Notes on the Mathematical Tripos

Sky Wilshaw

Part IB

University of Cambridge
2020–2024

Contents

I. Optimisation
Lectured in Easter 2021 by Dr. V. Jog **5**

II. Variational Principles
Lectured in Easter 2021 by Dr. M. Dunajski **63**

III. Markov Chains
Lectured in Michaelmas 2021 by Dr. P. Sousi **101**

IV. Analysis and Topology
Lectured in Michaelmas 2021 by Dr. V. Zsák **137**

V. Methods
Lectured in Michaelmas 2021 by Prof. E. P. Shellard **215**

VI. Quantum Mechanics
Lectured in Michaelmas 2021 by Dr. M. Ubiali **297**

VII. Linear Algebra
Lectured in Michaelmas 2021 by Prof. P. Raphael **347**

VIII. Groups, Rings and Modules
Lectured in Lent 2022 by Dr. R. Zhou **427**

IX. Complex Analysis
Lectured in Lent 2022 by Prof. N. Wickramasekera **485**

X. Geometry
Lectured in Lent 2022 by Prof. I. Smith **543**

XI. Statistics
Lectured in Lent 2022 by Dr. S. Bacallado **609**

I. Optimisation

Lectured in Easter 2021 by Dr. V. Jog

Many real-world problems involve finding optimal points of functions, for instance making the most valuable products given limited resources, or finding the optimal way to transport goods across a network. In this course, we study the theory behind optimisation, and produce various algorithms for computing optima in different environments.

An important class of functions is the convex functions. One can show that if a function is convex, we can use local behaviour to make conclusions about global minima and maxima. This helps guide our study of optimisation. Linear functions are convex, and the study of optimising linear functions is called linear programming. We show that linear programs can be solved computationally using the simplex method, allowing us to easily solve lots of real-world optimisation problems.

Contents

1. Introduction and convex functions 8
1.1. Outline and definitions 8
1.2. Convexity 8
1.3. Unconstrained optimisation 9
1.4. First-order conditions for convexity 9
1.5. Second-order conditions for convexity 10
2. Optimisation algorithms 12
2.1. Gradient descent 12
2.2. Smoothness assumption 12
2.3. Strong convexity assumption 15
2.4. Proving gradient descent 16
2.5. Rate of convergence 17
2.6. Condition numbers and oscillation 17
2.7. Newton's method 18
2.8. Barrier methods 18
3. Lagrange multipliers 20
3.1. Introduction and Lagrange sufficiency 20
3.2. Using Lagrange multipliers in general 21
3.3. Complementary slackness 22
3.4. Weak duality 24
3.5. Strong duality and the Lagrange method 25
3.6. Hyperplane condition for strong duality 25
3.7. Strong duality and convex functions 26
3.8. Shadow prices interpretation of Lagrange multipliers 27
4. Linear programming 29
4.1. Linear programs 29
4.2. Maximising convex functions 30
4.3. Basic solutions and basic feasible solutions 31
4.4. Extreme points of the feasible set in standard form 32
5. Duality in linear programming 34
5.1. Strong duality of linear programs 34
5.2. Duals of linear programs in standard form 34
5.3. Duals of linear programs in general form 35
5.4. Dual of dual program 35
5.5. Dual of arbitrary linear program 36
5.6. Optimality conditions 37

6. Simplex method **38**
6.1. Introduction 38
6.2. Feasibility of basic directions 39
6.3. Cost of basic directions 39
6.4. Moving to basic feasible solutions 40
6.5. Simplex method 41
6.6. Tableau implementation 41
7. Game theory **45**
7.1. Zero-sum games 45
7.2. Mixed strategies 46
7.3. Duality of mixed strategy problems 47
7.4. Finding optimal strategies 48
8. Network flows **50**
8.1. Minimum cost flow 50
8.2. Transport problem 50
8.3. Sufficiency of transport problem 51
8.4. Optimality conditions for transport problem 52
9. The transport algorithm **53**
9.1. Transportation tableaux 53
9.2. Updating the transportation tableau 55
10. Maximum flow, minimum cut **57**
10.1. Introduction 57
10.2. Cuts and flows 57
10.3. Max-flow min-cut theorem 58
10.4. Ford–Fulkerson algorithm 59
10.5. Termination of Ford–Fulkerson 61
10.6. Bipartite matching problem 61

1. Introduction and convex functions

1.1. Outline and definitions

An *optimisation problem* is a problem in which we want to minimise some function $f(\mathbf{x})$ such that $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$. We may have a set of *constraints* $h(\mathbf{x}) = \mathbf{b}$ where $h(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$. Note that we will only ever consider minimisation of functions since we can maximise a function by minimising its negative. Such a problem is often written with notation such as

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimise}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h(\mathbf{x}) = \mathbf{b} \end{aligned}$$

Definition. The following definitions will be used.

(i) The function f that we want to minimise is called the *objective function*.

(ii) The components of the vector $\mathbf{x}$ are called the *decision variables*.

(iii) A constraint of the form $h(\mathbf{x}) = \mathbf{b}$ is called a *functional constraint*.

(iv) A constraint of the form $\mathbf{x} \in \mathcal{X}$ is called a *regional constraint*.

(v) The set $\mathcal{X}(\mathbf{b}) = \{\mathbf{x} : \mathbf{x} \in \mathcal{X}, h(\mathbf{x}) = \mathbf{b}\}$ is called the *feasible set*.

(vi) If the feasible set is non-empty, the optimisation problem is called feasible. If the feasible set is empty, the problem is infeasible.

(vii) The problem is called *bounded* if the minimum on $\mathcal{X}(\mathbf{b})$ is bounded.

(viii) A point $\mathbf{x}^\star \in \mathcal{X}(\mathbf{b})$ is *optimal* if it minimises f over $\mathcal{X}(\mathbf{b})$. The value $f(\mathbf{x}^\star)$ is called the optimal cost.

We can convert an inequality constraint into an equality constraint with a regional constraint, for instance

$$h(\mathbf{x}) \leq b \longrightarrow h(\mathbf{x}) + s = b;\ s \geq 0$$

1.2. Convexity

Definition. A set $S \subseteq \mathbb{R}^n$ is *convex* if for all $\mathbf{x}, \mathbf{y} \in S$, the line segment from $\mathbf{x}$ to $\mathbf{y}$ lies entirely inside S. In other words, for all $\lambda \in [0, 1]$, $\mathbf{x}(1 - \lambda) + \mathbf{y}(\lambda) \in S$.

Definition. A function $f : S \to \mathbb{R}$ is convex if

- S is convex, and
- for all $\mathbf{x}, \mathbf{y} \in S$,

$$f((1 - \lambda)\mathbf{x} + \lambda\mathbf{y}) \leq (1 - \lambda)f(\mathbf{x}) + \lambda f(\mathbf{y})$$

So informally, for a convex function, if we take two inputs, the chord connecting their outputs lies above the function's curve. If the given inequality above is strict, the function is called *strictly* convex. f is (strictly) *concave* if $-f$ is (strictly) convex. Note that if f is linear, f is convex and concave, since f is linear in its input. Hence linear optimisation is a special case of convex optimisation.

1.3. Unconstrained optimisation

The *unconstrained* optimisation problem is simply to minimise $f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Convex functions allow you to generalise the behaviour of a function in a small neighbourhood to global behaviour, so it becomes easier to solve optimisation problems expressed in terms of convex functions.

1.4. First-order conditions for convexity

Suppose we have a tangent to a curve $f : \mathbb{R} \to \mathbb{R}$ at a given point x. If f is convex, then f must only touch the curve once, since if it touched twice we would contradict the definition of convexity. In particular, we have the following necessary and sufficient condition for convexity:

$$f(y) \geq f(x) + (y - x)f'(x)$$

In higher dimensions, we might guess that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x}) \cdot \nabla f(\mathbf{x})$$

Theorem. A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,\ f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x}) \cdot \nabla f(\mathbf{x}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x})$$

Remark. If $\nabla f(\mathbf{x}) = \mathbf{0}$ for some vector $\mathbf{x}$, then the first-order condition implies that $f(\mathbf{y}) \geq f(\mathbf{x})$, so $\mathbf{x}$ is the global minimum of f. This is an example of how we can use local properties (the gradient of the function at $\mathbf{x}$) to deduce global properties (the minimum value of the function).

