)(\omega_1 \otimes \omega_2)(\mathbf{V}_{\pi(1)}, ..., \mathbf{V}_{\pi(n+m)}) \tag{47.1.38}$$

$$= \frac{(-1)^n}{n!m!} \sum_{\pi \in S_{n+m}} \mathrm{sgn}(\pi)(\omega_1 \otimes \omega_2)(\mathbf{V}_{\pi(n+1)}, ... \mathbf{V}_{\pi(n)}, \mathbf{V}_{\pi(1)}, ... \mathbf{V}_{\pi(n+m)}) \tag{47.1.39}$$

$$\vdots$$

$$= \frac{(-1)^{nm}}{n!m!} \sum_{\pi \in S_{n+m}} \mathrm{sgn}(\pi)(\omega_1 \otimes \omega_2)(\mathbf{V}_{\pi(n+1)}, ..., \mathbf{V}_{\pi(n+m)}, \mathbf{V}_{\pi(1)}, ..., \mathbf{V}_{\pi(n)}) \tag{47.1.40}$$

$$= \frac{(-1)^{nm}}{n!} \frac{1}{m!} \sum_{\pi \in S_{n+m}} \mathrm{sgn}(\pi)(\omega_2 \otimes \omega_1)(\mathbf{V}_{\pi(1)}, ..., \mathbf{V}_{\pi(n+m)}) \tag{47.1.41}$$

$$= (-1)^{nm} \omega_2 \wedge \omega_1 \tag{47.1.42}$$

as desired. ∎

47.2 The exterior derivative

> **Definition** *(Exterior derivative)*
>
> We define the **exterior derivative**:
>
> $$d : \Omega^n(X) \to \Omega^{n+1}(X) \tag{47.2.1}$$
> $$\omega \mapsto d\omega \tag{47.2.2}$$
>
> where $\forall \mathbf{V}_\mu \in \Gamma(TX)$:
>
> $$(d\omega)(\mathbf{V}_1, ..., \mathbf{V}_{n+1}) = \sum_i (-1)^{i+1} \mathbf{V}_i(\omega(\mathbf{V}_1, ..., \widehat{\mathbf{V}_i}, ..., \mathbf{V}_{n+1})) \tag{47.2.3}$$
>
> $$+ \sum_{i \leq j} \omega([\mathbf{V}_i, \mathbf{V}_j], \mathbf{V}_1, ..., \widehat{\mathbf{V}_i}, ..., \widehat{\mathbf{V}_j}, ..., \mathbf{V}_{n+1}) \tag{47.2.4}$$
>
> where $\widehat{\mathbf{V}_i}$ means that the vector field $\mathbf{V}_i$ is omitted.

For example, we have that if $n = 1$ then the exterior derivative reads

$$d\omega(\mathbf{V}, \mathbf{W}) = \mathbf{V}(\omega(\mathbf{W})) - \mathbf{W}(\omega(\mathbf{V})) - \omega([\mathbf{V}, \mathbf{W}]) \tag{47.2.5}$$

Then we see that:

$$d\omega(\mathbf{V}, \mathbf{W}) = -d\omega(\mathbf{W}, \mathbf{V}) \tag{47.2.6}$$

by standard commutator rules. Moreover:

$$d\omega(\mathbf{V}, f\mathbf{W}) = \mathbf{V}(\omega(f\mathbf{W})) - f\mathbf{W}(\omega(\mathbf{V})) - \omega(\mathbf{V}(f\mathbf{W}) + f\mathbf{W}\mathbf{V}) \tag{47.2.7}$$

$$= f\mathbf{V}(\omega(\mathbf{W})) + \mathbf{V}(f)\omega(\mathbf{W}) - f\mathbf{W}(\omega(\mathbf{V})) - \omega(\mathbf{V}(f)\mathbf{W} + f\mathbf{V}\mathbf{W} - f\mathbf{W}\mathbf{V}) \tag{47.2.8}$$

The last term can be written as:

$$\mathbf{V}(f)\omega(\mathbf{W}) + f\omega([\mathbf{V}, \mathbf{W}]) \tag{47.2.9}$$

so that:

$$d\omega(\mathbf{V}, f\mathbf{W}) = f\mathbf{V}\omega(\mathbf{W}) - f\mathbf{W}(\omega(\mathbf{V})) - f\omega([\mathbf{V}, \mathbf{W}]) \tag{47.2.10}$$

$$= f d\omega(\mathbf{V}, \mathbf{W}) \tag{47.2.11}$$

as desired.

Now let's looka t how d acts on the wedge product.

Proposition *(Exterior derivative of wedge product)*
Let $\omega_1 \in \Omega^n(X)$ and $\omega_2 \in \Omega^m(X)$. Then:

$$d(\omega_1 \wedge \omega_2) = d\omega \wedge \omega_2 + (-1)^n \omega_1 \wedge d\omega_2 \tag{47.2.12}$$

Proof. ∎

Theorem *(Exterior derivative commutes with pull-back)*
Let $\phi : X \to Y$ be a smooth map between two manidolds. Then exterior derivative d commutes with the pull-back Φ^*:

$$\Phi^*(d\omega) = d(\Phi^*(\omega)), \ \forall \omega \in T^*Y \tag{47.2.13}$$

where ω is an n-form on Y.

Proof. ∎

Definition *(Anti-symmetrisation bracket)*
Let $A_{\mu_1 \dots \mu_n}$ be some object with n indices. Then we define:

$$A_{[\mu_1 \dots \mu_n]} = \frac{1}{n!} \sum_{\pi \in S_n} \mathrm{sgn}(\pi) A_{\pi(\mu_1) \dots \pi(\mu_n)} \tag{47.2.14}$$

The same goes for superscripts.

It follows from this definition is:

$$A_{\mu\nu}B^{[\mu\nu]} = \frac{1}{n!}\sum_{\pi\in S_n} \text{sgn}(\pi)A_{\mu\nu}B^{\pi(\mu_1)\pi(\mu_2)} \tag{47.2.15}$$

$$= \frac{1}{n!}\sum_{\pi\in S_n} \text{sgn}(\pi)A_{\pi^{-1}(\mu)\pi^{-1}(\nu)}B^{\mu\nu} \tag{47.2.16}$$

$$= \frac{1}{n!}\sum_{\pi\in S_n} \text{sgn}(\pi)A_{\pi(\mu)\pi(\nu)}B^{\mu\nu} \tag{47.2.17}$$

$$= A_{[\mu\nu]}B^{\mu\nu} \tag{47.2.18}$$

It also follows from this definition that:

> **Definition** *(Exact and closed forms)*
> Let $\omega \in \Omega^n(X)$. Then ω is **exact** if $\omega \in \text{Im}(d_n)$ and closed if $\omega \in \ker(d_{n+1})$. We let B^n denote the set of exact n-forms on X and Z^n denote the set of closed n-forms on X.

47.3 de Rham cohomology and Electromagnetism

> **Theorem** $(d^2 = 0)$
> Given any n-form $\omega \in \Omega^n(M)$, then the $n + 2$-form $d^2\omega \equiv (d \circ d)(\omega) = 0$ is closed.

Let $\omega in \Omega^n(X)$, then given a chart (U, ϕ) with induced local basis $\{d\phi^\mu\}$:

$$d\omega = (\partial_\nu \omega_{\mu_1,...,\mu_n})d\phi^\nu \wedge d\phi^{\mu_1} \wedge ... \wedge d\phi^{\mu_n} \tag{47.3.1}$$
$$\implies d^2\omega = (\partial_\alpha\partial_\nu \omega_{\mu_1,...,\mu_n})d\phi^\alpha d\phi^\nu \wedge d\phi^{\mu_1} \wedge ... \wedge d\phi^{\mu_n} \tag{47.3.2}$$

Note that the α, ν indices in the wedge product are anti-symmetric by definition. However, as long as $\omega_{\mu_1,...,\mu_n} \in C^2$ then α, ν are symmetric since partial derivatives can be commuted. The contraction o symmetric and anti-symmetric indices gives zero so we find that $d^2\omega = 0$ as desired.

Let's now look at the following sequence of maps:

$$\Omega^{n-1}(X) \xrightarrow{d_n} \Omega^n(X) \xrightarrow{d_{n+1}} \Omega^{n+1}(X) \tag{47.3.3}$$

Since $d^2 = 0$ it follows that $\text{Im}d_n \subseteq \ker d_{n+1}$. In other words **all exact forms are closed**:

$$B^n \subseteq Z^n \tag{47.3.4}$$

However, it is not generally true that B^n is equivalent to Z^n. One important case is when we are working in euclidean topologies:

> **Theorem** *(Poincare lemma)*
> If $X = \mathbb{R}^m$ then $B^n = Z^n$ for $n > 0$.

Let F be the electromagnetic field strength. Then we know that

$$F = dA \implies dF = 0 \tag{47.3.5}$$

for some $A \in \Omega^1(X)$. This reproduces the homogeneous Maxwell equations!

Definition *(de Rham cohomology groups)*

The n-**th deRham cohomology group** is the quotient vector space:

$$H^n(X) = Z^n/B^n \tag{47.3.6}$$

DG4 Connections and parallel transport

48.1 Covariant derivatives

Recall that a vector field $\mathbf{V} \in \Gamma(TM)$ can be used to produce a directional derivative $\mathbf{V}(f)$ of a smooth map $f \in C^\infty(X)$. For notational simplicity we will define:

$$\nabla_{\mathbf{V}} f \equiv \mathbf{V}(f) \tag{48.1.1}$$

One might wonder why we should have this cumbersome notation, especially sine df was already defined so that $df(\mathbf{V}) = \mathbf{V}(f)$. However, note that

$$X : C^\infty(X) \to C^\infty(X) \tag{48.1.2}$$
$$df : \Gamma(TX) \to C^\infty(X) \tag{48.1.3}$$

so $\mathbf{V}$ and df are different things. Moreover, $\nabla_{\mathbf{V}}$, which as of now only works on smooth functions $((0,0)$-tensors), will soon be extended so as to act on $C^\infty(X)$-tensor fields, thus requiring a different symbol.

A directional derivative should have a list of properties that we would like to be satisfied.

> **Definition** *(Connection)*
> A **connection** ∇ on a smooth manifold X is a map that maps a pair of a vector (field) $\mathbf{V}$ and a (p,q) tensor field T to (p,q)-tensor (field) $\nabla_{\mathbf{V}} T$, such that:
> (i) Extension rule: $\nabla_{\mathbf{V}} f = \mathbf{V}(f)$, $\forall f \in C^\infty(M)$
> (ii) Additivity 1 rule: $\nabla_{\mathbf{V}}(T + S) = \nabla_{\mathbf{V}} T + \nabla_{\mathbf{V}} S$, $\forall T, S \in \mathcal{T}_q^p(M)$
> (iii) Additivity 2 rule: $\nabla_{f\mathbf{V}+\mathbf{U}} T = f \nabla_{\mathbf{V}} T + \nabla_{\mathbf{U}} T$
> (iv) Leibniz rule: $\nabla_{\mathbf{V}} \underbrace{T(\omega, \mathbf{W})}_{\in C^\infty(X)} = (\nabla_{\mathbf{V}} T)(\omega, \mathbf{W}) + T(\nabla_{\mathbf{V}}\omega, \mathbf{W}) + T(\omega, \nabla_{\mathbf{V}}\mathbf{W})$ and analogously
> for any (p,q) tensor field T.

How many such connections are there, is the definition above enough to fix just one connection?

Let $\mathbf{V}, \mathbf{W}$ be vector fields, then working in a local basis:

$$\nabla_{\mathbf{V}}\mathbf{W} = \nabla_{V^\mu \partial_\mu} W^\nu \partial_\nu \tag{48.1.4}$$

$$= V^\mu \left[\left(\nabla_{\partial_\mu} W^\nu \right) \partial_\nu + W^\nu \nabla_{\partial_\mu} \partial^\nu \right] \tag{48.1.5}$$

$$= V^\mu \frac{\partial W^\nu}{\partial x^\mu} \frac{\partial}{\partial x^\nu} + V^\mu W^\nu \nabla_{\partial_\mu} \left(\frac{\partial}{\partial x^\nu} \right) \tag{48.1.6}$$

Now the second term is not fixed since we do not know what the action of the connection is on a vector. However, we note that the result of $\nabla_{\partial_\mu} \frac{\partial}{\partial x^\nu}$ will itself be a vector so it may be expanded as:

$$\nabla_{\partial_\mu} \left(\frac{\partial}{\partial x^\nu} \right) = \Gamma^\alpha_{\nu\mu} \frac{\partial}{\partial x^\alpha} \iff \Gamma^\alpha_{\nu\mu} = d\varphi^\alpha \left(\nabla_{\partial_\mu} \frac{\partial}{\partial x^\nu} \right) \tag{48.1.7}$$

where $\Gamma^\alpha_{\nu\mu}$ are the **connection coefficient functions**, and can be recognized as **Christoffel symbols**. Therefore:

$$\boxed{(\nabla_{\mathbf{V}}\mathbf{W})^\alpha = \mathbf{V}(W^\alpha) + \Gamma^\alpha_{\nu\mu} V^\mu W^\nu} \tag{48.1.8}$$

We are not done yet, we still have to see if given the connection coefficient functions, the action of $\nabla_{\mathbf{V}}$ on a 1-form is also fixed. By similar reasoning as before:

$$\nabla_{\mathbf{V}}\omega = \nabla_{V^\mu \partial_\mu} \omega_\nu d\varphi^\nu \tag{48.1.9}$$

$$= V^\mu \left[\left(\nabla_{\partial_\mu} \omega_\nu \right) d\varphi^\nu + \omega_\nu \nabla_{\partial_\mu} d\varphi^\nu \right] \tag{48.1.10}$$

$$= V^\mu \frac{\partial \omega_\nu}{\partial x^\mu} d\varphi^\nu + V^\mu \omega_\nu \nabla_{\partial_\mu} (d\varphi^\nu) \tag{48.1.11}$$

Note however that:

$$\nabla_{\partial_\mu} (d\varphi^\nu(\mathbf{e}_\alpha)) = \nabla_{\partial_\mu} (\delta^\nu_\alpha) = 0 \tag{48.1.12}$$

and using the Leibnitz rule:

$$\nabla_{\partial_\mu} (d\varphi^\nu(\mathbf{e}_\alpha)) = \nabla_{\partial_\mu} (d\varphi^\nu)(\mathbf{e}_\alpha) + d\varphi^\nu \left(\nabla_{\partial_\mu}(\mathbf{e}_\alpha) \right) = 0 \tag{48.1.13}$$

$$\implies \nabla_{\partial_\mu} (d\varphi^\nu)(\mathbf{e}_\alpha) = -d\varphi^\nu \left(\Gamma^\beta_{\alpha\mu} \frac{\partial}{\partial x^\beta} \right) \tag{48.1.14}$$

$$\implies \nabla_{\partial_\mu} (d\varphi^\nu)(\mathbf{e}_\alpha) = -\Gamma^\nu_{\alpha\mu} \iff \nabla_{\partial_\mu}(d\varphi^\nu) = -\Gamma^\nu_{\alpha\mu} d\varphi^\alpha \tag{48.1.15}$$

Consequently:

$$\boxed{(\nabla_{\mathbf{V}}\omega)_\alpha = \mathbf{V}(\omega_\alpha) - \Gamma^\nu_{\alpha\mu} V^\mu \omega_\nu} \tag{48.1.16}$$

Similarly we get that for a $(1,1)$-tensor T, the covariant derivative acts as:

$$(\nabla_{\mathbf{V}}T)^\alpha_\beta = \mathbf{V}(T^\alpha_\beta) + \Gamma^\alpha_{\nu\mu} V^\mu T^\nu_\beta - \Gamma^\nu_{\beta\mu} V^\mu T^\alpha_\nu \tag{48.1.17}$$

This can be seen by applying the Leibnitz rule to $T = \mathbf{T} \otimes \tau$, where $\mathbf{T} \in \Gamma(TX)$ and $\tau \in \Gamma(T^*X)$.

> **Definition** (*Divergence*)

Let $\mathbf{V}$ be a vector field on a manifold. The divergence of $\mathbf{V}$ is given by:

$$\text{div}\mathbf{V} = \left(\nabla_{\partial_\mu}\mathbf{V}\right)^\mu \tag{48.1.18}$$

We have not checked the change of the components of Γ when we perform a chart transition. Let (U, φ) and $(\tilde{U}, \tilde{\varphi})$ be two smooth charts with $U \cap \tilde{U} \neq \emptyset$. Then, starting from Γ' and transforming all $\tilde{x}$ to x:

$$\tilde{\Gamma}^\alpha_{\nu\mu} = d\tilde{\varphi}^\alpha\left(\nabla_{\tilde{\partial}_\mu}\frac{\partial}{\partial\tilde{x}^\nu}\right) = \frac{\partial\tilde{x}^\alpha}{\partial x^\beta}d\varphi^\beta\left(\nabla_{\tilde{\partial}_\mu x^\delta \partial_\delta}\frac{\partial x^\eta}{\partial\tilde{x}^\nu}\frac{\partial}{\partial x^\eta}\right) \tag{48.1.19}$$

$$= \frac{\partial\tilde{x}^\alpha}{\partial x^\beta}d\varphi^\beta\left\{\frac{\partial x^\delta}{\partial\tilde{x}^\mu}\left[\nabla_{\partial_\delta}\left(\frac{\partial x^\eta}{\partial\tilde{x}^\nu}\right)\frac{\partial}{\partial x^\eta} + \frac{\partial x^\eta}{\partial\tilde{x}^\nu}\nabla_{\partial_\delta}\left(\frac{\partial}{\partial x^\eta}\right)\right]\right\} \tag{48.1.20}$$

$$= \frac{\partial\tilde{x}^\alpha}{\partial x^\beta}\frac{\partial x^\delta}{\partial\tilde{x}^\mu}\left[\frac{\partial}{\partial x^\delta}\left(\frac{\partial x^\eta}{\partial\tilde{x}^\nu}\right)\delta^\beta_\eta + \frac{\partial x^\eta}{\partial\tilde{x}^\nu}\Gamma^\beta_{\eta\delta}\right] \tag{48.1.21}$$

$$\implies \boxed{\tilde{\Gamma}^\alpha_{\nu\mu} = \frac{\partial\tilde{x}^\alpha}{\partial x^\beta}\frac{\partial x^\delta}{\partial\tilde{x}^\mu}\frac{\partial x^\eta}{\partial\tilde{x}^\nu}\Gamma^\beta_{\eta\delta} + \frac{\partial\tilde{x}^\alpha}{\partial x^\beta}\frac{\partial}{\partial\tilde{x}^\mu}\left(\frac{\partial x^\beta}{\partial\tilde{x}^\nu}\right)} \tag{48.1.22}$$

So we see that in general the connection coefficient functions do not transform as tensor components. Note that the even if Γ is zero in some chart, it may not be zero in another chart.

Theorem *(Symmetric part of Γ vanish)*
Let $p \in X$ be a point in a smooth connection. Then one can always construct a chart (U, φ) containing p such that the symmetric part of the connection coefficient functions vanish at that point $\Gamma^\alpha_{(\nu\mu)} = 0$. The corresponding coordinates are known as **normal coordinates**.

Proof. Let (U, φ) be some smooth chart with $p \in V$, so that Γ does not necessarily vanish. We construct a new chart $(\tilde{U}, \tilde{\varphi})$ with $p \in \tilde{U}$ with chart transition map:

$$(\tilde{\varphi} \circ \varphi^{-1})^\alpha(x^1, ..., x^d) = x^\mu - \Gamma^\alpha_{(\nu\mu)}(p)x^\mu x^\nu \tag{48.1.23}$$

where $(x^1, ..., x^d) \in \mathbb{R}^d$ and $\Gamma^\alpha_{(\nu\mu)}$ is evaluated at the point p and is thus a constant. Then we see that:

$$\frac{\partial\tilde{x}^\beta}{\partial x^\nu} = \frac{\partial}{\partial x^\nu}(\tilde{\varphi}^\beta \circ \phi) = \delta^\mu_\nu - \Gamma^\beta_{(\nu\mu)}(p)x^\mu \tag{48.1.24}$$

$$\implies \frac{\partial}{\partial x^\mu}\frac{\partial\tilde{x}^\beta}{\partial x^\nu} = -\Gamma^\beta_{(\nu\mu)}(p) \tag{48.1.25}$$

Consequently the new connection coefficient function at p takes the form:

$$\tilde{\Gamma}^\alpha_{(\nu\mu)}(p) = 0 \tag{48.1.26}$$

as desired. $\blacksquare$

48.2 Parallel transport

Suppose you are on an expedition in the North pole, with your nose pointing out of this page, and start walking down until you reach the equator. You then move east along one fourth of the

equator, and up back to the initial point, always keeping your nose pointing in the same direction relative to your body. It is clear that your nose at the end will be pointing in a different direction to when it started! This is an example of parallel transport, the movement of a vector field $\mathbf{W}$ along a path γ on a smooth manifold so that it stays parallel to the connection ∇ we impose on it.
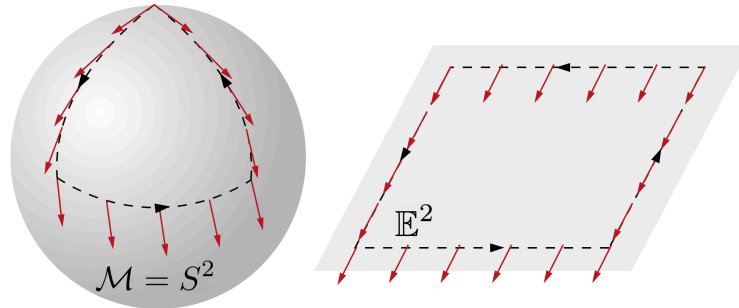
Definition *(Parallel transport)* Let X be a smooth manifold with a connection ∇ and let $\mathbf{W}$ be a vector field on X. Then we say that $\mathbf{W}$ is **parallely transported** along a smooth curve γ on X if:

$$V^\mu \nabla_\mu \mathbf{W} = \nabla_{\mathbf{V}} \mathbf{W} = 0 \qquad (48.2.1)$$

Consider the three curves and vectors fields below:
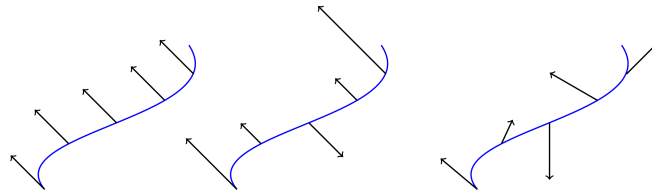
In the first case the vector field is parallely transported along γ. In the second case the vector field is just parallel to γ, but since its magnitude changes along it there is no parallel transports. Finally, in the third case the curve is neither parallel nor parallely transported.

Definition *(Autoparallel transport)*

Let $\gamma : \mathbb{R} \to X$ be a smooth curve on a smooth manifold X with tangent vector $\mathbf{V}$. Then γ is **autoparallelly transported** if:

$$\nabla_{\mathbf{V}} \mathbf{V} = 0 \qquad (48.2.2)$$

Such curves γ are knows as **geodesics**.

Suppose we choose a chart $(U, \phi) \in \mathcal{A}_{C^\infty}$ and consider the portion of γ in U. Working in the local

chart-induced basis, we can write $\nabla_\mu \equiv \nabla_{\partial_\mu}$ as long as we are working just in one chart:

$$\nabla_{\mathbf{V}}\mathbf{V} = V^\mu \nabla_\mu(V^\nu \partial_\nu) \tag{48.2.3}$$

$$= V^\mu\left((\nabla_\mu V^\nu)\frac{\partial}{\partial x^\nu} + V^\nu \Gamma^\sigma_{\nu\mu}\frac{\partial x}{\partial x^\sigma}\right) \tag{48.2.4}$$

$$= \left(V^\mu \frac{\partial V^\sigma}{\partial x^\mu} + V^\mu V^\nu \Gamma^\sigma_{\nu\mu}\right)\frac{\partial x}{\partial x^\sigma} \tag{48.2.5}$$

Recalling that if γ is parametrised by t then $V^\mu = \frac{dx^\mu}{dt}$, we see that $V^\mu \frac{\partial V^\sigma}{\partial x^\mu} = \frac{d^2 x^\sigma}{dt^2}$, giving the following, very important equation:

$$\boxed{\frac{d^2 x^\sigma}{dt^2} + \Gamma^\sigma_{\mu\nu}\frac{dx^\mu}{dt}\frac{dx^\nu}{dt} = 0} \tag{48.2.6}$$

known as the **Geodesic equation**. This is the condition for a curve γ with coordinates x^μ in a given chart to be autoparallelly transported.

For example, in Euclidean space $\mathbb{R}^2$ with standard topology, and where Γ^i_{jl} by definition, we would have that straight lines:

$$\frac{d^2 x^\sigma}{dt^2} = 0 \implies x^\sigma(t) = A^\sigma t + B^\sigma \tag{48.2.7}$$

are the geodesics.

Consider a universe with at least two particles interacting gravitationally. Here Newton's first law is completely useless since there is no particle such that the force acting upon it is zero. To salvage the first law one could envisage gravity not as a force but a curvature of space-time. Then we could have a particle with no force acting upon it (gravity is not a force anymore), and the path that it takes would be the geodesics described by the geodesic equation which is mathematically equivalent to the path given by the second law. This is qualitatively the same description given by General relativity, as we shall see in the subsection on Newtonian space-time.

> **Definition** *(Torsion)*
> The **torsion** of a connection ∇ is:
>
> $$T(\omega, \mathbf{V}, \mathbf{W}) = \omega\big(\nabla_{\mathbf{V}}\mathbf{W} - \nabla_{\mathbf{W}}\mathbf{V} - [\mathbf{V}, \mathbf{W}]\big) \tag{48.2.8}$$
>
> where
>
> $$[\mathbf{V}, \mathbf{W}](f) = \mathbf{V}(\mathbf{W}(f)) - \mathbf{W}(\mathbf{V}(f)) \tag{48.2.9}$$

Proof. We must check that this is a tensor field. Linearity in the first argument is trivial, but for the other two some care is needed. Indeed, given $f \in C^\infty$:

$$T(\omega, f\mathbf{V}, \mathbf{W}) = \omega\big(\nabla_{f\mathbf{V}}\mathbf{W} - \nabla_{\mathbf{W}}(f\mathbf{V}) - [f\mathbf{V}, \mathbf{W}]\big) \tag{48.2.10}$$

$$= \omega(f\nabla_{\mathbf{V}}\mathbf{W} - (\mathbf{W}(f))\mathbf{V} - f\nabla_{\mathbf{W}}\mathbf{V} - f[\mathbf{V}, \mathbf{W}] + (\mathbf{W}(f))\mathbf{V} \tag{48.2.11}$$

$$= f\omega\big(\nabla_{\mathbf{V}}\mathbf{W} - \nabla_{\mathbf{W}}\mathbf{V} - [\mathbf{V}, \mathbf{W}]\big) = fT(\omega, \mathbf{V}, \mathbf{W}) \tag{48.2.12}$$

where to go from the second to third step we used the property:

$$[f\mathbf{V}, \mathbf{W}](g) = f\mathbf{V}(\mathbf{W}(g)) - \mathbf{W}(f\mathbf{V}(g)) \tag{48.2.13}$$
$$= f\mathbf{V}(\mathbf{W}(g)) - \mathbf{W}(f)\mathbf{V}(g) - f\mathbf{W}(\mathbf{V}(g)) \tag{48.2.14}$$
$$= f[\mathbf{V}, \mathbf{W}](g) - \mathbf{W}(f)\mathbf{V}(g) \tag{48.2.15}$$

Similarly we find that for additivity:

$$T(\omega, \mathbf{V}_1 + \mathbf{V}_2, \mathbf{W}) = \omega(\nabla_{\mathbf{V}_1 + \mathbf{V}_2}\mathbf{W} - \nabla_{\mathbf{W}}(\mathbf{V}_1 + \mathbf{V}_2) - [\mathbf{V}_1 + \mathbf{V}_2, \mathbf{W}]) \tag{48.2.16}$$
$$= \omega(\nabla_{\mathbf{V}_1}\mathbf{W} + \nabla_{\mathbf{V}_2}\mathbf{W} - \nabla_{\mathbf{W}}\mathbf{V}_1 - \nabla_{\mathbf{W}}\mathbf{V}_2 - [\mathbf{V}_1, \mathbf{W}] - [\mathbf{V}_2, \mathbf{W}] \tag{48.2.17}$$
$$= T(\omega, \mathbf{V}_1, \mathbf{W}) + T(\omega, \mathbf{V}_2, \mathbf{W}) \tag{48.2.18}$$

as desired. Linearity in the second argument follows from the anti-linearity of the tensor field in $\mathbf{V}$ and $\mathbf{W}$. ∎

> **Definition** *(Torsion-free manifold)*
> A manifold with connection ∇ is torsion free if $T = 0$.

Working in a chart, the condition for a manifold with connection to be torsion free is that:

$$T^\alpha{}_{\mu\nu} = T(dx^\alpha, \partial_\mu, \partial_\nu) = dx^\mu(\nabla_\mu\partial_\nu - \nabla_\nu\partial_\mu) = \Gamma^\alpha_{\nu\mu} - \Gamma^\alpha_{\mu\nu} = 0 \implies \Gamma^\alpha_{[\mu\nu]} = 0 \tag{48.2.19}$$

so the Christoffel symbols must have vanishing anti-symmetric components. Note that for a torsion-free manifold we can always locally set the Christoffel symbols to zero since we can make both its symmetric and anti-symmetric components equal to zero.
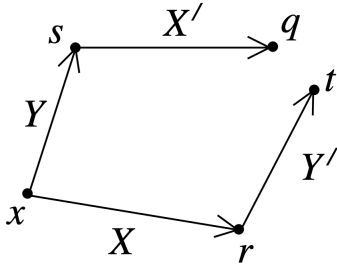
We can visualize torsion as the failure of parallelograms to close. Indeed let us take two vectors fields $\mathbf{X}$ and $\mathbf{Y}$ in $T_x M$. It is clear that the points r and s will have coordinates

$$r : x^\mu + \epsilon X^\mu, \; s : x^\mu + \epsilon Y^\mu \tag{48.2.20}$$

where ϵ is infinitesimal. Suppose we parallelly transport $\mathbf{X}$ along $\mathbf{Y}$ giving a new vector

$$X'^\mu = X^\mu - \epsilon \Gamma^\rho_{\mu\nu} X^\mu Y^\nu \tag{48.2.21}$$

The corresponding point q then has coordinates

$$q : x^\mu + \epsilon(X^\mu + Y^\mu) - \epsilon^2 \Gamma^\rho_{\mu\nu} X^\mu Y^\nu \tag{48.2.22}$$

On the other hand the point t has coordinates

$$t : x^\mu + \epsilon(X^\mu + Y^\mu) - \epsilon^2 \Gamma^\rho_{\mu\nu} Y^\mu X^\nu \tag{48.2.23}$$

In general, unless torsion vanishes we will find that q and t do not coincide, in other words the parallelogram formed by parallelly transporting two vectors does not necessarily close. It is clear how for most physical applications we can safely take torsion to vanish as we do not want any openings in our space-time manifold.

48.3   Riemannian curvature

> **Definition** *(Riemann curvature)*
> The **Riemannian curvature** of a connection ∇ is the tensor field:
>
> $$R(\omega, \mathbf{U}, \mathbf{V}, \mathbf{W}) = \omega\big(\nabla_{\mathbf{V}}\nabla_{\mathbf{W}}\mathbf{U} - \nabla_{\mathbf{W}}\nabla_{\mathbf{V}}\mathbf{U} - \nabla_{[\mathbf{V},\mathbf{Y}]}\mathbf{U}\big) \qquad (48.3.1)$$

Working in one chart:

$$R^{\alpha}{}_{\sigma\mu\nu} = dx^{\alpha}(\nabla_{\mu}\nabla_{\nu}\partial_{\sigma} - \nabla_{\nu}\nabla_{\mu}\partial_{\sigma} - \nabla_{\partial_{\mu}\partial_{\nu}}\partial_{\sigma} + \nabla_{\partial_{\nu}\partial_{\mu}}\partial_{\sigma}) \qquad (48.3.2)$$

$$= dx^{\alpha}[\nabla_{\mu}(\Gamma^{\alpha}_{\sigma\nu}\partial_{\alpha}) - \nabla_{\nu}(\Gamma^{\alpha}_{\sigma\mu}\partial_{\alpha}) - \nabla_{\partial_{\mu}\partial_{\nu}}\partial_{\sigma} + \nabla_{\partial_{\nu}\partial_{\mu}}\partial_{\sigma}] \qquad (48.3.3)$$

$$= dx^{\alpha}[\Gamma^{\tau}_{\sigma\nu}\Gamma^{\gamma}_{\tau\mu}\partial_{\gamma} - \Gamma^{\tau}_{\sigma\mu}\Gamma^{\gamma}_{\tau\nu)\partial_{\gamma}} + (\partial_{\mu}\Gamma^{\tau}_{\sigma\nu})\partial_{\tau} - (\partial_{\nu}\Gamma^{\tau}_{\sigma\mu})\partial_{\mu}] \qquad (48.3.4)$$

$$\implies \boxed{R^{\alpha}{}_{\sigma\mu\nu} = \Gamma^{\tau}_{\sigma\nu}\Gamma^{\alpha}_{\tau\mu} - \Gamma^{\tau}_{\sigma\mu}\Gamma^{\alpha}_{\tau\nu} + \partial_{\mu}\Gamma^{\alpha}_{\sigma\nu} - \partial_{\nu}\Gamma^{\alpha}_{\sigma\mu}} \qquad (48.3.5)$$

Geometrically, the Riemann curvature tensor captures the failure of vector field **U** to be parallelly transported to the same vector along **V** or **W** on a smooth manifold X. The more curved our manifold is, the greater the discrepancies between the different parallel transports. Indeed, consider the figure beside. We take two vector fields **X** and **Y** along which we transport another vector field **Z** taking two different paths, $p \to q \to r$ and $p \to u \to r$ on a torsion-free Riemannian manifold. We can thus take x^{μ} to be normal coordinates in which the Christoffel symbols evaluated at p vanish. We perform all approximations up to second order.
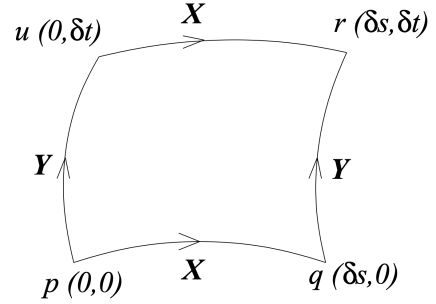
To go from p to q it is clear that:

$$(\nabla_{\mathbf{X}}\mathbf{Z})^{\rho} = X^{\nu}\frac{dZ^{\mu}}{dx^{\nu}} + \Gamma^{\rho}_{\mu\nu}Z^{\mu}X^{\nu} = 0 \qquad (48.3.6)$$

Letting $X^{\mu}\frac{\partial}{\partial x^{\mu}} = \frac{\partial}{\partial s}$ then

$$\frac{d^2 Z^{\rho}}{ds^2} = -X^{\sigma}\partial_{\sigma}(\Gamma^{\rho}_{\mu\nu}Z^{\mu}X^{\nu}) \qquad (48.3.7)$$

We can now perform a Taylor expansion about p:

$$Z^{\rho}(q) = Z^{\rho}(p) + \frac{dZ^{\mu}}{ds}\bigg|_{p}\delta s + \frac{1}{2}\frac{d^2 Z^{\mu}}{ds^2}\bigg|_{p}(\delta s)^2 \qquad (48.3.8)$$

but it is clear that $\frac{dZ^{\mu}}{ds}\big|_{p}$ must vanish so we find that:

$$Z^{\rho}(q) - Z^{\rho}(p) = -\frac{1}{2}(X^{\sigma}Z^{\mu}X^{\nu}\partial_{\sigma}\Gamma^{\rho}_{\mu\nu})\big|_{p}(\delta s)^2 \qquad (48.3.9)$$

Now we go from q to r:

$$Z^{\rho}(r) - Z^{\rho}(q) = \frac{dZ^{\mu}}{ds}\bigg|_{q}\delta t + \frac{1}{2}\frac{d^2 Z^{\mu}}{ds^2}\bigg|_{q}(\delta t)^2 \qquad (48.3.10)$$

$$= -(\Gamma^{\rho}_{\mu\nu}Z^{\mu}Y^{\nu})|_{q}\delta t - \frac{1}{2}[X^{\sigma}\partial_{\sigma}(\Gamma^{\rho}_{\mu\nu}Z^{\mu}Y^{\nu})]\big|_{q}(\delta t)^2 \qquad (48.3.11)$$

but recall that expanding $(\Gamma^\rho_{\mu\nu}Z^\mu Y^\nu)|_q$ to first order (since we already have a δt)

$$(\Gamma^\rho_{\mu\nu}Z^\mu Y^\nu)|_q = \overbrace{(\Gamma^\rho_{\mu\nu}Z^\mu Y^\nu)|_p}^{0} + X^\sigma \partial_\sigma (\Gamma^\rho_{\mu\nu}Z^\mu Y^\nu)\big|_p \delta s \tag{48.3.12}$$

$$= X^\sigma Z^\mu Y^\nu \partial_\sigma \Gamma^\rho_{\mu\nu}|_p \delta s \tag{48.3.13}$$

hence

$$Z^\rho(r) - Z^\rho(q) = -X^\sigma Z^\mu Y^\nu \partial_\sigma \Gamma^\rho_{\mu\nu}|_p \delta s \delta t - \frac{1}{2}[Y^\sigma \partial_\sigma (\Gamma^\rho_{\mu\nu}Z^\mu Y^\nu)]\big|_q (\delta t)^2 \tag{48.3.14}$$

Similarly we may expand $[X^\sigma \partial_\sigma (\Gamma^\rho_{\mu\nu}Z^\mu Y^\nu)]\big|_q$, this time to just zeroth order since it comes with $(\delta t)^2$. We can thus evaluate this term at p and simplify:

$$Z^\rho(r) - Z^\rho(q) = -Y^\sigma Z^\mu Y^\nu \partial_\sigma \Gamma^\rho_{\mu\nu}|_p \delta s \delta t - \frac{1}{2}(Y^\sigma Z^\mu Y^\nu \partial_\sigma \Gamma^\rho_{\mu\nu})\big|_p (\delta t)^2 \tag{48.3.15}$$

Thus, we find that:

$$(Z^\rho(r) - Z^\rho(p))_{pqr} = -\frac{1}{2}\partial_\sigma \Gamma^\rho_{\mu\nu}(X^\sigma X^\nu (\delta s)^2 + Y^\sigma Y^\nu (\delta t)^2 + 2X^\sigma Y^\nu \delta s \delta t)Z^\mu|_p \tag{48.3.16}$$

By simply replacing $\mathbf{X} \leftrightarrow \mathbf{Y}$ and $\delta s \leftrightarrow \delta t$ we find that

$$(Z^\rho(r) - Z^\rho(p))_{pur} = -\frac{1}{2}\partial_\sigma \Gamma^\rho_{\mu\nu}(Y^\sigma Y^\nu (\delta t)^2 + X^\sigma X^\nu (\delta s)^2 + 2Y^\sigma X^\nu \delta s \delta t)Z^\mu|_p \tag{48.3.17}$$

We can now compute the difference between $Z^\rho(r)$ transported along γ' : pqr and along γ : pur

$$\frac{Z^\rho_{(\gamma')} - Z^\rho_{(\gamma)}}{\delta s \delta t} = \partial_\sigma \Gamma^\rho_{\mu\nu}(Y^\sigma X^\nu - X^\sigma Y^\nu)Z^\mu|_p \tag{48.3.18}$$

$$= Z^\mu X^\nu Y^\sigma (\partial_\sigma \Gamma^\rho_{\mu\nu} - \partial_\nu \Gamma^\rho_{\mu\sigma})|_p \tag{48.3.19}$$

$$\implies \boxed{\frac{Z^\rho_{(\gamma')} - Z^\rho_{(\gamma)}}{\delta s \delta t} = (Z^\mu Y^\sigma X^\nu R^\rho_{\mu\sigma\nu})|_p} \tag{48.3.20}$$

where we used the fact that at p, the Riemann curvature tensor reads $R^\rho_{\mu\sigma\nu} = \partial_\sigma \Gamma^\rho_{\mu\nu} - \partial_\nu \Gamma^\rho_{\mu\sigma}$ in normal coordinates. Note also that this result is not an artifact of working within a special chart. Indeed our end result is a relation between tensor so there is no coordinate dependence.

So when we parallelly transport a vector field along two different paths their final differences is encoded in the Riemann curvature tensor.

DG5 Metrics and Curvature

49.1 The metric tensor

We are missing a notion of length and angles in our formulation of differential geometry, which we can establish by imposing a metric on our manifold.

> **Definition** *(Metric tensor)*
> Let X be a smooth manifold, then a **metric** g on X is a $(0, 2)$-tensor field such that:
>   (i) symmetry: $g(\mathbf{V}, \mathbf{W}) = g(\mathbf{W}, \mathbf{V})$
>   (ii) non-degeneracy: define the map $\sharp : \Gamma(TX) \to \Gamma(T^*X)$ such that $(\sharp(\mathbf{V}))(\mathbf{W}) = g(\mathbf{V}, \mathbf{W})$.
>   Then $\flat$ is a C^∞-isomorphism.

We not discuss a bit of convention. we can denote the components of the $\sharp$-map acting on $\mathbf{V}$ as simply:

$$(\sharp(\mathbf{V}))_\mu \equiv V_\mu = g_{\mu\nu} V^\nu \tag{49.1.1}$$

where the last equality was established by definition.

> **Definition** *(Inverse metric)*
> The symmetric $(2, 0)$-tensor field g^{-1} related to the metric tensor field g is defined by:
>
> $$g^{-1} : \Gamma(T^*X) \times \Gamma(T^*X) \to C^\infty(X) \tag{49.1.2}$$
> $$(\omega, \tau) \mapsto \omega(\sharp^{-1}(\tau)) \tag{49.1.3}$$

This is not really the inverse of g since the domains and images don't match up, but can be naively viewed as such when looking at it purely as a matrix. Indeed we have that

$$\sharp(\partial_\nu) = g_{\mu\nu} dx^\mu \implies \sharp^{-1}(g_{\mu\nu} dx^\mu) = \partial_\nu \tag{49.1.4}$$

and hence

$$(g^{-1})^{\mu\nu} g_{\nu\sigma} = dx^\mu(\sharp^{-1}(dx^\nu)) g_{\nu\sigma} = dx^\mu(\sharp^{-1}(g_{\nu\sigma} dx^\nu)) = dx^\mu(\partial_\sigma) = \delta^\mu_\sigma \tag{49.1.5}$$

Hence, since $\sharp$ is an isomorphism it is invertible we may write:

$$(\sharp^{-1}(\omega)) \equiv \omega^\mu = \omega_\nu \sharp^{-1}((g^{-1})^{\nu\mu} g_{\mu\sigma} dx^\sigma) = (g^{-1})^{\mu\nu} \omega_\nu \partial_\mu \tag{49.1.6}$$
$$\implies \omega^\mu = (g^{-1})^{\mu\nu} \omega_\nu \tag{49.1.7}$$

For simplicity we will often write $g^{-1})^{\mu\nu} \equiv g^{\mu\nu}$ since the indices being up already indicate that we are taking components of g^{-1} and not g. Hence we have found that we can raise and lower the operators of covectors and vectors using the metric:

$$\boxed{V_\mu = g_{\mu\nu}V^\nu, \; \omega_\mu = g^{\mu\nu}\omega_\nu} \tag{49.1.8}$$

We claim that a metric tensor can always be reduced to a diagonal form where the entries are either $+1, -1, 0$. If this diagonal form of the metric contains p 1's and q -1's then we say that the metric has (p, q) signature.

Definition *(Metric types)*
A metric is said to be:
  (i) **Riemannian** if the signature is $(+ + ... +)$.
  (ii) **pseudo-Riemannian** otherwise, with the special case of a **Lorentzian** metric if the signature is $(+ - ... -)$
A manifold equipped with a Riemannian metric is a Riemannian manifold.

Definition *(Curve length)*
Let X be a Riemannian manifold with metric g. Then the speed of a smooth curve $\gamma(0, 1) \mapsto X$ parametrised by t at $p = (\phi \circ \gamma)(t_p) \in X$ is given by

$$s(t_p) = \sqrt{g(\mathbf{V}_{\gamma,p}, \mathbf{V}_{\gamma,p})} \tag{49.1.9}$$

The length of γ is then given by:

$$L[\gamma] = \int_0^1 dt \; s(t) = \int_0^1 dt \sqrt{g(\mathbf{V}_\gamma, \mathbf{V}_\gamma)} \tag{49.1.10}$$

Of course, the length of a curve should not depend on the way it is parametrised. This is indeed the case:

Theorem *(Curve reparametrisation)*
Let $\gamma : (0, 1) \to X$ be a curve and let $\sigma : (0, 1) \to (0, 1)$ be a smooth, increasing bijection. Then:
$$L[\gamma] = L[\gamma \circ \sigma] \tag{49.1.11}$$

Proof. We have that:

$$V_{\gamma \circ \sigma, t_p} = \frac{d}{dt}f(\gamma(\sigma(t)))\Big|_{t_p} = \frac{d\sigma(t)}{dt}\frac{d}{d\sigma(t)}f(\gamma(\sigma(t)))\Big|_{t_p} = \frac{d\sigma(t)}{dt}\frac{d}{dt}f(\gamma(t))\Big|_{\sigma(t_p)} \tag{49.1.12}$$

$$= \frac{d\sigma}{dt}V_{\gamma, \sigma(t_p)} \tag{49.1.13}$$

This then gives

$$L[\gamma \circ \sigma] = \int_0^1 dt \sqrt{g(\mathbf{V}_{\gamma \circ \sigma, t}, \mathbf{V}_{\gamma \circ \sigma, t})} \tag{49.1.14}$$

$$= \int_0^1 dt \sqrt{g(\mathbf{V}_{\gamma, \sigma(t)}, \mathbf{V}_{\gamma, \sigma(t)})} \tag{49.1.15}$$

$$= \int_0^1 dt \frac{d\sigma(t)}{dt} \sqrt{g(\mathbf{V}_{\gamma, \sigma(t)}, \mathbf{V}_{\gamma, \sigma(t)})} \tag{49.1.16}$$

We now use change variables $s = \sigma(t) \implies ds = \frac{d\sigma}{dt} dt$. The bounds of integration remain the same since σ is bijective and increasing, implying that $\sigma(0) = 0$, $\sigma(1) = 1$. Thus:

$$L[\gamma \circ \sigma] = \int_0^1 ds \sqrt{g(\mathbf{V}_{\gamma, s}, \mathbf{V}_{\gamma, s})} = L[\gamma] \tag{49.1.17}$$

as desired. ∎

In the context of metric manifolds, we should define geodesics as curves that have the shortest possible length defined by the metric.

> **Definition** *(Geodesic)*
> A curve γ is a **geodesic** on a Riemannian manifold if it is a stationary curve of L.

We would of course like this geodesic to be the same geodesic defined by autoparallel transports, thus requiring the connection on a manifold to be determined by the metric we impose on it. Of course we can express the minimum length principle of geodesics using the Euler-Lagrange equation:

$$\frac{\partial \mathcal{L}}{\partial x^\mu(t)} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}^\mu(t)} \tag{49.1.18}$$

using the Lagrangian:

$$\mathcal{L} : TX \to \mathbb{R} \tag{49.1.19}$$

$$\mathbf{V} \mapsto \sqrt{g(\mathbf{V}, \mathbf{V})} \tag{49.1.20}$$

so that $\mathcal{L}(x^\mu, \dot{x}^\mu) = \sqrt{g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu}$. Letting $\sqrt{g(\mathbf{V}, \mathbf{V})} = g$ we find that

$$\frac{\partial \mathcal{L}}{\partial x^\alpha} = \frac{1}{2g} \partial_\alpha g_{\mu\nu}(x) \dot{x}^\mu \dot{x}^\nu \tag{49.1.21}$$

Also:

$$\frac{\partial \mathcal{L}}{\partial \dot{x}^\alpha} = \frac{1}{2g} 2 g_{\alpha\mu}(x) \dot{x}^\mu = \frac{1}{g} g_{\alpha\mu}(x) \dot{x}^\mu \tag{49.1.22}$$

so that

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}^\alpha} = \frac{d}{dt} \left(\frac{1}{g} \right) g_{\alpha\mu}(x) \dot{x}^\mu + \frac{1}{g} \left(g_{\alpha\mu}(x) \ddot{x}^\mu + \dot{x}^\nu \frac{\partial}{\partial x^\nu}(g_{\alpha\mu}(x)) \dot{x}^\mu \right) \tag{49.1.23}$$

Now recall that we can re-parametrise our curve without affecting the length functional. Thus we can choose a parametrisation such that $g(\mathbf{V}, \mathbf{V}) = \text{cnst}$, that is such that the speed of the curve is constant (note that the metric can never give zero so this can always be done). The first term then

vanishes, and hence:

$$g_{\alpha\mu}\ddot{x}^{\mu} + \partial_{\nu}(g_{\alpha\mu})\dot{x}^{\mu}\dot{x}^{\nu} - \frac{1}{2}\partial_{\alpha}g_{\mu\nu}\dot{x}^{\mu}\dot{x}^{\nu} = 0 \qquad (49.1.24)$$

We can now use the inverse matrix to raise the α index, and use $g^{\alpha\beta}g_{\beta\rho} = \delta^{\alpha}_{\rho}$:

$$\ddot{x}^{\sigma} + g^{\alpha\sigma}\left(\partial_{\nu}g_{\alpha\mu} - \frac{1}{2}\partial_{\alpha}g_{\mu\nu}\right)\dot{x}^{\mu}\dot{x}^{\nu} = 0 \qquad (49.1.25)$$

Since the i, j indices are symmetric, we may double the term in parenthesis by writing its copy with i, j exchanged, and hence write that:

$$\boxed{\frac{d^2x^{\sigma}}{dt^2} + \frac{1}{2}g^{\alpha\sigma}\left(\partial_{\nu}g_{\alpha\mu} + \partial_{\mu}g_{\alpha\nu} - \partial_{\alpha}g_{\mu\nu}\right)\frac{dx^{\mu}}{dt}\frac{dx^{\nu}}{dt} = 0} \qquad (49.1.26)$$

Now suppose we wish to establish a connection on our Riemannian manifold, such that the autoparallel geodesics are exactly the geodesics described above. Comparing (48.2.6) with (49.1.26) it follows that the Christoffel symbols must be defined so that:

$$\boxed{\Gamma^{\sigma}_{\mu}\nu = \frac{1}{2}g^{\alpha\sigma}\left(\partial_{\nu}g_{\alpha\mu} + \partial_{\mu}g_{\alpha\nu} - \partial_{\alpha}g_{\mu\nu}\right)} \qquad (49.1.27)$$

The connection with these Christoffel symbols is known as the **Levi-Civita** connection. There are several useful curvatures that one may define with the Levi-Civita connection.

> **Definition** *(Important curvatures)*
> On a Riemannian manifold with the Levi-Civita connection, the
> (i) **Riemann Christoffel curvature** is defined as
>
> $$R_{\rho\sigma\mu\nu} = g_{\rho\alpha}R^{\alpha}{}_{\sigma\mu\nu} \qquad (49.1.28)$$
>
> (ii) **Ricci curvature** is defined as:
> $$R_{\mu\nu} = R^{\sigma}_{\mu\sigma\nu} \qquad (49.1.29)$$
>
> (iii) **Scalar curvature**:
> $$R = g^{\mu\nu}R_{\mu\nu} \qquad (49.1.30)$$
>
> (iv) **Einstein curvature**:
> $$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R \qquad (49.1.31)$$

49.2 Lie derivatives and symmetry

We saw how we could push forward vectors between tangent spaces, and pull back 1-forms between dual tangent spaces, but having seen metrics it is reasonable to wonder whether the same can be done with a metric tensor.

If we start with a metric manifold N and embed a smaller manifold M in it, then we can define a metric on M by pushing forward vectors on M using the embedding and use the metric for N on them. This defines the so-called induced metric.

> **Definition** *(Induced metric)*
> Suppose we have a metric g on a manifold N, and suppose ϕ is an embedding (injective map) of a manifold M into N. We define the **induced metric** g_M on M as:
>
> $$g_M(\mathbf{V}, \mathbf{W}) = g(\phi_*(\mathbf{V}), \phi_*(\mathbf{W})) \tag{49.2.1}$$
>
> or alternatively
>
> $$(g_M)_{\mu\nu} = g_{\alpha\beta} \frac{\partial X^\alpha}{\partial x^\mu} \frac{\partial X^\beta}{\partial x^\nu} \tag{49.2.2}$$
>
> where recall that $X^\alpha = y^\alpha \circ \phi$.

> **Definition** *(Integral curve)*
> Let M be a smooth manifold and let $\mathbf{W}$ be a vector field on M. A curve γ is an **integral curve** of $\mathbf{W}$ if the value of the vector field on the curve coincides everywhere with the tangent vectors $\mathbf{V}$ along the curve:
> $$\mathbf{V}_{\gamma, \gamma(t)} = \mathbf{W}_{\gamma(t)} \tag{49.2.3}$$
>
> Furthermore, $\mathbf{W}$ is complete if all its integral curves have a domain that can be extended to $\mathbb{R}$.

It would also be nice if given some complete vector field, we could create a family of integral curves passing through all the points on the manifold. This would generate a flow of the vector field, with every point flowing along it by some distance t.

> **Definition** *(Vector field flow)*
> The **flow of a complete vector field $\mathbf{W}$** on M is:
>
> $$h^{\mathbf{W}} : \mathbb{R} \times M \to M \tag{49.2.4}$$
> $$(t, p) \mapsto \gamma_p(t) \tag{49.2.5}$$
>
> where $\gamma_p : \mathbb{R} \to M$ is an integral curve of $\mathbf{W}$ traversing p at $t = 0$. This flow is also a **one parameter group of diffeomorphisms**.

We can fix t in the flow map to get $h_t^{\mathbf{W}}$. Its action is to take every point p on the manifold and map it to the point $\gamma_p(t)$ on the integral curve passing through it at $t = 0$, thus shifting it a parameter distance t.

The differential equation that the coordinates $y^\mu(t)$ of an integral curve must satisfy is thus

$$\frac{dy^\mu}{dt} = X^\mu, \ y^\mu(0) = x^\mu \tag{49.2.6}$$

where x_μ^0 are the coordinates of p. For infinitesimal flows then we see that to leading order in t

$$y^\mu(h_t^{\mathbf{W}}) = x^\mu + tW^\mu \tag{49.2.7}$$

> **Theorem** *(Vector fields form Lie algebra)*

Let $\Gamma(TM)$ denote the set of all vector fields on a manifold M, and let $[\cdot,\cdot]$ be the commutator:

$$[\mathbf{V},\mathbf{W}]f = \mathbf{V}(\mathbf{W}(f)) - \mathbf{W}(\mathbf{V}(f)) \qquad (49.2.8)$$

which satisfies the following properties:
  (i) $[\mathbf{V},\mathbf{W}] = -[\mathbf{W},\mathbf{V}]$
  (ii) $[\lambda\mathbf{V} + \mathbf{U},\mathbf{W}] = \lambda[\mathbf{V},\mathbf{W}] + [\mathbf{U},\mathbf{W}]$
  (iii) $[\mathbf{U},[\mathbf{V},\mathbf{W}]] + [\mathbf{W},[\mathbf{U},\mathbf{V}]] + [\mathbf{V},[\mathbf{W},\mathbf{U}]] = 0$
Then $(\Gamma(TM),[\cdot,\cdot])$ is a **Lie algebra**.

Proof. We prove Jacobi's identity (iii) only as the first two are trivial. ∎

Definition *(Lie subalgebra)*
Let $\{\mathbf{V}_i\}_{i=0,1,\dots}$ be vector fields on X such that:

$$[\mathbf{V}_i,\mathbf{V}_j] = C_{ij}^k \mathbf{V}_k \qquad (49.2.9)$$

where C_{ij}^k are known as **structure constants** of the **Lie subalgebra** $(\mathrm{span}_{\mathbb{R}}\{\mathbf{V}_i\},[\cdot,\cdot]$.

Definition *(Symmetry)*
A Lie subalgebra L on a smooth metric manifold is a **symmetry** of a metric g if for any vector fields $\mathbf{V}$ in L [a] then:

$$((h_t^{\mathbf{V}})^*g)(\mathbf{X},\mathbf{Y}) \equiv g((h_t^{\mathbf{V}})_*(\mathbf{X}),(h_t^{\mathbf{V}})_*(\mathbf{Y})) = g(\mathbf{X},\mathbf{Y}),\ \forall\mathbf{X},\mathbf{Y}\in\Gamma(TM),\ \forall t \qquad (49.2.10)$$

[a] which are guaranteed to be complete by a very nice theorem assuming they are compactly supported

This definition makes intuitively sense, it says that if L is a symmetry, then moving the metric (backwards) along the flows generated by any of its vector fields has no effect at all. We can take the metric of two vector fields at some point, move them along the flows generated by a vector field in the symmetry, and get the same result back.

Despite the intuitiveness behind this definition it is often very difficult to check for symmetries by looking at the pull-back of the metric, it is a very tedious task. A much faster approach involves defining the Lie derivative.

Definition *(Lie derivative)*
We define the **Lie derivative** of a tensor T along some vector field $\mathbf{V}$ in a Lie sub-algebra L as

$$\mathcal{L}_{\mathbf{V}}T \equiv \lim_{t\to 0} \frac{((h_{-t}^{\mathbf{V}})_*T)_p - T_p}{t} \qquad (49.2.11)$$

so that if the Lie subalgebra L is a symmetry then $\mathcal{L}_{\mathbf{V}}g = 0,\ \forall\mathbf{V}\in L$.

This definition is quite close to how we define normal derivatives. Indeed a naive guess would be to simply write

$$\mathcal{L}_{\mathbf{V}}T = \lim_{t\to\infty} \frac{T_{p+t} - T_p}{t} \qquad (49.2.12)$$

but there are two problems with this definition. Firstly, the notion of adding points on a manifold makes no sense, $p + t$ is not a thing. We need an alternative notion that is physically equivalent, that of flowing the point p along a $\mathbf{V}$ by an amount t, which is precisely what the flow map does.
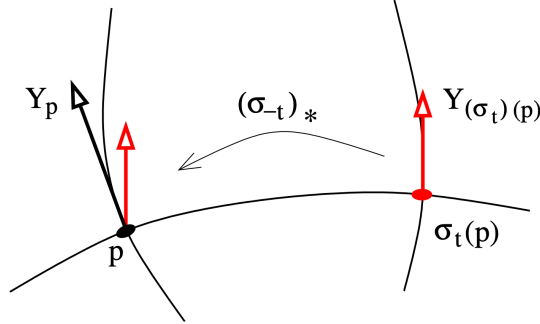
Secondly, the tensors $T_{h_t(p)}$ and T_p do not belong to the same tensor space so it doesn't make sense to simply take their difference (this is like taking the difference of tangent vectors at different points). We can resolve this problem by pushing forward $T_{h_t(p)}$ to the tangent space at p, so that the difference may be evaluated. The end result is the definition we provided previously

We can define special coordinates to write the Lie derivative in a nice way. Suppose that we have generated the integral curves of some vector field $\mathbf{V}$ in a n-manifold. Let $p \in M$ and let us create a $n - 1$ dimensional hypersurface Σ of M that is not tangent to the integral curve through p. Then we can let the parametrisation variable t be the nth coordinate together with the x^i coordinates of Σ.

It follows that the push-forward of the flow h_t^* maps a point p with coordinates (t_p, x_p^i) to the point q with coordinates $(t_p + t, x_p^i)$. Therefore

$$((h_t)_*)^\mu{}_\nu = \frac{\partial y^\mu}{\partial x^n u} = \delta^\mu_\nu \tag{49.2.13}$$

implying that

$$((h^{\mathbf{V}}_{-t})_* T^{\mu_1 \cdots \mu_n}{}_{\nu_1 \dots \nu_n})_p = \frac{\partial y^{\mu_1}}{\partial x^{\sigma_1}} \cdots \frac{\partial y^{\mu_n}}{\partial x^{\sigma_n}} \frac{\partial x^{\rho_1}}{\partial y^{\nu_1}} \cdots \frac{\partial x^{\rho_n}}{\partial y^{\nu_n}} (T^{\sigma_1 \cdots \sigma_n}{}_{\rho_1 \dots \rho_n})_{\phi(p)} \tag{49.2.14}$$

$$= (T^{\mu_1 \cdots \mu_n}{}_{\nu_1 \dots \nu_n})(t_p + t, x^i) \tag{49.2.15}$$

Hence we get that

$$(\mathcal{L}_{\mathbf{V}} T)^{\mu_1 \cdots \mu_n}{}_{\nu_1 \dots \nu_n} = \frac{\partial}{\partial t} T^{\mu_1 \cdots \mu_n}{}_{\nu_1 \dots \nu_n} \bigg|_{(t_p, x^i)} \tag{49.2.16}$$

so the Lie derivative in this particular coordinate frame is just the partial derivative of the tensor components.

It can be shown that this limit definition of the Lie derivative is equivalent to the following algebraic definition:

> **Definition** *(Lie derivative)*
> The **Lie derivative** $\mathcal{L}$ on a smooth manifold maps a vector field and a (p, q) tensor field to

another (p, q) tensor field such that:

(i) $\mathcal{L}_{\mathbf{V}} f = \mathbf{V} f$

(ii) $\mathcal{L}_{\mathbf{V}} \mathbf{W} = [\mathbf{V}, \mathbf{W}]$

(iii) $\mathcal{L}_{\mathbf{V}}(T + S) = \mathcal{L}_{\mathbf{V}} T + \mathcal{L}_{\mathbf{V}} S$

(iv) $\mathcal{L}_{\mathbf{V}} T(\omega, \mathbf{W}) = (\mathcal{L}_{\mathbf{V}} T)(\omega, \mathbf{W}) + T(\mathcal{L}_{\mathbf{V}} \omega, \mathbf{W}) + T(\omega, \mathcal{L}_{\mathbf{V}} \mathbf{W})$ and similarly for any other tensor T

(v) $\mathcal{L}_{\mathbf{V} + \mathbf{W}} T = \mathcal{L}_{\mathbf{V}} T + \mathcal{L}_{\mathbf{W}} T$

It may seem like the Lie derivative is pretty much like a covariant derivative, but with even less structure. Indeed the Lie derivative has no C^∞ linearity which can be verified by looking at (ii) and evaluating $[f\mathbf{V}, \mathbf{W}] \neq f[\mathbf{V}, \mathbf{W}]$. We could not have used this Lie derivative in place of the covariant derivative for exactly this reason, we need the linearity to differentiate tensors and talk about parallel transport, geodesics etc...

Theorem *(Lie derivative components)*

In a local basis, we have that:

$$(\mathcal{L}_{\mathbf{V}} \mathbf{W})^\mu = V^\nu \frac{\partial W^\mu}{\partial x^\nu} - \frac{\partial V^\mu}{\partial x^\nu} W^\nu \tag{49.2.17}$$

Proof. The proof is immediate from the definition of the commutator

$$(\mathcal{L}_{\mathbf{V}} \mathbf{W}) f = \mathbf{V}(\mathbf{W}(f)) - \mathbf{W}(\mathbf{V}(f)) = V^\nu \frac{\partial}{\partial x_\nu}\left(W^\mu \frac{\partial}{\partial x^\mu} f\right) - W^\mu \frac{\partial}{\partial x^\mu}\left(V^\nu \frac{\partial}{\partial x^\nu} f\right) \tag{49.2.18}$$

$$= \left(V^\nu \frac{\partial W^\mu}{\partial x^\nu} - \frac{\partial V^\mu}{\partial x^\nu} W^\nu\right)\frac{\partial}{\partial x^\mu} f \tag{49.2.19}$$

as desired. ∎

Compare this to the definition of the covariant derivative:

$$(\nabla_{\mathbf{V}} \mathbf{W})^\mu = V^\mu \frac{\partial W^\nu}{\partial x^\mu} - \Gamma^\mu_{\rho\nu} V^\nu W^\rho \tag{49.2.20}$$

then we see that while ∇ does not require information about the vector fields outside of the point it is evaluated at, $\mathcal{L}$ does due to the partial derivatives.

We can use the Leibniz rule to write:

$$(\mathcal{L}_{\mathbf{V}} T)^\mu_\nu = V^\alpha \frac{\partial}{\partial x^\alpha} T^\mu_\nu + \frac{\partial V^\alpha}{\partial x^\nu} T^\mu_\alpha - \frac{\partial V^\mu}{\partial x^\alpha} T^\alpha_\nu \tag{49.2.21}$$

There is a second, more intuitive definition of the Lie derivative which is equivalent to the first, and can be used in treating symmetries.

DG6 Integration on manifolds

Let us take a smooth manifold M and consider a function $f : M \to \mathbb{R}$. Working in a chart (U, ϕ) we integrate $f \circ \phi$ in $\phi(U)$, we want to transition to another chart (U, φ) and integrate $f \circ \phi$ in $\varphi(U)$. This situation is summarized in the following commutative diagram.

where we defined $f_\phi = f \circ \phi^{-1}$ and $f_\varphi = f \circ \varphi^{-1}$. We can integrate in $\mathbb{R}^d$ and use the typical rule of change of variables to find:

$$\int_{\varphi(U)} d^n \tilde{x} f_\varphi(\tilde{x}) = \int_{\phi(U)} d^n x \big| \det\big(\partial_a(\varphi \circ \phi^{-1})^b\big)(x)\big| (f_\varphi \circ (\varphi \circ \phi^{-1}))(x) \tag{50.0.1}$$

$$= \int_{\phi(U)} d^n x \left| \det\left(\frac{\partial y^b}{\partial x^a}\right)(\phi^{-1}(x))\right| f_\phi(x) \tag{50.0.2}$$

$$\neq \int_{\phi(U)} d^n x f_\phi(x) \tag{50.0.3}$$

so if we want to define integration so that it is not hart-dependent then we must insert a factor that takes care of the Jacobian. One object which we know transforms inversely to the Jacobian are differential forms. Let for example α be a differential form which we expand in some local basis as:

$$\alpha = a(x^1, ..., x^n) d\varphi^1 \wedge ... \wedge d\varphi^n \tag{50.0.4}$$

Then if we define

$$\int_U \alpha = \int_{\varphi(U)} a(x^1, ..., x^n) d\varphi^1 ... d\varphi^n \tag{50.0.5}$$

we find that working in some other chart $(U, \tilde{\varphi})$:

$$\int_U \alpha = \int_{\tilde{\varphi}(U)} \tilde{a}(\tilde{x}^1, ..., \tilde{x}^n) d\tilde{\varphi}^1 ... d\tilde{\varphi}^n \tag{50.0.6}$$

$$= \int_{\varphi(U)} \det\left(\frac{\partial x^\mu}{\partial \tilde{x}^\nu}\right) a(x^1, ..., x^n) \det\left(\frac{\partial \tilde{x}^\mu}{\partial x^\nu}\right) d\varphi^1 ... d\varphi^n \tag{50.0.7}$$

$$= \int_{\varphi(U)} a(x^1, ..., x^n) d\varphi^1 ... d\varphi^n \tag{50.0.8}$$

as desired. It is therefore clear that we can use special differential forms, known as volume forms, to clear up the mess from the Jacobians.

Definition *(Volume form)*
Consider a smooth, n-dimensional manifold M. Then a $(0, n)$-tensor field Ω such that:
  (a) Ω does not vanish anywhere
  (b) Ω is totally antisymmetric
is known as a **volume form**.

Definition *(Metric volume form)*
On a metric manifold we can always construct a **metric volume form** in a chart (U, ϕ) from the metric g:
$$\Omega^\phi = \sqrt{\det\left(g^\phi_{\mu\nu}\right)} d\phi^1 \wedge ... \wedge d\phi^n \tag{50.0.9}$$

Proof. Our definition is seemingly chart-dependent so we must make sure that transformation under chart transitions are well-defined. Suppose we have two charts (U, ϕ) and (U, φ) with local bases x and $\tilde{x}$ respectively. We then have that:

$$\Omega^\varphi = \sqrt{\det(g^\varphi_{\mu\nu})} d\varphi^1 \wedge ... \wedge d\varphi^n \tag{50.0.10}$$

$$= \sqrt{\det\left(g^\phi_{\alpha\beta}\right) \det\left(\frac{\partial x^\alpha}{\partial \tilde{x}^\mu} \frac{\partial x^\beta}{\partial \tilde{x}^\nu}\right)} \det\left(\frac{\partial \tilde{x}}{\partial x}\right) d\phi^1 \wedge ... \wedge d\phi^n \tag{50.0.11}$$

$$= \sqrt{\det\left(g^\phi_{\alpha\beta}\right)} \left|\det\left(\frac{\partial x}{\partial \tilde{x}}\right)\right| \det\left(\frac{\partial \tilde{x}}{\partial x}\right) d\phi^1 \wedge ... \wedge d\phi^n \tag{50.0.12}$$

$$= \sqrt{\det\left(g^\phi_{\alpha\beta}\right)} \mathrm{sgn}\left(\det\left(\frac{\partial x}{\partial \tilde{x}}\right)\right) d\phi^1 \wedge ... \wedge d\phi^n \tag{50.0.13}$$

so this is only equal to Ω^ϕ if the Jacobian is positive, that is we must impose

$$\mathrm{sgn}\left(\det\left(\frac{\partial x}{\partial \tilde{x}}\right)\right) = 1 \tag{50.0.14}$$

In this case then we do indeed find that $\Omega^\phi = \Omega^\varphi$ so we may simply write the volume form as Ω. ■

Definition *(Integration on a metric manifold)*

Let f be a function on a chart domain U. Then:

$$\int_U f \equiv \int_{\phi(U)} d^n x \sqrt{\det(g_{\mu\nu})} f_\phi(x) \tag{50.0.15}$$

Part VII

Complex analysis

CA1 Complex numbers

51.1 What are complex numbers

> **Definition** *(Complex numbers and functions)*
> Let $x, y \in \mathbb{R}$ and let us denote by i the quantity such that $i^2 = -1$. Then we say that $z = x + iy$ is a **complex number** expressed in **Cartesian form**, with $\operatorname{Re} z = x$ representing the **real part** of z, and $\operatorname{Im} z = y$ representing the **imaginary part** of z. The set of all complex numbers is $\mathbb{C}$ and has the structure of a field with the operations:
>
> $$+ : \mathbb{C} \times \mathbb{C} \to \mathbb{C} \tag{51.1.1}$$
> $$(z_1, z_2) \mapsto z_1 + z_2 \tag{51.1.2}$$
>
> and
>
> $$\cdot : \mathbb{C} \times \mathbb{C} \to \mathbb{C} \tag{51.1.3}$$
> $$(z_1, z_2) \mapsto z_1 \cdot z_2 \tag{51.1.4}$$
>
> where if $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ then:
>
> $$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2) \tag{51.1.5}$$
> $$z_1 \cdot z_2 = x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1) \tag{51.1.6}$$

> **Definition** *(Complex conjugate)*
> Let $z = x + iy \in \mathbb{C}$. Then its **complex conjugate** is defined as $\overline{z} = x - iy$.

There are several useful properties that come with the complex conjugate:

> **Proposition** *(Complex conjugate properties)*
> Let $z_1, z_2 \in \mathbb{C}$, then:
> (i) $z_1 + \overline{z_1} = 2 \operatorname{Re} z_1$
> (ii) $z_1 - \overline{z_1} = 2i \operatorname{Im} z_1$
> (iii) $\overline{(\overline{z_1})} = z_1$
> (iv) $\overline{z_1 \pm z_2} = \overline{z_1} \pm \overline{z_2}$
> (v) $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$
> (vi) $\overline{\left(\frac{z_1}{z_2}\right)} = \frac{\overline{z_1}}{\overline{z_2}}$

Sice the arithmetic of real numbers applies equally well to complex numbers, it is not so surprising that the Binomial theorem holds in $\mathbb{C}$.

Theorem *(Complex binomial theorem)*

Let $z_1, z_2 \in \mathbb{C}$ and $n \in \mathbb{N}$, then:

$$(z_1 + z_2)^n = \sum_{k=0}^{n} \binom{n}{k} (z_1)^{n-k} z_2^k \tag{51.1.7}$$

Proof. Identical to the $\mathbb{R}$ counterpart. ∎

Similarly, the following result from real analysis still holds:

Theorem *(Geometric series)*

Let $z_1, z_2 \in \mathbb{C}$ and $n \in \mathbb{N}$, then:

$$z_1^n - z_2^n = (z_1 - z_2)(z_1^{n-1} + z_1^{n-2} z_2 + ... + z_2^{n-1}) \tag{51.1.8}$$

Due to the isomorphism between $\mathbb{C}$ and $\mathbb{R}^2$, we can represent complex numbers in a two-dimensional plane, known as the **complex plane**, with the x-axis replaced by the real axis, and the y-axis replaced by the imaginary axis. Recall also that in real analysis, the modulus of a real number gave the distance between the origin and this point. Similarly, we may define the complex counterpart as follows

Definition *(Complex modulus)*

Let $z = x + iy \in \mathbb{C}$ be a complex number. Then the **complex modulus** of z is defined as:

$$|z| = \sqrt{x^2 + y^2} \tag{51.1.9}$$

We list some important properties of the modulus.

Proposition *(Complex modulus properties)*

Let $z \in \mathbb{C}$, then:
  (i) $|z| \geq 0$, with equality only *iff* $z = 0$
  (ii) $|\bar{z}| = |z|$ and $|-z| = |z|$
  (iii) $|z|^2 = z\bar{z}$
  (iv) $|z_1 z_2| = |z_1||z_2|$
  (v) $\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}$

Note that the modulus of a complex number is not enough to fully specify it (unless it is zero). Indeed the set of all numbers in $\mathbb{C}$ with the same modulus r form a circle of radius r. It follows that the second piece of information needed to specify a complex number is the angle it makes with the real axis, known as the argument.

Definition *(Argument)*

Let $z = x + iy \in \mathbb{C}$. Then we define its **argument** $\theta \in \mathbb{R}$ so that:

$$\sin \theta = \frac{y}{r}, \ \cos \theta = \frac{x}{r} \tag{51.1.10}$$

where $r = |z|$.

Recall that by convention, positive angles are measured anti-clockwise. Also, note that a given number there has an infinite number of arguments. Indeed if θ is an argument then so is $\theta + 2k\pi$ for any $k \in \mathbb{Z}$.

Definition *(Principal argument)*
The argument θ of $z \in \mathbb{C}$ satisfying $-\pi < z \leq \pi$ is known as the **principal argument** $\arg z$.

Since the modulus and argument of z are enough to fully specify it, we may define a new form other than the Cartesian form in which complex numbers may be expressed.

Definition *(Polar form)*
Let z be a complex number with $|z| = r, \arg z = \theta$. Then z may be expressed in the form

$$z = r(\cos \theta + i \sin \theta) = re^{i\theta} \tag{51.1.11}$$

known as its **polar form**.