Proof. First, we will prove that convexity implies the first-order condition. By convexity, we have

$$f((1 - \lambda)\mathbf{x} + \lambda\mathbf{y}) \leq (1 - \lambda)f(\mathbf{x}) + \lambda f(\mathbf{y})$$

Initially, let $n = 1$ so that we have the one-dimensional case. We have

$$f(y) \geq f(x) + \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} = f(x) + \frac{f(x + \lambda(y - x)) - f(x)}{\lambda(y - x)}(y - x)$$

Hence, taking the limit as $\lambda \to 0$, we have

$$f(y) \geq f(x) + f'(x)(y - x)$$

For the general case, we define a function g such that $g(\lambda) = f((1-\lambda)\mathbf{x} + \lambda\mathbf{y})$. Since f is convex, so is g. We can calculate

$$g'(\lambda) = \nabla f((1-\lambda)\mathbf{x} + \lambda\mathbf{y}) \cdot (\mathbf{y} - \mathbf{x})$$

Since $g : [0,1] \to \mathbb{R}$ is convex, by the above argument for $n = 1$ we have

$$\begin{aligned} g(1) &\geq g(0) + g'(0)(1-0) \\ f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \end{aligned}$$

Now we must prove the converse; if the first-order condition holds, then f is convex. Let

$$\mathbf{x}_\lambda = (1-\lambda)\mathbf{x} + \lambda\mathbf{y}$$

The first-order condition shows that

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{x}_\lambda) + \nabla f(\mathbf{x}_\lambda) \cdot (\mathbf{x} - \mathbf{x}_\lambda) \\ f(\mathbf{y}) &\geq f(\mathbf{x}_\lambda) + \nabla f(\mathbf{x}_\lambda) \cdot (\mathbf{y} - \mathbf{x}_\lambda) \end{aligned}$$

Multiplying the first equation by $1-\lambda$ and multiplying the second equation by λ, we get

$$\begin{aligned} (1-\lambda)f(\mathbf{x}) + \lambda f(\mathbf{y}) &\geq f(\mathbf{x}_\lambda) + \nabla f(\mathbf{x}_\lambda) \cdot [(\mathbf{x} - \mathbf{x}_\lambda)(1-\lambda) + (\mathbf{y} - \mathbf{x}_\lambda)\lambda] \\ &= f(\mathbf{x}_\lambda) + \nabla f(\mathbf{x}_\lambda) \cdot [(\mathbf{x} - (1-\lambda)\mathbf{x} - \lambda\mathbf{y})(1-\lambda) + (\mathbf{y} - (1-\lambda)\mathbf{x} - \lambda\mathbf{y})\lambda] \\ &= f(\mathbf{x}_\lambda) + \nabla f(\mathbf{x}_\lambda) \cdot [(\lambda\mathbf{x} - \lambda\mathbf{y})(1-\lambda) + ((1-\lambda)\mathbf{y} - (1-\lambda)\mathbf{x})\lambda] \\ &= f(\mathbf{x}_\lambda) + \nabla f(\mathbf{x}_\lambda) \cdot 0 \\ &= f(\mathbf{x}_\lambda) \end{aligned}$$

Hence f really is convex. □

1.5. Second-order conditions for convexity

When $n = 1$, we suspect that $f''(x) \geq 0$ is the condition for convexity. In higher dimensions, the analogous operator to the double derivative is the Hessian matrix.

$$\nabla^2 f(\mathbf{x}) = \mathbf{H}_f = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

Definition. An $n \times n$ matrix A is *positive semidefinite* if for all $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{x}^\intercal A \mathbf{x} \geq 0$. Equivalently, all eigenvalues of A are non-negative. If A is positive semidefinite, we write $A \succeq 0$.

Note that the higher-dimensional analogue of the Taylor expansion of $f(\mathbf{y})$ is

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\intercal \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \cdots$$

Theorem. A twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\nabla^2 f(\mathbf{x}) \succeq 0$ at all $\mathbf{x}$. The converse also holds, but it is not important for this course, so it will not be proven.

Proof. Using the Taylor expansion of f, we have

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\intercal \nabla^2 f(\mathbf{z})(\mathbf{y}-\mathbf{x})$$

where $\mathbf{z} = (1-\lambda)\mathbf{x} + \lambda\mathbf{y}$ for some $\lambda \in [0,1]$. The rightmost term is positive, hence

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x})$$

So the first-order conditions are satisfied, which imply convexity. □

2. Optimisation algorithms

2.1. Gradient descent

Consider minimising $f(x)$ such that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Recall that a local minimum of f is also the global minimum. Consider the following 'greedy' method:

- Start at a point $\mathbf{x}_0$.
- Search for close points around $\mathbf{x}_0$ whose values of f are smaller than $f(\mathbf{x}_0)$.
  - If such a point exists, let this be $\mathbf{x}_1$. Repeat the algorithm.
  - If such a point does not exist, we have found a local minimum, which is the global minimum.

We can find such $\mathbf{x}_1$ points by considering the Taylor series expansion of f around a point.

$$f(\mathbf{x} - \varepsilon\nabla f(\mathbf{x})) \approx f(\mathbf{x}) - \varepsilon\nabla f(\mathbf{x})^\intercal \cdot \nabla f(\mathbf{x}) = f(\mathbf{x}) - \varepsilon\|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x})$$

Hence $-\nabla f(\mathbf{x})$ is called a descending direction. Although the gradient of the function is the most natural way of decreasing a function, any $\mathbf{v}$ with $f(\mathbf{x}) \cdot \mathbf{v} < 0$ is a descending direction. This gives us the *gradient descent* algorithm.

Algorithm 1: Gradient Descent Algorithm

Result: Global minimum of $f(\mathbf{x})$
start at a point $\mathbf{x}_0$;
$t \leftarrow 0$;
repeat
  find a descending direction $\mathbf{v}_t$, e.g. $-\nabla f(\mathbf{x})$;
  choose a step size η_t;
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta_t \mathbf{v}_t$;
until $\nabla f(\mathbf{x}) = 0$ *or t is large enough*;

Different choices of $\mathbf{v}_t$ and η_t give rise to many different qualities of algorithm.

2.2. Smoothness assumption

Some restrictions must be applied to a function to let us prove that gradient descent works.

Definition. A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *β-smooth* if ∇f is a β-Lipschitz function:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta\|\mathbf{x} - \mathbf{y}\|$$

In the following sections, we assume all functions f are β-smooth. Further, if f is twice differentiable (i.e. the Hessian exists everywhere), then the β-smoothness assumption is equivalent to

$$\nabla^2 f(\mathbf{x}) \preceq \beta I$$

so all eigenvalues of $\nabla^2 f(\mathbf{x})$ have $\lambda \leq \beta$. Also,

$$\mathbf{u}^\intercal \nabla^2 f(\mathbf{x})\mathbf{u} \leq \mathbf{u}^\intercal(\beta I)\mathbf{u} = \beta\|\mathbf{u}\|^2$$

Definition. The *linear approximation* to f at $\mathbf{x}$ is

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x})$$

We might assume that the linear approximation is close to the actual function in a small neighbourhood around x.

Claim. If f is β-smooth, then

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Note that

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y})$$

since f is convex, so this claim would show that f really is close to the actual function, deviating by an arbitrarily small amount as we let $\mathbf{x}$ approach $\mathbf{y}$.