This polar form allows us to express the product/quotient of two complex numbers with the same ease we express the sum/difference of two complex numbers in cartesian form. Let $z_1 = r_1 e^{i\theta_1}$, $z_2 = r_2 e^{i\theta_2}$, then

$$z_1 z_2 = r_1 r_2 e^{i(\theta_1 + \theta_2)} = r_1 r_2 (\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)) \tag{51.1.12}$$

$$\frac{z_1}{z_2} = \frac{r_1}{r_2} e^{i(\theta_1 - \theta_2)} = \frac{r_1}{r_2} (\cos(\theta_1 - \theta_2) + i \sin(\theta_1 - \theta_2)) \tag{51.1.13}$$

This provides an alternative derivation of the identities $|z_1 z_2| = |z_1||z_2|$, $\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}$. Also, note that while the sum of an argument of z_1 and z_2 is an argument of $z_1 z_2$, the same does not hold for the principal argument. Indeed the sum of two principal arguments could be larger than π, and thus not qualify. Instead we have the weaker result that

$$\arg(z_1 z_2) = \arg z_1 + \arg z_2 + 2n\pi, \ n = -1, 0, 1 \tag{51.1.14}$$

In the special case where $z_1 = 1 = e^{i \cdot 0}$ then letting $z_2 = z = re^{i\theta}$ we find that

$$\frac{1}{z} = \frac{1}{r} e^{-i\theta} = \overline{\left(\frac{1}{r} e^{i\theta}\right)} = \frac{\overline{z}}{|z|^2} \tag{51.1.15}$$

from which it follows that:

$$\arg \overline{z} = \arg z^{-1} = -\arg z \tag{51.1.16}$$

Theorem *(De Moivre's theorem)*
It holds that if $\theta \in \mathbb{R}$ and $n \in \mathbb{N}$ then:

$$(\cos\theta + i\sin\theta)^n = \cos(n\theta) + i\sin(n\theta) \tag{51.1.17}$$

Proof. We notice that $\cos\theta + i\sin\theta = e^{i\theta}$ so:

$$(\cos\theta + i\sin\theta)^n = (e^{i\theta})^n = e^{i(n\theta)} = \cos(n\theta) + i\sin(n\theta) \tag{51.1.18}$$

as desired. ∎

Theorem *(Complex roots)*
Let $w = \rho(\cos\phi + i\sin\phi)$ be a non-zero complex number, and define the solutions to $z^n = w$ to be the n**th roots** of w. These roots are given by:

$$z_m = \rho^{1/n}\left[\cos\left(\frac{\phi}{n} + \frac{2m\pi}{n}\right) + i\sin\left(\frac{\phi}{n} + \frac{2m\pi}{n}\right)\right], \ m \in \mathbb{Z}_n \tag{51.1.19}$$

Proof. We need to solve $z^n = w$, so let $z = r(\cos\theta + i\sin\theta)$. Using DeMoivre's theorem yields

$$z^n = r^n(\cos(n\theta) + i\sin(n\theta)) = \rho(\cos\phi + i\sin\phi) \tag{51.1.20}$$

It is then clear that $r^n = \rho$ by equating moduli. Moreover, arguments can only differ by a multiple of 2π so $n\theta = \phi + 2m\pi$ where $m \in \mathbb{Z}$.Consequently solutions are of the form:

$$z_m = \rho^{1/n}\left[\cos\left(\frac{\phi}{n} + \frac{2m\pi}{n}\right) + i\sin\left(\frac{\phi}{n} + \frac{2m\pi}{n}\right)\right], \ m \in \mathbb{Z}_n \tag{51.1.21}$$

Note that these are not infinitely many solutions, one for each m. However, if m_1 and m_2 differ by an integer multiple kn then:

$$\frac{\phi}{n} + \frac{2m_2\pi}{n} = \frac{\phi}{n} + \frac{2m_1\pi}{n} + 2k\pi \tag{51.1.22}$$

so $z_{m_1} = z_{m_2}$. Consequently the only distinct solutions can be formed from $k = 0, 1, 2, ..., n-1$. Note that the angle between z_k and z_{k+1} is the same for all k, and their moduli are all the same. Consequently these roots lie on a circle and form the vertices of a regular n-polygon. ∎

Complex inequalities work the same way as real inequalities. We repeat the standard rules for rearranging inequalities for completeness:

(i) $a < b \iff 0 < b - a$

(ii) $a < b \iff a + c < b + c$

(iii) $a < b \iff ac < bc$ for $c > 0$ and $a < b \iff ac > bc$ for $c < 0$

(iv) $a > b \iff \frac{1}{a} < \frac{1}{b}$ for $a, b > 0$

(v) $a < b \iff a^p < b^p$ if $a, b \geq 0$ and $p > 0$

(vi) $|a| < b \iff -b < a < b$

(vii) $a < b$ and $b < c \implies a < c$

(viii) $a < b$ and $c < d \implies a + c < b + d$

(ix) $a < b$ and $c < d \implies ac < bd$ if $a, c \geq 0$

The triangle inequalities are also quite useful:

> **Proposition** *(Triangle inequalities)*
> If $z_1, z_2 \in \mathbb{C}$ then:
>
> $$|z_1 + z_2| \leq |z_1| + |z_2| \tag{51.1.23}$$
> $$||z_1| - |z_2|| \leq |z_1 - z_2| \tag{51.1.24}$$

Proof.

$$|z_1 + z_2|^2 = (z_1 + z_2)\overline{(z_1 + z_2)} \tag{51.1.25}$$
$$= (z_1 + z_2)(\overline{z_1} + \overline{z_2}) \tag{51.1.26}$$
$$= |z_1|^2 + z_1\overline{z_2} + \overline{z_1}z_2 + |z_2|^2 \tag{51.1.27}$$
$$= |z_1|^2 + 2\operatorname{Re}(z_1\overline{z_2}) + |z_2|^2 \tag{51.1.28}$$
$$\leq |z_1|^2 + 2|z_2\overline{z_2}| + |z_2|^2 \tag{51.1.29}$$
$$= (|z_1| + |z_2|)^2 \tag{51.1.30}$$

Applying one of the rules for rearranging inequalities gives (51.1.23).

Also, note that:
$$|z_1| \leq |z_1 - z_2| + |z_2| \implies |z_1| - |z_2| \leq |z_1 - z_2| \tag{51.1.31}$$

and similarly:
$$|z_2| \leq |z_2 - z_1| + |z_1| \implies |z_2| - |z_1| \leq |z_2 - z_1| \tag{51.1.32}$$

Consequently:
$$-|z_1 - z_2| \leq |z_2| - |z_1 \leq |z_1 - z_2| \implies ||z_2| - |z_1|| \leq |z_1 - z_2| \tag{51.1.33}$$

■

The geometrical interpretation of the first triangle inequality is that the diagonal of a parallelogram cannot be longer than the sum of the lengths of its two adjacent sides. The geometrical interpretation of the second triangle inequality is that the distance between two points on two concentric circles cannot be larger than the difference in the circles' radii.

As a consequence of the fact that $\mathbb{C}$ is not a complex field, we cannot write an inequality between two complex numbers $z_1 \leq z_2$, it simply does not make sense. Nevertheless one can use the modulus and argument of complex numbers to construct meaningful inequalities with geometric interpretations.

Definition *(Subsets of $\mathbb{C}$)*
A **line** is a set of the type $\{z : a\operatorname{Re}z + b\operatorname{Im}z > c\}$ where $a, b, c \in \mathbb{R}$ with the first two not both equal to zero. An **open half-plane** is a set of the type $\{z : a\operatorname{Re}z + b\operatorname{Im}z > c\}$ and a **closed half-plane** is a set of the type $\{z : a\operatorname{Re}z + b\operatorname{Im}z \geq c\}$.
The **circle** centered at $\alpha \in \mathbb{C}$ with radius $r > 0$ is the set $\{z : |z - \alpha| = r\}$. Analogously, an **open disc** is a set of the type $\{z : |z - \alpha| < r\}$ and a **closed disc** is a set of the type $\{z : |z - \alpha| < r\}$.
A **half-line** is a set of the form $\{z : \arg(z - \alpha) = \theta\}$ where $\alpha \in \mathbb{C}$ and $-\pi < \theta \leq \pi$. Hence an **open sector** is a set of the form $\{z : a < \arg(z-\alpha) < b\}$ or $\{z : \arg(z-\alpha) < a \vee \arg(z-\alpha) > b\}$.

51.2 Complex functions

Definition *(Complex functions)*
Let $A, B \subseteq \mathbb{C}$. Then we define the map:

$$f : A \to B \tag{51.2.1}$$
$$z \mapsto f(z) \tag{51.2.2}$$

to be a **complex function**. The **domain** of f is A, the **co-domain** of f is B, and the image of f is:

$$f(A) \equiv \{f(z) : z \in A\} \tag{51.2.3}$$

If $f(A) = B$ then f is **surjective**, while if $f(z_1) = f(z_2) \implies z_1 = z_2$ then f is **injective**.

Example. Consider for example $f(z) = \frac{3z+1}{z+i}$. It is implied that the domain of f is defined to be the subset of $\mathbb{C}$ where its rule is applicable. In this case, the domain of f is thus $A = \mathbb{C} \setminus \{-i\}$ and the codomain is $\mathbb{C}$. Consequently:

$$f(A) = \left\{ w = \frac{3z+1}{z+i} : z \in \mathbb{C} \setminus \{i\} \right\} \tag{51.2.4}$$

Now note that:

$$w = \frac{3z+1}{z+i} \implies w(z^2+1) = 3z^2 + (1-3i)z - i \tag{51.2.5}$$
$$\implies (w-3)z^2 - (1-3i)z + (w+i) = 0 \tag{51.2.6}$$

which has a well defined solution:

$$z = \frac{(1-3i) \pm \sqrt{(1-3i)^2 - 4(w-3)(w+i)}}{2(w-3)} = \tag{51.2.7}$$

as long as $w \neq 3$. Thus $f(A) = \mathbb{C} \setminus \{3\}$, and due to this f is not surjective. However it is

injective, since:

$$f(z_1) = f(z_2) \equiv \frac{3z_1 + 1}{z_1 + i} = \frac{3z_2 + 1}{z_2 + i} \implies 3z_1 z_2 + i + 3iz_1 + z_2 = 3z_2 z_1 + i + 3iz_2 + z_1 \tag{51.2.8}$$

$$\implies 3i(z_1 - z_2) = z_1 - z_2 \implies z_1 = z_2 \tag{51.2.9}$$

as desired. ◀

The standard operations of sums, products/quotients and compositions can be performed on complex functions just like with real functions.

Example. Let $f(z) = \frac{1}{z}$ with domain $z \in \mathbb{C} \setminus \{0\}$ and $g(z) = \frac{z+3i}{z^2 - z}$ with domain $z \in \mathbb{C} \setminus \{0, 1\}$. We see that the complex functions:

$$(f + g)(z) = \frac{2z - 1 + 3i}{z^2 - z}, \ (fg)(z) = \frac{z + 3i}{z(z^2 - z)} \tag{51.2.10}$$

have domain $\mathbb{C} \setminus \{0, 1\}$. Similarly, we have that:

$$\frac{f}{g} = \frac{z^2 - z}{z^2 + 3iz} = \frac{z - 1}{z + 3i} \tag{51.2.11}$$

has domain $\mathbb{C} \setminus \{0, 1, -3i\}$. Also, we can compose these functions to obtain:

$$(g \circ f)(z) = \frac{\frac{1}{z} + 3i}{\frac{1}{z^2} - \frac{1}{z}} = \frac{3iz^2 + z}{1 - z} \tag{51.2.12}$$

Its domain is:

$$\{z \in \mathbb{C} \setminus \{0\} : f(z) \in \mathbb{C} \setminus \{0, 1\}\} = \mathbb{C} \setminus \{0, 1\} \tag{51.2.13}$$

Similarly, we obtain that

$$(f \circ g)(z) = \frac{z^2 - z}{z + 3i} \tag{51.2.14}$$

whose domain is

$$\{z \in \mathbb{C} \setminus \{0, 1\} : g(z) \in \mathbb{C} \setminus \{0\}\} = \mathbb{C} \setminus \{0, 1, -3i\} \tag{51.2.15}$$

◀

Definition *(Inverse of complex function)*
Let $f : A \to B$ be a complex injective function. Then the inverse function of f is defined as:

$$f^{-1} : B \to B \tag{51.2.16}$$

$$z \mapsto f^{-1}(z) \tag{51.2.17}$$

such that:

$$f^{-1}(w) = z \iff f(w) = z \tag{51.2.18}$$

It follows that the existence of an inverse function boils down to whether or not one can find for every $w \in B$ a $z \in A$ such that $f(z) = w$. In many cases this cannot be done for a given B, but can be done if we somehow restrict B to some new codomain B', redefining our function to $f|_B$. Thus, a function is invertible if for all $w \in f(A)$ there exists a unique $z \in A$ such that $f(z) = w$.

Example. Let $A = \{0\} \cup \{z : -\pi/3 < \arg z \le \pi/3\}$ be the domain of $f(z) = z^3$. This function has an inverse. Indeed, note that the images set of A is:

$$f(A) = \{z^3 : z \in \{0\} \cup \{z : -\pi/3 < \arg z \le \pi/3\}\} \tag{51.2.19}$$
$$= \{z^3 : z = re^{i\theta}, -\pi/3 < \theta \le \pi/3, r \ge 0\} \tag{51.2.20}$$
$$= \{r^3 e^{3i\theta} : -\pi/3 < \theta \le \pi/3, r \ge 0\} \tag{51.2.21}$$
$$= \mathbb{C} \tag{51.2.22}$$

Hence, given $w \in \mathbb{C}$ we need to find a unique $z \in A$ such that $z^3 = w$. If $w = 0$ then clearly $f(0) = 0 \implies f^{-1}(0) = 0$. If instead $w \ne 0$ then:

$$z^3 = w = re^{i\theta}, \ r > 0, -\pi < \theta \le \pi \tag{51.2.23}$$

but using De Moivre's theorem we have that:

$$z = r^{1/3} e^{i(\theta + 2k\pi)/3}, \ \forall k = 0, 1, 2 \tag{51.2.24}$$

Since we need $z \in A$, the only possible choice is $k = 0$. Consequently, the inverse of f is:

$$f^{-1}(0) = 0, \ f^{-1}(w) = r^{1/3} e^{i\theta/3}, w \ne 0 \tag{51.2.25}$$

with domain $\mathbb{C}$ and co-domain A. ◀

Definition *(Real-valued functions)*
Given a complex function f, we say that it is **real valued** if $f(A) \subseteq \mathbb{R}$ and a **real function** if $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$.

Note that any complex function can be decomposed into real functions. For example consider $f(z) = \frac{1}{z}$. Then:
$$(\operatorname{Re} f)(z) = \operatorname{Re}(f(z)) = \operatorname{Re}\left(\frac{1}{z}\right) = \frac{\operatorname{Re} z}{|z|^2}, \ (z \in \mathbb{C} \setminus \{0\}) \tag{51.2.26}$$

and similarly:
$$(\operatorname{Im} f)(z) = \operatorname{Im}(f(z)) = \operatorname{Im}\left(\frac{1}{z}\right) = -\frac{\operatorname{Im} z}{|z|^2}, \ (z \in \mathbb{C} \setminus \{0\}) \tag{51.2.27}$$

so that:
$$f(z) = (\operatorname{Re} f)(z) + i(\operatorname{Im} f)(z) = \frac{\operatorname{Re} z - i \operatorname{Im} z}{|z|^2} = \frac{z^*}{|z|^2} \ (z \in \mathbb{C} \setminus \{0\}) \tag{51.2.28}$$

as expected.

Definition *(Complex function parametrisation)*
A **path** $\Gamma \subseteq \mathbb{C}$ is the image set of a continuous function $\gamma : I \to \mathbb{C}$ where $I \subseteq \mathbb{R}$, known as

the **parametrisation** of Γ.

Consider for example the unit circle Γ in the complex plane. It can be parametrised by:

$$\gamma(t) = \cos t + i \sin t, \ t \in [0, 2\pi] \tag{51.2.29}$$

which traverses the circle anti-clockwise. Consider the effect of letting z vary on Γ following the parametrisation γ, and seeing the effects on $f(z) = \frac{1}{z}$. We saw that letting $z = x + iy$ and $w = f(z) = u + iv$ then

$$u = \frac{x}{x^2 + y^2}, \ v = -\frac{y}{x^2 + y^2} \tag{51.2.30}$$

Thus, letting $x = \cos t$ and $y = \sin t$ then we find that:

$$u = \cos t, \ v = -\sin t, \ t \in [0, 2\pi] \implies z = u + iv = \cos(-t) + i \sin(-t), \ t \in [0, 2\pi] \tag{51.2.31}$$

Consequently, the unit circle in the w plane will be traversed clockwise as z traverses the unit circle anti-clockwise.

51.3 Mappings under complex functions

51.4 Special complex functions

We now extend notions of exponential, logarithmic, trigonometric and hyperbolic functions to the complex domain.

Definition *(Complex exponential)*
Let $z \in \mathbb{C}$. Then we have define:

$$e^z = e^{\operatorname{Re} z}(\cos(\operatorname{Im} z) + i \sin(\operatorname{Im} z)) \tag{51.4.1}$$

Proposition *(Exponential identities)*
Let $z_1, z_2 \in \mathbb{C}$, then:
  (i) $e^{z_1 + z_2} = e^{z_1} e^{z_2}$
  (ii) $|e^z| = e^{\operatorname{Re} z}$
  (iii) $e^{-z} = \frac{1}{e^z}$
  (iv) $e^{z + 2\pi i} = e^z$

Proof.   (i) We find that

$\blacksquare$

Example. Note that for $z \in \mathbb{C}$ then

$$|e^z| = \left| e^{\operatorname{Re} z}(\cos(\operatorname{Im} z) + i \sin(\operatorname{Im} z)) \right| = e^{\operatorname{Re} z} \tag{51.4.2}$$

and since $\operatorname{Re} z \leq |z|$ we have that $|e^z| \leq e^{|z|}$ by the monotonicity of the real exponential function. ◀

Interestingly, we have that $e^{z+2in\pi} = e^z$ for all $n \in \mathbb{Z}$, so going up the complex plane in steps of $2i\pi$ brings you back to your starting point under complex exponentiation. This suggests that we take the line $\Gamma : \gamma(t) = a + it$, and look at how it is mapped by $w = f(z) = e^z$. We have that if $z = x + iy$ then:

$$w = u + iv = e^x(\cos y + i \sin y) \implies u = e^x \cos y, \ v = e^x \sin y \tag{51.4.3}$$

thus:

$$u = e^a \cos t, \ v = e^a \sin t \tag{51.4.4}$$

so that $u^2 + y^2 = e^{2a}$. As expected, the straight line gets mapped to a circle. Similarly, taking the path $\Gamma : \gamma(t) = t + ib$ then:

$$w = u + iv = e^t(\cos b + i \sin b) \implies u = e^t \cos b, \ v = e^t \sin b \tag{51.4.5}$$

thus:

$$u = e^t \cos b, \ v = e^t \sin b \tag{51.4.6}$$

which is a half-line with $\arg z = b$. Hence if we take a grid of $x = $ cnst and $y = $ cnst then it will get mapped to outward radial lines with concentric circles. A rectangular region gets mapped to a quarter annulus. Points in the left half-plane with $\operatorname{Re} z < 0$ get mapped to points inside the unit circle while points in the right halfplane get mapped outside the unit circle as a result of $|e^z| = e^{\operatorname{Re} z}$. Finally, note that the image of the strip $\{x + iy : -\pi < y \leq \pi\}$ under $f(z) = e^z$ is $\mathbb{C} \setminus \{0\}$.

We have seen that $f(z) = e^z$ is not injective on $\mathbb{C}$, but we can make it injective by restricting its domain, for example, to a strip of width 2π (with one edge removed), such as:

$$A = \{x + iy : -\pi < y \leq \pi\} \tag{51.4.7}$$

Then we can show that $f(z) = e^z$ has an inverse on this domain.

> **Example.** We firstly look at the image set of $f(z)$:
>
> $$f(A) = \{w = e^{x+iy} : -\pi < y \leq \pi\} = \{w = e^x e^{iy} : -\pi < y \leq \pi\} = \{w = \rho e^{i\theta} : \rho > 0, -\pi < \theta \leq \pi\} \tag{51.4.8}$$
>
> $$= \mathbb{C} \setminus \{0\} \tag{51.4.9}$$
>
> Now, given any $w \in \mathbb{C} \setminus \{0\}$ then $w = \rho e^{i\theta}$ with $\rho > 0$ and $-\pi < \theta \leq pi$. We claim that $z = \log(\rho) + i\theta \in A$ is the required inverse (well defined since $\rho \neq 0$):
>
> $$f(z) = e^z = e^{\log(\rho)} e^{i\theta} = \rho e^{i\theta} = w \tag{51.4.10}$$
>
> as desired. We could have also chosen $z = \log(\rho) + i\theta + 2in\pi$ but this would not be in A. Consequently, f is an injective function with image set $\mathbb{C} \setminus \{0\}$ and inverse:
>
> $$f^{-1}(z) = \log|z| + i \arg z \tag{51.4.11}$$
>
> with domain $\mathbb{C} \setminus \{0\}$. ◄

Much like how a complex number can have infinitely many arguments, its logarithm is also multiply valued. However we can define the principal logarithm by restricting the argument of w to its principal argument.

Definition *(Principal logarithm)*

Let $z \in \mathbb{C} \setminus \{0\}$. The **principal logarithm** of z is defined as:

$$\log(z) = \log|z| + i \arg z \tag{51.4.12}$$

We collect some useful properties of the principal logarithm below:

Proposition *(Logarithm identities)*

For $\arg z_1, \arg z_2 \in (-\pi/2, \pi/2]$ then:

$$\log(z_1 z_2) = \log z_1 + \log z_2 \tag{51.4.13}$$

and for $\arg z \in (-\pi, \pi)$ then:

$$\log\left(\frac{1}{z}\right) = -\log z \tag{51.4.14}$$

Proof. Suppose $\arg z_1, \arg z_2 \in (-\pi/2, \pi/2]$ so that $\arg(z_1 z_2) = \arg z_1 + \arg z_2 \in (-\pi, \pi]$. Then:

$$\log(z_1 z_2) = \log|z_1 z_2| + i \arg(z_1 z_2) = \log|z_1| + \log|z_2| + i \arg z_1 + i \arg z_2 = \log(z_1) + \log(z_2) \tag{51.4.15}$$

Similarly, if $\arg z \in (-\pi, \pi)$ then $\arg \frac{1}{z} = -\arg z$ so that:

$$\log\left(\frac{1}{z}\right) = \log\left|\frac{1}{z}\right| + i \arg \frac{1}{z} = -\log|z| - i \arg z = -\log z \tag{51.4.16}$$

as desired. $\blacksquare$

We can view the geometric effect of the complex logarithm in the same (or inverse) way for the complex exponential. Going back to the complex exponential, we can invert Euler's identity to express $\cos z$ and $\sin z$ as follows:

$$\begin{cases} e^{i\theta} = \cos\theta + i\sin\theta \\ e^{-i\theta} = \cos\theta - i\sin\theta \end{cases} \implies \begin{cases} \cos\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) \\ \sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) \end{cases} \tag{51.4.17}$$

This motivates us to make the following definitions:

Definition *(Complex trigonometric functions I)*

We define for all $z \in \mathbb{C}$:

$$\cos z = \frac{1}{2}(e^{iz} + e^{-iz}), \ \sin z = \frac{1}{2i}(e^{iz} - e^{-iz}) \tag{51.4.18}$$

Interestingly, the zeros of the complex trigonometric functions are the same as their real counterparts. Indeed:

$$\sin z = 0 \implies e^{iz} = e^{-iz} \implies e^{2iz} = 1 = e^0 \implies 2z = 2n\pi \implies z = n\pi \forall n \in \mathbb{Z} \tag{51.4.19}$$

and similarly:

$$\cos z = 0 \implies e^{iz} = -e^{-iz} \implies e^{2iz} = -1 = e^{i\pi} \tag{51.4.20}$$

$$\implies 2z = \pi + 2n\pi \implies z = \left(n + \frac{1}{2}\right)\pi, \ \forall n \in \mathbb{Z} \tag{51.4.21}$$

Thus, we may define the tangent, secant, cotangent and cosecant as follows:

Definition *(Complex trigonometric functions II)*
We define for all $z \in \mathbb{C} \setminus \{z = (n + 1/2)\pi : n \in \mathbb{Z}\}$:

$$\tan z = \frac{\sin z}{\cos z}, \ \sec z = \frac{1}{\cos z} \tag{51.4.22}$$

and for all $z \in \mathbb{C} \setminus \{z = n\pi : n \in \mathbb{Z}\}$:

$$\cot z = \frac{\cos z}{\sin z}, \ \csc z = \frac{1}{\sin z} \tag{51.4.23}$$

Proposition *(Trigonometric identities)*
  (i) Addition:

$$\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \cos z_1 \sin z_2 \tag{51.4.24}$$
$$\cos(z_1 + z_2) = \cos z_1 \cos z_2 - \sin z_1 \sin z_2 \tag{51.4.25}$$
$$\tan(z_1 + z_2) = \frac{\tanh z_1 + \tanh z_2}{1 - \tanh z_1 \tanh z_2} \tag{51.4.26}$$

  (ii) Squares:

$$\cos^2 z + \sin^2 z = 1 \tag{51.4.27}$$
$$\sec^2 z = 1 + \tan^2 z \tag{51.4.28}$$
$$\csc^2 z = 1 + \cot^2 z \tag{51.4.29}$$

Proof. The proof is identical to the real case. ∎

Suppose we remove the i in the exponentials defining $\sin z$ and $\cos z$. We know from real analysis that the result are hyperbolic functions:

$$\sinh x = \frac{1}{2}(e^x - e^{-x}), \ \cosh x = \frac{1}{2}(e^x + e^{-x}) \tag{51.4.30}$$

Thus, extending these definitions to $\mathbb{C}$, we get:

Definition *(Complex hyperbolic functions)*
We define for all $z \in \mathbb{C}$:

$$\cosh z = \frac{1}{2}(e^z + e^{-z}), \ \sinh z = \frac{1}{2}(e^z - e^{-z}) \tag{51.4.31}$$

We define for all $z \in \mathbb{C} \setminus \{z = i(n + 1/2)\pi : n \in \mathbb{Z}\}$:

$$\tanh z = \frac{\sinh z}{\cosh z}, \ \operatorname{sech} z = \frac{1}{\cosh z} \tag{51.4.32}$$

and for all $z \in \mathbb{C} \setminus \{z = in\pi : n \in \mathbb{Z}\}$:

$$\coth z = \frac{\cosh z}{\sinh z}, \ \operatorname{csch} z = \frac{1}{\sinh z} \tag{51.4.33}$$

Proposition *(Trigonometric functions are hyperbolic)*
For all $z \in \mathbb{C}$:

$$\cosh(iz) = \cos z, \ \sinh(iz) = i \sin z \tag{51.4.34}$$

Proof. We find:

$$\cosh(iz) = \frac{1}{2}(e^{iz} + e^{-iz}) = \cos z \tag{51.4.35}$$

$$\sinh(iz) = i\frac{1}{2i}(e^{iz} - e^{-iz}) = i \sin z \tag{51.4.36}$$

as desired. ∎

Proposition *(Hyperbolic identities)*
(i) Addition:

$$\sinh(z_1 + z_2) = \sinh z_1 \cosh z_2 + \cosh z_1 \sinh z_2 \tag{51.4.37}$$
$$\cosh(z_1 + z_2) = \cosh z_1 \cosh z_2 + \sinh z_1 \sinh z_2 \tag{51.4.38}$$
$$\tanh(z_1 + z_2) = \frac{\tanh z_1 + \tanh z_2}{1 + \tanh z_1 \tanh z_2} \tag{51.4.39}$$

(ii) Squares:

$$\cosh^2 z - \sinh^2 z = 1 \tag{51.4.40}$$
$$\operatorname{sech}^2 z = 1 - \tanh^2 z \tag{51.4.41}$$
$$\operatorname{csch}^2 z = \coth^2 z - 1 \tag{51.4.42}$$

Proof. Can be found from the trigonometric identities using the previous proposition's correspondence between $\sin z, \sinh z$ and $\cos z, \cosh z$. ∎

CA2 Continuity of complex functions

52.1 Complex sequences

We begin our study of continuity by extending some results from the rReal analysis of sequences to $\mathbb{C}$.

> **Definition** *(Convergence in $\mathbb{C}$)*
> The complex sequence (z_n) **converges** to α if $\forall \epsilon > 0$ there exists an integer N such that
>
> $$|z_n - \alpha| < \epsilon, \ \forall n > N \tag{52.1.1}$$
>
> If the sequence converges to zero then it is a **null sequence**. Equivalently, (z_n) converges to α if $(z_n - \alpha)$ is null. If a sequence is not convergent then it is **divergent**.

The geometrical intuition behind the epsilon-delta definition is clear. If (z_n) converges to α then given a circle centered at α with finite radius, all terms after a given number will fall within this circle. This implies that every convergent sequence is bounded.

> **Proposition** *(Bounded $\impliedby$ convergent)*
> Every convergent sequence is bounded.

Proof. Let (z_n) converge to α. Applying the triangle inequality to the epsilon-delta definition with $\epsilon = 1$ we get
$$|z_n - \alpha| \leq |z_n| - |\alpha| < 1 \implies |z_n| < 1 + |\alpha|, \ \forall n > N \tag{52.1.2}$$
Choosing $M = \max\{z_1, z_2, ..., z_N, 1 + |\alpha|\}$ it is clear that $|z_n| < M$ for all n. ∎

> **Example.** We want to prove that the sequence (z_n) with $z_n = \frac{1+i}{n}$ is convergent. We claim that it converges to 0. Indeed let $\epsilon > 0$, we need to find an integer N such that
>
> $$\left| \frac{1+i}{n} \right| < \epsilon \implies \frac{\sqrt{2}}{n} < \epsilon, \ \forall n > N \tag{52.1.3}$$
>
> which is clearly satisfied if $N > \frac{\sqrt{2}}{\epsilon}$. Thus (z_n) is a null sequence. ◄

As usual most sequence converges theorems from real analysis still apply. One important example is the Squeeze rule.

Theorem *(Squeeze rule)*
If (a_n) is a null non-negative sequence and if (z_n) is a sequence such that

$$|z_n| \leq a_n, \ n = 1, 2, \dots \tag{52.1.4}$$

then (z_n) is a null sequence.

Example. Let us prove that (z_n) with $z_n = \left(\frac{i}{2}\right)^n$ is a null sequence. We have that

$$|z_n| = \frac{1}{2^n} \leq \frac{1}{n} \tag{52.1.5}$$

but $\left(\frac{1}{n}\right)$ is a null sequence. Hence by the Squeeze rule (z_n) is null. ◀

Theorem *(Standard null sequences)*
The following sequences are null
  (a) $\left(\frac{1}{n^p}\right)$ for $p > 0$
  (b) α^n for $|\alpha| < 1$

Example. Let us find the limit of (z_n) with $z_N = \frac{(3+i)^n + (2+2i)^n}{(1+2i)n + 2(3+i)^n}$. We can guess that the dominant term will be $(3 + i)^n$ so we divide everything by it:

$$z_n = \frac{1 + (2 + 2i)^n/(3 + i)^n}{2 + (1 + 2i)^n/(3 + i)^n} \tag{52.1.6}$$

We have that $\frac{(2+2i)^n}{(3+i)^n}$ is a null sequence since

$$\left|\frac{(2 + 2i)^n}{(3 + i)^n}\right| = \left(\sqrt{\frac{4}{5}}\right)^n \tag{52.1.7}$$

Similarly $\frac{(1+2i)^n}{(3+i)^n}$ is a null sequence since

$$\left|\frac{(1 + 2i)^n}{(3 + i)^n}\right| = \frac{1}{\sqrt{2}^n} \tag{52.1.8}$$

Consequently we find that

$$\lim_{n \to \infty} z_n = \frac{1}{2} \tag{52.1.9}$$

so z_n converges to $\frac{1}{2}$. ◀

Proposition *(Limit properties)*
If $\lim_{n \to \infty} z_n = \alpha$ then
  (a) $\lim_{n \to \infty} |z_n| = |\alpha|$
  (b) $\lim_{n \to \infty} \overline{z_n} = \overline{\alpha}$

(c) $\lim_{n\to\infty} \operatorname{Re}(z_n) = \operatorname{Re}(\alpha)$

(d) $\lim_{n\to\infty} \operatorname{Im}(z_n) = \operatorname{Im}(\alpha)$

Proof. (a) We have that $||z_n| - |\alpha|| \le |z_n - \alpha| \le \epsilon$ by definition for all $n > N$. Hence for a given $\epsilon > 0$, a value of N which proves the convergence of z_n will also prove the convergence of $|z_n|$. In other words, the result follows from the Squeeze rule.

(b) Immediate from $|\overline{z_n} - \overline{\alpha}| = |z_n - \alpha|$

(c) Immediate from $|\operatorname{Re}(z_n) - \operatorname{Re}(\alpha)| = |\operatorname{Re}(z_n - \alpha)| \le |z_n - \alpha|$ and the application of the Squeeze rule.

(d) Immediate from $|\operatorname{Im}(z_n) - \operatorname{Im}(\alpha)| = |\operatorname{Im}(z_n - \alpha)| \le |z_n - \alpha|$ and the application of the Squeeze rule.

■

We can visualize the above proposition in the following figure
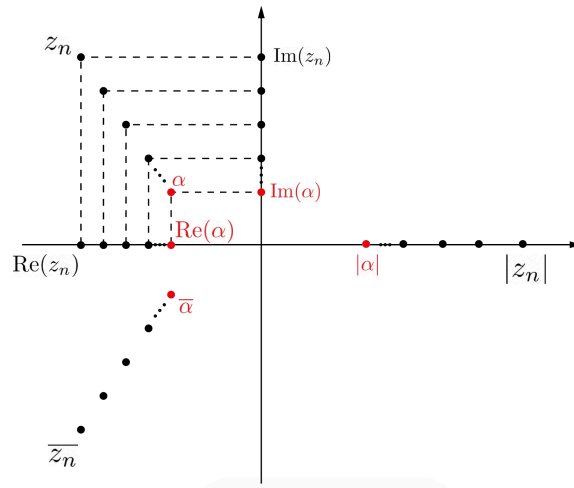
Finally, it is also useful to be able to combine different limits using the standard operations of addition, multiplication and division.

Proposition *(Limit combinations)*

Let $\lim_{n\to\infty} z_n = \alpha$ and $\lim_{n\to\infty} w_n = \beta$. Then

(i) $\lim_{n\to\infty}(z_n + w_n) = \alpha + \beta$

(ii) $\lim_{n\to\infty}(\lambda z_n) = \lambda\alpha$ for $\lambda \in \mathbb{C}$

(iii) $\lim_{n\to\infty}(z_n w_n) = \alpha\beta$

(iv) $\lim_{n\to\infty} \frac{z_n}{w_n} = \frac{\alpha}{\beta}$ if $\beta \neq 0$

Proof. Since z_n converges to α and w_n converges to β we have that for all $\epsilon_z, \epsilon_w > 0$ there exists N_z and N_w such that

$$|z_n - \alpha| < \epsilon_z, \ \forall n > N_z \tag{52.1.10}$$
$$|w_n - \beta| < \epsilon_w, \ \forall n > N_w \tag{52.1.11}$$

(i) Let $\epsilon > 0$ and set $\epsilon_z = \epsilon_w = \frac{1}{2}\epsilon$. Then, letting $N = \max(N_z, N_w)$ we have that

$$|(z_n + w_n) - (\alpha + \beta)| \le |z_n - \alpha| + |w_n - \beta| < \epsilon, \ \forall n > N \tag{52.1.12}$$

as desired.

(ii) Let $\epsilon > 0$ and set $|\epsilon_z| = \frac{\epsilon}{|\lambda|}$. Then, letting $N = N_z$ we have that

$$|\lambda z_n - \lambda \alpha| = |\lambda||z_n - \alpha| < |\lambda|\epsilon_z = \epsilon, \ \forall n > N \tag{52.1.13}$$

as desired.

(iii) Let $\epsilon > 0$. We have that

$$\lim_{n \to \infty} (z_n w_n - \alpha\beta) = \lim_{n \to \infty} [z_n(w_n - \beta) + \beta(z_n - \alpha)] \tag{52.1.14}$$

The second term vanishes by the convergence of z_n and the scalar multiple property we proved in (ii). Furthermore, since z_n is convergent it is also bounded by some real number M. Hence

$$z_n w_n - \alpha\beta \le M(w_n - \beta) \implies \lim_{n \to \infty} (z_n w_n - \alpha\beta) = 0 \tag{52.1.15}$$

as desired.

(iv) Let $\epsilon > 0$. We have that

$$\left| \frac{z_n}{w_n} - \frac{\alpha}{\beta} \right| = \frac{|\beta||z_n - \alpha| - |\alpha||w_n - \beta|}{|\beta||w_n|} \tag{52.1.16}$$

Our main worry is that $|w_n| = 0$ for some n making this quantity ill-defined. However, $|w_n|$ must eventually be positive, since

$$|\beta| - |w_n| < |w_n - \beta| < |\frac{1}{2}|\beta| \implies |w_n| > \frac{1}{2}|\beta| \ \forall n > N \tag{52.1.17}$$

Consequently

$$\left| \frac{z_n}{w_n} - \frac{\alpha}{\beta} \right| < \frac{2}{|\beta|^2}(|\beta||z_n - \alpha| - |\alpha||w_n - \beta|) \tag{52.1.18}$$

and since the RHS defines a null sequence, the LHS must also be null by the Squeeze rule. ∎

We have given a brief rundown of convergent complex sequences. Now let us look at the opposite case, that of divergent sequences. One very common family of divergent sequences are those that tend to infinity which we define below.

> **Definition** *(Infinite sequence)*
> A sequence (z_n) **tends to infinity** if $\forall M > 0$ there exists an integer N such that $|z_n| >$

M, $\forall n > N$.

In other words a sequence converges to infinity if its points in the complex plane cannot be enclosed by a circle centered at the origin of finite radius. This is opposite to the case of a null sequence, where all terms after a certain point fall into any given circle of finite radius. Recalling that taking the reciprocal of a complex number is equivalent to scaling it by its norm and reflecting it in the real axis, it seems like the reciprocal of a null sequence tends to infinity and vice-versa.