Proof. By Taylor's theorem,

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\intercal \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \\ &\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\intercal(\beta I)(\mathbf{y} - \mathbf{x}) \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^2 \end{aligned}$$

□

Corollary. If we move by a step size of $\frac{1}{\beta}$, we will descend by at least $\frac{1}{2\beta}\|\nabla f(x)\|^2$.

$$f\left(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})\right) \leq f(\mathbf{x}) - \frac{1}{2\beta}\|\nabla f(x)\|^2$$

Proof. Consider

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

as a function of $\mathbf{y}$, and try to minimise it for a fixed $\mathbf{x}$.

$$\nabla_{\mathbf{y}}\left(f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|^2\right) = \nabla f(\mathbf{x}) + \beta(\mathbf{y} - \mathbf{x}) = 0$$

Hence,

$$\frac{\nabla f(\mathbf{x})}{\beta} = \mathbf{x} - \mathbf{y}$$
$$\mathbf{y} = \mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})$$

Substituting into the claim above, we have

$$\begin{aligned} f\left(x - \frac{1}{\beta}\nabla f(\mathbf{x})\right) &\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal\left(\frac{-1}{\beta}\nabla f(\mathbf{x})\right) + \frac{\beta}{2}\left\|\frac{1}{\beta}\nabla f(\mathbf{x})\right\|^2 \\ &= f(\mathbf{x}) - \frac{1}{\beta}\|\nabla f(\mathbf{x})\|^2 + \frac{1}{2\beta}\|\nabla f(\mathbf{x})\|^2 \\ &= f(\mathbf{x}) - \frac{1}{2\beta}\|\nabla f(\mathbf{x})\|^2 \end{aligned}$$

□

Claim (Improved first order condition)**.**

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{1}{2\beta}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

Proof. For any $\mathbf{z}$, by the standard first order condition and the corollary above we have

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{z} - \mathbf{x}) \leq f(\mathbf{z}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\intercal(\mathbf{z} - \mathbf{y}) + \frac{\beta}{2}\|\mathbf{z} - \mathbf{y}\|^2$$

This then implies

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^\intercal(\mathbf{x} - \mathbf{z}) + \nabla f(\mathbf{y})^\intercal(\mathbf{z} - \mathbf{y}) + \frac{\beta}{2}\|\mathbf{z} - \mathbf{y}\|^2$$

The left hand side is not dependent on $\mathbf{z}$, so by minimising $\mathbf{z}$ we get the best bound for the left hand side. We set the gradient of $\mathbf{z}$ to zero.

$$-\nabla f(\mathbf{x}) + \nabla f(y) + \beta(\mathbf{z} - \mathbf{y}) = 0$$
$$\implies \mathbf{z} = \frac{\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})}{\beta} + \mathbf{y}$$

Substituting back, we have

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^\intercal(\mathbf{x} - \mathbf{y}) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

□

2.3. Strong convexity assumption

In general, a small gradient does not imply that we are close to the optimum value of the function. We must therefore add an additional assumption in order to justify gradient descent.

Definition. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *α-strongly convex* if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

If f is twice differentiable, then its Hessian satisfies

$$\nabla^2 f(\mathbf{x}) \geq \alpha I$$

for all $\mathbf{x}$.

Claim. Let f be α-strongly convex. Let $p^\star$ be the optimal cost; i.e. the minimum value of f. Then for any $\mathbf{x}$ we have

$$p^\star \geq f(\mathbf{x}) - \frac{1}{2\alpha}\|\nabla f(\mathbf{x})\|^2$$

Remark. If $\|\nabla f(\mathbf{x})\| \leq \sqrt{2\alpha\varepsilon}$, then

$$p^\star \leq f(\mathbf{x}) \leq p^\star + \varepsilon$$

So a small gradient means we are close to the optimum.

Proof. The α-strong convexity assumption gives

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Taking the minimum over $\mathbf{y}$ of both sides, the left hand side becomes $p^\star$. Setting the gradient of the right hand side to zero,

$$\begin{aligned} \nabla f(\mathbf{x}) - \alpha(\mathbf{x} - \mathbf{y}) &= 0 \\ \frac{\nabla f(\mathbf{x})}{\alpha} &= (\mathbf{x} - \mathbf{y}) \end{aligned}$$

This gives

$$\begin{aligned} p^\star &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal\left(-\frac{\nabla f(\mathbf{x})}{\alpha}\right) + \frac{\alpha}{2}\left\|\frac{\nabla f(\mathbf{x})}{\alpha}\right\|^2 \\ &= f(\mathbf{x}) - \frac{\|\nabla f(\mathbf{x})\|^2}{\alpha} + \frac{\|\nabla f(\mathbf{x})\|^2}{2\alpha} \\ &= f(\mathbf{x}) - \frac{\|\nabla f(\mathbf{x})\|^2}{2\alpha} \end{aligned}$$

□

Claim. Let $\mathbf{x}^\star$ be the minimising value, i.e. $f(\mathbf{x}^\star) = p^\star$. Then

$$\|\mathbf{x} - \mathbf{x}^\star\| \leq \frac{2}{\alpha}\|\nabla(\mathbf{x})\|$$

So if a function is strongly convex, we can find a region in which we know the global maximum lies.

Proof. By the Cauchy–Schwarz inequality,

$$\begin{aligned} f(\mathbf{x}^\star) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{x}^\star - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}^\star - \mathbf{x}\|^2 \\ &\geq f(\mathbf{x}) - \|\nabla f(\mathbf{x})\|\|\mathbf{x}^\star - \mathbf{x}\| + \frac{\alpha}{2}\|\mathbf{x}^\star - \mathbf{x}\|^2 \end{aligned}$$

Since $f(\mathbf{x}^\star) \leq f(\mathbf{x})$, we have

$$0 \geq f(\mathbf{x}^\star) - f(\mathbf{x}) \geq -\|\nabla f(\mathbf{x})\|\|\mathbf{x}^\star - \mathbf{x}\| + \frac{\alpha}{2}\|\mathbf{x}^\star - \mathbf{x}\|^2$$

Hence,

$$\begin{aligned} \|\nabla f(\mathbf{x})\|\|\mathbf{x}^\star - \mathbf{x}\| &\geq \frac{\alpha}{2}\|\mathbf{x}^\star - \mathbf{x}\|^2 \\ \|\nabla f(\mathbf{x})\| &\geq \frac{\alpha}{2}\|\mathbf{x}^\star - \mathbf{x}\| \end{aligned}$$

□

2.4. Proving gradient descent

Let f be a β-smooth and α-strongly convex, where $0 < \alpha < \beta$. Then

$$\alpha I \preceq \nabla^2 f(\mathbf{x}) \preceq \beta I$$

Theorem. Gradient descent with step size $\frac{1}{\beta}$ satisfies

$$\begin{aligned} f(\mathbf{x}_T) - f(\mathbf{x}^\star) &\leq \left(1 - \frac{\alpha}{\beta}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star)) \\ &\leq e^{-\frac{\alpha T}{\beta}} (f(\mathbf{x}_0) - f(\mathbf{x}^\star)) \\ &\leq e^{-\frac{\alpha T}{\beta}} \frac{\beta}{2}\|\mathbf{x}^\star - \mathbf{x}_0\|^2 \end{aligned}$$

Proof.

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) &\leq f(\mathbf{x}_t) - f(\mathbf{x}^\star) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_t)\|^2 \\ &\leq f(\mathbf{x}_t) - f(\mathbf{x}^\star) - \frac{\alpha}{\beta}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \\ &\leq \left(1 - \frac{\alpha}{\beta}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \end{aligned}$$

Hence by induction,

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\alpha}{\beta}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))$$

The second line of the theorem is a consequence of the properties of the exponential function. The last inequality in the theorem can be shown by β-smoothness.

$$f(\mathbf{x}_0) \leq f(\mathbf{x}^\star) + \nabla f(\mathbf{x}^\star)^\intercal(\mathbf{x}_0 - \mathbf{x}^\star) + \frac{\beta}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$
$$f(\mathbf{x}_0) - f(\mathbf{x}^\star) \leq \frac{\beta}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

□

2.5. Rate of convergence

For example, suppose that we would like $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq 0.1$, and it takes k steps to reach this tolerance. Then, it would take around $2k$ steps to reach a tolerance of 0.01, since the $\left(1 - \frac{\alpha}{\beta}\right)^T$ power might increase by a factor of 2. In general, the number of steps needed to ensure that the error is less than ε is

$$T = \frac{\beta}{\alpha}\log\left(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^\star)}{\varepsilon}\right)$$

This $\log(1/\varepsilon)$ term is called 'linear convergence', since for each extra order of magnitude of accuracy, we need a linear amount of computation steps. Linear convergence is very fast, and such algorithms are very useful.

2.6. Condition numbers and oscillation

Note that

$$1 - \frac{\alpha}{\beta}$$

is the term which controls the convergence of gradient descent. We call β/α the *condition number* of f. Such a number is always greater than 1. If the condition number is very close to 1, the convergence is fast. Consider the function

$$f(x_1, x_2) = \frac{1}{2}(x_1^2 + 100x_2^2)$$

The Hessian of f at any point is

$$\nabla^2 f(x_1, x_2) = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}$$

Hence, $\alpha = 1, \beta = 100$ giving a condition number of 100. This function would optimise very slowly, and we may continually overshoot in the x_2 direction since the gradient points so strongly in this direction. We may like to prevent this oscillation between over-guessing and under-guessing certain coordinate components.