Theorem *(Reciprocal rule)*
Let (z_n) b a sequence. Then (z_n) tends to infinity *iff* $\left(\frac{1}{z_n}\right)$ is null.

Proof. Suppose that (z_n) is null and let $\epsilon > 0$ so that $|z_n| \leq \epsilon$, $\forall n > N$ for some integer N. Then

$$\left|\frac{1}{z_n}\right| = \frac{1}{|z_n|} > \epsilon, \ \forall n > N \tag{52.1.19}$$

So given $\epsilon > 0$ then by choosing the N which satisfies the epsilon-delta convergence of (z_n) we also find the N required to prove the divergence of $\left(\frac{1}{z_n}\right)$. The converse proof is identical. ∎

Example. We have that $z_n = n^3 - in^2 + (1+i)n$ tends to infinity since

$$\frac{1}{z_n} = \frac{1}{n^3 - in^2 + (1+i)n} = \frac{1}{n^3} \frac{1}{1 - i/n + (1+i)/n^2} \to 0 \cdot 1 = 0 \tag{52.1.20}$$

◀

Theorem *(Subsequence rules)*
First subsequence rule: the sequence (z_n) diverges if it has two convergent subsequences with different limits.
Second subsequence rule: the sequence (z_n) diverges if (z_n) has a subsequence that tends to infinity.

Proof. (i) Trivial by applicaation of epsilon-delta definition.

(ii) Trivial by boundedness of convergent sequences.

∎

Example. Consider the sequence (z_n) with $z_n = n^2 \sin(n\pi/3)$. The subsequence $z_{3k+1} = \frac{\sqrt{3}}{2}(3k+1)^2$ clearly diverges. Indeed its reciprocal

$$\frac{1}{z_{3k+1}} = \frac{2}{\sqrt{3}} \frac{1}{(3k+1)^2} \leq \frac{2}{\sqrt{3}} \frac{1}{k^2} \tag{52.1.21}$$

is a null sequence. Hence by the second subsequence rule (z_n) also diverges. ◀

> **Proposition** *(Subsequence decomposition)*
> If a sequence (z_n) can be decomposed into two subsequences converging to the same value then (z_n) also converges to the same value.

Proof. ∎

52.2 Continuity of complex functions

> **Definition** *(Sequential continuity)*
> Let $f : A \to \mathbb{C}$ and $\alpha \in A$, then f is continuous at α if for each sequence (z_n) in A such that $z_n \to \alpha$:
> $$f(z_n) \to f(\alpha) \tag{52.2.1}$$
> If f is continuous at all $\alpha \in A$ then we say that it is continuous on A. If f is not continuous at α then it is discontinuous.

Example. We consider some examples
  (i) Let $f(z) = \overline{z}$ whose trivial domain is $\mathbb{C}$. We have that if (z_n) is a sequence on $\mathbb{C}$ converging to some $\alpha \in \mathbb{C}$ then $\lim_{n\to\infty} f(z_n) = \lim_{n\to\infty} \overline{z_n} = \overline{\alpha} = f(\alpha)$ so f is indeed continuous.
  (ii) Let $f(z) = \mathrm{Im}(z)$ whose trivial domain is $\mathbb{C}$. We have that if (z_n) is a sequence on $\mathbb{C}$ converging to some $\alpha \in \mathbb{C}$ then $\lim_{n\to\infty} f(z_n) = \lim_{n\to\infty} \mathrm{Im}(z_n) = \mathrm{Im}(\alpha) = f(\alpha)$ so f is indeed continuous.
  (iii) Let $f(z) = |z|$ whose trivial domain is $\mathbb{C}$. We have that if (z_n) is a sequence on $\mathbb{C}$ converging to some $\alpha \in \mathbb{C}$ then $\lim_{n\to\infty} f(z_n) = \lim_{n\to\infty} |z| = |\alpha| = f(\alpha)$ so f is indeed continuous.
  (iv) Consider $f(z) = \arg(z)$, defined on $\mathbb{C}$, and let $z_n = e^{i(\pi+1/n)}$ for $n = 1, 2, \dots$ Then we have that $f(\lim_{n\to\infty} z_n) = \arg(e^{i\pi}) = \pi$ while $\lim_{n\to\infty} f(z_n) = \lim_{n\to\infty}\left(\frac{1}{n} - \pi\right) = -\pi$. Hence $f(z)$ is discontinuous at $z = -1$. Consequently by the multiple rule for complex limits, $f(z)$ is discontinuous on the negative real line.

◀

In general we can combine continuous functions to get another continuous function back.

> **Theorem** *(Combination rules)*
> Let f, g be continuous functions at α and let $\lambda \in \mathbb{C}$, then
>   (i) $f + g$ is continuous at α
>   (ii) λf is continuous at α
>   (iii) fg is continuous at α
>   (iv) f/g is continuous at α provided $g(\alpha) \neq 0$

Proof. These are immediate from the combination of limit theorems. ∎

We can also compose continuous functions to get other continuous functions.

> **Theorem** *(Composition rule)*
> Let f be continuous at α and let g be continuous at $f(\alpha)$. Then $g \circ f$ is continuous at α.

Proof. We have that for any sequence (z_n) on $\mathbb{C}$, if $z_n \to \alpha$ then $f(z_n) \to f(\alpha)$. Furthermore, g is continuous at $f(\alpha)$ so given any sequence $w_n \to \beta$, we must have that $g(w_n) \to g(\beta)$. In our case we consider the sequence $f(z_n)$, then $g(f(z_n)) \to g(f(\alpha))$. It thus follows that for any sequence $(z_n) \to \alpha$, $(g \circ f)(z_n) \to (g \circ f)(\alpha)$ as desired. ∎

> **Proposition** *(Restriction rule)*
> Let f, g be functions defined on A, B respectively such that $A \subseteq B$. If
> (i) $f(z) = g(z)$ for all $z \in A$
> (ii) g is continuous at $\alpha \in A$
> then f is continuous at α.

Proof. Let z_n be a sequence on A, and thus also in B such that $z_n \to \alpha$. Then we know that $g(z_n) = f(z_n) \to g(\alpha) = f(\alpha)$. Hence f is continuous at α. ∎

> **Example.** We consider some examples
> (i) Consider $f(x) = \frac{x^2+i}{x^2-i}$ defined for $x \in \mathbb{R}$. This is a restriction of $g(z) = \frac{z^2+i}{z^2-i}$ over $\mathbb{C} \setminus \{e^{i\pi/4}\}$. We have that $z^2 - i \neq 0$ for $z \in \mathbb{R}$ so by the quotient combination rule we have that $g(z)$ is continuous for all $z \in \mathbb{R}$, and thus so is $f(x)$.
> (ii) Consider $f(z) = \log|z|$ defined over $\mathbb{C}$. We have that $g(z) = |z|$ is a continuous function (this was proven earlier), and we know from real analysis that $h(x) = \log x$ is continuous for $x > 0$. Since $|z| > 0$ for all $z \in \mathbb{C}$, it follows from the composition rule that $f(z)$ is continuous.
>
> ◄

We now provide an equivalent definition of continuous functions.

> **Definition** *($\epsilon - \delta$ continuity)*
> Let $f : A \to \mathbb{c}$ and $\alpha \in A$. Then f is continuous at α if for all $\epsilon > 0$ there exists a $\delta > 0$ such that
> $$|f(z) - f(\alpha)| < \epsilon, \ \forall z \in A \text{ s.t. } |z - \alpha| < \delta \tag{52.2.2}$$

Geometrically, this means that given any open disc of radius ϵ around $f(\alpha)$, there exists an open disc of some radius δ around α such that its image under f lies within the disc of radius ϵ.

> **Theorem** *(Sequential and $\epsilon - \delta$ continuity are equivalent)*
> A function $f(z)$ is continuous according to the sequential definition *iff* it is continuous according to the $\epsilon - \delta$ definition.

Proof. Let $\epsilon > 0$. Suppose that $|f(z) - f(\alpha)| < \epsilon$ for all z such that $|z - \alpha| < \delta$, where $\delta > 0$. Let (z_n) be a sequence on A with $z_n \to \alpha$, so there exists N such that $\forall n > N, |z_n - \alpha| < \delta$. This however implies that $|f(z_n) - f(\alpha)| < \epsilon$, proving that f is continuous according to the $\epsilon - \delta$ definition.

Now suppose that there is some $\epsilon > 0$ such that for all $\delta > 0$, $|f(z) - f(\alpha)| \geq \epsilon$ for z with $|z - \alpha| < \delta$. So setting $\delta = \frac{1}{n}$ then $|f(z_n) - d(\alpha)| \geq \epsilon$ for $z_n \in A$ with $|z_n - \alpha| < \frac{1}{n}$. Clearly $z_n \to \alpha$ but $f(z_n) \not\to f(\alpha)$. ∎

> **Example.** Let us show that $f(z) = \arg(z)$ is continuous on $A = \mathbb{C} \setminus \{x \in \mathbb{R} : x \leq 0\}$. Let $\alpha \in A$, and let $\epsilon > 0$, then we need to find $\delta > 0$ such that
>
> $$|z - \alpha| < \delta \implies |f(z) - f(\alpha)| < \epsilon \tag{52.2.3}$$
>
> We choose δ so that the disc $\{z \in \mathbb{C} : |z - \alpha| < 0\}$ does not cross the negative real axis. This is important to ensure that the disc lies entirely in A. We also choose it so that the angular size of the circle from the origin is less than 2ϵ, which is equivalent to saying that we choose $\delta > 0$ such that $\sin^{-1}(\delta/|\alpha|) < \epsilon$.
> Then we find that if z is in this disc then
>
> $$|f(z) - f(\alpha)| = |\arg(z) - \arg(\alpha)| < \epsilon \tag{52.2.4}$$
>
> as desired. ◀

52.3 Limits of complex functions

> **Definition** *(Accumulation point)*
> Let $A \subseteq \mathbb{C}$. Suppose there exists a sequence (z_n) in $A \setminus \{\alpha\}$ such that $z_n \to \alpha$. Then α is an accumulation point of A.

> **Definition** *(Limit of a function)*
> Let α be an accumulation point of A and let $f(z)$ be defined on A. Then we say that $\lim_{z \to \alpha} f(z) \to \beta$ if for any sequence z_n in $A \setminus \{\alpha\}$ such that $z_n \to \alpha$, we have that $f(z_n) \to \beta$.

> **Example.**
> (i) Consider the following limit
>
> $$\lim_{z \to i} \frac{z^3 + i}{z - i} \tag{52.3.1}$$
>
> The function $f(z) = (z^3 + i)/(z - i)$ is defined on $\mathbb{C} \setminus \{i\}$, and i is a limit point in this set. Indeed consider $z_n = i - \frac{1}{n}$, which lies entirely in the given region, it is clear that $z_n \to i$ as $n \to \infty$.
> Therefore, if z_n is a sequence in $\mathbb{C} \setminus \{i\}$ such that $z_n \to i$ then we have that
>
> $$f(z_n) = \frac{z_n^3 + i}{z_n - i} = (z_n^2 + iz_n - 1) \to -3 \tag{52.3.2}$$

so

$$\lim_{z \to i} \frac{z^3 + i}{z - i} = -3 \tag{52.3.3}$$

(ii) Consider the limit

$$\lim_{z \to 0} \frac{z}{\text{Re}(z)} \tag{52.3.4}$$

Note that $f(z) = \frac{z}{\text{Re}(z)}$ is defined on $\mathbb{C} \setminus \{z : \text{Re}(z) = 0\}$ and thus 0 is a limit point in this domain. Consider $x \in \mathbb{R} \setminus \{0\}$ we have that $\frac{x}{\text{Re}(x)} = 1$. Instead $\frac{ix}{\text{Re}(ix)}$ is ill-defined. Consequently let us consider the sequence $z_n = \frac{1}{n} \to 0$. We have that

$$f(z_n) = \frac{1/n}{1/n} \to 1 \tag{52.3.5}$$

On the other hand, consider the sequence $w_n = \frac{1}{n} + \frac{i}{n}$. Then we have that

$$f(w_n) = \frac{(1+i)/n}{1/n} \to 1 + i \tag{52.3.6}$$

so there is no β such that $f(z_n)$ converges to the same limit for all sequences z_n.

◀

Theorem *(Continuity with limits)*
Let f be a function defined on A and let $\alpha \in A$ be a limit point of A. Then f is continuous at α iff $\lim_{z \to \alpha} f(z) = f(\alpha)$.

Proof. Suppose that f be continuous at α. Then given any sequence $z_n \to \alpha$ we have that $f(z_n) \to f(\alpha)$, which is the definition of $\lim_{z \to \alpha} f(z) = f(\alpha)$.

Now suppose that $\lim_{z \to \alpha} f(z) = f(\alpha)$. Then let (z_n) be a sequence on $A \setminus \{\alpha\}$ such that $z_n \to \alpha$. Then we can separate (z_n) into a subsequence (z_{n_k}) such that $f(z_{n_k}) = f(\alpha)$ for all k, and a subsequence (z_{m_k}) such that $f(z_{m_k}) \neq f(\alpha)$. Clearly $f(z_{n_k}) \to f(\alpha)$ and $f(z_{m_k}) \to f(\alpha)$ by assumption, so we have that f is continuous at α. ∎

Finally, to facilitate computations with limits we have the typical combination rules which can be proven using the sequence limit rules

Theorem *(Combination rules)*
Let f, g be functions with domains A, B respectively, and let α be a limit point of $A \cup B$ so that

$$\lim_{z \to \alpha} f(z) = \beta, \ \lim_{z \to \alpha} g(z) = \gamma \tag{52.3.7}$$

Then
  (i) $\lim_{z \to \infty} (f(z) + g(z)) = \beta + \gamma$
  (ii) $\lim_{z \to \infty} (\lambda f(z)) = \lambda \beta$ for $\lambda in \mathbb{C}$
  (iii) $\lim_{z \to \infty} (f(z)g(z)) = \beta \gamma$

(iii) $\lim\limits_{z \to \infty} \frac{f(z)}{g(z)} = \frac{\beta}{\gamma}$ if $\gamma \neq 0$

52.4 Topology on $\mathbb{C}$

Definition *(Open set)*

A set $A \subseteq \mathbb{C}$ is open if $\forall \alpha \in A$, there exists an open disc $D_r(\alpha) = \{z \in \mathbb{C} : |z - \alpha| < r\}$ of some radius $r > 0$ that lies entirely in A, that is $D_r(\alpha) \subseteq A$.

Example.

(i) The set $A = \mathbb{C} \setminus \{0\}$ is open. Let $\alpha \in A$, then clearly $D_{|\alpha|}(\alpha) = \{z \in \mathbb{C} : |z - \alpha| < |\alpha|\}$ lies entirely in A.

(ii) The set $A = \{z : -2 < \mathrm{Re}(z) < 2, -1 < \mathrm{Im}(z) < 1\}$ is open. Let $\alpha \in A$, then let $r = \min\{\mathrm{Re}(\alpha) - 2, \mathrm{Re}(\alpha) + 2, \mathrm{Im}(\alpha) - 1, \mathrm{Im}(\alpha) + i\}$ which is non-zero since $\alpha = \pm 2, \pm i$. Then $D_r(\alpha)$ is entirely contained in A.

(iii) The set $A = \{z : 1 < |z| < 2\}$ is open. Let us define for all $\alpha in A$ the following quantity

$$r = \begin{cases} |\alpha| - 1, & |\alpha| < \frac{3}{2} \\ 2 - |\alpha|, & |\alpha| > \frac{3}{2} \end{cases} \tag{52.4.1}$$

Then $D_r(\alpha)$ is entirely contained in A.

(iv) Let $A = \{z : \pi/3 < \arg z < 2\pi/3\}$, and let $\alpha \in A$. Let $\theta = \arg(\alpha)$, and define

$$r = \begin{cases} |\alpha| \sin\left(\frac{2\pi}{3} - \theta\right), & \alpha \leq 0 \\ |\alpha| \sin\left(\theta - \frac{\pi}{3}\right), & \alpha > 0 \end{cases} \tag{52.4.2}$$

Then clearly $D_r(\alpha)$ lies entirely in A.

◀

Theorem *(Combination of open sets)*

If A_1 and A_2 are closed sets then

(i) $A_1 \cup A_2$

(ii) $A_1 \cap A_2$

are also open.

Proof.

(i) Let $\alpha \in A_1 \cup A_2$, and suppose $\alpha \in A_1$ wlog. Then, there exists some $r > 0$ such that $D_r(\alpha) \subseteq A_1 \subseteq A_1 \cup A_2$, as desired.

(ii) Let $\alpha \in A_1 \cap A_2$, so there exist $r_1, r_2 > 0$ such that $D_{r_1}(\alpha) \subseteq A_1$ and $D_{r_2}(\alpha) \subseteq A_2$. Choosing $r = \min\{r_1, r_2\}$ then $D_r(\alpha) \subseteq A_1 \cap A_2$.

■

> **Definition** *(Closed set)*
> A set $E \subseteq \mathbb{C}$ is closed if its complement $\overline{E} \equiv \mathbb{C} \setminus E$ is open.

The combination rules for open sets apply equally to closed sets.

> **Theorem** *(Combination of closed sets)*
> If E_1 and E_2 are open sets then
>   (i) $E_1 \cup E_2$
>   (ii) $E_1 \cap E_2$
> are also open.

Proof.

(i) Since E_1 and E_2 are closed, their complements $\mathbb{C} \setminus E_1$ and $\mathbb{C} \setminus E_2$ are open. Consequently

$$\mathbb{C} \setminus (E_1 \cup E_2) = (\mathbb{C} \setminus E_1) \cap (\mathbb{C} \setminus E_2) \tag{52.4.3}$$

is open since it is the intersection of two open sets.

(ii) Similarly

$$\mathbb{C} \setminus (E_1 \cap E_2) = (\mathbb{C} \setminus E_1) \cup (\mathbb{C} \setminus E_2) \tag{52.4.4}$$

is open since it is the union of two open sets.

■

> **Definition** *(Interior, exterior and boundary)*
> Let $A \subseteq \mathbb{C}$ and let $\alpha \in \mathbb{C}$. Then α is an **interior point** of A if there is an open disc centered at α that lies in A. Similarly α is an **exterior point** of A if there is an open disc centered at α that lies outside of A. Finally, α is a **boundary point** of A any open disc centered at α contains at least one point in A and one point in $\overline{A}$.
> The set of interior points of A forms the interior int$A$, the set of exterior points forms the exterior ext$A$ and the set of boundary points forms the boundary ∂A.

Note that any point in an open set is an interior point by definition.

> **Definition** *(Connectedness)*
> A set $A \subseteq \mathbb{C}$ is (pathwise) connected if for any two distinct points $\alpha, \beta \in A$, there exists a path $\Gamma \subseteq A$ joining the two.

> **Example.** Let us prove that if A, B are both connected and $A \cap B \neq \emptyset$ then $A \cup B$ is connected.
> Let $\alpha \in A \cup B$, and assume wlog that $\alpha \in A$ and $\beta \in B$ (if both points lie in A then the proof is trivial). Since $A \cap B \neq \emptyset$ we may assume that there exists at least some $\delta \in A \cap B$. Since $\delta \in A$ there exists a path Γ_A connecting α with δ. Since $\delta \in B$ there exists a path Γ_B connecting δ with β. Consequently $\Gamma_1 \cup \Gamma_2$ connects α with β as desired. ◀

Proposition *(Continuous functions preserve connectedness)*
Let f be a continuous function on a connected domain A. Then $f(A)$ is also connected.

Proof. Since A is connected, given $\alpha, \beta \in A$ we can find a curve Γ parametrised continuously by a function

$$\gamma : [a, b] \to A, \ \gamma(a) = \alpha, \ \gamma(b) = \beta \tag{52.4.5}$$

Since f is continuous on A, it follows from the composition rule that $f \circ \gamma$ is continuous on $[a, b]$. Moreover, $(f \circ \gamma)(a) = f(\alpha)$ and $(f \circ \gamma)(b) = \beta$, so $f \circ \gamma$ parametrised a curve $f(\Gamma) = (f \circ \gamma)[a, b] \in f(A)$. $\blacksquare$

Definition *(Region)*
A region $\mathcal{R}$ is a non-empty, open, connected subset of $\mathbb{C}$.

Theorem *(Point-removal of regions)*
If $\mathcal{R}$ is a region and $\alpha_0 \in \mathcal{R}$ then $\mathcal{R} \setminus \{\alpha_0\}$ is also a region.

Proof. Since $\mathcal{R}$ is a region, and thus open, it cannot be a singleton and contains more than one element. Consequently $\mathcal{R} \setminus \{\alpha_0\}$ is not empty. Now note that $\mathcal{R} \setminus \{\alpha_0\} = \mathcal{R} \cap (\mathbb{C} \setminus \{\alpha_0\})$ is the intersection of two open sets, and so must also be open. Finally, let $\alpha, \beta \in \mathcal{R} \setminus \{\alpha_0\}$. Since $\alpha, \beta \in \mathcal{R}$ they can be joined by a curve Γ. If $\alpha_0 \notin \Gamma$ then this curve will lie in $\mathcal{R} \setminus \{\alpha_0\}$. If instead $\alpha_0 \in \Gamma$, we consider a disc $D_r(\alpha_0) \in \mathcal{R}$ and deform the curve so as to avoid α_0. $\blacksquare$

Note that the above proof can be extended to the removal of a finite number of points, but breaks down for infinitely many points. Indeed the removal of infinitely many points should of course not necessarily read to a non-empty set, so the point-removal theorem is in general false for infinitely many points.

Example. Let us prove that $\mathcal{R} = \mathbb{C} \setminus \{(n + \frac{1}{2})\pi : n \in \mathbb{Z}\}$ is a region. It is obviously non-empty.
Unfortunately we cannot use the point-removal theorem since we are removing an infinite number of points, we go back to the original definition. Firstly note that $\mathcal{R}$ is non-empty, since for example $i \in \mathcal{R}$. Now let $\alpha \in \mathcal{R}$, and let m be the integer such that $(m + \frac{1}{2})\pi$ is the closest removed point to α. Then letting $r = |\alpha - (m + \frac{1}{2})\pi|$ we have that $D_r(\alpha)$ lies entirely in $\mathcal{R}$. Finally, let $\alpha, \beta \in \mathcal{R}$ and let Γ be the straight segment in $\mathbb{C}$ joining them. If no removed point lies on Γ then this curve also lies in $\mathcal{R}$. If one of the removed point lies on Γ then we can take a neighborhood of the point and deform it so as to skip around it. In the worst case scenario only finitely many such deformations will be necessary, and the curve will lie entirely in $\mathcal{R}$. $\blacktriangleleft$

We make one final definition that will be useful in the next section.

Definition *(Compact sets)*
A set $E \in \mathbb{C}$ that is contained in some closed disc is bounded. Otherwise it is unbounded.

> If E is both closed and bounded then it is compact.

52.5 Extreme value theorem

Theorem *(Extreme value theorem)*
Let f be a function that is continuous on a compact set E. Then $\exists \alpha, \beta \in E$ such that

$$|f(\beta)| \leq |f(z)| \leq |f(\alpha)|, \ \forall z \in E \tag{52.5.1}$$

Proof. We begin by proving the following lemma

Lemma. If E is a closed set and (z_n) is convergent in E to α, then $\alpha \in E$.

Indeed suppose that $\alpha \in \overline{E}$, which is open and thus contains an open disc $D_r(\alpha)$. Then if $z_n \to \alpha$ then $D_r(\alpha) \subseteq \overline{E}$ contains all but a finite number of points of z_n. This contradicts the fact that z_n is in E.

We will also need the following result

Theorem *(Nested rectangles theorem)*
Let $R_0, R_1, R_2, ...$ be a sequence of closed rectangular regions whose sides are parallel to the real and imaginary axes, and whose diagonals' lengths form a sequence $s_0, s_1, s_2, ...$ such that

$$R_0 \supseteq R_1 \supseteq R_2 \supseteq ..., \ \lim_{n \to \infty} s_n = 0 \tag{52.5.2}$$

Then

$$\bigcap_i R_i = \{\alpha\} \tag{52.5.3}$$

for a unique complex number α. Furthermore, given $\epsilon > 0$ then there is an integer N such that

$$R_n \subseteq D_\epsilon(\alpha), \ \forall n > N \tag{52.5.4}$$

Proof. Let the nth rectangle be defined as

$$R_n = \{x + iy : a_n \leq x \leq c_n, b_n \leq y \leq d_n\} \tag{52.5.5}$$

so that since the rectangles are nested within each other

$$a_0 \leq a_1 \leq a_2 \leq ... \leq c_2 \leq c_1 \leq c_0 \tag{52.5.6}$$
$$b_0 \leq b_1 \leq b_2 \leq ... \leq d_2 \leq d_1 \leq d_0 \tag{52.5.7}$$

Since a_n is an increasing, bounded sequence ($a_n \leq c_0$ for all n) it follows from the monotone convergence theorem that a_n converges. Similarly b_n also converges. Let $\lim_{n \to \infty} a_n = a$ and $\lim_{n \to \infty} b_n = b$.

We also have that

$$0 \leq c_n - a_n \leq s_n, \ 0 \leq d_n - b_n \leq s_n \tag{52.5.8}$$

implying that $c_n \to a$ and $d_n \to b$ by the Squeeze theorem. Let $\alpha = a + ib$, then since by the monotone convergence theorem $a_n \leq a \leq c_n$ and $b_n \leq b \leq d_n$ for all n it follows that α is contained in all the rectangles:

$$\alpha \in \bigcap_i R_i \tag{52.5.9}$$

Now let $z \in R_n$, so that $|z - \alpha| \leq s_n$ since any two points in a rectangle are at most a diagonal length apart. By assumption $s_n \to 0$ so for a given $\epsilon > 0$ there exists N such that $s_n \leq \epsilon$ for all $n > N$ and thus such that $|z_n - \alpha| < \epsilon$. It follows that

$$R_n \subseteq \{z : |z - \alpha| < \epsilon\}, \ \forall n > N \tag{52.5.10}$$

Consequently, suppose that there is another complex number β satisfying

$$\beta \in \bigcap_i R_i \tag{52.5.11}$$

but then there for any ϵ we have that $|\beta - \alpha| < \epsilon$. If we let $\epsilon = |\beta - \alpha| > 0$ then we obtain a contradiction, so α is the only element in the intersection of nested rectangles. ∎

We can now tackle the Extreme value theorem. We begin by proving that there exists an $\alpha \in E$ such that $|f(z)| \leq |f(\alpha)|$ for all $z \in E$.

Since E is compact it can be surrounded by a rectangle R_0 of diagonal length s_0. This rectangle can be subdivided into four equivalent rectangles T_1, T_2, T_3, T_4 of diagonal length $\frac{s_0}{2}$. Then at least one of the rectangles, T_j, is such that for all $z \in E$ there exists $w \in E \cap T_j$ such that $|f(z)| \leq f(w)|$. In other words, there is at least one rectangle containing the value of z which bounds f from above. Indeed if this were false, then there is a complex $z_k \in E$ with $|f(z_k)| > |f(w)|$ for all $w \in E \cap T_k$. But then

$$\max\{|f(z_1)|, |f(z_2)|, |f(z_3)|, |f(z_4)|\} > |f(w)|, \ \forall w \in E \tag{52.5.12}$$

which is false since $>$ should be replaced with a $\geq$.

Now let $R_1 = T_j$ and repeat the process of subdividing and finding the quadrant containing the maximal value of f. Iterating this indefinitely we will get a sequence R_n of nested closed rectangular regions with the property that

$$R_{n+1} \subseteq R_n, \ s_n = \frac{1}{2^n} s_0 \to 0 \tag{52.5.13}$$

and for each $z \in E \cap R_n$ there exists some $w \in E \cap R_{n+1}$ such that $|f(z)| < |f(w)|$. Using the nested rectangle theorem we know that only one complex number α will lie in all rectangles, and furthermore for a given $\epsilon > 0$ there is a N such that $R_n \subseteq D_\epsilon(\alpha)$ for all $n > N$. We now claim that the number α must lie in E and that

$$f(z) \leq f(\alpha), \ \forall z \in E \tag{52.5.14}$$

To see why, let $z_0 \in E$ and construct from it a sequence such that

$$z_n \in E \cap R_n, \ (z_0 \in E \cap R_0 = E \text{ is trivially satisfied}) \tag{52.5.15}$$

and

$$|f(z_n)| \leq |f(z_{n+1})| \tag{52.5.16}$$

Since z_n is a sequence on a closed set E converging to α it follows that $\alpha \in E$, as desired. Moreover, by the continuity of f on E it follows that

$$\lim_{n \to \infty} |f(z_n)| = |f(\alpha)| \qquad (52.5.17)$$

implying that

$$|f(z_0)| \leq |f(z_1)| \leq ... \leq |f(\alpha)| \qquad (52.5.18)$$

since $|f(z_0)|$ is an increasing sequence. Notice that this argument applies to any $z_0 \in E$ so it follows that $|f(z)| \leq |f(\alpha)|$ for any z.

We have proven the first part of the theorem, now we prove that there is β such that $|f(\beta)| \leq |f(z)|$ for all $z \in E$.

If $f(w) = 0$ for some w then $\beta = w$ will do the job. Otherwise f does not vanish on E, and thus $g(z) = \frac{1}{f(z)}$ is a continuous function by the quotient rule. We can apply the argument used to prove the first part of the theorem to prove that such a complex number β must exist. ∎

> **Proposition** *(Map of compact set is compact)*
> Let f be a function that is continuous on a compact set E. Then $f(E)$ is also compact.

Proof. We know that $f(E)$ must be bounded by the Extreme value theorem.

To prove that $f(E)$ is closed, we prove that $A = \mathbb{C} \setminus f(E)$ is open. Let $\alpha \in A$ and define

$$g(z) = f(z) - \alpha \qquad (52.5.19)$$

By the extreme value theorem there is some $\beta \in E$ such that

$$r \equiv |g(\beta)| \leq |g(z)|, \; \forall z \in E \qquad (52.5.20)$$

implying that

$$|f(z) - \alpha| \geq r, \; \forall z \in E \qquad (52.5.21)$$

Therefore if $|f(z) - \alpha| < r$ then $z \notin E$ so $f(z) \notin f(E)$. Consequently we have that $D_r(\alpha) \subseteq A$, as desired. ∎

CA3 Differentiating complex functions

53.1 Derivatives of complex functions

Having discussed the concept of a limit of a complex function and its relation to continuity, we are now ready to introduce the concept of differentiating complex functions.

> **Definition** *(Differentiability)*
> Let f be a complex function on A. Then it is said to be **differentiable** at α if the limit
>
> $$\lim_{h \to 0} \frac{f(\alpha + h) - f(\alpha)}{h} \tag{53.1.1}$$
>
> exists. This limit is known as the **derivative of f at α**. If f is differentiable on every point in a set B then it is differentiable on B. A differentiable function is differentiable on its domain. A function differentiable on $\mathbb{C}$ is defined to be **entire**.

Note that for a function to be differentiable at a point we need that point to be a limit point of the function's domain.

> **Example.** Let $f(z) = \frac{1}{z}$ be defined on $\mathbb{C} \setminus \{0\}$. Letting $\alpha \in \mathbb{C} \setminus \{0\}$ then we find that
>
> $$\lim_{z \to \alpha} \frac{f(z) - f(\alpha)}{z - \alpha} = \lim_{z \to \alpha} \frac{(\alpha - z)/(z\alpha)}{z - \alpha} = -\lim_{z \to \alpha} \frac{1}{z\alpha} = -\frac{1}{\alpha^2} \tag{53.1.2}$$
>
> so the derivative of $f(z) = \frac{1}{z}$ is $f'(z) = -\frac{1}{z^2}$, and is defined on $\mathbb{C} \setminus \{0\}$. ◀

Note that while $f(z)$ is not entire, it is still differentiable on a region. This property is useful and deserves its own name, such functions are known as analytic functions.

> **Definition** *(Analyticity)*
> A function $f(z)$ that is differentiable on a region $\mathcal{R}$ is said to be **analytic on $\mathcal{R}$**. If this region is the domain of f then it is simply **analytic**. Also f is **analytic at α** if there exists a region containing α on which it is differentiable.

Notice that if a function is differentiable at a point then it is not necessarily analytic at that point, although the converse is true.

As usual we can also combine differentiable functions

Theorem *(Derivative properties)* Let f, g have domains A and B, and be differentiable at α. Then

(i) $(f + g)'(\alpha) = f'(\alpha) + g'(\alpha)$.

(ii) $(\lambda f)'(\alpha) = \lambda f'(\alpha)$ for $\lambda \in \mathbb{C}$

(iii) $(fg)'(\alpha) = f'(\alpha)g(\alpha) + f(\alpha)g'(\alpha)$

(iv) $\left(\dfrac{f}{g}\right)'(\alpha) = \dfrac{f'(\alpha)g(\alpha) - f(\alpha)g'(\alpha)}{(g(\alpha))^2}$ provided $g(\alpha) \neq 0$.

Theorem *(Differentiability $\implies$ continuity)*

Let f be differentiable at α, then it must be continuous at α.

Proof. Suppose that $z \to \alpha$, then

$$\lim_{z \to \alpha} (f(z) - f(\alpha)) = \lim_{z \to \alpha} \frac{f(z) - f(\alpha)}{z - \alpha} \cdot \lim_{z \to \alpha} (f(z) - f(\alpha)) = f'(\alpha) \cdot 0 = 0 \tag{53.1.3}$$

so $f(z) \to f(\alpha)$ as desired. $\blacksquare$

Consider for example the principal logarithm function $\log z = \log |z| + i\arg(z)$. We know that at no point on the negative real axis is this function continuous, so it is also not differentiable there.

A more striking example is that of the modulus function $f(z) = |z|$. In real analysis this function is differentiable at all non-zero points, but in complex analysis we shall see that this no longer holds.

The proof that $f(z) = |z|$ is discontinuous at $z = 0$ is the same as in real analysis, so let $\alpha \neq 0$ and consider the circle centered at the origin passing through α. Assume Let us define (z_n) to be the sequence

$$z_n = |\alpha|\exp\left(i\arg(\alpha) + \frac{i}{n}\right) \tag{53.1.4}$$

and similarly let us define (w_n) so that

$$w_n = \left(|\alpha| + \frac{1}{n}\right)e^{i\arg(\alpha)*} \tag{53.1.5}$$

Clearly $z_n \to \alpha$ and $w_n \to \alpha$. However, note that

$$\lim_{n \to \infty} \frac{f(z_n) - f(\alpha)}{z_n - \alpha} = \lim_{n \to \infty} \frac{|\alpha| - |\alpha|}{z_n - \alpha} = 0 \tag{53.1.6}$$

while

$$\lim_{n \to \infty} \frac{f(w_n) - f(\alpha)}{w_n - \alpha} = \lim_{n \to \infty} \frac{1/n}{e^{i\arg(\alpha)}(|\alpha| + 1/n - |\alpha|)} = e^{-i\arg(\alpha)} \tag{53.1.7}$$

Consequently the limit

$$\lim_{z \to \alpha} \frac{f(z) - f(\alpha)}{z - \alpha} \tag{53.1.8}$$

does not exist, and $f(z) = |z|$ is not differentiable anywhere on $\mathbb{C}$.

Example. Consider the function $f(z) = \bar{z}$ defined over $\mathbb{C}$. There are no points where $f(z)$ is differentiable. Indeed let $\alpha \in \mathbb{C}$ and consider the sequence $z_n = \alpha - \frac{1}{n}$. Then

$$\lim_{n \to \infty} \frac{f(z_n) - f(\alpha)}{z_n - \alpha} = \lim_{n \to \infty} \frac{-1/n}{-1/n} = 1 \qquad (53.1.9)$$

Now consider the sequence $w_n = \alpha - \frac{i}{n}$. Then

$$\lim_{n \to \infty} \frac{f(z_n) - f(\alpha)}{z_n - \alpha} = \lim_{n \to \infty} \frac{i/n}{-i/n} = -1 \qquad (53.1.10)$$

Since these two limits are not equal the function f is not differentiable at α which we assumed to be any complex number. ◀

53.2 Cauchy-Riemann equations

Theorem *(Cauchy-Riemann theorem)*

Let $f(x+iy) = u(x,y) + iv(x,y)$ be a complex function defined on a region $\mathcal{R}$ which contains $a + ib$. If f is differentiable at $a + ib$ then the following **Cauchy-Riemann equations** are satisfied:

$$\left. \frac{\partial u}{\partial x} \right|_{(a,b)} = \left. \frac{\partial v}{\partial y} \right|_{(a,b)}, \qquad \left. \frac{\partial v}{\partial x} \right|_{(a,b)} = -\left. \frac{\partial u}{\partial y} \right|_{(a,b)} \qquad (53.2.1)$$

Proof. Let that z_n be any sequence on $\mathcal{C} \setminus \{\alpha\}$ where $\alpha = a + ib$ and let $z_n = x_n + iy_n$ for all n. Firstly we observe that

$$\frac{f(z_n) - f(\alpha)}{z_n - \alpha} = \frac{u(x_n, y_n) - u(a,b)}{(x_n - a) + i(y_n - b)} + i\frac{v(x_n, y_n) - v(a,b)}{(x_n - a) + i(y_n - b)} \qquad (53.2.2)$$

Consequently, let us choose x_n to be a sequence on $\mathcal{R} \setminus \{a\}$ and define $z_n = x_n + ib$ which converges to α. Then we see that

$$\frac{f(z_n) - f(\alpha)}{z_n - \alpha} = \frac{u(x_n, b) - u(a,b)}{(x_n - a)} + i\frac{v(x_n, b) - v(a,b)}{(x_n - a)} \qquad (53.2.3)$$

so after taking the limit we find

$$\left. \frac{df}{dz} \right|_{\alpha} = \left. \frac{\partial u}{\partial x} \right|_{(a,b)} + i\left. \frac{\partial u}{\partial x} \right|_{(a,b)} \qquad (53.2.4)$$

Now let us choose y_n to be a sequence on $\mathcal{R} \setminus \{b\}$ and define $z_n = a + iy_n$ which converges to α. Then we see that

$$\frac{f(z_n) - f(\alpha)}{z_n - \alpha} = \frac{u(a, y_n) - u(a,b)}{i(y_n - b)} + i\frac{v(a, y_n) - v(a,b)}{i(y_n - b)} \qquad (53.2.5)$$

so after taking the limit we find

$$\left. \frac{df}{dz} \right|_{\alpha} = -i\left. \frac{\partial u}{\partial y} \right|_{(a,b)} + \left. \frac{\partial v}{\partial y} \right|_{(a,b)} \qquad (53.2.6)$$

Since f is differentiable at α we require (**??**) and (**??**) to be equal to each other. This gives the desired Cauchy-Riemann equations. ■

This theorem is extremely useful as it also tells us that if the Cauchy-Riemann equations fail then the function is not differentiable. Let us apply it to $f(x + iy) = x - iy$ which we have previously determined to not be differentiably anywhere on $\mathbb{C}$. Now we see that

$$\frac{\partial u}{\partial x}\bigg|_{(a,b)} = 1 \neq \frac{\partial v}{\partial y}\bigg|_{(a,b)} = -1 \tag{53.2.7}$$

as desired.