2.7. Newton's method

In gradient descent, we have

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$$

In Newton's method, we replace this formula with

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}_t)$$

Note that the second order approximation for f is

$$f(\mathbf{x}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\intercal + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\intercal \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t)$$

So if we instead try to minimise the right hand side of the second-order approximation with respect to $\mathbf{x}$, we have

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}_t)$$

as given by Newton's method. This ideally allows us to deal with 'badly-proportioned' coordinates independently, by scaling each coordinate using the Hessian rather than by a constant. Essentially, Newton's method iteratively approximates the function with a parabola, and then moves to the minimum point of this parabola. We can show that Newton's method converges according to

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq c\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

when $\mathbf{x}_t - \mathbf{x}^\star$ is small enough. We can see here that the squared term provides very fast convergence once we are in the neighbourhood of the optimum. Newton's method can also be used to find a root of a function. Suppose $f : \mathbb{R} \to \mathbb{R}$, and define $f' = g$.

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - \frac{g(x_t)}{g'(x_t)}$$

So we can find the root of g by computing the stationary point of f. We are essentially taking a linear approximation at a point, and setting this linear approximation to zero.

2.8. Barrier methods

Suppose we impose a constraint on an optimisation problem, for instance minimising $f(\mathbf{x})$ such that $f_i(\mathbf{x}) \leq 0$ for $1 \leq i \leq m$. We can transform such a constrained problem into an unconstrained problem. Let us minimise

$$f(\mathbf{x}) + \sum_{i=1}^{m} \phi(f_i(\mathbf{x}))$$

where $\phi(y_i) = +\infty$ outside the feasible set, and $\phi(y_i) = 0$ inside the feasible set. However, this ϕ function is not differentiable, so this introduces even more problems. We instead consider a *logarithmic barrier function*. Let us minimise the unconstrained problem

$$tf(\mathbf{x}) - \sum_{i=1}^{m} \log(-f_i(\mathbf{x})) \implies \phi(x) = -\log(-x)$$

This barrier function is infinite for negative x, and gradually rises as $x \to 0$. When t is chosen to be very large, the optimum of this problem is very close to the optimum of the original problem.

Algorithm 2: Barrier Method

Result: Global minimum of $f(\mathbf{x})$
start at a point $\mathbf{x}$ inside the feasible set;
set t to be a positive real number;
repeat
  solve the minimiser of $tf(\mathbf{x}) - \sum_{i=1}^{m} \log(-f_i(\mathbf{x}))$ with $\mathbf{x}$ as the initial point using Newton's method giving $\mathbf{x}^\star$;
  $\mathbf{x} \leftarrow \mathbf{x}^\star$;
  $t \leftarrow \alpha t$ for some fixed $\alpha > 1$;
until *t is large enough*;

3. Lagrange multipliers

3.1. Introduction and Lagrange sufficiency

Consider minimising $f(\mathbf{x})$ subject to $\mathbf{x} \in \mathcal{X}$, $h(\mathbf{x}) = \mathbf{b}$ where $h : \mathbb{R}^n \to \mathbb{R}^m$. The *Lagrangian* associated with this problem is

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal (h(\mathbf{x}) - \mathbf{b})$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the vector of Lagrange multipliers. We want to instead minimise $L(\mathbf{x}, \boldsymbol{\lambda})$, $x \in \mathcal{X}$.

Theorem (Lagrange Sufficiency)**.** Suppose we can find a $\boldsymbol{\lambda}^\star$ such that

(i) $\min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}^\star) = L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$

(ii) $\mathbf{x}^\star \in \mathcal{X}(\mathbf{b}) = \{\mathbf{x} : \mathbf{x} \in \mathcal{X}, h(\mathbf{x}) = \mathbf{b}\}$

Then $\mathbf{x}^\star$ is optimal for the original constrained problem, i.e.

$$\min_{\mathbf{x} \in \mathcal{X}(\mathbf{b})} f(\mathbf{x}) = f(\mathbf{x}^\star)$$

Proof. First, note that condition (ii) states that $f(\mathbf{x}^\star) \geq \min_{\mathbf{x} \in \mathcal{X}(\mathbf{b})} f(\mathbf{x})$, because $\mathbf{x}^\star$ is feasible. Then,

$$\begin{aligned}
\min_{\mathbf{x} \in \mathcal{X}(\mathbf{b})} f(\mathbf{x}) &= \min_{\mathbf{x} \in \mathcal{X}(\mathbf{b})} f(\mathbf{x}) - \underbrace{(\boldsymbol{\lambda}^\star)^\intercal (h(\mathbf{x}) - \mathbf{b})}_{0 \text{ when } \mathbf{x} \in \mathcal{X}(\mathbf{b})} \\
&\geq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) - (\boldsymbol{\lambda}^\star)^\intercal (h(\mathbf{x}) - \mathbf{b}) \\
&= \min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}^\star) \\
&= L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \\
&= f(\mathbf{x}^\star) - (\boldsymbol{\lambda}^\star)^\intercal (h(\mathbf{x}^\star) - \mathbf{b}) \\
&= f(\mathbf{x}^\star)
\end{aligned}$$

□

Example.

$$\begin{aligned}
\underset{x \in \mathbb{R}^3}{\text{minimise}} \quad & -x_1 - x_2 + x_3 \\
\text{subject to} \quad & x_1^2 + x_2^2 = 4 \\
& x_1 + x_2 + x_3 = 1
\end{aligned}$$

In this problem, we have

$$h(\mathbf{x}) = \begin{pmatrix} x_1^2 + x_2^2 \\ x_1 + x_2 + x_3 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

Taking Lagrange multipliers, we have

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= (-x_1 - x_2 + x_3) - \lambda_1(x_1^2 + x_2^2 - 4) - \lambda_2(x_1 + x_2 + x_3 - 1) \\ &= (-(1+\lambda_2)x_1 - \lambda_1 x_1^2) + (-(1+\lambda_2)x_2 - \lambda_1 x_2^2) + (1-\lambda_2)x_3 + 4\lambda_1 + \lambda_2 \end{aligned}$$

We want to fix a value of $\boldsymbol{\lambda}$ and minimise L, only considering solutions such that $\mathbf{x}^\star$ is finite. Note that if $\lambda_1 > 0$, then the first bracket can be made as small as we like by picking very small values of x_1; this bracket would diverge to negative infinity so we cannot choose such a λ_1. If $\lambda_2 \neq 1$, the infimum is also negative infinity by considering the x_3 term. So let us consider $\lambda_1 \leq 0, \lambda_2 = 1$. Setting the derivative of the first term to zero, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x_1}\left(-(1+\lambda_2)x_1 - \lambda_1 x_1^2\right) = -(1+\lambda_2) - 2\lambda_1 x_1 &= 0 \\ \implies x_1 = \frac{-1-\lambda_2}{2\lambda_1} \\ = \frac{-2}{2\lambda_1} \\ = \frac{-1}{\lambda_1} \end{aligned}$$

Setting the derivative of the second term to zero,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x_1}\left(-(1+\lambda_2)x_2 - \lambda_1 x_2^2\right) = -(1+\lambda_2) - 2\lambda_1 x_2 &= 0 \\ \implies x_2 = \frac{-1}{\lambda_1} \end{aligned}$$

We now want to choose λ_1 such that x_1, x_2, x_3 satisfy the constraints.

$$x_1^2 + x_2^2 = 4 \implies x_1^2 = x_2^2 = 2 \implies x_1 = x_2 = \sqrt{2}$$

Note that $x_1, x_2 > 0$ since $\lambda_1 \leq 0$, and correspondingly $\lambda_1 = \frac{-1}{\sqrt{2}}$. Further, we can now find $x_3 = 1 - 2\sqrt{2}$. This solution optimises the original problem.

3.2. Using Lagrange multipliers in general

Consider the problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimise}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h(\mathbf{x}) \leq \mathbf{b} \end{aligned}$$

We can solve this problem using the following steps.