Note that the converse of the Cauchy-Riemann theorem is not true, if a function satisfies (53.2.2) then it is not necessarily differentiable. If we add some extra conditions however a converse-like theorem can be proven.

Theorem *(Cauchy-Riemann converse theorem)*
Let $f(x + iy) = u(x, y) + iv(x, y)$ be defined on a region $\mathcal{R}$ and let $\alpha = a + ib \in \mathcal{R}$. If the derivatives $\partial_x u, \partial_y u, \partial_x v, \partial_y v$
  (i) exist at (x, y) for all $x + iy \in \mathcal{R}$
 (ii) are continuous at (a, b)
(iii) satisfy the Cauchy-Riemann equations
then f is differentiable at $a + ib$ and its derivative is given by

$$f'(a + ib) = \frac{\partial u}{\partial x}\bigg|_{(a,b)} + i\frac{\partial v}{\partial x}\bigg|_{(a,b)} \tag{53.2.8}$$

Example. Let us show that $f(z) = \sin z$ is entire. Indeed we have that

$$\sin(x + iy) = \sin x \cosh y + i \sinh y \cos x \tag{53.2.9}$$

so that $u(x, y) = \sin x \cosh y$ and $v(x, y) = \sinh y \cos x$. Consequently

$$\frac{\partial u}{\partial x} = \cos x \cosh y \qquad \frac{\partial v}{\partial y} = \cosh y \cos x \tag{53.2.10}$$

$$\frac{\partial u}{\partial y} = \sin x \sinh y \qquad \frac{\partial v}{\partial x} = -\sin x \sinh y \tag{53.2.11}$$

These derivatives exist and are continuous everywhere on $\mathbb{C}$. Also they satisfy the Cauchy-Riemann equations:

$$\frac{\partial u}{\partial x}\bigg|_{(a,b)} = \cos a \cosh b \frac{\partial v}{\partial y}\bigg|_{(a,b)} \qquad \frac{\partial v}{\partial x}\bigg|_{(a,b)} = \sin a \sinh b = -\frac{\partial u}{\partial y}\bigg|_{(a,b)} \tag{53.2.12}$$

Since all criteria of the converse Cauchy-Riemann theorem are satisfied, we find that

$$f'(a + ib) = \cos a \cosh b - i \sin a \sinh b = \cos(a + ib) \tag{53.2.13}$$

◀

53.3 Derivative rules

We now prove the chain rule and the inverse function rule. They take the same form as in real analysis. These properties are particularly useful in the computation of the derivatives of more complex functions.

> **Theorem** *(Chain rule)*
> Let f, g be complex functions, let D be the domain of $g \circ f$ and let α be a limit point in D. If f is differentiable at α and g at $f(\alpha)$ then $g \circ f$ is differentiable at α and
>
> $$(g \circ f)'(\alpha) = g'(f(\alpha))f'(\alpha) \tag{53.3.1}$$

Proof. Let us define the following function with the same domain as g

$$h(w) = \begin{cases} \frac{g(w)-g(\beta)}{w-\beta}, \ w \neq \beta \\ g'(\beta), \ w = \beta \end{cases} \tag{53.3.2}$$

where $w = f(z)$ and $\beta = f(\alpha)$. Notice that since g is differentiable at β, $g'(\beta) = \lim_{w \to \beta} \frac{g(w)-g(\beta)}{w-\beta}$ so $h(w)$ is continuous at β. Then we see that

$$\frac{g(f(z)) - g(f(\alpha))}{z - \alpha} = h(f(z))\left(\frac{f(z) - f(\alpha)}{z - \alpha}\right) \tag{53.3.3}$$

for all $z \neq \alpha$ in the domain of f. Indeed if $w = \beta$ then both sides vanish, while if $w \neq \beta$ then

$$h(f(z))\left(\frac{f(z) - f(\alpha)}{z - \alpha}\right) = \frac{g(f(z)) - g(\beta)}{f(z) - f(\alpha)} \frac{f(z) - f(\alpha)}{z - \alpha} \tag{53.3.4}$$

Taking the limit as $z \to \alpha$ we find

$$\lim_{z \to \alpha} \frac{g(f(z)) - g(f(\alpha))}{z - \alpha} = \lim_{z \to \alpha} h(f(z)) \lim_{z \to \alpha} \frac{f(z) - f(\alpha)}{z - \alpha} = g'(\beta)f'(\alpha) \tag{53.3.5}$$

since h is continuous at $f(\alpha)$, g differentiable at β and f is differentiable at α. ∎

> **Theorem** *(Inverse function derivative)*
> Let $f : A \to B$ be invertible, and suppose f^{-1} is continuous at $\beta \in B$. If f' is non-zero at $f^{-1}(\beta) \in A$ then f^{-1} is differentiable at β with
>
> $$(f^{-1})'(\beta) = \frac{1}{f'(f^{-1}(\beta))} \tag{53.3.6}$$

Proof. We know that $\alpha = f^{-1}(\beta)$ is a limit point of A so let z_n be a sequence in $A - \{\alpha\}$ converging to α. Then $f(z_n)$ will be a sequence in $B - \{\beta\}$ since $f(z_n) \neq \alpha$ (by injectivity). Since f is continuous at α, $f(z_n)$ converges to β, thus proving that β is a limit point of B.

Thus let w_n be a sequence in $B - \{\beta\}$ converging to β. Also, f^{-1} is continuous at β so if we define $z_n = f^{-1}(w_n)$ then $z_n \to f^{-1}(\beta) = \alpha$. Consequently

$$\lim_{n \to \infty} \frac{f^{-1}(w_n) - f^{-1}(\beta)}{w_n - \beta} = \lim_{n \to \infty} \frac{z_n - \alpha}{f(z_n) - f(\alpha)} \qquad (53.3.7)$$

and since f is injective, $f(z_n) \neq f(\alpha)$ showing that

$$(f^{-1})'(\beta) = \frac{1}{f'(f^{-1}(\beta))} \qquad (53.3.8)$$

as desired. ∎

53.4 Smooth paths

Definition *(Path)*
Let $I \subseteq \mathbb{R}$ be an interval and let $\gamma : I \to \Gamma \subseteq \mathbb{C}$ be a continuous function. The set Γ is known as the **path** parametrised by γ.
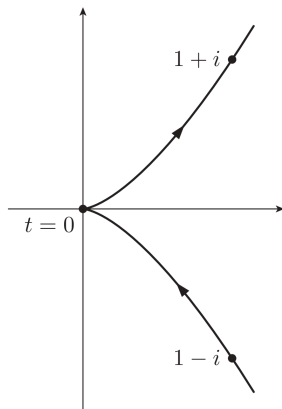
It is often helpful, given a parametrisation $\gamma(t)$ of a path Γ, to express it in cartesian form

$$\gamma(t) = \phi(t) + i\psi(t), \ t \in I \qquad (53.4.1)$$

where ϕ and ψ are real functions on I. It is clear that the differentiability of both ϕ and ψ is equivalent to the differentiability of γ.

Differentiability however is not enough to guarantee smoothness. For example, the following parametrisation

$$\gamma(t) = t^2 + it^3, \ t \in \mathbb{R} \qquad (53.4.2)$$

is certainly differentiable on $\mathbb{R}$, but as we can see below Γ contains a kink at the origin where the slope is zero.

Definition *(Smooth parametrisation)*
Let $\gamma : I \to \mathbb{R}$ be a parametrisation such that
  (i) γ is differentiable
  (ii) γ' is continuous
  (iii) γ' is non-zero
on I. Then γ is a **smooth parametrisation** and the associated path Γ is **smooth**.

The notion of paths allows us to provide a geometrical interpretation of derivatives. Indeed let f be analytic on $\mathcal{R}$ and let Γ be a smooth path in $\mathcal{R}$ parametrised by $\gamma : I \to \mathbb{C}$. The image of Γ under f, $f(\Gamma)$ is then parametrised by $f(\gamma(t))$. We may view the derivatives of $\gamma(t)$ and $f(\gamma(t))$ at $t = c$ as tangent vectors to Γ and $f(\Gamma)$ at $\gamma(c)$ and $f(\gamma(c))$ respectively.

Consider $\alpha = \gamma(c) \in \Gamma$ for some $c \in I$. Then $f(\alpha) \in f(\Gamma)$, and since Γ and $f(\Gamma)$ are smooth it

follows that

$$(f \circ \gamma)'(c) = f'(\alpha)\gamma'(c) \tag{53.4.3}$$

So the derivatives to the two paths are related by a scaling factor of $f'(\alpha)$. In other words, one should rotate the tangent vector at α by $\mathrm{Arg}(f'(\alpha))$ and scaled by $|f'(\alpha)|$.
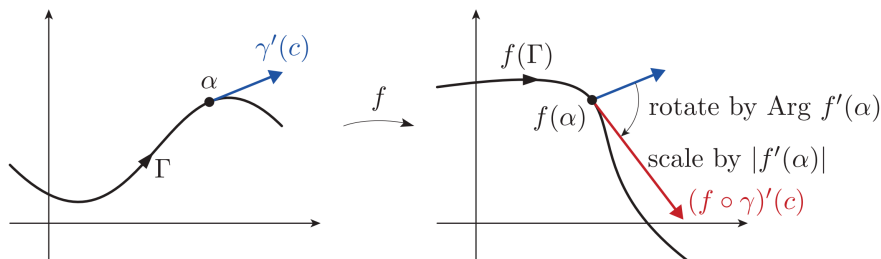
Viewing the derivative of the parametrisation as tangent vectors allows us to define the angle between two curves at a point. Suppose Γ_1 and Γ_2 parametrised by γ_1 (I_1) and γ_2 (I_2) intersect at $\alpha = \gamma_1(t_1) = \gamma_2(t_2)$. Note that

$$\gamma_2'(t_2) = \frac{\gamma_2'(t_2)}{\gamma_1'(t_1)}\gamma_1'(t_1) \tag{53.4.4}$$

so the tangent vector to one path is given by the tangent vector to the other path rotated by the argument of $\frac{\gamma_2'(t_2)}{\gamma_1'(t_1)}$.

Definition *(Angle between paths)*

Let Γ_1 and Γ_2 be smooth paths with parametrisations γ_1 and γ_2 respectively, intersecting at $\alpha = \gamma_1(t_1) = \gamma_2(t_2)$. Then the angle from Γ_1 to Γ_2 at α is

$$\theta = \mathrm{Arg}\left(\frac{\gamma_2'(t_2)}{\gamma_1'(t_1)}\right) \tag{53.4.5}$$
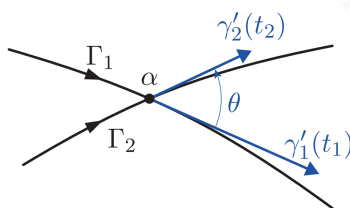
There are special types of analytic functions which conserve the angle between any two paths in its domain. This angle-preserving property is known as conformality.

Definition *(Conformal functions)*

Let f be analytic at α. Then it is **conformal** at α if the angle from any smooth path through α to any other smooth path through α is preserved under f.

Suppose f is analytic at α. Then the tangent vectors to $f(\Gamma)$ at $f(\alpha)$ are given by the tangent vectors to Γ at α through a rotation by $\mathrm{Arg}(f'(\alpha))$, as long as $f'(\alpha) \neq 0$. Consequently the angle between two paths at a point is preserved by analytic functions whose derivative does not vanish at that point.

We will later prove the converse of this result, namely that conformal functions at a point have non-zero derivative at that point.

> **Theorem** *(Support of conformal function derivatives)*
> Let f be analytic at α, then f is conformal at α *iff.* $f'(\alpha) \neq 0$.

CA4 Integrating complex functions

54.1   Introduction to complex integration

Unlike in real analysis, where the integral from a to b of some function $f(x)$ was well-defined since it could only be taken along one path, in complex analysis the integral between two points must also specify a smooth path to integrate over.

Letting $\Gamma : \gamma(t)$ for $t \in [a, b]$ be a smooth path with $\gamma(a) = \alpha$ and $\gamma(b) = \beta$ then we say that

$$\int_\Gamma f(z)\,dz \tag{54.1.1}$$

is the integral of f along Γ. There are two ways one could calculate such an integral. One way, inspired by real analysis, is to

(i) choose a partition $P = \{\Gamma_1, ..., \Gamma_n\}$ of Γ.

(ii) calculate the Riemann sum

$$R(f, P) = \sum_{k=1}^n f(z_k)\delta z_k \tag{54.1.2}$$

where $\delta z_k = z_k - z_{k-1}$.

(iii) Evaluate

$$\int_\Gamma f(z)\,dz \equiv \lim_{n \to \infty} R(f, P_n) \tag{54.1.3}$$

where (P_n) is a sequence of partitions such that $||P_n|| \to 0$ as $n \to \infty$.

One can then show that this limit exists and is independent of (P_n) when f is continuous. A quicker approach is to treat the complex integral as two real integrals, just like how we treated the complex derivative as two real derivatives.

Let $\{t_0, t_1, ..., t_n : a = t_0 < t_1 < ... < t_n = b\}$ be any set of parameter values, these define a partition $P = \{\Gamma_1, \Gamma_2, ..., \Gamma_n\}$ where Γ_i is the path from $\gamma(t_{i-1}) = z_{i-1}$ to $\gamma(t_i) = z_i$. The Riemann sum for this partition is then

$$R(f, P) = \sum_{k=1}^n f(z_k)\delta z_k \tag{54.1.4}$$

Unlike before however, note that

$$\delta z_k = z_k - z_{k-1} = \frac{\gamma(t_k) - \gamma(t_{k-1})}{t_k - t_{k-1}} \cdot (t_k - t_{k-1}) \tag{54.1.5}$$

and if $\delta t_k = t_k - t_{k-1}$ is very small then we find that

$$\delta z_k \approx \gamma'(t_k)\delta t_k \tag{54.1.6}$$

Thus if $\max\{\delta t_1, ..., \delta t_n\}$ is chosen to be very small then we can write the Riemann sum as

$$R(f, P) \approx \sum_{k=1}^{n} f(\gamma(t_k))\gamma'(t_k)\delta t_k \tag{54.1.7}$$

and taking the limit as $n \to \infty$ then we get that

$$\int_\Gamma f(z)\, dz = \int_a^b f(\gamma(t))\gamma'(t)\, dt \tag{54.1.8}$$

> **Definition** *(Complex integral)*
> Let $\Gamma : \gamma(t)$ for $t \in [a, b]$ be a smooth path in $\mathbb{C}$ and let f be continuous on Γ. We then define
>
> $$\int_\Gamma f(z)\, dz = \int_a^b f(\gamma(t))\gamma'(t)\, dt \tag{54.1.9}$$
>
> If we let $f(\gamma(t))\gamma'(t) = u(t) + iv(t)$ where u and v are real valued functions then
>
> $$\int_\Gamma f(z)\, dz = \int_a^b u(t)dt + i \int_a^b v(t)dt \tag{54.1.10}$$

Note that since Γ is smooth, its parametrisation $\gamma(t)$ and its derivative $\gamma'(t)$ is continuous on $[a, b]$. Also, f is continuous on $[a, b]$, so the integral exists.

> **Example.** Let's calculate $\int_\Gamma \frac{1}{z}\, dz$ where Γ is the unit circle $\{z : |z| = 1\}$ traversed anti-clockwise. A parametrisation for the unit circle is $\Gamma : \gamma(t) = e^{it}$ with $t \in [0, 2\pi)$, so we see that
>
> $$\int_\Gamma \frac{1}{z}\, dz = \int_0^2 \pi e^{-it} \cdot ie^{it}\, dt = 2\pi i \tag{54.1.11}$$
>
> Let us instead use the parametrisation $\Gamma : \gamma(t) = e^{3it}$ with $t \in [0, 2\pi/3)$. Then we see that
>
> $$\int_\Gamma \frac{1}{z}\, dz = \int_0^{2\pi/3} e^{-3it} \cdot 3ie^{3it}\, dt = 2\pi i \tag{54.1.12}$$
>
> just as we found before. It is not just a coincidence that this integral is independent of the chosen parametrisation, indeed we calculated the integral of the same function over the same circle, we just went around the circle with a different speed. ◀

We have seen in the previous example that the integral $\int_\Gamma \frac{1}{z}\, dz$ is parametrisation-independent. Let's prove this in the more general case.

> **Theorem** *(Parametrisation independence of integration)*
> Let γ_1 and γ_2 be two smooth parametrisations of a path Γ, and let f be continuous on Γ.

Then the integral of f over Γ

$$\int_\Gamma f(z)\, dz \tag{54.1.13}$$

is independent of the used parametrisation so that

$$\int_{a_1}^{b_1} f(\gamma_1(t))\gamma_1'(t)\, dt = \int_{a_2}^{b_2} f(\gamma_2(t))\gamma_2'(t)\, dt \tag{54.1.14}$$

Proof. Suppose that γ_1 and γ_2 are both injective (if they intersect themselves, split Γ into parts where they are injective), and let $h = \gamma_2^{-1} \circ \gamma_1$ which by the inverse function and composition rule is differentiable with a continuous derivative. We see that

$$\int_{a_1}^{b_1} f(\gamma_1(t))\gamma_1'(t)\, dt = \int_{a_1}^{b_1} f(\gamma_2(h(t)))\gamma_2'(h(t))h'(t)\, dt \tag{54.1.15}$$

$$= \int_{h(a_1)}^{h(a_2)} f(\gamma_2(s))\, \gamma_2'(s)\, ds \tag{54.1.16}$$

$$= \int_{a_2}^{b_2} f(\gamma_2(s))\gamma_2'(s)\, ds \tag{54.1.17}$$

as desired. ∎

Up until now we have only dealt with smooth parametrisations and smooth paths. However, sometimes we may be asked to integrate over a non-smooth (but continuous) path which can be decomposed into finitely many smooth paths.

> **Definition** (*Contour*)
> Let $\Gamma_1, \Gamma_2, ..., \Gamma_n$ be smooth paths in $\mathbb{C}$ with matching endpoints. Then their oriented union $\Gamma = \tilde{\bigcup}_i \Gamma_i$ is a **contour**, and the **contour integral of f along Γ** is given by
>
> $$\int_\Gamma f(z)\, dz = \int_{\Gamma_1} f(z)\, dz + \int_{\Gamma_2} f(z)\, dz + ... + \int_{\Gamma_1} f(z)\, dz \tag{54.1.18}$$

Example. Let us evaluate

$$\int_\Gamma \overline{z}\, dz \tag{54.1.19}$$

where Γ is a contour starting at $z = -1$, traversing the real axis until $z = 1$ and then returning to -1 along a semi-circular arc in the upper-complex plane. A suitable piece-wise parametrisation for Γ is

$$\gamma(t) = \begin{cases} t, \ t \in [-1, 1] \\ e^{it}, \ t \in [0, \pi] \end{cases} \tag{54.1.20}$$

so that

$$\int_\Gamma \overline{z}\, dz = \int_{-1}^1 t\, dt + \int_0^\pi e^{-it} \cdot ie^{it} dt = i\pi \tag{54.1.21}$$

◀

> **Definition** *(Reverse path)*
> Let $\Gamma : \gamma(t)$ with $t \in [a, b]$ be a smoth path. Its **reverse path** $\tilde{\Gamma}$ is defined as the path with parametrisation
> $$\tilde{\gamma}(t) = \gamma(a + b - t), \ t \in [a, b] \tag{54.1.22}$$

> **Proposition** *(Reverse contour theorem)*
> Let Γ be a contour and let f be continuous on Γ with reverse contour $\tilde{\Gamma}$. Then we have that
> $$\int_{\Gamma} f(z) \, dz = - \int_{\tilde{\Gamma}} f(z) \, dz \tag{54.1.23}$$

Proof. Initially, let us suppose that Γ is a smooth path with parametrisation $\gamma(t)$. Then $\tilde{\gamma}(t) = \gamma(a + b - t)$ so that

$$\int_{\tilde{\Gamma}} f(z) \, dz = \int_{b}^{a} f(\tilde{\gamma}(t))\tilde{\gamma}'(t) \, dt = - \int_{a}^{b} f(\gamma(a + b - t))\gamma'(a + b - t) \, dt \tag{54.1.24}$$

$$= - \int_{b}^{a} f(\gamma(s))\gamma'(s)(-ds) = - \int_{a}^{b} f(\gamma(t))\gamma'(t) \, dt \tag{54.1.25}$$

$$= - \int_{\tilde{\Gamma}} f(z) \, dz \tag{54.1.26}$$

as desired. Now suppose that $\Gamma = \Gamma_1 + ... + \Gamma_2$ is a contour made of smooth paths Γ_i for $i = 1, 2...$ Then we find that

$$\int_{\tilde{\Gamma}} f(z) \, dz = \int_{\tilde{\Gamma_1}} f(z) \, dz + \int_{\tilde{\Gamma_2}} f(z) \, dz + ... + \int_{\tilde{\Gamma_n}} f(z) \, dz \tag{54.1.27}$$

$$= - \left(\int_{\Gamma_1} f(z) \, dz + \int_{\Gamma_2} f(z) \, dz + ... \int_{\Gamma_n} f(z) \, dz \right) \tag{54.1.28}$$

$$= - \int_{\Gamma} f(z) \, dz \tag{54.1.29}$$

$\blacksquare$

> **Example.** Let's calculate
> $$\int_{\tilde{\Gamma}} \frac{1}{z} \, dz \tag{54.1.30}$$
>
> where Γ is the unit circle. According to our proposition this should be $-2i\pi$ so let's verify this. We choose the parametrisation $\tilde{\gamma}(t) = e^{-it}$, $t \in [0, 2\pi)$ and find that
>
> $$\int_{\tilde{\Gamma}} \frac{1}{z} \, dz = \int_{0}^{2\pi} e^{it} \cdot (-ie^{-it}) \, dt = -2i\pi \tag{54.1.31}$$
>
> as expected. ◀

54.2   Calculating complex integrals

Theorem *(Combination rules for integrals)* Let Γ be a contour and let f, g be continuous on Γ. Then

$$\int_\Gamma (f + g)(z) \, dz = \int_\Gamma f(z) \, dz + \int_\Gamma g(z) \, dz \tag{54.2.1}$$

$$\int_\Gamma (\lambda f)(z) \, dz = \lambda \int_\Gamma f(z) \, dz \tag{54.2.2}$$

for any $\lambda \in \mathbb{C}$.

Proof. This follows immediately from the combination rules of Riemann integration. ∎

Definition *(Primitive)*
Let f be a function on a region $\mathcal{R}$, then the **primitive of f on** $\mathcal{R}$ is the analytic function F on $\mathcal{R}$ such that $F'(z) = f(z)$ for all $z \in \mathcal{R}$.

Theorem *(Fundamental theorem of Calculus)*
Let Γ be a contour in a region $\mathcal{R}$ starting at α and ending at β, and let f be continuous on $\mathcal{R}$ with primitive F. Then

$$\int_\Gamma f(z) \, dz = F(\beta) - F(\alpha) \tag{54.2.3}$$

Proof. Again, let us firstly prove this result for smooth paths, and then generalize to any contour.

Let Γ be a smooth path in $\mathcal{R}$ with parametrisation $\gamma(t)$, and let f be analytic on $\mathcal{R}$ with primitive F. Then

$$\int_\Gamma f(z) \, dz = \int_a^b f(\gamma(t))\gamma'(t) \, dt = \int_a^b F'(\gamma(t)) \, dt \tag{54.2.4}$$

Let $F'(\gamma(t)) = u(t) + iv(t)$, then we get

$$\int_\Gamma f(z) \, dz = \int_a^b [u(t) + iv(t)]' \, dt = u(b) - (a) + i[v(b) - v(a)] = F(\beta) - F(\alpha) \tag{54.2.5}$$

where we used the fundamental theorem of calculus from real analysis.

Now suppose that Γ is a contour made up of smooth paths $\Gamma_1, ..., \Gamma_n$ where Γ_k has left endpoint α_k and right endpoint β_k. Due to the continuity of Γ we must have that $\beta_k = \alpha_{k+1}$ for $k = 1, 2...$ and thus

$$\int_\Gamma f(z) \, dz = \int_{\Gamma_1} f(z) \, dz + \int_{\Gamma_2} f(z) \, dz + ... + \int_{\Gamma_n} f(z) \, dz \tag{54.2.6}$$

$$= [F(\beta_1) - F(\alpha_1)] + [F(\beta_2) - F(\alpha_2)] + ... + [F(\beta_n) - F(\alpha_n)] \tag{54.2.7}$$

$$= F(\beta_n) - F(\alpha_1) \tag{54.2.8}$$

as desired. ∎

It follows immediately from this theorem that if Γ_1 and Γ_2 are both smooth paths in a region with the same starting and ending points then the integral of any continuous function with a primitive will be the same over both paths. In other words under suitable conditions a contour integral will only depend on the endpoints of the contour.

Example.

(i) Let's evaluate $\int_\Gamma e^{3iz}\, dz$ where Γ is a path from 2 to -2. We have that $f(z) = e^{3iz}$ is continuous on $\mathbb{C}$, and has a primitive $F(z) = -\frac{1}{3}ie^{3iz}$ on $\mathbb{C}$. Consequently we have that

$$\int_\Gamma e^{3iz} = \left[-\frac{1}{3}ie^{3iz} \right]_2^{-2} = \frac{1}{3}i(e^{6i} - e^{-6i}) = -\frac{2}{3}\sin(6) \qquad (54.2.9)$$

(ii) Let's evaluate $\int_\Gamma \sinh z\, dz$ where Γ is any contour from 2 to $2i + \frac{1}{3}$. We know that $f(z) = \sinh z$ is continuous on $\mathbb{C}$ (since it is made up of exponentials) and that it has a primitive $F(z) = \cosh z$. Consequently

$$\int_\Gamma \sinh z\, dz = \left[\cosh z \right]_2^{2i + \frac{1}{3}} = \cosh\left(\frac{1}{3} + 2i \right) - \cosh(2) \qquad (54.2.10)$$

(iii) Let's evaluate $\int_\Gamma \frac{\sin z}{\cos^2 z}\, dz$ where Γ is a contour from 0 to π lying in $\mathbb{C} \setminus \{(n + 1/2)\pi : n \in \mathbb{Z}\}$. Since $\cos z = 0$ only has roots at $z = (n + 1/2)\pi$, $n \in \mathbb{Z}$, it follows from the reciprocal rule of continuity that $f(z) = \frac{\sin z}{\cos^2 z}$ is continuous on the region the contour is specified in. Furthermore, it has a primitive $F(z) = \frac{1}{\cos^2 z}$ so that

$$\int_\Gamma \frac{\sin z}{\cos^2 z}\, dz = \left[\frac{1}{\cos z} \right]_0^\pi = -2 \qquad (54.2.11)$$

◀

There are some important immediate applications of the fundamental theorem of calculus.

Proposition *(Corollary of Fundamental theorem)*
Let f be continuous and with a primitive F on $\mathcal{R}$, then

$$\int_\Gamma f(z)\, dz \qquad (54.2.12)$$

is independent of contour Γ and vanishes if Γ is closed.

Indeed by the fundamental theorem of calculus, the integral over a contour only depends on the endpoint (as long as that path lies in a region where f is continuous. Furthermore, if Γ is closed then its endpoints α and β are equal, implying that $F(\alpha) = F(\beta)$, and that the integral vanishes.

Proposition *(Integration by parts)*
Let f, g be analytic on $\mathcal{R}$ and let their derivatives f' and g' be continuous on $\mathcal{R}$. Then

$$\int_\Gamma f(z)g'(z)\, dz = [f(z)g(z)]_\alpha^\beta - \int_\Gamma f'(z)g(z)\, dz \qquad (54.2.13)$$

for any smooth path $\Gamma \subseteq \mathcal{R}$ starting at α and ending at β.

Proof. We define $H(z) = f(z)g(z)$ and $h(z) = H'(z) = f'(z)g(z) + f(z)g'(z)$. By assumption h must be continuous on $\mathcal{R}$ and has primitive $H(z)$ since the latter is analytic on $\mathcal{R}$. Using the fundamental theorem of Calculus we find that

$$\int_\Gamma h(z)dz = [H(z)]_a^b \tag{54.2.14}$$

or equivalently

$$\int_\Gamma f(z)g'(z)\,dz = [f(z)g(z)]_\alpha^\beta - \int_\Gamma f'(z)g(z)\,dz \tag{54.2.15}$$

as desired. ∎

> **Example.** Let us evaluate $\int_\Gamma \mathrm{Log}(z)\,dz$ where Γ is a contour from 1 to i in $\mathcal{R} = \mathbb{c} \setminus \{x \in \mathbb{R} : x \le 0\}$. We can write this integral as
>
> $$\int_\Gamma (z)' \mathrm{Log}(z)\,dz = [z\mathrm{Log}(z)]_1^i - \int_\Gamma dz = -\frac{\pi}{2} - i + 1 \tag{54.2.16}$$
>
> where we used integration by parts since $g(z) = z$ and $f(z) = \mathrm{Log}(z)$ are analytic on $\mathcal{R}$. ◄

54.3 Estimating contour integrals

Sometimes (very often actually) it may not be possible to compute a contour integral. In such cases an approximation, such as an upper/lower bound can be found. To do so it is useful to define the length of a contour first.

> **Definition** *(Contour length)*
> Let Γ be a contour, then its length if the sum of the **lengths of the constistuent paths** Γ_i : $\gamma_i(t),\ t \in [a, b]$, given by
>
> $$L(\Gamma_i) = \int_{a_i}^{b_i} |\gamma_i'(t)|\,dt \tag{54.3.1}$$

In real analysis, if a function is bounded above by some finite value M, then we should expect its integral over an interval of length L to be at most $M \cdot L$. In complex analysis we can prove a similar result

> **Theorem** *(Estimation theorem)*
> Let f be continuous on a contour Γ and such that $|f(z)| \le M,\ \forall z \in \Gamma$. Then
>
> $$\left| \int_\Gamma f(z)\,dz \right| \le M \cdot L(\Gamma) \tag{54.3.2}$$

Proof. We begin by proving the following Lemma

Lemma. Let $f : [a, b] \to \mathbb{C}$ be a continuous function, then

$$\left| \int_a^b f(t) \, dt \right| \leq \int_a^b |f(t)| \, dt \tag{54.3.3}$$

Indeed if $\int_a^b f(t) \, dt = 0$ then this result holds trivially. Suppose the integral does not vanish, so that it may be written in polar form as $\int_a^b f(t) \, dt = re^{i\theta}$. Then we find that

$$\left| \int_a^b f(t) \, dt \right| = r \tag{54.3.4}$$

but note that

$$r = \int_a^b e^{-i\theta} f(t) \, dt \int_a^b \text{Re}(e^{-i\theta} f(t)) \, dt \leq \int_a^b |e^{-i\theta} f(t)| \, dt = \int_a^b |f(t)| \, dt \tag{54.3.5}$$

where we used the monotonicity inequality of real integrals. Consequently

$$\left| \int_a^b f(t) \, dt \right| \leq \int_a^b |f(t)| \, dt \tag{54.3.6}$$

as desired. Now that we have proven this Lemma the estimation theorem follows immediately. Suppose Γ is a smooth path, then

$$\left| \int_\Gamma f(z) \, dz \right| = \left| \int_a^b f(\gamma(t)) \gamma'(t) \, dt \right| \leq \int_a^b |f(\gamma(t))||\gamma'(t)| \, dt \tag{54.3.7}$$

We can then use the Monotonicity inequality of real integrals and the fact that $|f(z)| < M$ for $z \in \Gamma$ implies $|f(\gamma(t))|$ for $t \in [a, b]$

$$\left| \int_\Gamma f(z) \, dz \right| \leq \int_a^b |f(\gamma(t))| \, dt \cdot \int_a^b |\gamma'(t)| \, dt \leq M \cdot L(\Gamma) \tag{54.3.8}$$

Now suppose Γ is a contour made up of smooth paths $\{\Gamma_i\}$ with lengths $L(\Gamma_i)$. Then

$$\int_\Gamma f(z) \, dz = \sum_i \int_{\Gamma_i} f(z) \, dz \leq \sum_i ML(\Gamma_i) = ML \tag{54.3.9}$$

as desired. $\blacksquare$

> **Example.**
>   (i) Let us find an upper bound for $\left| \int_\Gamma \frac{3z-4}{2z-5} \, dz \right|$ where Γ is the circle $\{z \in \mathbb{C} : |z| = 3\}$. Note that $f(z) = \frac{3z-4}{2z-5}$ is continuous on $\mathbb{C} \setminus \{5/2\}$ and thus on Γ. We see that
>
> $$|3z - 4| \leq |3z| + |4| = 3|z| + 4 = 13 \tag{54.3.10}$$
>
>   and
>
> $$|2z - 5| \geq |2z| - |5| = 2|z| - 5 = 1 \tag{54.3.11}$$

so that

$$|f(z)| \leq 13 \implies \left| \int_\Gamma \frac{3z-4}{2z-5} \, dz \right| \leq 13 \cdot 6\pi = 78\pi \tag{54.3.12}$$

(ii) Let us find an upper bound for $\left| \int_\Gamma \frac{e^{2iz}}{z^2-9} \, dz \right|$ where Γ is the upper-half of the circle $\{z \in \mathbb{C} : |z| = 4\}$ from 4 to -4. Note that $f(z) = \frac{e^{2iz}}{z^2-9}$ is continuous on $\mathbb{C} \setminus \{\pm 3\}$ and thus on Γ. We see that

$$|e^{2iz}| = |e^{2ix} \cdot e^{-2y}| = |e^{-2y}| \leq 1 \tag{54.3.13}$$

and

$$|z^2 - 9| \geq |z^2| - |9| = 7 \tag{54.3.14}$$

so that

$$|f(z)| = \frac{1}{7} \implies \left| \int_\Gamma \frac{e^{2iz}}{z^2-9} \, dz \right| \leq \frac{4}{7}\pi \tag{54.3.15}$$

(iii) Lets find an upper bound for $\left| \int_\Gamma \frac{\sin z}{1+z^2} \right|$ where Γ is the circle $\{z : |z| = 3\}$. We have that

$$|\sin z| = \frac{1}{2}|e^{iz} - e^{-iz}| \leq \frac{1}{2}(|e^y| + |e^{-y}|) = e^3 \tag{54.3.16}$$

$$|z^2 + 1| \geq |z^2| - |1| = 8 \tag{54.3.17}$$

and therefore

$$|f(z)| \leq \frac{1}{8} \implies \left| \int_\Gamma \frac{\sin z}{1+z^2} \right| = \frac{e^3}{8} \tag{54.3.18}$$

◀

54.4 Cauchy's theorem

In this section we will prove a very important result in complex analysis, Cauchy's theorem. It will simplify a lot of closed-contour integrals for us, but before we must introduce some terminology.

> **Definition** *(Path classification)*
> Let $\Gamma : \gamma(t),\ t \in [a, b]$ be a path. If γ is injective on $[a, b]$ then Γ is **simple**, if $\gamma(a) = \gamma(b)$ then Γ is **closed**.

The following simply result is surprisingly difficult to prove, so we state it without proof.

> **Theorem** *(Jordan Curve Theorem)*
> Let Γ be a simple-closed curve, then $\mathbb{C} \setminus \Gamma$ is the disjoint union of a bounded region in(Γ), known as the **inside of** Γ, and an unbounded region out(Γ), known as the **outside of** Γ.

There is a simple way to determine if a point lies inside a curve or not. Simply take any line from infinity which passes through the given point, and count the number of intersections with the curve before the curve crosses the given point. If this number is even then the point is outside the curve, while if the number is odd then the point is inside the curve.

The definition of paths is instrumental in defining the concept of simply-connected regions, informally they are regions with no "holes".

Definition *(Simply-connected)* A region $\mathcal{R}$ is **simply-connected** if, given any simple-closed curve $\Gamma \subset \mathcal{R}$, the inside of Γ is also a subset of $\mathcal{R}$.

We are now ready to state Cauchy's Theorem.

Theorem *(Cauchy's theorem)*
Let f be analytic on a simply connected region $\mathcal{R}$. Then for any closed contour $\Gamma \in \mathcal{R}$ we have that

$$\int_\Gamma f(z)\,dz = 0 \tag{54.4.1}$$

Proof. The proof of Cauchy's theorem is quite difficult but very important, so we will give it an entire section. ∎

Proposition *(Contour independence)*
Let f be analytic on a simply-connected region $\mathcal{R}$. Suppose Γ_1, Γ_2 are contours in $\mathcal{R}$ with the same initial and final points, then

$$\int_{\Gamma_1} f(z)\,dz = \int_{\Gamma_2} f(z)\,dz \tag{54.4.2}$$

Proof. Take Γ_1 to be a contour from α to β, and let $\tilde{\Gamma}_2$ be the reverse contour of Γ_2. Clearly $\Gamma = \Gamma_1 \tilde{\cup} \tilde{\Gamma}_2$ is a closed contour, so using Cauchy's theorem

$$\int_\Gamma f(z)\,dz = \int_{\Gamma_1} f(z)\,dz + \int_{\tilde{\Gamma}_2} f(z)\,dz = 0 \tag{54.4.3}$$

and therefore

$$\int_{\Gamma_1} f(z)\,dz = \int_{\Gamma_2} f(z)\,dz \tag{54.4.4}$$

as desired. We begin by proving Cauchy's theorem for rectangular contours. Let f be analytic on a simply connected region $\mathcal{R}$ and let Γ be a rectangular contour in $\mathcal{R}$ with perimeter l. We dissect Γ into four smaller rectangular contours, $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$, as shown below ∎

Theorem *(Contour shrinking)*
Let f be analytic on $\mathcal{R} \setminus \{\alpha\}$ where $\mathcal{R}$ is a simply connected region containing the point α. Suppose $\Gamma \subseteq \mathcal{R}$ is a simple closed contour, then

$$\int_\Gamma f(z)\,dz = \int_C f(z)\,dz \tag{54.4.5}$$

where $C \in \mathcal{R}$ is any circle centered at α.

Proof. Let L_1, L_2 be two line segments connecting C and Γ, partitioning the former into C_1, C_2 and latter into P_1, P_2 as shown
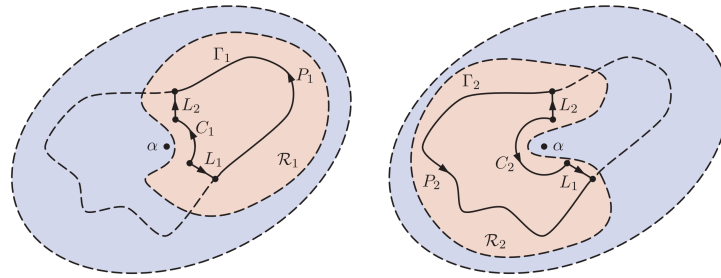
Figure 54.1. Contours Γ_1 and Γ_2.

This allows us to define two new contours, $\Gamma_1 = L_1 \breve{\cup} P_1 + \breve{\cup} \tilde{L}_2 \breve{\cup} \tilde{C}_1$ and $\Gamma_2 = L_2 \breve{\cup} P_2 + \breve{\cup} \tilde{L}_1 \breve{\cup} \tilde{C}_2$, each contained within the simply connected regions $\mathcal{R}_1$ and $\mathcal{R}_2$ respectively. Note that $\alpha \notin \mathcal{R}_1, \mathcal{R}_2$ so f is analytic on both these regions, so we can use Cauchy's theorem:

$$\int_{L_1} f(z)\, dz + \int_{P_1} f(z)\, dz + \int_{\sim L_2} f(z)\, dz + \int_{\sim C_1} f(z)\, dz = 0 \tag{54.4.6}$$

$$\int_{L_2} f(z)\, dz + \int_{P_2} f(z)\, dz + \int_{\sim L_1} f(z)\, dz + \int_{\sim C_2} f(z)\, dz = 0 \tag{54.4.7}$$

Adding these two we find that

$$\int_{P_1} f(z)\, dz + \int_{P_2} f(z)\, dz = \int_{C_1} f(z)\, dz + \int_{C_2} f(z)\, dz \tag{54.4.8}$$

or in other words

$$\int_\Gamma f(z)\, dz = \int_C f(z)\, dz \tag{54.4.9}$$

as desired. ∎

54.5 Cauchy's integral formula

We not present another important result given by Cauchy which is useful in both evaluating integrals and the values of special functions.

> **Theorem** *(Cauchy's integral formula)*
> Let Γ be a simple-closed contour on a simply-connected region $\mathcal{R}$, over which the function f is analytic. Then
> $$f(\alpha) = \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{z - \alpha}\, dz, \ \forall \alpha \in \text{in}(\Gamma) \tag{54.5.1}$$

Proof. Firstly, note that since $\text{in}(\Gamma)$ is open, we can find a circle C centered at α lying entirely inside Γ, and thus by the contour shrinking theorem we can set

$$\int_\Gamma \frac{f(z)}{z - \alpha}\, dz = \int_C \frac{f(z)}{z - \alpha}\, dz \tag{54.5.2}$$

for any such circle C. Now we make use of the fact that

$$\int_C \frac{1}{z - \alpha} \, dz = \int_0^{2\pi} \frac{ire^{it}}{(\alpha + re^{it}) - \alpha} \, dt = 2\pi i \qquad (54.5.3)$$

so that

$$I = \int_C \frac{f(z)}{z - \alpha} \, dz - 2\pi i f(\alpha) = \int_C \frac{f(z) - f(\alpha)}{z - \alpha} \, dz \qquad (54.5.4)$$

Our goal is now to prove that $I = 0$. Using the fact that f is differentiable at α, given any $\epsilon > 0$ one can find $\delta > 0$ such that

$$|z - \alpha| < \delta \implies |f(z) - f(\alpha)| < \epsilon \qquad (54.5.5)$$

Letting $r < \delta$ for a given ϵ, then we see that

$$\left| \frac{f(z) - f(\alpha)}{z - \alpha} \right| < \frac{\epsilon}{r} \qquad (54.5.6)$$

Using the Estimation theorem for contour integrals we then find that

$$|I| \leq 2\pi r \cdot \frac{\epsilon}{r} = 2\pi \epsilon \qquad (54.5.7)$$

Since this holds for any arbitrary $\epsilon > 0$ it follows that $I = 0$, as desired. ∎

Example.

(i) Consider

$$I = \int_\Gamma \frac{e^{3z}}{z^2 - 4} \, dz, \ \Gamma = \{z : |z - 1| = 2\} \qquad (54.5.8)$$

Seeing as the inside of Γ does not include $z = -2$, we choose $f(z) = \frac{e^{3z}}{z+2}$ which is analytic on any region excluding $z = -2$. Then we see that

$$I = \int_\Gamma \frac{f(z)}{z - 2} \, dz = f(2) = 2\pi i \frac{e^6}{4} = \frac{e^6 \pi i}{2} \qquad (54.5.9)$$

where we used the Cauchy integral formula with a region $\mathcal{R}$ containing Γ but not $z = -2$ (e.g. $\mathcal{R} = \{z : \operatorname{Re}(z) > -2\}$.

(ii) Consider

$$I = \int_\Gamma \frac{\cos 2z}{z(z^2 + 4)} \, dz, \ \Gamma = 1 + i \to -1 + i \to -1 - i \to 1 - i \qquad (54.5.10)$$

The inside of Γ does not include $\pm 2i$ so we should choose $f(z) = \frac{\cos 2z}{z^2 + 4}$. This function is analytic on any region excluding $\pm 2i$, so for example a disc centered at 0 of radius $\sqrt{2} < r < 2$. Then we see that

$$I = \int_\Gamma \frac{f(z)}{z} \, dz = 2\pi i \cdot \frac{1}{4} = \frac{\pi i}{2} \qquad (54.5.11)$$

(iii) Consider

$$I = \int_\Gamma \frac{\cos 3z}{z^2 - z} \, dz, \ \Gamma = \{z : |z - 1| = 2\} \qquad (54.5.12)$$

It is not immediately obvious what $f(z)$ to choose, since both denominator zeros $z = 0$ and $z = 1$ are included inside Γ. We solve this problem by using the partial fraction expansion of $\frac{1}{z(z-1)}$:

$$\frac{1}{z(z-1)} = \frac{A}{z} + \frac{B}{z-1} \implies A + B = 0, -A = 1 \implies A = -1, B = 1 \qquad (54.5.13)$$

Therefore

$$I = \int_\Gamma \left(\frac{\cos 3z}{z-1} - \frac{\cos 3z}{z} \right) dz = 2\pi i(\cos 3 - 1) \qquad (54.5.14)$$

◀

An immediate corollary of the Cauchy integral formula is **Gauss' Mean Value Theorem**

$$f(\alpha) = \frac{1}{2\pi} \int_0^{2\pi} f(\alpha + re^{it})\, dt \qquad (54.5.15)$$

where r is the radius of the circle C contained in a simply-connected region $\mathcal{R}$ on which f is analytic.

Cauchy's integral formula is really useful in the proof of the following remarkable result.

Theorem *(Louiville's theorem)*
Every bounded entire function is constant (i.e. an entire non-constant function is always unbounded)

Proof. Let f be a bounded, entire function, and let $\alpha \in \mathbb{C}$. Define C to be a circle centered at the origin with radius $r > |\alpha|$ so that $\alpha \in \text{in}(C)$. Since f is analytic on $\mathbb{C}$, a simply-connected region, we can apply Cauchy's theorem to find that

$$f(\alpha) = \frac{1}{2\pi i} \int_C \frac{f(z)}{z - \alpha}\, dz \qquad (54.5.16)$$

and thus

$$f(\alpha) - f(0) = \frac{\alpha}{2\pi i} \int_C \frac{f(z)}{z(z-\alpha)}\, dz \qquad (54.5.17)$$

By assumption, f is bounded so there exists $K > 0$ such that $|f(z)| \leq K$, $\forall z \in \mathbb{C}$. Similarly, we also have that $|z| \leq r$, $\forall z \in \text{in}(C)$ so that $|z(z - \alpha)| > r(r - |\alpha|)$. Therefore we can find an upper bound for $|f(\alpha) - f(0)|$:

$$|f(\alpha) - f(0)| \leq \frac{|\alpha|}{2\pi} \frac{K}{r(r - |\alpha|)} \cdot 2\pi r = \frac{|\alpha|}{r - |\alpha|} K \qquad (54.5.18)$$

This result holds for any $r > |\alpha|$, so given any $\epsilon > 0$ we can always find $r > |\alpha|$ such that

$$|f(\alpha) - f(0)| < \epsilon \qquad (54.5.19)$$

by setting $r > |\alpha|(1 + K/\epsilon)$. It then follows that $f(\alpha) = f(0)$, so f is a constant function. ■

For example, let's consider $f(z) = e^{i|z|}$, which is clearly a non-constant function, and also bounded since $|f(z)| = 1$, $\forall z \in \mathbb{C}$. Consequently, we see that f cannot be entire, it is not differentiable somewhere in the complex plane (e.g. the origin).

54.6 Cauchy's derivative formula

One can derive a theorem similar to Cauchy's integral formula which calculates $f'(\alpha)$.

Theorem *(Cauchy's first derivative formula)*
Let f be analytic on a simply connected region $\mathcal{R}$ containing a simple-closed contour Γ, then

$$f'(\alpha) = \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{(z-\alpha)^2}\, dz, \ \forall \alpha \in \mathcal{R} \tag{54.6.1}$$

Proof. Using the contour shrinking theorem we can write

$$I = \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{(z-\alpha)^2}\, dz = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z-\alpha)^2}\, dz \tag{54.6.2}$$

where C is a circle centered at α with radius r, which for our purposes is contained inside Γ.
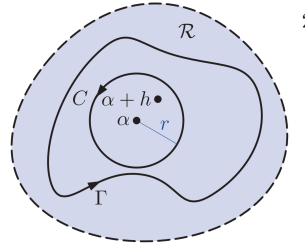
On the other hand, note that

$$f'(\alpha) = \lim_{h\to 0} \frac{f(\alpha+h) - f(\alpha)}{h} = \lim_{h\to 0} \frac{1}{2\pi i h}\left(\int_C \frac{f(z)}{z-\alpha-h}\, dz - \int_C \frac{f(z)}{z-\alpha-h} \right) dz \tag{54.6.3}$$

$$= \frac{1}{2\pi i} \lim_{h\to 0} \int_C \frac{f(z)}{(z-\alpha)(z-\alpha-h)}\, dz \tag{54.6.4}$$

implying that

$$f'(\alpha) - I = \lim_{h\to 0} \frac{1}{2\pi i} \int_C \frac{h}{(z-\alpha)^2(z-\alpha-h)} f(z)\, dz \tag{54.6.5}$$

Now we will find an upper bound to the modulus of the integrand and show that it can be made arbitrarily small. We have that

$$|z-\alpha| = r, \ |z-\alpha-h| \geq r - |h| \tag{54.6.6}$$

and since f is analytic on $\mathcal{R}$ and thus inside C, it will be bounded. Thus there exists $K > 0$ such that $|f(z)| < K$, $\forall z \in \text{in}(C)$. The upper bound for the integrand can now be written as

$$\left| \frac{1}{2\pi i} \int_C \frac{h}{(z-\alpha)^2(z-\alpha-h)} f(z)\, dz \right| \leq \frac{|h|K}{r(r-|h|)} < \frac{K}{r^2}|h| \tag{54.6.7}$$

so that

$$0 < |f'(\alpha) - I| < \lim_{h \to 0} \frac{K}{r^2}|h| = 0 \implies f'(\alpha) = I \tag{54.6.8}$$

as desired. ∎

> **Example.** In the following, let $\Gamma = \{z \in \mathbb{C} : |z| = 3\}$.
>
> (i) Using the Cauchy derivative formula we see that
>
> $$\int_\Gamma \frac{e^{2z}}{(z+1)^2}\, dz = 2\pi i \frac{d}{dz}(e^{2z})\Big|_{z=-1} = \frac{4\pi i}{e^2} \tag{54.6.9}$$
>
> (ii) Consider the following
>
> $$I = \int_\Gamma \frac{4\cos z}{z(z+2)^2}\, dz \tag{54.6.10}$$
>
> At a first glance it looks like setting $f(z) = \frac{4\cos z}{z}$ should work, but unfortunately this function is not analytic at $z = 0$ which is inside Γ. We instead use the method of partial fraction expansions and write
>
> $$\frac{1}{z(z+2)^2} = \frac{A}{z} + \frac{B}{z+2} + \frac{Cz}{(z+2)^2} \tag{54.6.11}$$
>
> where A, B, C need to be determined. They are easily found to be $A = \frac{1}{4}, B = -\frac{1}{4}, C = \frac{1}{2}$ so that
>
> $$I = \int_\Gamma \frac{\cos z}{z}\, dz - \int_\Gamma \frac{\cos z}{z+2} + \int_\Gamma \frac{2\cos z}{(z+2)^2} \tag{54.6.12}$$
>
> The first two integral are an application of Cauchy's integral formula with $f(z) = \cos z$ which is analytic on $\mathcal{R} = \mathbb{C}$:
>
> $$\int_\Gamma \frac{\cos z}{z}\, dz = 2\pi i, \quad \int_\Gamma \frac{\cos z}{z+2} = 2\pi i\cos(2) \tag{54.6.13}$$
>
> The third integral instead requires us to use Cauchy's first derivative formula
>
> $$\int_\Gamma \frac{2\cos z}{(z+2)^2} = -4\pi i \sin(-2) \tag{54.6.14}$$
>
> so that
>
> $$\int_\Gamma \frac{4\cos z}{z(z+2)^2}\, dz = 2\pi i(1 - z\cos(2) - 2\sin(2)) \tag{54.6.15}$$
>
> ◀

Using induction it is not hard to generalise the Cauchy derivative formulas to nth order

> **Theorem** *(Cauchy derivative formula)*
> Let f be analytic on a simply-connected region $\mathcal{R}$ containing a simple-closed path Γ. Then
>
> $$f^{(n)}(\alpha) = \frac{n!}{2\pi i}\int_\Gamma \frac{f(z)}{(z-\alpha)^n}\, dz, \ \forall \alpha \in \text{in}(\Gamma) \tag{54.6.16}$$

Example. Let $\Gamma = \{z : |z| = 2\}$.

(i) Consider

$$\int_\Gamma \frac{\cosh 2z}{(z+i)^3} \, dz \tag{54.6.17}$$

Note that $\cosh 2z$ is analytic on $\mathbb{C}$, so applying Cauchy's third derivative formula with $\alpha = i \in \text{in}(\Gamma)$ we find that

$$\int_\Gamma \frac{\cosh 2z}{(z+i)^3} \, dz = \frac{2\pi i}{2} \cdot 4 \cosh(-2i) = 4\pi i \cos(2) \tag{54.6.18}$$

(ii) Consider

$$I = \int_\Gamma \frac{e^{2z}}{z^3(z+1)} \, dz \tag{54.6.19}$$

It seems like there is no immediately obvious choice for $f(z)$, since both roots of the denominator, $z = 0, -1$ lie inside Γ. So we use the method of partial fractions:

$$\frac{1}{z^3(z+1)} = \frac{1}{z^3} - \frac{1}{z^2} + \frac{1}{z} - \frac{1}{z+1} \tag{54.6.20}$$

Therefore

$$I = \int_\Gamma \frac{e^{2z}}{z^3} \, dz - \int_\Gamma \frac{e^{2z}}{z^2} \, dz + \int_\Gamma \frac{e^{2z}}{z} \, dz - \int_\Gamma \frac{e^{2z}}{z+1} \, dz \tag{54.6.21}$$

The first integral is $4\pi i$, the second integral is $4\pi i$, the third integral is $2\pi i$ and the third integral is $\frac{2\pi i}{e^2}$. Finally we get that

$$\int_\Gamma \frac{e^{2z}}{z^3(z+1)} \, dz = 2\pi i \left(1 - \frac{1}{e^2}\right) \tag{54.6.22}$$

◀

Theorem *(Analyticity of derivatives)*
Let f be analytic on $\mathcal{R}$, then $f^{(n)}$ for $n = 1, 2....$ is also analytic on $\mathcal{R}$.

Proof. Let $\alpha \in \mathcal{R}$ and let D be an open disc centered at α lying in $\mathcal{R}$ (which exists since $\mathcal{R}$ is open). Let Γ be a circle centered at α lying entirely in D, then $f^{(n)}(\alpha)$ exists and is given by Cauchy's formula applied to α with contour Γ in the region D (not $\mathcal{R}$, which may not be simply-conencted). This also implies analyticity. ∎

54.7 Proof of Cauchy's theorem

Let us begin by proving Cauchy's theorem for rectangular contours.

Proposition *(Cauchy's theorem for rectangular contours)*

Let f be analytic on a simply connected region $\mathcal{R}$ and let Γ be a rectangular contour in $\mathcal{R}$. Then

$$\int_\Gamma f(z)\,dz = 0 \tag{54.7.1}$$

Proof. Let f be analytic on a simply connected region $\mathcal{R}$ and let Γ be a rectangular contour in $\mathcal{R}$ with perimeter l. We dissect Γ into four smaller rectangular contours, $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$, as shown below.
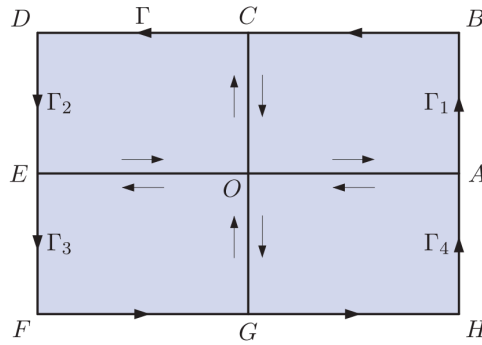
It is easy to see that

$$I = \int_\Gamma f(z)\,dz = \int_{\Gamma_1} f(z)\,dz + \int_{\Gamma_2} f(z)\,dz + \int_{\Gamma_3} f(z)\,dz + \int_{\Gamma_4} f(z)\,dz \tag{54.7.2}$$

implying that

$$|I| \le 4\left|\int_{\Delta_1} f(z)\,dz\right| \tag{54.7.3}$$

where Δ_1 is one of $\{\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4\}$ which gives the largest contour integral. We know repeat this procedure for Δ_1, dissecting it into four smaller pieces and finding the upper bound:

$$|I| \le 4^2\left|\int_{\Delta_2} f(z)\,dz\right| \tag{54.7.4}$$

We reiterate this process ad infinitum thus creating a sequence $\{\Delta_i\}_{i=1,\dots}$ of contours with perimeter $L(\Delta_n) = \frac{l}{2^n}$ and with the property that

$$|I| \le 4^n\left|\int_{\Delta_n} f(z)\,dz\right| \tag{54.7.5}$$

Let R_n by the rectangle formed by the inside of Γ_n and its boundary. These rectangles are nested and their diagonal satisfies

$$0 \le s_n \le \frac{l}{2^n} \implies \lim_{n\to\infty} s_n = 0 \tag{54.7.6}$$

so by the nested rectangles theorem there is a complex number α that lies in all the rectangles, and thus in all the contours. Let's taylor expand f around α:

$$f(z) = f(\alpha) + (z-\alpha)f'(\alpha) + e(z) \tag{54.7.7}$$

where $e(z)$ is analytic on $\mathcal{R}$ due to the analyticity of f, and define

$$r(z) = \begin{cases} \frac{e(z)}{z-\alpha}, & z \neq \alpha \\ 0, & z = \alpha \end{cases} \to 0 \text{ as } z \to \alpha \tag{54.7.8}$$

Note that $\lim_{z \to \alpha} = r(\alpha)$ so $r(z)$ is a continuous function on $\mathcal{R}$. We can then integrate f as follows

$$\int_{\Delta_n} f(z)\, dz = \int_{\Delta_n} (f(\alpha) + (z - \alpha)f'(\alpha))\, dz + \int_{\Delta_n} (z - \alpha)r(z)\, dz \tag{54.7.9}$$

The first integral vanishes by the Closed Contour Theorem, leaving

$$\int_{\Delta_n} f(z)\, dz = \int_{\Delta_n} (z - \alpha)r(z)\, dz \tag{54.7.10}$$

Let $\epsilon > 0$, by the continuity of r there exists $\delta > 0$ such that

$$|z - \alpha| < \delta \implies |r(z)| < \epsilon \tag{54.7.11}$$

Therefore let us choose N such that for all $n > N$, the contour Δ_n lies entirely within the disc $\{z \in \mathbb{C} : |z - \alpha| < \delta\}$, which can be done by the nested triangles theorem. We then have that

$$|(z - \alpha)r(z)| < \frac{l\epsilon}{2^n} \tag{54.7.12}$$

so that

$$|I| \leq 4^n \cdot \frac{l\epsilon}{2^n} \cdot \frac{l}{2^n} = l^2\epsilon \tag{54.7.13}$$

Since this inequality holds for any $\epsilon > 0$ we must have that $I = 0$ as desired. ∎

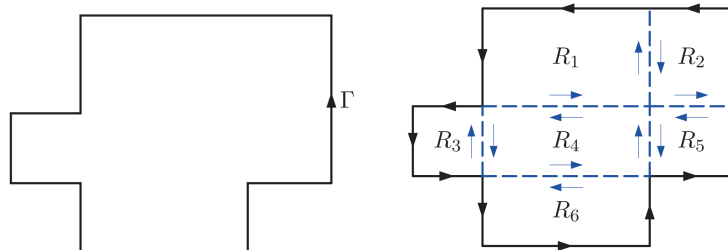We can apply Cauchy's theorem for rectangular theorems to prove the more general case of closed grid paths.

> **Proposition** *(Cauchy's theorem for closed grid paths)*
> Let f be analytic on a simply connected region $\mathcal{R}$ and let Γ be a closed grid path in $\mathcal{R}$. Then
> $$\int_\Gamma f(z)\, dz = 0 \tag{54.7.14}$$

Proof. Consider a simple-closed grid path as shown below.

We can again dissect it into smaller rectangular contours such that the shared edges cancel out, and thus using Cauchy's theorem for rectangular contours we see that

$$\int_{\Gamma} f(z)\, dz = \sum_{i=1}^{6} \int_{R_i} f(z)\, dz = 0 \tag{54.7.15}$$

Now consider a closed grid path. In general it can be decomposed into a finite number of simple-closed grid paths and a finite number of line segments which must be traversed twice. Contour integrals along both such paths must vanish, proving the desired result. ∎

Note that Cauchy's theorem for closed grid paths implies that the contour integral along a grid path from α to β is actually independent of the chosen grid path. As long as it starts at α and ends in β any grid path will yield the same contour integral.

We now prove another important result which will allow us to generalise Cauchy's theorem to any closed contour.

Theorem *(Primitive theorem)*
Let f be analytic on a simply connected region $\mathcal{R}$, then it has a primitive on $\mathcal{R}$.

Proof. We begin by proving the following lemma: **Lemma.** Let f be continuous on $\mathcal{R}$ and such that

$$\int_{\Gamma} f(z)\, dz = 0 \tag{54.7.16}$$

for all closed grid paths $\Gamma \subset \mathcal{R}$. Then f has a primitive on $\mathcal{R}$.

Proof. Let

$$\int_{\alpha}^{\beta} f(z)\, dz \tag{54.7.17}$$

be the contour integral of f along any grid path Γ from α to β. This integral must be independent of the chosen contour by Cauchy's theorem for closed grid paths.

Let $\alpha \in \mathcal{R}$, then

$$F(z) = \int_{\alpha}^{z} f(w)\, dw \tag{54.7.18}$$

is the primitive of f. Indeed, note that

$$F'(z) = \lim_{h \to 0} \frac{1}{h} \left(\int_{\alpha}^{z+h} f(w)\, dw - \int_{\alpha}^{z} f(w)\, dz \right) = \lim_{h \to 0} \frac{1}{h} \int_{z}^{z+h} f(w)\, dw \tag{54.7.19}$$

where we used a grid path that goes through both z and $z + h$. Consequently we have that

$$F'(z) - f(z) = \lim_{h \to 0} \frac{1}{h} \left(\int_{\alpha}^{z+h} f(w)\, dw - \int_{\alpha}^{z} f(w)\, dz \right) = \lim_{h \to 0} \frac{1}{h} \int_{z}^{z+h} (f(w) - f(z))\, dw \tag{54.7.20}$$
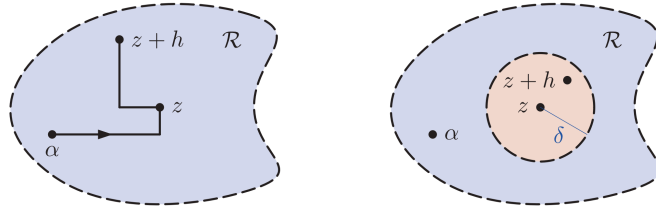
Now let $\epsilon > 0$, then for all $z \in \mathcal{R}$ the continuity of f implies that there exists $\delta > 0$ such that

$$|w - z| < \delta \implies |f(w) - f(z)| < \epsilon \ (z \in \mathcal{R}) \tag{54.7.21}$$

Note that if we shrink δ then the inequality must still hold, so let us choose δ to be small enough so that $\{w \in \mathbb{C} : |w - z| < \delta\} \subset \mathcal{R}$. Let h be a non-zero complex number such that $|h| < \delta$, and let Γ be the two-segment contour from z to $z + h$.

Then $L(\Gamma) \leq 2|h|$, and since $w \in \Gamma$

$$|w - z| < \delta \implies |f(w) - f(z)| < \epsilon \implies \left| \frac{1}{h} \int_z^{z+h} (f(w) - f(z))\, dw \right| \leq 2\epsilon \tag{54.7.22}$$

In other words, we have that

$$|h| < \delta \implies \left| \frac{1}{h} \int_z^{z+h} (f(w) - f(z))\, dw \right| \leq 2\epsilon \tag{54.7.23}$$

proving that the limit on the RHS of (54.7.20) vanishes, and therefore that $F'(z) = f(z)$. ∎

Let $\mathcal{R}$ be simply connected, then for any closed grid path $\Gamma \in \mathcal{R}$, Cauchy's theorem for closed grid path yields

$$\int_\Gamma f(z)\, dz = 0 \implies f \text{ has a primitive} \tag{54.7.24}$$

proving the desired theorem. ∎

Finally, let us prove Cauchy's theorem.

Theorem *(Cauchy's theorem)*

Let f be analytic on a simply connected region $\mathcal{R}$. Then for any closed contour $\Gamma \in \mathcal{R}$ we have that

$$\int_\Gamma f(z)\, dz = 0 \tag{54.7.25}$$

Proof. Since f is analytic on a simply-connected region $\mathcal{R}$., it has a primitive F. Using the Closed Contour theorem, we have that

$$\int_\Gamma f(z)\, dz = 0 \tag{54.7.26}$$

for any closed contour Γ in $\mathcal{R}$. ∎

We prove a final result that is linked to Cauchy's theorem.

Theorem *(Morera's theorem)*

Let f be continuous on a region $\mathcal{R}$ such that

$$\int_{\Gamma} f(z)\, dz = 0 \qquad\qquad (54.7.27)$$

for all rectangular contours $\Gamma \in \mathcal{R}$. Then f is analytic on $\mathcal{R}$.

Proof. The Lemma in the proof of the Primitive theorem implies that f has a primitive F that is analytic on $\mathcal{R}$. Moreover, the analyticity of derivatives implies that $f = F'$ must also be analytic on $\mathcal{R}$. ∎

CA5 Taylor and Laurent series

55.1   Series

> **Theorem** *(Convergence of series $\implies$ null sequence)*
> If $\sum_{n=1}^{\infty} z_n$ converges then (z_n) is a null sequence.

Proof. Let $s_n = \sum_{k=1}^{n} z_k$ form the partial sum sequence of $\sum_{N=1}^{\infty} z_n$. Since the sum converges we must have that $s_n \to s$, therefore

$$z_n = s_n - s_{n-1} \to s - s = 0 \tag{55.1.1}$$

so z_n is null. ∎

> **Proposition** *(Geometric series)*
> Consider $\sum_{n=0}^{\infty} az^n$ for $a, z \in \mathbb{C}$. If $|z| < 1$ then the series converges to $\frac{a}{1-z}$ while if $|z| \geq 1$ and $a \neq 0$ then the series diverges.

Proof. Suppose that $z \neq 1$. Then

$$s_n = a + az + az^2 + ... + az^n = a\frac{1 - z^{n+1}}{1 - z} \tag{55.1.2}$$

and since (z^n) is a basic null sequence for $|z| \leq 1$ we find that

$$\lim_{n \to \infty} s_n = \frac{a}{1 - z}, \quad |z| < 1 \tag{55.1.3}$$

If instead $|z| \geq 1$ then $|z^n|$ diverges so (z_n) cannot converge to 0. Moreover if $a \neq 0$ then (az^n) does not converge to 0 so the series must diverge, ∎

> **Proposition** *(Harmonic series)*
> The p-series
> $$\sum_{n=1}^{\infty} \frac{1}{n^p} = 1 + \frac{1}{2^p} + \frac{1}{3^p} + ... \tag{55.1.4}$$
> converges if $p > 1$ and diverges if $p \leq 1$.

Proof. If $p \leq 0$ then each term in the series is larger than or equal to one, so the partial sum sequence cannot be null. It follows that the series then diverges.

Now suppose that $0 < p \leq 1$. We construct a set of n rectangles of width 1 starting at $x = 1$, and height $\frac{1}{x^p}$ where x is the left endpoint of the rectangle. This is a very crude Riemann integral, so we should expect the sum of the areas of each rectangle to be bounded below by the actual area under the curve $y = \frac{1}{x^p}$:

$$\sum_{m=1}^{n} \frac{1}{m^p} = s_n \geq \int_1^{n+1} \frac{dx}{x} = \log(n+1) \tag{55.1.5}$$

Since $(\log(n+1))$ is a non-null sequence (divergent) it follows that s_n is non-null. Hence the p-series diverges for $p \leq 1$.

By a similar argument, for $p > 1$ we can construct n rectangles of width 1 starting at $x = 0$, and height $\frac{1}{x^p}$ where x is now the right endpoint of the rectangle. This time we expect the partial sum to underestimate the actual area under the curve $y = \frac{1}{x^p}$ so we can write

$$s_n - 1 \leq \int_1^n \frac{dx}{x^p} = \frac{n^{1-p} - 1}{1-p} = \frac{1}{p-1}\left(1 - \frac{1}{n^{p-1}}\right) < \frac{1}{p-1} \tag{55.1.6}$$

Since s_n is bounded above, it follows from the monotone convergence theorem that (s_n) converges, and thus the p-series must also converge, for $p > 1$. $\blacksquare$

From the combination rules for sequences we can derive the following combination rules for series

> **Proposition** *(Combination rules)*
> Suppose that $\sum_{n=1}^{\infty} z_n$ and $\sum_{n=1}^{\infty} w_n$ are both convergent series. Then
>
> $$\sum_{n=1}^{\infty}(z_n + w_n) = \sum_{n=1}^{\infty} z_n + \sum_{n=1}^{\infty} w_n, \ \sum_{n=1}^{\infty} \lambda z_n = \lambda \sum_{n=1}^{\infty} z_n, \ \forall \lambda \in \mathbb{C} \tag{55.1.7}$$

It follows immediately that if both the real and imaginary parts of $\sum_{n=1}^{\infty} z_n$ converge, then so will

$$\sum_{n=1}^{\infty} z_n$$

. The converse result of this can also be proven.

> **Proposition** *(Real and imaginary part)*
> The series $\sum_{n=1}^{\infty} z_n$ is convergent iff. both its real and imaginary parts converge, and
>
> $$\sum_{n=1}^{\infty} z_n = \sum_{n=1}^{\infty} \text{Re}(z_n) + \sum_{n=1}^{\infty} \text{Im}(z_n) \tag{55.1.8}$$

Proof. Follows from the analogous result for series. $\blacksquare$

Example. Let's calculate the following sum

$$\sum_{n=0}^{\infty} \frac{1}{2^n} \cos(nx), \ x \in \mathbb{R} \tag{55.1.9}$$

Note that $\frac{\cos(nx)}{2^n} = \mathrm{Re}\left(\left(\frac{e^{ix}}{2}\right)^n\right)$. Therefore

$$\sum_{n=0}^{\infty} \frac{1}{2^n} \cos(nx) = \mathrm{Re}\left(\sum_{n=0}^{\infty} \frac{e^{inx}}{2^n}\right) \tag{55.1.10}$$

$$= \mathrm{Re}\left(\frac{2}{2 - e^{ix}}\right) \tag{55.1.11}$$

$$= \mathrm{Re}\left(\frac{2(2 - e^{-ix})}{5 - 4\sin x}\right) = \frac{4 - 2\cos x}{5 - 4\sin x} \tag{55.1.12}$$

◀

Theorem *(Comparison test)*
If $\sum_{n=1}^{\infty} a_n$ is a convergent real series of non-negative terms, and $|z_n| \leq a_n$ then $\sum_{n=1}^{\infty} z_n$ converges.

Example. Consider the following series

$$\sum_{n=1}^{\infty} \frac{\cos n}{n\sqrt{n}} \tag{55.1.13}$$

Note that $|\cos n| \leq 1$ implying that

$$\left| frac{\cos n}{n\sqrt{n}} \right| \leq \frac{1}{n^{3/2}} \tag{55.1.14}$$

and since $\sum_{n=1}^{\infty} \frac{1}{n^{3/2}}$ converges, it follows from the Comparison test that $\sum_{n=1}^{\infty} z_n$ must also converge. ◀

Definition *(Absolute convergence)*
The complex series $\sum_{n=1}^{\infty} z_n$ is absolutely convergent if the real series $\sum_{n=1}^{\infty} |z_n|$ is convergent.

Theorem *(Absolute convergence theorem)* Let $\sum_{n=1}^{\infty} z_n$ be an absolutely convergent complex series. Then $\sum_{n=1}^{\infty} z_n$ is also convergent.

Example.
  (i) Consider the following series
$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \tag{55.1.15}$$

Since $\left|\frac{(-1)^{n+1}}{n}\right| = \frac{1}{n}$, and the harmonic series diverges, the series cannot be absolutely convergent.

(ii) Consider the following series

$$\sum_{n=0}^{\infty} \frac{(-1)^{n+1}(1+i)^n}{2^n} \tag{55.1.16}$$

We find that

$$\left|\frac{(-1)^{n+1}(1+i)^n}{2^n}\right| = \left(\frac{1}{\sqrt{2}}\right)^n \tag{55.1.17}$$

and since $\left(\frac{1}{\sqrt{2}}\right)^n$ is a null sequence it follows that the series converges absolutely.

◄

Theorem *(Ratio test)*

Let $\sum_{n=1}^{\infty} z_n$ be a complex series such that $|z_{n+1}/z_n| \to l$ as $n \to \infty$.
  (i) if $0 \le l < 1$ then $\sum_{n=1}^{\infty} z_n$ converges absolutely, and thus converges.
  (ii) if $l > 1$ then $\sum_{n=1}^{\infty} z_n$ diverges.

Example.

(i) Consider the following series

$$\sum_{n=1}^{\infty} \frac{n^2}{3^n + i} \tag{55.1.18}$$

Let $z_n = \frac{n^2}{3^n + i}$, we find that

$$\left|\frac{z_{n+1}}{z_n}\right| = \left|\frac{(n+1)^2}{3^{n+1} + i} \frac{3^n + i}{n^2}\right| = \left|\frac{3^n + i}{3 \cdot 3^n + i}\right| \left(1 + \frac{1}{n}\right)^2 \to \frac{1}{3} \tag{55.1.19}$$

so from the ratio test we see that the series converges absolutely.

(ii) Consider the following series

$$\sum_{n=1}^{\infty} \frac{z^n}{n!}, \ z \in \mathbb{C} \tag{55.1.20}$$

Let $z_n = \frac{z^n}{n!}$, then we find that

$$\left|\frac{z_{n+1}}{z_n}\right| = \left|\frac{z^{n+1}}{(n+1)!} \frac{n!}{z^n}\right| = \frac{|z|}{n+1} \to 0 \tag{55.1.21}$$

so we find that the series converges absolutely.

◄

55.2 Complex power series

> **Theorem** *(Radius of convergence)*
> Given any power series $\sum_{n=0}^{\infty} a_n(z-\alpha)^n$ then it either
>   (i) converges for all $z \in \mathbb{C}$
>   (ii) converges only if $z = \alpha$
>   (iii) converges for all $|z - \alpha| < R$ and diverges otherwise, for some **radius of convergence** R.