(1) Add a slack variable $\mathbf{s}$ to transform the problem to

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimise}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h(\mathbf{x}) + \mathbf{s} = \mathbf{b} \\ & \mathbf{s} \geq 0 \end{aligned}$$

(2) Calculate the Lagrangian,

$$L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) + \mathbf{s} - \mathbf{b})$$

(3) Let

$$\boldsymbol{\Lambda} = \left\{\boldsymbol{\lambda} : \inf_{\mathbf{x}\in\mathcal{X};\, \mathbf{s}\geq 0} L(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}) > -\infty\right\}$$

(4) For each $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, find $\mathbf{x}^\star(\boldsymbol{\lambda}), \mathbf{s}^\star(\boldsymbol{\lambda})$ such that

$$\min_{\mathbf{x}\in\mathcal{X};\, \mathbf{s}\geq 0} L(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}) = L(\mathbf{x}^\star(\boldsymbol{\lambda}), \mathbf{s}^\star(\boldsymbol{\lambda}), \boldsymbol{\lambda})$$

(5) Find $\boldsymbol{\lambda}^\star \in \boldsymbol{\Lambda}$ such that $(\mathbf{x}^\star(\boldsymbol{\lambda}), \mathbf{s}^\star(\boldsymbol{\lambda}))$ is feasible, i.e.

$$h(\mathbf{x}^\star(\boldsymbol{\lambda}^\star)) = \mathbf{b}; \quad \mathbf{s}^\star(\boldsymbol{\lambda}^\star) \geq 0$$

3.3. Complementary slackness

In step (4) above, we want to minimise the Lagrangian, i.e.

$$\begin{aligned} &\underset{\mathbf{x}\in\mathcal{X}}{\text{minimise}} && f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b}) - \boldsymbol{\lambda}^\intercal\mathbf{s} \\ &\text{subject to} && \mathbf{s} \geq 0 \end{aligned}$$

Suppose, for a particular value of $\boldsymbol{\lambda}$, that we solve this problem and arrive at $\mathbf{x}^\star(\boldsymbol{\lambda}), \mathbf{s}^\star(\boldsymbol{\lambda})$. Let

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix}$$

If $\lambda_i > 0$, then for some large $\mathbf{s}$ we can make $f \to -\infty$, hence $\boldsymbol{\lambda} \notin \boldsymbol{\Lambda}$. Hence, given $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, we must have $\lambda_i \leq 0$. Now, if $\lambda_i < 0$ for some i, we would want to choose $s_i = 0$ to minimise the increase to the function caused by the slack variable. If $\lambda_i = 0$, then s_i can be chosen arbitrarily since it will have no increase on the value of f. With these choices of s_i, we can make $\boldsymbol{\lambda}^\intercal\mathbf{s} = 0$, thus making the slack variable not impact the value of f. So either

- $h(\mathbf{x})_i = b_i$ and $\lambda_i \leq 0$, or
- $h(\mathbf{x})_i \geq b_i$ and $\lambda_i = 0$.

Alternatively (less precisely),

$$\lambda_i s_i = 0$$

In other words, either the constraint inequality is tight (defined by an equality) and the Lagrange multipliers are slack (defined by an inequality), or the constraint inequality is slack and the Lagrange multipliers are tight.

Example.

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^2}{\text{minimise}} \quad & x_1 - 3x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 \leq 4 \\ & x_1 + x_2 \leq 2 \end{aligned}$$

Adding slack variables, we have

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^2}{\text{minimise}} \quad & x_1 - 3x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 + s_1 = 4 \\ & x_1 + x_2 + s_2 = 2 \\ & s_1 \geq 0 \\ & s_2 \geq 0 \end{aligned}$$

Taking the Lagrangian,

$$\begin{aligned} L(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}) &= (x_1 - 3x_2) - x_1(x_1^2 + x_2^2 + s_1 - 4) - \lambda_2(x_1 + x_2 + s_2 - 2) \\ &= \left((1-\lambda_2)x_1 - \lambda_1 x_1^2\right) + \left((-3-\lambda_2)x_2 - \lambda_1 x_2^2\right) - \lambda_1 s_1 - \lambda_2 s_2 + (4\lambda_1 + 2\lambda_2) \end{aligned}$$

We must have $\lambda_1, \lambda_2 \leq 0$ by considering the slack variable. By complementary slackness,

$$\lambda_1 s_1 = \lambda_2 s_2 = 0 \text{ at the optimum}$$

Minimising each term independently, we have

$$\begin{aligned} 1 - \lambda_2 - 2\lambda_1 x_1 &= 0 \\ -3 - \lambda_2 - 2\lambda_1 x_2 &= 0 \end{aligned}$$

If $\lambda_1 = 0$, the above two equations are contradictory. Hence $\lambda_1 < 0$, giving $s_1 = 0$. If $\lambda_2 < 0$, then $s_2 = 0$ by complementary slackness, so

$$\begin{aligned} 1 - \lambda_2 - 2\lambda_1 x_1 &= 0 \\ -3 - \lambda_2 - 2\lambda_1 x_2 &= 0 \\ x_1^2 + x_2^2 &= 4 \\ x_1 + x_2 &= 2 \end{aligned}$$

Solving the lower two equations give

$$(x_1, x_2) = (0, 2), (2, 0)$$

If $(x_1, x_2) = (0, 2)$, solving the first two equations gives $(\lambda_1, \lambda_2) = (1, -3)$ which is impossible since λ_1 must be negative. Similarly, if $(x_1, x_2) = (2, 0)$, solving the first two equations gives

$(\lambda_1, \lambda_2) = (-1, 1)$ which is impossible again. We have ruled out every case apart from $\lambda_1 < 0, \lambda_2 = 0$. In this case,

$$\begin{aligned} 1 - 2\lambda_1 x_1 &= 0 \\ -3 - 2\lambda_1 x_2 &= 0 \\ x_1^2 + x_2^2 &= 4 \\ x_1 + x_2 + s_2 &= 2 \end{aligned}$$

The first two equations give

$$x_1 = \frac{1}{2\lambda_1}; \quad x_2 = \frac{-3}{2\lambda_1}$$

Substituting into the third equation,

$$\lambda_1^2 = \frac{5}{8} \implies \lambda_1 = -\sqrt{\frac{5}{8}}$$

Hence,

$$(x_1, x_2) = \left(-\sqrt{\frac{2}{5}}, -3\sqrt{\frac{2}{5}}\right)$$

which is feasible using the fourth equation. By Lagrange sufficiency, this is the optimum for the original problem.

3.4. Weak duality

We would like to solve a problem

$$\begin{aligned} \underset{\mathbf{x} \in \mathcal{X}}{\text{minimise}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h(\mathbf{x}) = \mathbf{b} \end{aligned}$$

by constructing the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal (h(\mathbf{x}) - \mathbf{b})$$

We now define the quantity

$$g(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda})$$

Theorem (Weak duality theorem)**.** If $\mathbf{x} \in \mathcal{X}(\mathbf{b})$ and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, then $f(\mathbf{x}) \geq g(\boldsymbol{\lambda})$. In particular,

$$\inf_{\mathbf{x} \in \mathcal{X}(\mathbf{b})} f(\mathbf{x}) \geq \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} g(\boldsymbol{\lambda})$$

Proof.

$$\begin{aligned}\inf_{\mathbf{x}\in\mathcal{X}(\mathbf{b})} f(\mathbf{x}) &= \inf_{\mathbf{x}\in\mathcal{X}(\mathbf{b})} f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b}) \\ &\geq \inf_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b}) \\ &= \inf_{\mathbf{x}\in\mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) \\ &= g(\boldsymbol{\lambda})\end{aligned}$$

□

Using this weak duality property, if $\inf_{\mathbf{x}\in\mathcal{X}(\mathbf{b})} f(\mathbf{x})$ is difficult to solve, we can first attempt $\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} g(\boldsymbol{\lambda})$. The problem

$$\begin{aligned}&\text{maximise} && g(\boldsymbol{\lambda}) \\ &\text{subject to} && \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\end{aligned}$$

is called the *dual problem*. The original is called the *primal problem*. The optimal cost of the primal problem is always greater than or equal to the optimal cost of the dual problem. The *duality gap* is the difference:

$$\inf_{\mathbf{x}\in\mathcal{X}(\mathbf{b})} f(\mathbf{x}) - \sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} g(\boldsymbol{\lambda})$$

If the duality gap is zero, then we say that *strong duality* holds. This strengthens the inequality into an equality.