Proof. If $z = \alpha$ then the power series converges, so suppose that $z \neq \alpha$.

Let

$$\lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right| = R \tag{55.2.1}$$

where R exists and is finite. We now note that the ratio of consecutive terms in the power series is given by

$$\left| \frac{a_{n+1}(z-\alpha)^{n+1}}{a_n(z-\alpha)^n} \right| \to \frac{|z-\alpha|}{R} \tag{55.2.2}$$

so by the ratio test, the power series converges if $|z - \alpha| < R$ and diverges if $|z - \alpha| > R$. ∎

> **Example.**
>   (i) Let's find the radius of convergence of
>
> $$\sum_{n=0}^{\infty} (\alpha z)^n \tag{55.2.3}$$
>
> Of course if $\alpha = 0$ then the series converges for all $z \in \mathbb{C}$. Suppose that $\alpha \neq 0$, and let $z_n = (\alpha z)^n$. Then
>
> $$\left| \frac{z_{n+1}}{z_n} \right| = |\alpha||z| \to |\alpha||z| \text{ as } n \to \infty \tag{55.2.4}$$
>
> We see that if $|\alpha||z| < 1 \implies |z| < \frac{1}{|\alpha|}$ the series converges by the ratio test. If instead $|z| > \frac{1}{\alpha}$ then it diverges. Finally, if $|z| = |\alpha|$, then the series converges for $|\alpha| < 1$ and diverges otherwise. The radius of converge of the given series is thus $R = \frac{1}{|\alpha|}$.
>   (ii) Let's find the radius of convergence of
>
> $$\sum_{n=0}^{\infty} (n + 2^{-n})(z-1)^n \tag{55.2.5}$$
>
> We find that
>
> $$\left| \frac{z_{n+1}}{z_n} \right| = \left| \frac{n+1+2^{-n-1}}{(z-1)^{n+1}} \frac{(z-1)^n}{n+2^{-n}} \right| = |z-1| \left| \frac{1+1/n+2^{-n-1}/n}{1+2^{-n}/n} \right| \to |z-1| \tag{55.2.6}$$
>
> so it is clear that z converges for $|z - 1| < 1$.
>
> ◀

Proposition *(Differentiating and integrating power series)*

A power series and its derivative both have the same radius of convergence. Furthermore if the power series converges in some disc so that

$$f(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n, \ z \in D_R(\alpha) = \{z : |z - \alpha| < R\} \tag{55.2.7}$$

then f is analytic on $D_R(\alpha)$ and its derivative is

$$f'(z) = \sum_{n=1}^{\infty} n a_n(z - \alpha)^{n-1}, \ z \in D_R(\alpha) = \{z : |z - \alpha| < R\} \tag{55.2.8}$$

Similarly, if (55.2.7) holds then

$$F(z) = b_0 + \sum_{n=0}^{\infty} \frac{a_n}{n+1}(z - \alpha)^{n+1}, \ b_0 \in \mathbb{C} \tag{55.2.9}$$

is a primitive of f on $D_R(\alpha)$.

Example. Consider the following series

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z}, \ \forall z \in \{z \in \mathbb{C} : |z| \leq 1\} \tag{55.2.10}$$

Differentiating term by term with respect to z we find that

$$\sum_{n=1}^{\infty} n z^{n-1} = \frac{1}{(1-z)^2}, \ z \in \{z \in \mathbb{C} : |z| \leq 1\} \tag{55.2.11}$$

Differentiating once more we find that

$$\sum_{n=2}^{\infty} n(n-1)z^{n-2} = \sum_{n=0}^{\infty} (n+1)(n+2)z^n = \frac{2}{(1-z)^3}, \ z \in \{z \in \mathbb{C} : |z| \leq 1\} \tag{55.2.12}$$

Finally, integrating the original series term by term with respect to z yields

$$b_0 + \sum_{n=1}^{\infty} \frac{z^n}{n} = -\log(1-z), \ z \in \{z \in \mathbb{C} : |z| \leq 1\} \tag{55.2.13}$$

Setting $z = 0$ it is clear that $b_0 = 0$. ◀

55.3 Taylor's theorem

Theorem *(Taylor's theorem)*

Let f be analytic on the open disc $D_r(\alpha)$. Then

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\alpha)}{n!}(z - \alpha)^n \qquad (55.3.1)$$

is a unique, convergent power series representation of f on $D_r(\alpha)$.

Note that sometimes it is easier to use Cauchy's nth derivative formula to find the Taylor coefficients

$$\frac{f^{(n)}(\alpha)}{n!} = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z - \alpha)^{n+1}} \, dz, \ C \subset D_r(\alpha) \qquad (55.3.2)$$

Example. Let's find the taylor series about $z = 0$ for $f(z) = \cosh z$. We find that

$$f^{(2n-1)}(z) = \sinh z \implies f^{(2n-1)}(0) = 0 \qquad (55.3.3)$$
$$f^{(2n)}(z) = \cosh z \implies f^{(2n)}(0) = 1 \qquad (55.3.4)$$

for $n \in \mathbb{N}$ and $z \in \mathbb{C}$. Consequently

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} \qquad (55.3.5)$$

By a similar argument

$$\sinh z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n + 1)!} \qquad (55.3.6)$$

Let's find the Taylor series about $z = 0$ for $f(z) = \frac{1}{(1+z)^3}$. This function is analytic on $\mathbb{C} \setminus \{1\}$ so we can apply Taylor's theorem to find a convergent power series representation on $\mathbb{D}_1(0)$. We find that

$$f'(z) = -\frac{3}{(1 + z)^4} \implies f'(0) = -3 \qquad (55.3.7)$$

$$f''(z) = \frac{12}{(1 + z)^5} \implies f''(0) = 12 \qquad (55.3.8)$$

$$\vdots \qquad (55.3.9)$$

$$f^{(n)}(z) = \frac{(n + 2)!}{2(1 + z)^n} \implies f^{(n)}(0) = \frac{(n + 2)!}{2} \qquad (55.3.10)$$

which can easily be proven by induction. Consequently

$$\frac{1}{(1 + z)^3} = \sum_{n=0}^{\infty} \frac{(n + 2)!}{2} z^n, \ |z| < 1 \qquad (55.3.11)$$

◀

An immediate application of Taylor's theorem to $f(z) = (1 + z)^\alpha$ yields the following famous result.

Proposition *(Binomial series)*

For $\alpha \in \mathbb{C}$ the binomial series is equal to

$$(1+z)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} z^n, \ |z| < 1 \tag{55.3.12}$$

Theorem *(Combining power series)*

Define the following power series

$$f(z) = \sum_{n=0}^{\infty} a_n(z-\alpha)^n, \ |z-\alpha| < R \tag{55.3.13}$$

$$g(z) = \sum_{n=0}^{\infty} b_n(z-\alpha)^n, \ |z-\alpha| < R' \tag{55.3.14}$$

$$\tag{55.3.15}$$

Then

(i) **Sum**: if $r = \min\{R, R'\}$ then

$$(f+g)(z) = \sum_{n=0}^{\infty} (a_n + b_n)(z-\alpha)^n, \ z \in D_r(\alpha) \tag{55.3.16}$$

(ii) **Multiple**: for $\lambda \in \mathbb{C}$

$$(\lambda f)(z) = \sum_{n=0}^{\infty} \lambda a_n(z-\alpha)^n \tag{55.3.17}$$

(iii) **Product**: if $R = \min\{R, R'\}$ then

$$(fg)(z) = \sum_{n=0}^{\infty} c_n(z-\alpha)^n, \ |z-\alpha| < r \tag{55.3.18}$$

where we defined

$$c_n = \sum_{m=0}^{n} a_m b_{n-m} \tag{55.3.19}$$

(iv) **Substitution**: let $w = \lambda z^k$ with $\lambda \neq 0$ and $k \in \mathbb{N}$. Then the corresponding power series $f(w)$ has radius of convergence $(R \cdot |\lambda|)^k$.

Theorem *(Composition rule)*

Define

$$f(z) = \sum_{n=0}^{\infty} a_n(z-\alpha)^n, \ |z-\alpha| < R \tag{55.3.20}$$

$$g(w) = \sum_{n=0}^{\infty} b_n(w-\beta)^n, \ |w-\beta| < R' \tag{55.3.21}$$

$$\tag{55.3.22}$$

Suppose that $\beta = f(\alpha)$, then

$$g(f(z)) = \sum_{n=0}^{\infty} c_n(z - \alpha)^n, \ |z - \alpha| < r \tag{55.3.23}$$

for some $r > 0$ and c_n is the coefficient of

$$c_n = \sum_{k=0}^{n} b_k \left(\sum_{l=1}^{n} a_l(z - \alpha)^l \right)^k \tag{55.3.24}$$

Example.

(i) Let's find the Taylor series for $h(z) = e^z \sin z$ to third order in z. Using the product rule we find that

$$e^z \sin z = \left(z - \frac{z^3}{3!} \right)\left(1 + z + \frac{z^2}{2!} \right) \tag{55.3.25}$$

$$= z + z^2 + \frac{z^3}{2!} - \frac{z^3}{3!} + o(z^4) \tag{55.3.26}$$

$$= z + z^2 + \frac{z^3}{3} + o(z^4) \tag{55.3.27}$$

(ii) Let's find the Taylor series for $h(z) = e^{\sin z}$ about 0, up to fifth order in z. We know that

$$f(z) = \sin z = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{2n+1} = z - \frac{z^3}{3!} + \frac{z^5}{5!}, \ z \in \mathbb{C} \tag{55.3.28}$$

$$g(z) = e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!}, \ z \in \mathbb{C} \tag{55.3.29}$$

so using the composition rule

$$e^{\sin z} = 1 + \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} \right) + \frac{1}{2!}\left(z - \frac{z^3}{3!} \right)^2 + \frac{1}{3!}\left(z - \frac{z^3}{3!} \right)^3 \tag{55.3.30}$$

$$+ \frac{1}{4!}\left(z - \frac{z^3}{3!} \right)^4 + \frac{1}{5!}\left(z - \frac{z^3}{3!} \right)^5 + o(z^6)$$

$$= 1 + z - \frac{z^3}{3!} + \frac{z^5}{5!} + \frac{1}{2}\left(z^2 - \frac{z^4}{3} \right) + \frac{1}{3!}\left(z^3 - \frac{3z^5}{3!} \right) \tag{55.3.31}$$

$$+ \frac{1}{4!}z^4 + \frac{1}{5!}z^5 + o(z^6)$$

$$= 1 + z + \frac{1}{2}z^2 - \frac{1}{8}z^4 - \frac{z^5}{15}, \ |z| < r \tag{55.3.32}$$

for some $r > 0$.

(iii) Let's find the Taylor series for $h(z) = \log(1 + z)$ about $z = 2$ to fourth order in z. Knowing that the Taylor series for the same function about $z = 0$ is

$$\log(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + o(z^5) \tag{55.3.33}$$

we need to find a suitable transformation of z that will give the Taylor series about $z = 2$. Note that

$$\log(1 + z) = \log(3 + (z - 2)) = \log(3(1 + (z - 2)/3)) \qquad (55.3.34)$$

and

$$\log(1 + (z - 2)/3) = (z - 2) - \frac{(z - 2)^2}{3^2 \cdot 2} + \frac{(z - 2)^3}{3^3 \cdot 3} - \frac{(z - 2)^4}{3^4 \cdot 4} + o(z^5), \ |z - 2| < 3$$
$$(55.3.35)$$

Therefore

$$\log(1 + z) = \log(3) + \frac{z - 2}{3} - \frac{(z - 2)^2}{3^2 \cdot 2} + \frac{(z - 2)^3}{3^3 \cdot 3} - \frac{(z - 2)^4}{3^4 \cdot 4} + o(z^5), \ |z - 2| < 3$$
$$(55.3.36)$$

(iv) Let's find the Taylor series for $h(z) = \sin^{-1}(z)$ up to seventh order. Note that h is analytic on $\mathbb{C} - \{x \in \mathbb{R} : |x| \geq 1\}$ and

$$\frac{d}{dz}(\sin^{-1} z) = \frac{1}{\sqrt{1 - z^2}}, \ z \in \mathbb{C} - \{x \in \mathbb{R} : |x| \geq 1\} \qquad (55.3.37)$$

So if we find the Taylor series for $\frac{1}{\sqrt{1-z^2}}$ then we can integrate term by term to find the desired Taylor series. The binomial coefficients for $\alpha = -\frac{1}{2}$ are

$$\binom{-\frac{1}{2}}{n} = \frac{(-1/2) \cdot (-3/2) \cdot \ldots \cdot (-n + 1/2)}{n!} \qquad (55.3.38)$$

so that for $|z| < 1$

$$(1 - z^2)^{-1/2} = 1 + \frac{1}{2}z^2 + \frac{3}{8}z^4 + \frac{5}{16}z^6 + o(z^8) \qquad (55.3.39)$$

The integral rule then yields

$$\sin^{-1} z = z + \frac{1}{6}z^3 + \frac{3}{40}z^5 + \frac{5}{112}z^7 + o(z^9), \ |z| < 1 \qquad (55.3.40)$$

◀

Proposition *(Radius of convergence and boundedness)*
Let f be an analytic and unbounded function on $D_R(\alpha)$. Then D is the disc of convergence of the Taylor series about α for f.

55.4   Uniqueness theorem

In this section we will apply the formalism of Taylor series to prove an important result on when two functions are identical on a region.

> **Definition** *(kth order zero)*
> Let f be analytic at α. Then we say that f has a zero of kth order at α if
> $$f^{(n)}(\alpha) = 0 \ \text{ for } n = 0, 1, ..., k - 1 \qquad (55.4.1)$$

Often, when a function is difficult to differentiate it may be hard to apply this definition to find the order of a zero. However, intuitively it is easy to see that if the term $(z - \alpha)^k$ containing the zero α is of order k then that zero should be of the same order.

> **Theorem** *(Zeros of a function)*
> Let f be analytic at α. Then α is a zero of order k *iff.*
> $$f(z) = (z - \alpha)^k g(z), \ |z - \alpha| < r \qquad (55.4.2)$$
> for some $r > 0$ and function g analytic and non-vanishing at α.

Proof. Since f is analytic at α and has a kth order zero at α, by Taylor's theorem we can express it as a Taylor series

$$f(z) = \sum_{n=k}^{\infty} \frac{f^{(n)}(\alpha)}{n!}(z - \alpha)^n = (z - \alpha)^k g(z), \ |z - \alpha| < r \qquad (55.4.3)$$

for some $r > 0$. Note that g is analytic at α by the differentiation rule.

Conversely, now suppose that

$$f(z) = (z - \alpha)^k g(z), \ |z - \alpha| < r \qquad (55.4.4)$$

where g is non-zero and analytic at α. Again we can expand it as a Taylor series of the form

$$g(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n, \ a_0 \neq 0, \ |z - \alpha| < r \qquad (55.4.5)$$

It follows that

$$f(z) = \sum_{n=k}^{\infty} \frac{f^{(n)}(\alpha)}{n!}(z - \alpha)^n, \ |z - \alpha| < r \qquad (55.4.6)$$

implying that f has a kth order zero at α. ∎

> **Example.**
> (i) Let's find the order of the zero $z = 0$ of $f(z) = z^2(\cos z - 1)$. Note that f is analytic at $z = 0$, and expanding $\cos z - 1$ as a Taylor series:
> $$\cos z - 1 = \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2n)!} z^{2n} - 1 = z^2 \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2(n+1))!} z^{2n} \qquad (55.4.7)$$

We thus find that

$$f(z) = z^4 \left(\frac{1}{2} - \frac{z^2}{24} + \dots \right) \tag{55.4.8}$$

so $z = 0$ is a fourth order zero.

(ii) Let's locate the zeros of $f(z) = z \sin z$ and their orders. We expect $\{z \in \mathbb{C} : z = k\pi, \ k \in \mathbb{Z}\}$ to be the set of zeros. The Taylor series of $\sin z$ about 0 is

$$\sin z = \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2n+1)!} z^{2n+1} = z \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2n+1)!} z^{2n} \tag{55.4.9}$$

Also, note that

$$\sin(z - k\pi) = (-1)^k \sin z \tag{55.4.10}$$

so the Taylor series of $z \sin z$ about $k\pi$ is

$$z \sin z = (-1)^k z (z - k\pi) \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2n+1)!} (z - k\pi)^{2n}, \ z \in \mathbb{C} \tag{55.4.11}$$

It is clear that if $k \neq 0$ then $z = k\pi$ is a first order zero. On the other hand, for the zero at $z = 0$ we find

$$z \sin z = (-1)^k z^2 \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{(2n+1)!} z^{2n}, \ z \in \mathbb{C} \tag{55.4.12}$$

so it is a second order zero.

◀

Theorem *(Finiteness of zeros)*
If f is analytic and not identically zero on a region $\mathcal{R}$ then any zero of f in $\mathcal{R}$ is of finite order.

It remains to be understood exactly when a zero is not of finite order. The following definition will help us.

Definition *(Isolated zero)*
A zero of a function is **isolated** if there is a disc centered at it that contains no other zeros. A zero of finite order is isolated.

It follows that a zero cannot be isolated and thus of finite order if it is the limit point of a set of zeros. But this implies that the function must be identically zero. Therefore

Proposition *(Zero functions)*
Let f be analytic on $\mathcal{R}$ containing a set S of zeros of f and their limit point. Then f is vanishes on $\mathcal{R}$.

An immediate consequence is the following uniqueness theorem.

> **Theorem** *(Uniqueness theorems)*
> Let f, g be analytic on $\mathcal{R}$ and suppose that $f(z) = g(z)$ for all $z \in S \subseteq \mathcal{R}$. Assume also that S has a limit point in $\mathcal{R}$. Then $f = g$ throughout all of $\mathcal{R}$.

Proof. Use the previous result with $f - g$. ∎

> **Example.** Suppose f is entire and
>
> $$f(i/n) = -\frac{1}{n^2} \quad \text{for } n = 1, 2, \dots \tag{55.4.13}$$
>
> Then let's prove that $f(z) = z^2$.
> Let $S = \{i/n : n \in \mathbb{N}\}$ be a subset of $\mathbb{C}$ with limit point 0 (geometrically we have a set of points on the imaginary axis approaching the origin). Since $g(z) = z^2$ agrees with $f(z)$ for all $z \in S$ it follows that $f(z) = g(z)$ on $\mathbb{C}$. ◀

55.5 Singularities

> **Definition** *(Isolated singularity)*
> An **isolated singularity** of a function f is a point α such that f is analytic on $\{z : 0 < |z - \alpha| < r\}$ for some $r > 0$, but not at α.

> **Example.**
> (i) Consider the function
>
> $$f(z) = \frac{3z - i}{z^3} \sin\left(\frac{1}{z+1}\right) \tag{55.5.1}$$
>
> This function is analytic everywhere except at $z = 0, -1$ which are singularities.
> (ii) Consider the function
>
> $$g(z) = \frac{1}{\sin(1/z)} \tag{55.5.2}$$
>
> This function is analytic everywhere except when $\sin(1/z) = 0$ or $z = 0$, implying that $z = \frac{1}{n\pi}$ for $n \in \mathbb{Z}$ or $z = 0$. The former is clearly a singularity as one can always find a due to the finite distance between each of these points. It follows that $z = 0$ is not a singularity since there is no punctured neighborhood of 0 around which $\frac{1}{\sin(1/z)}$ is analytic (since the sequence of $\pm\frac{1}{m\pi}$ converges to 0, this follows from the epsilon-delta definition).
>
> ◀

> **Definition** *(Function tending to ∞)*
> Let f be defined on A and suppose α is a limit point of A. We say that $f \to \infty$ as $z \to \alpha$ if for any sequence (z_n) in $A \setminus \{\alpha\}$ converging to α, $f(z_n) \to \infty$. Equivalently, if for each $M > 0$

there is a $\delta > 0$ such that

$$|f(z)| > M, \ \forall z \in A, \ 0 < |z - \alpha| < \delta \tag{55.5.3}$$

Using the reciprocal rule for sequences one can easily prove that

Proposition *(Reciprocal rule for functions)*
Let f have a domain A with a limit point α. Then

$$f(z) \to \infty \text{ as } z \to \infty \iff \lim_{z \to \alpha} \frac{1}{f(z)} = 0 \tag{55.5.4}$$

Consider the function

$$f(z) = \frac{\sin z}{z} \tag{55.5.5}$$

which can easily be verified to be analytic on $\mathbb{C} \setminus \{0\}$ with a singularity at $z = 0$. The taylor series about 0 for $\sin z$ is

$$\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}, \ z \in \mathbb{C} \tag{55.5.6}$$

so it follows that

$$\frac{\sin z}{z} = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n+1)!}, \ z \in \mathbb{C} \setminus \{0\} \tag{55.5.7}$$

Note that while the LHS is ill-defined at $z = 0$, the RHS does not suffer from this pathology, so we may as well define a new function corresponding to the power series found above

$$g(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n+1)!}, \ z \in \mathbb{C} \tag{55.5.8}$$

with

$$\lim_{z \to 0} f(z) = g(0) = 1 \tag{55.5.9}$$

By defining the power series as a new function on its convergence integral we removed the singularity of f at 0.

Definition *(Removable singularity)*
Let f have a singularity at α. Then α is a **removable singularity** if there exists a function g, known as the **analytic extension** of f, that is analytic on $\{z : |z - \alpha| < r\}$ and such that

$$f(z) = g(z), \ 0 < |z - \alpha| < r \tag{55.5.10}$$

or equivalently

$$\lim_{z \to \alpha} f(z) = g(\alpha) \neq 0 \tag{55.5.11}$$

This statement effectively says that $f(z)$ behaves like $\frac{g(\alpha)}{z - \alpha}$ as $z \to \alpha$.

Example. Consider the function

$$f(z) = 3z \cot z = \frac{3z \cos z}{\sin z} \tag{55.5.12}$$

This function is not analytic when the denominator vanishes, that is when $z = m\pi$ for $m \in \mathbb{Z}$. These are all singularities since the discs $\{z : 0 < |z - m\pi| < \pi\}$ do not overlap, and thus form a sequence of punctured discs about each $z = m\pi$ over which f is analytic. For $z = 0$, note that

$$\lim_{z \to 0} \frac{3z \cos z}{\sin z} = 3 \tag{55.5.13}$$

so we can analytically extend f by taking

$$g(z) = \frac{3(1 - z^2/2 + x^4/24...)}{(1 - z^2/6 + z^4/120 - ...)}, \ 0 < |z| < \frac{\pi}{2} \tag{55.5.14}$$

The denominator is an entire function with $h(0) = 1$ so by continuity there is some $r > 0$ for which h is non-zero on $|z| < r$. Consequently $g(z)$ is analytic on $|z| < r$ and satisfies

$$f(z) = g(z), \ |z| < r \tag{55.5.15}$$

implying that $z = 0$ is a removable singularity.
For $z = m\pi \neq 0$, we find that

$$\lim_{z \to m\pi} \frac{3z \cos z}{\sin z} \to \infty \tag{55.5.16}$$

so these are non-removable singularities, since any attempt to analytically extend f will still be ill-defined at $z = m\pi$. ◄

Definition *(Simple pole)*
Let f be singular at α. We say that α is a **simple pole** if there is a function g analytic on $\{z : |z - \alpha| < r\}$ such that $g(\alpha) \neq 0$ and such that

$$f(z) = \frac{g(z)}{z - \alpha}, \ 0 < |z - \alpha| < r \tag{55.5.17}$$

It follows that

$$\lim_{z \to \alpha}(z - \alpha)f(z) = g(\alpha) \tag{55.5.18}$$

Example. Consider the function

$$f(z) = \frac{z}{\sin z} \tag{55.5.19}$$

It is easy to verify that this has a pole at $z = k\pi$ for $k \in \mathbb{Z}$. We know that $z = 0$ is a removable singularity. Using the fact that

$$\sin(z - k\pi) = \sin z \cos k\pi - \cos z \sin k\pi = (-1)^k \sin z \tag{55.5.20}$$

we can rewrite the function it a more accessible way as

$$f(z) = \frac{(-1)^k z}{\sin(z - k\pi)}, \ k \in \mathbb{Z} \tag{55.5.21}$$

Now note that

$$\sin(z - k\pi) = (z - k\pi) \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} (z - k\pi)^{2n}, \ z \in \mathbb{C} \tag{55.5.22}$$

so we can define the following function

$$g(z) = \frac{(-1)^k z}{h(z)}, \ h(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} (z - k\pi)^{2n} \tag{55.5.23}$$

Since $h(k\pi) = 0$ it follows that $h(z) \neq 0$ and thus $g(z)$ is analytic on $|z - k\pi| < \pi$. Consequently we find that

$$f(z) = \frac{g(z)}{z - k\pi}, \ 0 < |z - k\pi| < \pi \tag{55.5.24}$$

implying that $k\pi$ is a simple pole for $k \neq 0$. ◄

We can extend the definition of a simple pole as follows.

Definition *(kth order pole)*

Let f be singular at α. We say that α is a k**th order pole** if there is a function g analytic on $\{z : |z - \alpha| < r\}$ such that $g(\alpha) \neq 0$ and such that

$$f(z) = \frac{g(z)}{(z - \alpha)^k}, \ 0 < |z - \alpha| < r \tag{55.5.25}$$

It follows that

$$\lim_{z \to \alpha} (z - \alpha)^k f(z) = g(\alpha) \tag{55.5.26}$$

By this definition, a zeroth order pole is a removable singularity.

Example. Consider the function

$$f(z) = \frac{z + 2}{z^4 (z^2 - 4)^3}, \tag{55.5.27}$$

We can write this as

$$f(z) = \frac{1}{z^4 (z - 2)^3 (z + 2)^2} \tag{55.5.28}$$

This function has singularities at $z = 0, \pm 2$. Defining the following analytic function

$$g(z) = \frac{1}{(z - 2)^3 (z + 2)^2}, \ 0 < |z| < 2 \implies g(0) \neq 0 \tag{55.5.29}$$

we see that

$$f(z) = \frac{g(z)}{z^4}, \ 0 < |z| < 2 \tag{55.5.30}$$

Hence $z = 0$ is a fourth order pole. Similarly let's now define

$$g(z) = \frac{1}{z^4(z+2)^2}, \ 0 < |z - 2| < 2 \implies g(2) \neq 0 \tag{55.5.31}$$

then

$$f(z) = \frac{g(z)}{(z-2)^3}, \ 0 < |z - 2| < 2 \tag{55.5.32}$$

so $z = 2$ is a third order pole. Finally, let's define

$$g(z) = \frac{1}{z^4(z-2)^3}, \ 0 < |z + 2| < 2 \implies g(-2) \neq 0 \tag{55.5.33}$$

then

$$f(z) = \frac{g(z)}{(z+2)^2}, \ 0 < |z + 2| < 2 \tag{55.5.34}$$

so $z = -2$ is a second order pole. ◀

Finally, if the singularity is neither removable nor a pole then we define it as an essential singularity.

Definition *(Essential singularity)*
Let f be singular at α, then α is an **essential singularity** if it is neither removable nor a pole. Therefore if $f(z)$ does not tend to a finite limit or to ∞ as z approaches α then α is an essential singularity.

Example. Consider the function $f(z) = e^{1/z}$. It is evident that $z = 0$ is an essential singularity, for example by considering $z_n = \frac{1}{n\pi i}$ which is null, and thus as $z_n \to$ then

$$e^{1/z_n} = e^{n\pi i} = (-1)^n \tag{55.5.35}$$

which does not tend to a finite limit or to ∞. Consequently neither does $f(z)$ thus proving that $z = 0$ is an essential singularity. ◀

55.6 Laurent's theorem

Definition *(Extended power series)*
Let $z \in \mathbb{C}$, then an **Extended power series about** α is a power series of the form

$$\sum_{n=-\infty}^{\infty} a_n(z - \alpha)^n = \ldots + \frac{a_{-2}}{(z-\alpha)^2} + \frac{a_{-1}}{z-\alpha} + a_0 + a_1(z-\alpha) + \ldots \tag{55.6.1}$$

The extended power series for a given $z, a_n, \alpha \in \mathbb{C}, n \in \mathbb{Z}$ converges if both the analytic part

$$\sum_{n=0}^{\infty} a_n(z - \alpha)^n = a_0 + a_1(z-\alpha) + \ldots \tag{55.6.2}$$

and singular part

$$\sum_{n=0}^{\infty} a_{-n}(z-\alpha)^{-n} = \frac{a_{-1}}{z-\alpha} + \frac{a_{-2}}{(z-\alpha)^2} + \dots \tag{55.6.3}$$

For example, the function $f(z) = \frac{\sin z}{z^3}$ has analytic part

$$-\frac{1}{3!} + \frac{z^2}{5!} - \frac{z^4}{7!}\dots \tag{55.6.4}$$

and singular part $\frac{1}{z^2}$. We know that the analytic part converges on its disc of convergence $D_s(\alpha)$. On the other hand, the singular part can be viewed as a normal power series with $(z-\alpha)^{-1}$ as its argument, meaning that it converges when

$$|z-\alpha|^{-1} = |(z-\alpha)^{-1}| < s' \implies |z-\alpha| > r \tag{55.6.5}$$

where $r = \frac{1}{s'}$ for some s'. Consequently, the extended power series will converge on the following **annulus of convergence**

$$A = \{z : r < |z-\alpha| < s\} \tag{55.6.6}$$

which can either be

(i) an open annulus

(ii) a punctured open disc

(iii) a punctured plane

(iv) the outside of a disc

(v) the empty set

Example. Consider the following extended power series

$$\dots + \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 1 + z + z^2 + z^3 + \dots \tag{55.6.7}$$

We see that the singular part

$$\frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \dots \tag{55.6.8}$$

converges when $|1/z| < 1 \implies |z| > 1$. Similarly, the analytic part

$$1 + z + z^2 + z^3 + \dots \tag{55.6.9}$$

converges when $|z| < 1$. Therefore, the annulus of convergence of the extended power series must be the empty set, $A = \emptyset$. ◀

We know that ordinary power series can be used to define analytic functions on their disc of convergence. Moreover, Taylor's theorem tells us that any analytic function on a disc D can be written as an ordinary power series. Since extended power series can be used to define analytic functions on their annulus of convergence, one should expect that any analytic function on an annulus A can be expressed as an extended power series. This is known as **Laurent's theorem**.

> **Theorem** *(Laurent's theorem)* Let f be analytic on the open annulus $A = \{z : r < |z - \alpha| < s\}$ for $0 \leq r < s \leq \infty$. Then $f(z)$ defines a unique, extended power series, known as a **Laurent series**:
>
> $$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - \alpha)^n, \; z \in A \qquad (55.6.10)$$
>
> where
>
> $$a_n = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z - \alpha)^{n+1}} \, dz, \; n \in \mathbb{Z} \qquad (55.6.11)$$
>
> and where $C \in A$ is a circle centered at α.

It follows immediately from the uniqueness of the Laurent power series that if f is analytic on the disc $D_r(\alpha)$ then the Taylor series will be the same as the Laurent series about α for f on the punctured disc $D_r(\alpha) \setminus \{0\}$.

> **Theorem** *(Singularities using Laurent series)*
> Let f be an analytic function with a singularity at α with a Laurent series about this point given by
>
> $$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - \alpha)^n \qquad (55.6.12)$$
>
> We can classify α as
>   (i) a removable singularity *iff* $a_n = 0$, $\forall n < 0$.
> (ii) a kth order pole *iff* $a_{-k} \neq 0$ and $a_n = 0$, $\forall n < -k$.
> (iii) an essential singularity *iff* $a_n \neq 0$ for infinitely many, but not necessarily all n.

> **Example.** Consider the function $f(z) = z \sin \frac{1}{z^2}$. This function is analytic on $\mathbb{C} \setminus \{0\}$, so its Laurent series will coincide with the Taylor series of this function, which is given by
>
> $$z \sin \frac{1}{z} = z \left(\frac{1}{z} - \frac{1}{3!z^3} + \frac{1}{5!z^5} - \dots \right), \; z \neq 0 \qquad (55.6.13)$$
>
> $$= 1 - \frac{1}{3!z^2} + \frac{1}{5!z^4} - \dots, \; z \neq 0 \qquad (55.6.14)$$
>
> Consequently, since there are an infinite number of non-zero a_{-k} coefficients for $k > 0$, $z = 0$ must be an essential singularity. ◀

Often, it is quite difficult to compute the Laurent series coefficients using the Cauchy integral formula

$$a_n = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z - \alpha)^{n+1}} \, dz, \; n \in \mathbb{Z} \qquad (55.6.15)$$

For partial fractions, one can use a different method to compute the Laurent series, which we show below.

> **Example.** Consider the function
>
> $$f(z) = \frac{4}{(z - 1)(z + 3)} \qquad (55.6.16)$$

Let's evaluate its Laurent series on $A = \{z : |z| < 1\}$ and $B = \{z : 1 < |z| < 3\}$. Using the method of partial fractions, we find that

$$f(z) = \frac{4}{(z-1)(z+3)} = \underbrace{\frac{1}{z-1}}_{h(z)} - \underbrace{\frac{1}{z+3}}_{g(z)} \tag{55.6.17}$$

Now note that $g(z)$ is analytic on $G_1 = \{z : |z| < 3\}$ and $G_2 = \{z : |z| > 3\}$, while $h(z)$ is analytic on $H_1 = \{z : |z| < 1\}$ and $H_2 = \{z : |z| > 1\}$.

If we want to find the Laurent series on A, then since $A \subseteq G_1$ and $A \subseteq H_1$, it suffices to find the Laurent series for g on G_1 and h on H_1, and taking their difference. For $g(z)$, since $|z| < 3 \implies |z/3| < 1$ we find that

$$\frac{1}{z+3} = \frac{1}{3}\frac{1}{1+z/3} = \frac{1}{3}\left(1 - \frac{z}{3} + \frac{z^2}{9} - \frac{z^3}{27} + o(z^4)\right) \tag{55.6.18}$$

$$= \frac{1}{3} - \frac{z}{9} + \frac{z^2}{27} - \frac{z^3}{81} + o(z^4), \; z \in G_1 \tag{55.6.19}$$

Similarly for $h(z)$, since $|z| < 1$ we find that

$$\frac{1}{z-1} = -1 - z - z^2 - z^3 + o(z^4), \; z \in H_1 \tag{55.6.20}$$

Consequently

$$f(z) = (-1 - z - z^2 - z^3) - \left(\frac{1}{3} - \frac{z}{9} + \frac{z^2}{27} - \frac{z^3}{81}\right) + o(z^4) \tag{55.6.21}$$

$$= -\frac{4}{3} - \frac{8}{9}z - \frac{28}{27}z^2 - \frac{80}{81}z^3 + o(z^4), \; z \in A \tag{55.6.22}$$

To find the Laurent series on B, note that $B \in G_1$ and $B \in H_2$ so it suffices to find that Laurent series for g on G_1 (which we have already calculated) and h on H_2, and take their difference. For $h(z)$, since $|z| > 1 \implies |1/z| < 1$ it follows that

$$\frac{1}{z-1} = \frac{1}{z}\frac{1}{1-1/z} = \frac{1}{z}\left(1 + \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + o(z^{-4})\right) \tag{55.6.23}$$

$$= \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \frac{1}{z^4} + o(z^{-5}) \tag{55.6.24}$$

Thus the Laurent series for f on B is

$$f(z) = \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \frac{1}{z^4} + o(z^{-5}) - \frac{1}{3} + \frac{z}{9} - \frac{z^2}{27} + \frac{z^3}{81} + o(z^4) + o(z^{-5}), \; z \in B \tag{55.6.25}$$

◀

This method works well if the Laurent series is centered at $z = 0$. If this is not the case, the simplest approach is to make a substitution so that the Laurent series about 0 may be used.

Example.

(i) Let's find the Laurent series for

$$f(z) = \frac{\cos(z - 2)}{(z - 2)^2} \tag{55.6.26}$$

about $z = 2$. Note that $f(z)$ is analytic on $\mathbb{C} \setminus \{2\}$ so the function's Taylor series can be used to compute the Laurent series. Let $w = z - 2 \iff z = 2 + w$. Then using the Laurent series for $f(w)$ about $w = 0$ we see that

$$f(z) = \frac{\cos w}{w^2} = \frac{1}{w^2}\left(1 - \frac{w^2}{2!} + \frac{w^4}{4!} - \frac{w^6}{6!}\right) + o(w^8) \tag{55.6.27}$$

$$= \frac{1}{(z-2)^2} - \frac{1}{2!} + \frac{(z-2)^2}{4!} - \frac{(z-2)^4}{6!} + o((z-2)^8), \ z \neq 2 \tag{55.6.28}$$

(ii) Let's find the Laurent series for

$$f(z) = \frac{4}{(z-1)(z+3)} \tag{55.6.29}$$

about $z = -3$ on the annulus $A = \{z : 0 < |z - 3| < 4\}$. Letting $w = z + 3$ then this is equivalent to finding the Laurent series for

$$f(w) = \frac{4}{w(w-4)} = \underbrace{\frac{1}{w-4}}_{g(z)} - \underbrace{\frac{1}{w}}_{h(z)} \tag{55.6.30}$$

Note that $g(z)$ is defined on $G_1 = \{w : |w| < 4\}$ and $G_2 = \{w : |w| > 4\}$ while $h(z)$ is defined on $H = \{w : |w| > 0\}$. Since $A = \{w : 0 < |w| < 4\} \subseteq G_1$ and $A = \{w : 0 < |w| < 4\} \subseteq H$ it suffices to evaluate the Laurent series about $w = 0$ for g on G_1 and h on H, and take their difference. For g, since $|w/4| < 1$ we find that

$$\frac{1}{w-4} = -\frac{1}{4}\frac{1}{1 - w/4} = -\frac{1}{4}\left(1 + \frac{w}{4} + \frac{w^2}{4^2} + \frac{w^3}{4^3} + o(w^4)\right) \tag{55.6.31}$$

$$= -\frac{1}{4} - \frac{w}{4^2} - \frac{w^2}{4^3} - \frac{w^3}{4^4} + o(w^4) \tag{55.6.32}$$

Consequently

$$f(w) = -\frac{1}{w} - \frac{1}{4} - \frac{w}{4^2} - \frac{w^2}{4^3} - \frac{w^3}{4^4} + o(w^4), \ 0 < |w| < 4 \tag{55.6.33}$$

or equivalently

$$f(z) = -\frac{1}{z+3} - \frac{1}{4} - \frac{z+3}{16} - \frac{(z+3)^2}{64} - \frac{(z+3)^3}{256} + o((z+3)^4) \tag{55.6.34}$$

for $0 < |z - 3| < 4$.

◀

55.7 Behaviour near a singularity

Theorem *(Equivalences for singularities)*

Let f have a singularity at α, then
  (i) α is removable
 (ii) $\lim_{z \to \alpha} f(z)$ exists
(iii) f is bounded on $\{z : 0 < |z - \alpha| < r\}$ for $r > 0$
 (iv) $\lim_{z \to \alpha}(z - \alpha)f(z) = 0$
are all equivalent.
Similarly for poles
  (i) α is a kth order pole, $k \in \mathbb{N}$
 (ii) $\lim_{z \to \alpha}(z - \alpha)^k f(z)$ exists and is non-zero
(iii) $\frac{1}{f}$ has a removable singularity at α which when removed gives rise to a kth order zero
        at α
are all equivalent. It follows that f has a pole at α *iff* $f(z) \to \infty$ as $z \to \alpha$.

Example.

(i) Consider the function $f(z) = \frac{3z}{\tan 3z}$ which has a removable singularity at $\alpha = 0$. Indeed

$$\lim_{z \to 0} \frac{3z^2}{\sin 3z} \cos 3z = \lim_{z \to 0} \frac{3z^2}{3z - (3z)^3/6 + \ldots} = 0 \tag{55.7.1}$$

(ii) Consider the function $f(z) = \frac{z}{\sin^3 z}$, which has a third order pole at $z = 0$. We find that

$$\lim_{z \to 0} z^k f(z) = \lim_{z \to 0} \frac{z^{k+1}}{\sin^3 z} = 1 \ \text{ iff } k = 2 \tag{55.7.2}$$

so $f(z)$ has a second order pole at $z = 0$. Similarly f also has pole at $z = k\pi$ for $k \neq 0$, and we find that

$$\lim_{z \to k\pi}(z - \pi)^k f(z) = \lim_{z \to k\pi} \frac{(-1)^{3k} z(z - \pi)^k}{\sin^3(z - k\pi)} = (-1)^k k\pi \ \text{ iff } k = 3 \tag{55.7.3}$$

so the singularity at $z = k\pi$ for $k \neq 0$ is a third order pole.

(iii) Consider $f(z) = g(z)/z$ where g is entire. If $g(0) \neq 0$ then $z = 0$ is just a first order pole, since

$$\lim_{z \to 0} z f(z) = \lim_{z \to 0} g(z) = g(0) \neq 0 \tag{55.7.4}$$

since g is entire and thus continuous at $z = 0$. If instead $g(0) = 0$ we have a removable singularity since

$$\lim_{z \to 0} z f(z) = \lim_{z \to 0} g(z) = g(0) = 0 \tag{55.7.5}$$

Finally, letting $f(z) = 1/g(z)$ where g is entire and has a pole of order 2 at $z = 0$, then it follows that there exists some $l \neq 0$ such that

$$\lim_{z \to 0} z^2 g(z) = l \tag{55.7.6}$$

Thus $f(z)$ has a removable singularity at $z = 0$ since

$$\lim_{z \to 0} z f(z) = \lim_{z \to 0} \frac{z^3}{z^2 g(z)} = \frac{0}{l} = 0 \tag{55.7.7}$$

◄

55.8 Evaluating integrals using Laurent series

Often times one can reverse Laurent's theorem and use the Laurent series for a function to evaluate a given contour integral. For example, if a_n is the Laurent coefficient of the $(z - \alpha)^n$ term in the Laurent series of $f(z)$ about α on the punctured disc $D = \{z : 0 < |z - \alpha| < r\}$. It follows that

$$\oint_C \frac{f(z)}{(z - \alpha)^{n+1}} \, dz = 2\pi i a_n \tag{55.8.1}$$

where C is any circle in D centered at α. From this we find that

$$\oint_C f(z) \, dz = 2\pi i a_{-1} \tag{55.8.2}$$

This formula is extremely general, it can reduce the contour integral of any function around a circle into a Laurent series problem. The coefficient a_{-1} is so important that it is given a name.

> **Definition** *(Residue)*
> Let f be analytic on a punctured disc centered at α. Then the **residue Res(f, α) of f at α** is the coefficient a_{-1} of $\frac{1}{z - \alpha}$ in the Laurent series of f about α.

Example.

(i) Let's evaluate

$$\oint_C w^4 \sinh \frac{1}{w} \, dw \tag{55.8.3}$$

where C is any circle centered at the origin. It is clear that to find the residue we must find the $\frac{1}{w^5}$ term in the Laurent series of $\sinh \frac{1}{w}$.

$$\sinh \frac{1}{w} = \frac{1}{w} + \frac{1}{2!} \frac{1}{w^2} + ... + \frac{1}{5!} \frac{1}{w^5} + ... \tag{55.8.4}$$

so it follows that the residue of $w^4 \sinh \frac{1}{w}$ at $w = 0$ is just 1. Consequently

$$\oint_C w^4 \sinh \frac{1}{w} \, dw = \frac{2\pi i}{5!} = \frac{\pi i}{60} \tag{55.8.5}$$

(ii) Let's evaluate

$$\int_C \frac{\sin 2z}{z^4} \, dz \tag{55.8.6}$$

where $C = \{z : |z| = 5\}$. This time we need to find the third order term in the Laurent

series of $\sin 2z$ about 0. This is clearly

$$\sin 2z = \dots - \frac{(2z)^3}{3!} + \dots \tag{55.8.7}$$

implying that

$$\int_C \frac{\sin 2z}{z^4}\, dz = 2\pi i \cdot \left(-\frac{4}{3} \right) = -\frac{8\pi i}{3} \tag{55.8.8}$$

◀

CA6 Residues

56.1 Calculating residues

We saw in the last section that the residues of the Laurent series are quite important for evaluating contour integrals. There are a number of ways it can be computed, one way is to just calculate the Laurent series and obtain the coefficient of z^{-1}.

> **Example.** Let's find the residue of
>
> $$f(z) = \frac{ze^{iz}}{(z-\pi)^2} = \frac{e^{iz}}{z-\pi} + \frac{\pi e^{iz}}{(z-\pi)^2} \tag{56.1.1}$$
>
> about $z = \pi$. Note that the Taylor series for e^{iz} is
>
> $$e^{iz} = -e^{i(z-\pi)} = -1 - i(z-\pi) + \frac{(z-\pi)^2}{2} - \frac{i(z-\pi)^3}{6} - \dots \text{ for } z \in \mathbb{C} \tag{56.1.2}$$
>
> so we get that
>
> $$f(z) = -\frac{1}{z-\pi}\left(1 + i(z-\pi) - \frac{(z-\pi)^2}{2} + \dots\right) - \frac{\pi}{(z-\pi)^2}\left(1 + i(z-\pi) - \frac{(z-\pi)^2}{2} + \dots\right), \; z \neq \pi \tag{56.1.3}$$
>
> The residue $\operatorname{Res}(f, \pi)$ is given by the coefficient of $(z - \pi)^{-1}$, so
>
> $$\operatorname{Res}(f, \pi) = -1 - i\pi \tag{56.1.4}$$
>
> ◀

If the function is odd or even, then finding the residue at α will also quickly yield he residue at $-\alpha$ as the next proposition states.

> **Proposition** *(Residues of odd/even functions)*
> Let f be singular at α and $-\alpha$. If f is odd then $\operatorname{Res}(f, \alpha) = \operatorname{Res}(f, -\alpha)$ while if f is an even function then $\operatorname{Res}(f, -\alpha) = -\operatorname{Res}(f, \alpha)$.

Nevertheless, in many cases residues can be very challenging to find, for example finding the residue at the point $z = 2i$ of

$$f(z) = \frac{e^{iz}}{z(z-2i)^3(2z+3)} \tag{56.1.5}$$

would be a lengthy calculation by the methods of the previous chapter.

Suppose f is analytic on $\{z : 0 < |z - \alpha| < r\}$ with a simple pole at α. Then it follows that the Laurent series for $f(z)$ must take the form

$$f(z) = \frac{a_{-1}}{z - \alpha} + a_0 + a_1(z - \alpha) + \dots \ \text{ for } \ 0 < |z - \alpha| < r \qquad (56.1.6)$$

with $a_{-1} \neq 0$. Consequently

$$\lim_{z \to \alpha}(z - \alpha)f(z) = a_{-1} = \text{Res}(f, \alpha) \qquad (56.1.7)$$

n the case of simple poles the computation of residues is thus reduced to that of a limit. If instead we have a removable singularity then $a_{-1} = 0$ implying that

$$\text{Res}(f, \alpha) = 0 \qquad (56.1.8)$$

These results suggest that the following theorem holds.

Theorem *(Residues from limits)*
Let f have a singularity at α and assume $\lim_{z \to \alpha}(z - \alpha)f(z)$ exists, then

$$\text{Res}(f, \alpha) = \lim_{z \to \alpha}(z - \alpha)f(z) \qquad (56.1.9)$$

There are two useful results from this theorem that apply to rational functions.

Proposition *(Rational function residues)*
Let $f(z) = \frac{g(z)}{h(z)}$ where g and h are analytic at α, and where $h(\alpha) = 0$, $h'(\alpha) \neq 0$. Then

$$\text{Res}(f, \alpha) = \frac{g(\alpha)}{h'(\alpha)} \qquad (56.1.10)$$

If $h(z) = z - \alpha$ then

$$\text{Res}(f, \alpha) = g(\alpha) \qquad (56.1.11)$$

Example.

(i) Let's find the residue of

$$f(z) = \frac{1}{z^2(1 - z)(1 - 2z)(1 - 3z)} \qquad (56.1.12)$$

about $\alpha = \frac{1}{3}$. We find that

$$\text{Res}(f, 1/3) = \lim_{z \to 1/3}(z - 1/3)f(z) = -\frac{1}{3}\lim_{z \to 1/3}\frac{1}{z^2(1 - z)(1 - 2z)} = -\frac{27}{2} \qquad (56.1.13)$$

Imagine calculating this using Laurent series!

(ii) Let's find the residue of

$$f(z) = \frac{z + 9}{(z^2 + 1)(z^2 + 9)} \qquad (56.1.14)$$

about $\alpha = 3i$. We find that

$$\text{Res}(f, 3i) = \lim_{z \to 3i} \frac{z+9}{(z^2+1)(z+3i)} = \frac{9+3i}{-8 \cdot 6i} = \frac{-1+3i}{16} \tag{56.1.15}$$

(iii) Let's find the residue of

$$f(z) = \frac{z^3}{z^4+1} \tag{56.1.16}$$

about $\alpha = e^{(2n+1)i\pi/4}$ for $n \in \mathbb{Z}$. We find that

$$\text{Res}(f, \alpha) = \frac{\alpha^3}{4\alpha^3} = \frac{1}{4} \tag{56.1.17}$$

(iv) Let's find the residue of

$$f(z) = \frac{\pi \csc(\pi z)}{4z^2 - 1} \tag{56.1.18}$$

about $\alpha = \pm\frac{1}{2}$. Let $g(z) = \pi$ and $h(z) = (4z^2 + 1)\sin(\pi z)$ which are both analytic at α and such that $h(\alpha) = 0$. Then we see that

$$h'(z) = 8z\sin(\pi z) + (4z^2+1)\pi\cos(\pi z) \implies h'(\alpha) = 4 \neq 0 \tag{56.1.19}$$

It follows that

$$\text{Res}(f, \alpha) = \frac{\pi}{4} \tag{56.1.20}$$

(v) Let's find the residue of

$$f(z) = \frac{\pi \cot(\pi z)}{4z^2 + 1} \tag{56.1.21}$$

about $\alpha = \pm\frac{1}{2}i$. Let $g(z) = \pi\cos(\pi z)$ and $h(z) = (4z^2 - 1)\sin(\pi z)$ which are both analytic at α and such that $h(\alpha) = 0$. Then we see that

$$h'(z) = 8z\sin(\pi z) + (4z^2+1)\pi\cos(\pi z) \implies h'(\alpha) = -4\sinh(\pi/2) \neq 0 \tag{56.1.22}$$

It follows that

$$\text{Res}(f, \alpha) = -\frac{\pi}{4}\coth\frac{\pi}{2} \tag{56.1.23}$$

◀

When dealing with higher order poles then the rules we have described no longer apply. Nevertheless one can still prove a formula analogous to that for simple poles which evaluates the residue of a function.

Theorem *(Residues for higher order poles)*
Let f have a kth order pole at α. Then

$$\text{Res}(f, \alpha) = \frac{1}{(k-1)!} \lim_{z \to \alpha} \left[\frac{d^{k-1}}{dz^{k-1}}\left((z-\alpha)^k f(z) \right) \right] \tag{56.1.24}$$

Proof. Let f have a kth order pole at α. Then its Laurent series about α can be written as

$$f(z) = ... + \frac{a_{-k}}{(z-\alpha)^k} + ... \frac{a_{-1}}{z-\alpha} + a_0 + ..., \ z \in D \tag{56.1.25}$$

in some punctured disc D. Therefore

$$(z-\alpha)^k f(z) = ... + a_{-k} + a_{-k+1}(z-\alpha) + ...(z-\alpha)^{k-1}a_{-1} + a_0(z-\alpha)^k + ..., \ z \in D \tag{56.1.26}$$

Taking the $k-1$th derivative of this expression we find that

$$\frac{d^{k-1}}{dz^{k-1}}\left((z-\alpha)^k f(z)\right) = (k-1)![a_{-1} + a_0(z-\alpha) + ...] \tag{56.1.27}$$

Note that this function is analytic on D, thus taking the limit as $z \to \alpha$ we find that

$$\frac{1}{(k-1)!} \lim_{z \to \alpha} \left[\frac{d^{k-1}}{dz^{k-1}}\left((z-\alpha)^k f(z)\right) \right] = a_{-1} = \mathrm{Res}(f, \alpha) \tag{56.1.28}$$

as desired. ■

> **Example.** Let's find the residue of
>
> $$f(z) = \frac{ze^{iz}}{(z-\pi)^2} \tag{56.1.29}$$
>
> about $z = \pi$. Since f has a second order pole at π, we find that
>
> $$\frac{d^{k-1}}{dz^{k-1}}\left((z-\alpha)^k f(z)\right) = \frac{d}{dz}(ze^{iz}) = (1+iz)e^{iz} \tag{56.1.30}$$
>
> and consequently
>
> $$\mathrm{Res}(f, \pi) = \lim_{z \to \pi}[(1+iz)e^{iz}] = -(1+\pi i) \tag{56.1.31}$$
>
> just as we found previously. ◄

56.2   The Residue Theorem

We saw in the last chapter that if a function is analytic on a punctured disc $D = \{z : 0 < |z-\alpha| < r\}$ then

$$\int_C f(z) \, dz = 2\pi i \mathrm{Res}(f, \alpha) \tag{56.2.1}$$

for any circle $C \subset D$ centered at α. This result can be extended to contours containing multiple integrals, the only modification being that one must now take the sum of the residues, multiplied by $2\pi i$.

> **Theorem** *(Cauchy's residue theorem)*

Let $\mathcal{R}$ be simply connected and f analytic on $\mathcal{R}$ except for a finite number of singularities α_i. Let $\Gamma \in \mathcal{R}$ be a simple-closed contour that does not traverse any of these singularities, then

$$\int_\Gamma f(z) \, dz = 2\pi i \sum_{i \text{ contained}} \text{Res}(f, \alpha_i) \tag{56.2.2}$$

where the sum goes over all residues contained within Γ.

Example.

(i) Consider the integral

$$\int_\Gamma \frac{\sin z}{z^2 + 1} \, dz \tag{56.2.3}$$

where $\Gamma = \{z : |z| = 3\}$. Since $f(z) = \frac{\sin z}{z^2 - 1}$ contains singularities at $z = \pm i$, both contained within Γ, it follows from Cauchy's residue theorem that

$$\int_\Gamma \frac{\sin z}{z^2 + 1} \, dz = 2\pi i [\text{Res}(f, i) + \text{Res}(f, -i)] = 4\pi i \text{Res}(f, i) \tag{56.2.4}$$

where we used the fact that f is an odd function, and hence that the two residues are equal to each other. We see that

$$\text{Res}(f, i) = \frac{\sin(i)}{2i} = \frac{1}{2} \sinh(1) \tag{56.2.5}$$

and thus

$$\int_\Gamma \frac{\sin z}{z^2 + 1} \, dz = 2\pi i \left(e + \frac{1}{e} \right) \tag{56.2.6}$$

(ii) Let's evaluate

$$I = \int_\Gamma \frac{z + 2}{4z^2 + k^2} \, dz \tag{56.2.7}$$

where $\Gamma = \{z : |z - i| = 2\}$. The integrand is analytic on $\mathcal{R} = \mathbb{C}$ except for two simple poles at $4z^2 = -k^2 \implies z = \pm ik/2$. The residues are given by

$$\text{Res}(f, ik/2) = \lim_{z \to ik/2} \frac{z + 2}{4z + 2ik} = \frac{2 + ik/2}{4ik} = \frac{k - 4i}{8k} \tag{56.2.8}$$

and

$$\text{Res}(f, -ik/2) = \lim_{z \to ik/2} \frac{z + 2}{4z - 2ik} = \frac{2 - ik/2}{-4ik} = \frac{k + 4i}{8k} \tag{56.2.9}$$

Note that if $-2 < \pm k < 6$ then Γ will contain at least one of them. If $k = 1$ for example then Γ contains both singularities so

$$I = 2\pi i \left(\frac{1 - 4i}{8} + \frac{1 + 4i}{8} \right) = \frac{\pi i}{2} \tag{56.2.10}$$

If instead $k = 3$ then only the $z = -3i/2$ singularity is included so that

$$I = 2\pi i \frac{3 - 4i}{24} = \pi i \frac{3 - 4i}{12} \tag{56.2.11}$$

Finally if $k = 7$ then no singularities are included so that

$$I = 0 \tag{56.2.12}$$

◄

The residue theorem is also very useful when evaluating trigonometric integrals from 0 to 2π:

$$I = \int_0^{2\pi} \Phi(\cos t, \sin t)\, dt \tag{56.2.13}$$

Firstly we make the substitution

$$\cos t = \frac{1}{2}(z + z^{-1}),\ \sin t = \frac{1}{2}(z - z^{-1}),\ dz = iz dt \tag{56.2.14}$$

which corresponds to the parametrisation $\Gamma\ :\ \gamma(t) = e^{it}$ for $t \in [0, 2\pi]$. This transforms I into a contour integral $\oint_C f(z)\, dz$ around the unit circle C. Assuming that f is analytic with a finite number of singularities in a simply-connected region containing then one can apply the Residue theorem and evaluate the contour integral.

Example. Let's consider the integral

$$\int_0^{2\pi} \cos^n t\, dt \tag{56.2.15}$$

We find that making the required substitution the integral turns into a contour integral

$$\int_0^{2\pi} \cos^n t\, dt = \int_C \frac{1}{2^n}(z - z^{-1})^n \frac{dz}{iz} = -\frac{i}{2^n} \int_C \frac{(z^2 - 1)^n}{z^{n+1}}\, dz \tag{56.2.16}$$

The integrand is analytic on $\mathbb{C}$ except for $z = 0$ where it has a n-th order pole. Using the binomial expansion we see that

$$f(z) = \frac{(z^2 + 1)^n}{z^{n+1}} = \frac{1}{z^{n+1}} \sum_{m=0}^{n} \binom{n}{m} z^{2m} \tag{56.2.17}$$

so the residue of f at $z = 0$ is

$$\mathrm{Res}(f, 0) = \binom{n}{n/2} \text{ if } n \text{ is even} \tag{56.2.18}$$

and zero if n is odd, in which case there is no $\frac{1}{z}$ term in the Laurent series. Consequently we find that

$$\int_0^{2\pi} \cos^n t\, dt = \begin{cases} \frac{\pi}{2^{n-1}} \binom{n}{n/2}, & n \text{ even} \\ 0, & n \text{ is odd} \end{cases} \tag{56.2.19}$$

◄

56.3 Improper integrals

The residue theorem can also be applied to evaluate improper integrals. Consider for example

$$I = \int_{-\infty}^{\infty} \frac{dt}{t^2 + 1} = \lim_{r \to \infty} \int_{-r}^{r} \frac{dt}{t^2 + 1} \tag{56.3.1}$$

(i) We identify a suitable contour integral. In our case we consider

$$\int_{\Gamma} \frac{dz}{z^2 + 1} \tag{56.3.2}$$

where $\Gamma = \Gamma_1 + \Gamma_2$ is the semi-circular contour from $-r$ to r (we take the limit as $r \to \infty$ in the end). This contour encloses the simple pole $z = i$ and is made of a semi-circular arc which
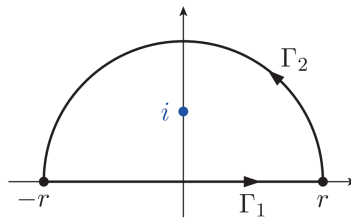
will extend to infinity as $r \to \infty$ and a straight path from $-\infty$ to ∞ which is precisely the integral I we are trying to evaluate. Generally, it is advantageous to choose a contour which allows for the Residue theorem to be applied and that is made up of one component which matches the interval of integration in the improper integral, and other components which vanish as $r \to \infty$.

(ii) Applying the residue theorem, we see that

$$\int_{\Gamma} \frac{dz}{z^2 + 1} = 2\pi i \frac{1}{2i} = \pi \tag{56.3.3}$$

This result is independent of r and a real number (this is crucial since as $r \to \infty$ we will show that this contour integral is equal to the improer integral which is real).

(iii) We split the contour integral as

$$\int_{\Gamma} \frac{dz}{z^2 + 1} = I + \int_{\Gamma_2} \frac{dz}{z^2 + 1} \tag{56.3.4}$$

(iv) We apply the estimation theorem to find an upper bound for $\left| \int_{\Gamma_2} \frac{dz}{z^2+1} \right|$. Indeed on Γ_2 we have that $|z| = r$ so

$$|f(z)| = \frac{1}{|z^2 + 1|} < \frac{1}{|z^2| - 1} = \frac{1}{r^2 - 1} \tag{56.3.5}$$

Therefore, since the length of Γ_2 is πr, and since f is continuous on Γ_2 the estimation theorem yields

$$\left| \int_{\Gamma_2} \frac{dz}{z^2 + 1} \right| \le \frac{\pi r}{r^2 - 1}, \ r > 1 \tag{56.3.6}$$

(v) We take the limit as $r \to \infty$. Then

$$\left| \lim_{r \to \infty} \int_{\Gamma_2} \frac{dz}{z^2 + 1} \right| \leq \lim_{r \to \infty} \frac{\pi r}{r^2 - 1} = 0 \qquad (56.3.7)$$

implying that

$$\int_{\Gamma_2} \frac{dz}{z^2 + 1} = 0 \qquad (56.3.8)$$

Consequently we conclude that $I = \pi$.

Example. Consider the improper integral

$$\int_{-\infty}^{\infty} \frac{dt}{t^4 + 1} \qquad (56.3.9)$$

We consider the contour

$$\int_{\Gamma} \frac{dz}{z^4 + 1} \qquad (56.3.10)$$

where Γ is the same semi-circular contour as in **??**. The integrand has four simple poles at $\alpha = \pm \sqrt{\pm i} = \pm e^{\pm i\pi/4}$. The poles $\alpha_1 = e^{i\pi/4}$ and $\alpha_2 = -e^{-i\pi/4}$ are contained inside Γ for $r > 1$. We therefore find that

$$\text{Res}(f, \alpha_1) = \lim_{z \to \alpha_1} \left(\frac{1}{(z + \alpha_1)(z^2 + i)} \right) = \frac{e^{-i\pi/4}}{4i} \qquad (56.3.11)$$

$$\text{Res}(f, \alpha_2) = \lim_{z \to \alpha_2} \left(\frac{1}{(z + \alpha_2)(z^2 - i)} \right) = \frac{e^{i\pi/4}}{4i} \qquad (56.3.12)$$

Since f is analytic on the region $\{z : \text{Im}(z) > 0\}$ except for $z = \alpha_1$ and $z = \alpha_2$, the Residue theorem yields

$$\int_{\Gamma} \frac{dz}{z^4 + 1} = 2\pi i \frac{e^{i\pi/4} + e^{-i\pi/4}}{4i} = \frac{\pi}{\sqrt{2}} \qquad (56.3.13)$$

As expected, this result is real and independent of the radius of the semi-circle. Now we split the integral into

$$\frac{\pi}{\sqrt{2}} = \int_{\Gamma} \frac{dz}{z^4 + 1} = I + \int_{\Gamma_2} \frac{dz}{z^4 + 1} \qquad (56.3.14)$$

Now note that on $\Gamma_2 = \{z : |z| = r\}$ we find that

$$|z^4 + 1| < |z|^4 + 1 = r^4 + 1 \implies \left| \frac{1}{z^4 + 1} \right| = \frac{1}{r^4 + 1} \qquad (56.3.15)$$

so that using the estimation theorem, which is applicable since $f(z)$ is continuous on Γ_2, yields

$$\left| \int_{\Gamma_2} \frac{dz}{z^4 + 1} \right| < \frac{\pi r}{r^4 + 1} \to 0 \text{ as } r \to \infty \qquad (56.3.16)$$

Finally, if we extend the contour to infinity we find that

$$I = \frac{\pi}{\sqrt{2}} \qquad (56.3.17)$$

◀

Suppose we tried to apply this strategy to evaluate

$$\int_{-\infty}^{\infty} \frac{t}{(t^2+1)(t-2)} \, dt \tag{56.3.18}$$

This time we can't use the semi-circular contour, since the integrand has a singularity at $z = 2$ which is crossed by the contour. To resolve this we can modify the contour in the vicinity of the singularity so as to skip over it.

Proposition *(Round-the-Pole)*
Let f be analytic on a punctured disc $\{z : 0 < |z - \alpha| < \delta\}$ with a simple pole at α. Let $\epsilon < \delta$, and let Γ be the upper-semicircle centered at α with radius ϵ. Then

$$\lim_{\epsilon \to 0} \int_{\Gamma} f(z) \, dz = \pi i \mathrm{Res}(f, \alpha) \tag{56.3.19}$$

Thus we see that by skipping over the pole gives us half the result from a normal application of the residue theorem.

From this we see that if p, q are rational functions such that the degree of q is larger than that of p by at least two, and that any poles of p/q on the real axis are simple, then

$$\int_{-\infty}^{\infty} \frac{p(t)}{q(t)} \, dt = \pi i(2S + T) \tag{56.3.20}$$

where S is the sum of the residues of p/q in the upper half plane, and T is the sum of the residues of p/q at the poles on the real axis.

Example. We evaluate

$$\int_{-\infty}^{\infty} \frac{dt}{(t^2 + a^2)(t^2 + b^2)} \tag{56.3.21}$$

We have four residues at $\pm ia, \pm ib$, of which only ia, ib lie in the upper half plane. We get that

$$\mathrm{Res}(f, ia) = \lim_{t \to ia} \frac{1}{(t + ia)(t^2 + b^2)} = \frac{1}{2ia} \frac{1}{b^2 - a^2} \tag{56.3.22}$$

$$\mathrm{Res}(f, ib) = \lim_{t \to ib} \frac{1}{(t + ib)(t^2 + a^2)} = -\frac{1}{2ib} \frac{1}{b^2 - a^2} \tag{56.3.23}$$

implying that

$$S = \frac{1}{2i(b^2 - a^2)} \left(\frac{1}{a} - \frac{1}{b} \right) = \frac{1}{2\pi ab(a + b)} \tag{56.3.24}$$

It follows that

$$\int_{-\infty}^{\infty} \frac{dt}{(t^2 + a^2)(t^2 + b^2)} = \frac{\pi}{ab(a + b)} \tag{56.3.25}$$

◄

We can actually extend the result by noting that exponentials e^{ikt} vanish in the circular portion of

the semi-circular contour for $k > 0$. Therefore

$$\int_{-\infty}^{\infty} \frac{p(t)}{q(t)} e^{ikt} \, dt = \pi i (2S + T) \tag{56.3.26}$$

where S and T ares the residues on the upper half complex plane and on the real line of $f(z) = \frac{p(z)}{q(z)} e^{ikz}$.

Let's evaluate

$$\int_{-\infty}^{\infty} \frac{e^{ikt}}{t} \, dt \tag{56.3.27}$$

We have that $f(z) = e^{ikz}/z$ has only one simple pole at $z = 0$, with a residue of 1. Therefore, using the Round-the-Pole proposition we obtain

$$int_{-\infty}^{\infty} \frac{e^{ikt}}{t} \, dt = i\pi \implies \int_{-\infty}^{\infty} \frac{cos(kt)}{t} \, dt = 0, \; int_{-\infty}^{\infty} \frac{\sin kt}{t} \, dt = \pi \tag{56.3.28}$$

◄

> **Theorem** *(Infinite sums of residues)*
> Let h be a even function analytic on $\mathbb{C}$ except at $\alpha_1, \alpha_2, ..., \alpha_k$ none of which are integers. Let S_N be the square contour with vertices at $(N + 1/2)(\pm 1 \pm i)$. Suppose also that $f(z) = \pi \cot(\pi z) h(z)$ is such that
> $$\lim_{N \to \infty} \int_{S_N} f(z) \, dz = 0 \tag{56.3.29}$$
> Then
> $$\sum_{n=1}^{\infty} h(n) = -\frac{1}{2} \left[\text{Res}(f, 0) + \sum_{j=1}^{k} \text{Res}(f, \alpha_j) \right] \tag{56.3.30}$$

Proof. The function $g(z) = \pi \cot(\pi z)$ has simple poles at $z = 0, \pm 1, \pm 2, ...$ which all have residue 1. Thus the singularities of f consist of all the integers and the singularities of h. Hence

$$\forall z \in \mathbb{Z}, \; \lim_{z \to n} (z - n) f(z) = h(n) = h(-n) = \lim_{z \to -n} (z + n) f(z) \tag{56.3.31}$$

which by assumption is a well-defined quantity. Now we evaluate the contour integral of f around S_N

$$\int_{S_N} f(z) \, dz = 2\pi i \left(\text{Res}(f, 0) + \sum_{j=1}^{k} \text{Res}(f, \alpha_j) + \sum_{n=1}^{N} [\text{Res}(f, n) + \text{Res}(f, -n)] \right) \tag{56.3.32}$$

$$= 2\pi i \left(\text{Res}(f, 0) + \sum_{j=1}^{k} \text{Res}(f, \alpha_j) + 2 \sum_{n=1}^{N} h(n) \right) \tag{56.3.33}$$

As we let $N \to \infty$, by assumption we have

$$\lim_{N \to \infty} \int_{S_N} f(z) \, dz = 0 \tag{56.3.34}$$

implying that

$$\sum_{n=1}^{N} h(n) = -\frac{1}{2}\left(\text{Res}(f,0) + \sum_{j=1}^{k} \text{Res}(f,\alpha_j) \right) \tag{56.3.35}$$

∎

It turns out that the condition in (56.3.34) is not too artificial in light of the following result.

Proposition *(Cotangent inequality)*
For $N \in \mathbb{N}$ we have $|\cot \pi z| \leq 2, \ \forall z \in S_N$.

We also state below the Laurent series for $\cot z$ which often proves useful to determine $\text{Res}(f,0)$

$$\cot z = \frac{1}{z} - \frac{1}{3}z - \frac{1}{45}z^3 - ..., \ 0 < |z| < r \tag{56.3.36}$$

Example.

(i) Let's solve the Basel problem which consists in finding

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \tag{56.3.37}$$

Let $h(z) = \frac{1}{z^2}$, so that

$$f(z) = \pi z \cot(\pi z) h(z) = \frac{\pi}{z}\left(\frac{1}{z} - \frac{1}{3}z - \frac{1}{45}z^3 - ... \right) \ 0 < |z| < r \tag{56.3.38}$$

and thus $\text{Res}(f,0) = -\frac{\pi}{3}$. Also, if $z \in S_N$ then $|z| \geq N + \frac{1}{2}$ and thus

$$L(S_N) \cdot \left| \frac{\pi \cot(\pi z)}{z} \right| \leq \frac{16\pi}{2N+1}, \ z \in S_N \implies \lim_{N \to \infty} \int_{S_N} f(z) \, dz = 0 \tag{56.3.39}$$

Consequently (56.3.34) holds and

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi}{6} \tag{56.3.40}$$

just as Euler had proven!

(ii) Similarly we find that

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi}{45} \tag{56.3.41}$$

◀

Theorem *(Infinite sums of residues)*
Let h be a even function analytic on $\mathbb{C}$ except at $\alpha_1, \alpha_2, ..., \alpha_k$ none of which are integers. Let S_N be the square contour with vertices at $(N + 1/2)(\pm 1 \pm i)$. Suppose also that $f(z) =$

$\pi \cot(\pi z) h(z)$ is such that

$$\lim_{N \to \infty} \int_{S_N} f(z)\, dz = 0 \tag{56.3.42}$$

Then

$$\sum_{n=1}^{\infty} h(n) = -\frac{1}{2}\left[\operatorname{Res}(f,0) + \sum_{j=1}^{k} \operatorname{Res}(f,\alpha_j) \right] \tag{56.3.43}$$

Proof. The function $g(z) = \pi \cot(\pi z)$ has simple poles at $z = 0, \pm 1, \pm 2, \dots$. The residue at $z = n$ is $(-1)^n$. Thus the singularities of f consist of all the integers and the singularities of h. Hence

$$\forall z \in \mathbb{Z}, \ \lim_{z \to n}(z - n)f(z) = (-1)^n h(n) = (-1)^n h(-n) = \lim_{z \to -n}(z + n)f(z) \tag{56.3.44}$$

which by assumption is a well-defined quantity. Now we evaluate the contour integral of f around S_N

$$\int_{S_N} f(z)\, dz = 2\pi i \left(\operatorname{Res}(f,0) + \sum_{j=1}^{k}\operatorname{Res}(f,\alpha_j) + \sum_{n=1}^{N}[\operatorname{Res}(f,n) + \operatorname{Res}(f,-n)] \right) \tag{56.3.45}$$

$$= 2\pi i\left(\operatorname{Res}(f,0) + \sum_{j=1}^{k}\operatorname{Res}(f,\alpha_j) + 2\sum_{n=1}^{N}(-1)^n h(n) \right) \tag{56.3.46}$$

As we let $N \to \infty$, by assumption we have

$$\lim_{N \to \infty} \int_{S_N} f(z)\, dz = 0 \tag{56.3.47}$$

implying that

$$\sum_{n=1}^{N}(-1)^n h(n) = -\frac{1}{2}\left(\operatorname{Res}(f,0) + \sum_{j=1}^{k}\operatorname{Res}(f,\alpha_j) \right) \tag{56.3.48}$$

∎

Once again the constraint in (56.3.47) is not very restrictive because of the following inequality

Proposition *(Cosecant inequality)* For $N \in \mathbb{N}$ we have $|\csc \pi z| \le 1, \ \forall z \in S_N$.

The Laurent series for $\csc z$ about $z = 0$ is

$$\csc z = \frac{1}{z} + \frac{1}{6}z + \frac{7}{360}z^3 + \dots \tag{56.3.49}$$

Example. Let's evaluate

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{4n^2 - 1} \tag{56.3.50}$$

Letting $h(z) = \frac{1}{4z^2 - 1}$ we can calculate $\operatorname{Res}(f, 1/2) = \operatorname{Res}(f, -1/2) = \pi/4$. Moreover since h

is analytic at $z = 0$ we get $\text{Res}(f, 0) = -1$.
Now note that if $z \in S_B$ then $|z| \geq N + 1/2$ so that

$$|4z^2 + 1| \geq 4\left(N + \frac{1}{2}\right)^2 - 1 \implies \left|\frac{\pi \csc \pi z}{4z^2 - 1}\right| \leq \frac{\pi}{4\left(N + \frac{1}{2}\right)^2 - 1}, \; z \in S_N \qquad (56.3.51)$$

Therefore

$$\left|\int_{S_N} f(z)\, dz\right| \leq \frac{\pi}{4(N + 1/2)^2 - 1} \cdot 4(2N + 1) \to 0 \text{ as } N \to \infty \qquad (56.3.52)$$

so

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{4n^2 - 1} = -\frac{1}{2}\left(-1 + \frac{\pi}{4} + \frac{\pi}{4}\right) = \frac{1}{2} - \frac{\pi}{4} \qquad (56.3.53)$$

◀

CA7 Zeros and extrema

CA8 Conformal mappings

CA9 Applications to fluid flows

CA10 The Mandelbrot set and complex dynamics

Part VIII

Calculus of Variations

CV1 Functionals and stationary values

Part IX

Fourier Analysis

FA1 Fourier series

62.1 Dirichlet conditions

The Fourier series technique may be employed to express functions that are not analytic (can't be expanded into a taylor series) as power series nonetheless. However, a number of conditions, known as **Dirichlet conditions** must still be met.

> **Definition** *(Dirichlet conditions)*
> The Dirichlet conditions are:
>   (i) the function has finite fundamental period
>   (ii) has at most a finite number of discontinuities
>   (iii) finite number of stationary points per fundamental period
>   (iv) the integral of $|f(x)|$ converges over a fundamental period
> where the **fundamental period** of a periodic functions is the smallest $T > 0$ such that $f(t + T) = f(t)$.

Fourier transforms

Convolutions

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).