3.5. Strong duality and the Lagrange method

If the Lagrange method works, then we know that

$$\inf_{\mathbf{x}\in\mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\mathbf{x}\in\mathcal{X}(\mathbf{b})} L(\mathbf{x}, \boldsymbol{\lambda})$$

So, taking such a $\boldsymbol{\lambda}$ in the proof above, we have equality instead of inequality. Hence the problem has strong duality. Conversely, if the duality gap is zero, then there exists a $\boldsymbol{\lambda}$ such that the inequality above is an equality. Hence, for this $\boldsymbol{\lambda}$,

$$\inf_{\mathbf{x}\in\mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\mathbf{x}\in\mathcal{X}(\mathbf{b})} L(\mathbf{x}, \boldsymbol{\lambda})$$

Hence this is the $\boldsymbol{\lambda}$ which will solve the Lagrange method. In summary, strong duality holds exactly when the Lagrange method works.

3.6. Hyperplane condition for strong duality

Definition. A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is said to have a *supporting hyperplane* at a point $\mathbf{b}$ if there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ such that for all $\mathbf{c} \in \mathbb{R}^m$,

$$\phi(\mathbf{c}) \geq \phi(\mathbf{b}) + \boldsymbol{\lambda}^\intercal(\mathbf{c} - \mathbf{b})$$

Pictorially, ϕ has a supporting hyperplane if there is a plane passing through $(\mathbf{b}, \phi(\mathbf{b}))$, where ϕ is always above the plane. This could be, for example, a tangent plane at $\mathbf{b}$.

Definition. We define a function $\phi : \mathbb{R}^m \to \mathbb{R}$ associated with the primal problem by

$$\phi(\mathbf{c}) = \inf_{\mathbf{x} \in \mathcal{X}(\mathbf{c})} f(\mathbf{x})$$

This ϕ can be thought of as the optimal cost of a family of optimisation problems with different functional constraint values $\mathbf{c}$. This is called the *value function*.

Theorem (Strong duality theorem)**.** Strong duality holds if and only if the value function ϕ has a supporting hyperplane at $\mathbf{b}$.

Proof. First, we show that a supporting hyperplane implies strong duality. We have $\boldsymbol{\lambda}$ such that

$$\phi(\mathbf{c}) \geq \phi(\mathbf{b}) + \boldsymbol{\lambda}^\intercal(\mathbf{c} - \mathbf{b})$$

Then, we have

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b}) \\ &= \inf_{\mathbf{c}} \inf_{\mathbf{x} \in \mathcal{X}(\mathbf{c})} f(\mathbf{x}) - \underbrace{\boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{c})}_{\text{zero since we are extremising}} - \boldsymbol{\lambda}^\intercal(\mathbf{c} - \mathbf{b}) \\ &= \inf_{\mathbf{c}} \phi(\mathbf{c}) - \boldsymbol{\lambda}^\intercal(\mathbf{c} - \mathbf{b}) \\ &\geq \phi(\mathbf{b}) \end{aligned}$$

By weak duality, we also have the reverse direction: $g(\boldsymbol{\lambda}) \leq \phi(\mathbf{b})$ Hence, $g(\boldsymbol{\lambda}) = \phi(\mathbf{b})$ and strong duality holds. Conversely, if strong duality holds, we want to show the existence of such a hyperplane. We have We have $\boldsymbol{\lambda}$ such that $g(\boldsymbol{\lambda}) = \phi(\mathbf{b})$. For such a $\boldsymbol{\lambda}$, we have

$$\begin{aligned} \phi(\mathbf{b}) = g(\boldsymbol{\lambda}) &= \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b}) \\ &= \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) - \boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{c}) - \boldsymbol{\lambda}^\intercal(\mathbf{c} - \mathbf{b}) \\ &\leq \phi(\mathbf{c}) - \boldsymbol{\lambda}^\intercal(\mathbf{c} - \mathbf{b}) \end{aligned}$$

The last inequality holds due to weak duality. So $\boldsymbol{\lambda}$ gives a supporting hyperplane. □

3.7. Strong duality and convex functions

We would now like to consider for which problems $\phi(\mathbf{b})$ has a supporting hyperplane. The following theorem is stated without proof.

Theorem. A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is convex if and only if every point $\mathbf{b} \in \mathbb{R}^m$ has a supporting hyperplane.

Now, for which problems do we have a convex value function?

Theorem. Consider a minimisation problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimise}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h(\mathbf{x}) \leq \mathbf{b} \end{aligned}$$

with value function ϕ. Then ϕ is convex if:

(i) $\mathcal{X}$ is convex;

(ii) f is convex;

(iii) h is convex.

This is proven in the example sheets.

3.8. Shadow prices interpretation of Lagrange multipliers

Suppose a factory owner produces n types of products from m types of raw materials. Suppose the owner produces $\mathbf{x} = (x_1, x_2, \dots, x_n)$ products, then the profit is some function $f(\mathbf{x})$. We then create $h_j(\mathbf{x})$ to be the amount of raw material j consumed when making products $\mathbf{x}$. The owner wants to maximise $f(\mathbf{x})$ subject to $h_i(\mathbf{x}) \leq b_i$ where b_i is the maximum amount of raw material i that is available.

Now, suppose a supplier offers some $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m)$ extra raw materials to the factory owner. We would like to calculate how much this $\boldsymbol{\varepsilon}$ is worth. The factory owner will try to maximise this new problem, replacing $\mathbf{b} \mapsto \mathbf{b}+\boldsymbol{\varepsilon}$. For a small enough $\boldsymbol{\varepsilon}$, this can be expressed easily using the value function.

$$\phi(\mathbf{b} + \boldsymbol{\varepsilon}) - \phi(\mathbf{b}) \approx \sum_{j=1}^{m} \frac{\partial \phi}{\partial b_j} \varepsilon_j$$

The quantity $\frac{\partial \phi}{\partial b_j}$ is the price of material j, and $\nabla\phi(\mathbf{b})$ is the vector of prices. These are called the 'shadow prices'; they are hidden to the outside world but depend on the internal state of the factory.

Theorem. If ϕ is differentiable at $\mathbf{b}$ and has a supporting hyperplane given by $\boldsymbol{\lambda}$, then

$$\boldsymbol{\lambda} = \nabla\phi(\mathbf{b})$$

Proof. Let $\mathbf{a} = (a_1, a_2, \dots, a_m)$ be an arbitrary vector. Then from the supporting hyperplane condition, for some small $\delta > 0$ we have

$$\frac{\phi(\mathbf{b} + \delta\mathbf{a})}{\delta} \geq \boldsymbol{\lambda}^\intercal \mathbf{a}$$

Since ϕ is differentiable, the limit can be taken to give

$$\nabla\phi(\mathbf{b}) \cdot \mathbf{a} \geq \boldsymbol{\lambda}^\intercal \mathbf{a}$$

But $\mathbf{a}$ was arbitrary. This can only hold if $\boldsymbol{\lambda} = \nabla\phi(\mathbf{b})$ as required. So the Lagrange multiplier $\boldsymbol{\lambda}$ at $\mathbf{b}$ is equal to the gradient vector of ϕ which is the gradient of partial derivatives and also the vector of shadow prices. □

Suppose that a particular raw material was not used up. Then there is a slack value in the inequality. The shadow price is zero in this instance, since we do not need more of this material. So the corresponding Lagrange multiplier is equal to zero. Conversely, if we are paying something for this material, then we must have used up all of that material. This is exactly the complementary slackness property seen earlier.

There is also an economics interpretation of the dual problem. Such a problem can be seen from the perspective of the raw material seller. This seller charges a certain price $\boldsymbol{\lambda}$ for their raw materials, and then buys the finished product from the factory. The profit of the raw material seller is

$$\underbrace{\boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b})}_{\text{cost of materials}} - \underbrace{f(\mathbf{x})}_{\text{buying products}}$$

For every choice of $\boldsymbol{\lambda}$, the factory owner will try to maximise their profit, that is, find an $\mathbf{x}^\star$ such that we maximise

$$\underbrace{f(\mathbf{x})}_{\text{selling products}} \quad \underbrace{-\boldsymbol{\lambda}^\intercal(h(\mathbf{x}) - \mathbf{b})}_{\text{cost of materials}}$$

4. Linear programming

4.1. Linear programs

A linear program is a specific case of a constrained optimisation problem in which the objective function and all constraints are linear functions. For instance, consider the problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^4}{\text{minimise}} \quad & 2x_1 - x_2 + 4x_3 \\ \text{subject to} \quad & x_1 + x_2 + x_4 \leq 2 \\ & 3x_2 - x_3 = 5 \\ & x_3 + x_4 \geq 3 \\ & x_1 \geq 0 \\ & x_3 \leq 0 \end{aligned}$$

A general linear program is of the form

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimise}} \quad & \mathbf{c}^\intercal \mathbf{x} \\ \text{subject to} \quad & \mathbf{a}_i^\intercal \mathbf{x} \geq b_i, i \in M_1 \\ & \mathbf{a}_i^\intercal \mathbf{x} \leq b_i, i \in M_2 \\ & \mathbf{a}_i^\intercal \mathbf{x} = b_i, i \in M_3 \\ & x_j \geq 0, j \in N_1 \\ & x_j \leq 0, j \in N_2 \end{aligned}$$

Note that we can convert the first inequalities to the other direction by inverting the sign of $\mathbf{a}$. We can convert the 'sign' constraints (the last two constraints) by letting $\mathbf{a}$ be a one-hot vector, thus writing them in terms of the first two inequality types. We call this process *reduction* to an equivalent form. Two linear programs are *equivalent* if any feasible solution for one problem can be converted into a feasible solution for the other, with the same cost. We can reduce any linear problem into the form

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimise}} \quad & \mathbf{c}^\intercal \mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \geq \mathbf{b} \end{aligned}$$

where

$$A = \begin{pmatrix} \cdots & \mathbf{a}_1^\intercal & \cdots \\ \cdots & \mathbf{a}_2^\intercal & \cdots \\ & \vdots & \\ \cdots & \mathbf{a}_m^\intercal & \cdots \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

This is known as the *general form* of a linear programming problem. We could alternatively use a 'less-than' inequality, or simply an equality using a slack variable vector. A linear problem is said to be in *standard form* if it is written as

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimise}} \quad & \mathbf{c}^\intercal\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned}$$

This is a special case of the general form. However, we can always reduce any general-form problem into a standard-form problem. First, we add slack variables to convert the inequality into an equality. Then we can convert each variable x_i into the sum of $x_j^+ - x_j^-$, where $x_j^+, x_j^- \geq 0$. Then, we have the problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimise}} \quad & \mathbf{c}^\intercal(\mathbf{x}^+ - \mathbf{x}^-) \\ \text{subject to} \quad & A(\mathbf{x}^+ - \mathbf{x}^-) = \mathbf{b} \\ & \mathbf{x}^+, \mathbf{x}^- \geq 0 \end{aligned}$$

Then by concatenating the vectors $\mathbf{x}^+, \mathbf{x}^-$ into a larger vector $\mathbf{z} \in [0,\infty)^{2n}$, we have the standard form as required.

4.2. Maximising convex functions

Solving linear programs can be seen as a special case of maximising a convex function, since we can maximise $\mathbf{c}^\intercal\mathbf{x}$. Consider the problem

$$\begin{aligned} \text{minimise} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & x \in C, C \text{ convex} \\ & \mathbf{x} \geq 0 \end{aligned}$$

where f is a convex function. Since C is convex, if $\mathbf{z} = (1-\lambda)\mathbf{x} + \lambda\mathbf{y}$ we have

$$f(\mathbf{z}) \leq (1-\lambda)f(\mathbf{x}) + \lambda f(\mathbf{y}) \leq \max\{f(\mathbf{x}), f(\mathbf{y})\}$$

If we wish to maximise f over C, we might guess that we only need to consider points on the boundary. After all, any point not on the boundary can be written as the weighted average of two points on the boundary. Considering those points will give a greater (or equal) value for f.

Definition. A point $\mathbf{x}$ in a convex set C is an *extreme point* if it cannot be written as a convex combination of two distinct points in C; that is,

$$(1-\delta)\mathbf{y} + \delta\mathbf{z}$$

for $\delta \in (0,1)$ and $\mathbf{y} \neq \mathbf{z}$.

So, more precisely, convex functions on convex sets are maximised at extreme points.

4.3. Basic solutions and basic feasible solutions

Consider a linear problem in standard form.

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimise}} \quad & \mathbf{c}^\intercal \mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned}$$

where $A \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$.

Definition. A vector $\mathbf{x}$ is said to be a *basic solution* if it satisfies $A\mathbf{x} = \mathbf{b}$ (that is, it is a solution) and $\mathbf{x}$ has at most m nonzero entries. If also the nonzero entries are positive, then this is called a *basic feasible solution*, since it lies in the feasible set $\mathbf{x} \geq 0$.

We will start the analysis of basic solutions by making three assumptions (one is defined later).

A: All m rows of A are linearly independent. That is, $\{\mathbf{a}_1^\intercal, \dots, \mathbf{a}_m^\intercal\}$ is a linearly independent set. This assumption can be made without loss of generality since we can simply remove linearly dependent constraints.

B: Every set of m columns of A is linearly independent. That is, any m-subset of the set of columns $\{A_1, \dots, A_n\}$ is a linearly independent set. This can also be made without loss of generality by removing the linearly dependent variables.

To find a basic solution, we will start by choosing the coordinates $B(1), B(2), \dots, B(m)$ to be the indices of $\mathbf{x}$ that are allowed to be nonzero. Now, $A\mathbf{x}$ is

$$\begin{pmatrix} \vdots & & \vdots \\ A_1 & \cdots & A_n \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ x_{B(1)} \\ \vdots \\ x_{B(2)} \\ \vdots \\ x_{B(m)} \\ \vdots \\ 0 \end{pmatrix} = \underbrace{\begin{pmatrix} \vdots & & \vdots \\ A_{B(1)} & \cdots & A_{B(m)} \\ \vdots & & \vdots \end{pmatrix}}_{B} \begin{pmatrix} x_{B(1)} \\ x_{B(2)} \\ \vdots \\ x_{B(m)} \end{pmatrix}$$

By setting $A\mathbf{x} = \mathbf{b}$, using the above assumptions, we can invert the matrix on the left-hand side to get

$$\begin{pmatrix} x_{B(1)} \\ x_{B(2)} \\ \vdots \\ x_{B(m)} \end{pmatrix} = \begin{pmatrix} \vdots & & \vdots \\ A_{B(1)} & \cdots & A_{B(m)} \\ \vdots & & \vdots \end{pmatrix}^{-1} \mathbf{b} = B^{-1}\mathbf{b}$$

We call B the basis matrix. The indices $x_{B(1)}, \dots, x_{B(m)}$ are called the basic variables. The indices $B(i)$ are called the basic indices. The columns $A_{B(i)}$ are called the basic columns. If $B^{-1}\mathbf{b} \geq 0$, we have found a basic feasible solution. We now need to specify one further assumption in order to continue to analyse basic solutions.

C: Every basic solution has *exactly m* nonzero entries. This assumption is known as the non-degeneracy assumption. This assumption cannot be created without loss of generality, but it is far simpler to discuss problems with this assumption met. Throughout this course, we will keep this assumption to be true.

4.4. Extreme points of the feasible set in standard form

Consider a linear program in standard form.

Theorem. $\mathbf{x}$ is an extreme point (of the set $\{\mathbf{x} : A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$) if and only if $\mathbf{x}$ is a basic feasible solution.

Remark. Since a linear program is optimised at extreme points, we only need to consider the basic feasible solutions in order to solve the original problem. We will pick all possible m columns of A (there are $\binom{n}{m}$ such choices) to find all basic solutions. Filter to consider only basic feasible solutions, then evaluate $\mathbf{c}^\intercal\mathbf{x}$ to find the $\mathbf{x}$ which has the least cost. This algorithm will always work, but the amount of choices to evaluate in higher dimensions becomes too inefficient for real-world use.

Proof. First, suppose we know $\mathbf{x}$ is a basic feasible solution, and there exist feasible $\mathbf{y}, \mathbf{z}$ such that $\mathbf{x} = (1-\delta)\mathbf{y} + \delta\mathbf{z}$ and $\delta \in (0,1)$. We know $\mathbf{x} \geq 0$ and $\mathbf{x}$ has at most m nonzero entries. Since $\mathbf{y}, \mathbf{z}$ are positive, then $\mathbf{y}, \mathbf{z}$ must be zero in every index that $\mathbf{x}$ must be zero. Specifically, $y_j = z_j = 0$ for $j \notin \{B(1), \dots, B(m)\}$. Now, we define

$$\mathbf{y}_B = \begin{pmatrix} y_{B(1)} \\ \vdots \\ y_{B(m)} \end{pmatrix}; \quad \mathbf{z}_B = \begin{pmatrix} z_{B(1)} \\ \vdots \\ z_{B(m)} \end{pmatrix}$$

We then have $B\mathbf{y}_B = \mathbf{b}; B\mathbf{z}_B = \mathbf{b}$ because $A\mathbf{y} = A\mathbf{z} = \mathbf{b}$. Hence, $\mathbf{y}_B = \mathbf{z}_B = B^{-1}\mathbf{b}$ and so $\mathbf{x} = \mathbf{y} = \mathbf{z}$.

Conversely, suppose $\mathbf{x}$ is not a basic feasible solution. We wish to show it is not an extreme point. Then $\mathbf{x}$ has an amount of nonzero indices greater than m. Let such indices be $i_1, \dots, i_r$ where $r > m$. Consider the columns $A_{i_1}, \dots, A_{i_r}$. Since the rank of A is only m, these columns form a linearly dependent set. Hence, we can find some weights, not all of which are zero, which give zero when multiplied by the columns.

$$w_{i_1}A_{i_1} + w_{i_2}A_{i_2} + \cdots + w_{i_r}A_{i_r} = 0$$

We now define the vector $\mathbf{w}$ by

$$w_i = \begin{cases} 0 & i \notin \{i_1, \dots, i_r\} \\ w_{i_j} & i = i_j \end{cases}$$

So we have a nonzero vector $\mathbf{w}$ with $A\mathbf{w} = 0$. We can consider the two points $\mathbf{x} \pm \varepsilon\mathbf{w}$, which satisfy $A(\mathbf{x} \pm \varepsilon\mathbf{w}) = 0$. Such perturbed points only change the nonzero indices of $\mathbf{x}$. So we

can find an ε small enough such that both of $\mathbf{x} \pm \varepsilon \mathbf{w}$ are in the feasible set, that is, $\mathbf{x} \pm \varepsilon \mathbf{w} \geq 0$. We therefore can express $\mathbf{x}$ as the midpoint of these two points, hence $\mathbf{x}$ is not an extreme point. □

5. Duality in linear programming

5.1. Strong duality of linear programs

Theorem. If a linear program is bounded and feasible, then strong duality holds.

Proof. This is true since the value function is convex. □

5.2. Duals of linear programs in standard form

Consider a linear program in standard form:

$$\begin{aligned} \text{minimise} \quad & \mathbf{c}^\intercal \mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned}$$

The dual problem is therefore

$$\begin{aligned} \text{maximise} \quad & g(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \in \boldsymbol{\Lambda} \end{aligned}$$

The function g is given by

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\mathbf{x} \geq 0} \mathbf{c}^\intercal \mathbf{x} - \boldsymbol{\lambda}^\intercal (A\mathbf{x} - \mathbf{b}) \\ &= \inf_{\mathbf{x} \geq 0} (\mathbf{c}^\intercal - \boldsymbol{\lambda}^\intercal A)\mathbf{x} + \boldsymbol{\lambda}^\intercal \mathbf{b} \end{aligned}$$

This is only bounded below where $\mathbf{c}^\intercal - \boldsymbol{\lambda}^\intercal A \geq 0$. Hence

$$\boldsymbol{\Lambda} = \{\boldsymbol{\lambda} : \ \boldsymbol{\lambda}^\intercal A \leq \mathbf{c}^\intercal\}$$

Further, the minimum value of g for $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ is $\boldsymbol{\lambda}^\intercal \mathbf{b}$. Therefore, the dual problem is

$$\begin{aligned} \text{maximise} \quad & \boldsymbol{\lambda}^\intercal \mathbf{b} \\ \text{subject to} \quad & \boldsymbol{\lambda}^\intercal A \leq \mathbf{c}^\intercal \end{aligned}$$

The dual of a linear program in standard form is a linear problem, but no longer in standard form.

5.3. Duals of linear programs in general form

Consider a linear program in general form:

$$\begin{aligned} &\text{minimise} && \mathbf{c}^\intercal \mathbf{x} \\ &\text{subject to} && A\mathbf{x} \geq \mathbf{b} \end{aligned}$$

We can introduce a slack variable $\mathbf{s}$ and write equivalently

$$\begin{aligned} &\text{minimise} && \mathbf{c}^\intercal \mathbf{x} \\ &\text{subject to} && A\mathbf{x} - \mathbf{s} = \mathbf{b} \\ & && \mathbf{s} \geq 0 \end{aligned}$$

To calculate the dual, we need to calculate $g(\boldsymbol{\lambda})$.

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\mathbf{x},\mathbf{s}\geq 0} \mathbf{c}^\intercal \mathbf{x} - \boldsymbol{\lambda}^\intercal (A\mathbf{x} - \mathbf{s} - \mathbf{b}) \\ &= \inf_{\mathbf{x},\mathbf{s}\geq 0} (\mathbf{c}^\intercal - \boldsymbol{\lambda}^\intercal A)\mathbf{x} + \boldsymbol{\lambda}^\intercal \mathbf{s} + \boldsymbol{\lambda}^\intercal \mathbf{b} \end{aligned}$$

In this case, since $\mathbf{x}$ may be any value, we must have $\mathbf{c}^\intercal - \boldsymbol{\lambda}^\intercal A = 0$. Further, since the slack variable can be any positive value, $\boldsymbol{\lambda}^\intercal \geq 0$. The infimum is $\boldsymbol{\lambda}^\intercal \mathbf{b}$ since $\mathbf{s}$ may be set to zero. Thus, the dual is

$$\begin{aligned} &\text{maximise} && \boldsymbol{\lambda}^\intercal \mathbf{b} \\ &\text{subject to} && \boldsymbol{\lambda}^\intercal A = \mathbf{c}^\intercal \\ & && \boldsymbol{\lambda} \geq 0 \end{aligned}$$

The dual of a general linear program is a linear program in standard form.

5.4. Dual of dual program

The dual of a dual problem is the primal problem. Suppose the primal problem is in standard form:

$$\begin{aligned} &\text{minimise} && \mathbf{c}^\intercal \mathbf{x} \\ &\text{subject to} && A\mathbf{x} = \mathbf{b} \\ & && \mathbf{x} \geq 0 \end{aligned}$$

We know the dual is

$$\begin{aligned} &\text{maximise} && \boldsymbol{\lambda}^\intercal \mathbf{b} \\ &\text{subject to} && \boldsymbol{\lambda}^\intercal A \leq \mathbf{c}^\intercal \end{aligned}$$

Equivalently,

$$\begin{array}{ll} -\text{ minimise} & -\boldsymbol{\lambda}^\intercal \mathbf{b} \\ \text{subject to} & -\boldsymbol{\lambda}^\intercal A \geq -\mathbf{c}^\intercal \end{array}$$

Defining $\tilde{\boldsymbol{\lambda}} = -\boldsymbol{\lambda}^\intercal$, we have

$$\begin{array}{ll} -\text{ minimise} & \tilde{\boldsymbol{\lambda}} \mathbf{b} \\ \text{subject to} & \tilde{\boldsymbol{\lambda}} A \geq -\mathbf{c}^\intercal \end{array}$$

We can find the dual of this problem using the solution above.

$$\begin{array}{ll} -\text{ maximise} & -\theta^\intercal \mathbf{c} \\ \text{subject to} & \theta^\intercal A^\intercal = \mathbf{b}^\intercal \\ & \theta \geq 0 \end{array}$$

This is equivalent to the primal problem.

5.5. Dual of arbitrary linear program

Consider the problem

$$\begin{array}{lll} \text{minimise} & \mathbf{c}^\intercal \mathbf{x} & \\ \text{subject to} & \mathbf{a}_i^\intercal \mathbf{x} \geq \mathbf{b}_i & i \in M_1 \\ & \mathbf{a}_i^\intercal \mathbf{x} \leq \mathbf{b}_i & i \in M_2 \\ & \mathbf{a}_i^\intercal \mathbf{x} = \mathbf{b}_i & i \in M_3 \\ & x_j \geq 0 & j \in N_1 \\ & x_j \leq 0 & j \in N_2 \\ & x_j \text{ free} & j \in N_3 \end{array}$$

The dual of this problem is

$$\begin{array}{lll} \text{maximise} & \mathbf{p}^\intercal \mathbf{b} & \\ \text{subject to} & p_i \geq 0 & i \in M_1 \\ & p_i \leq 0 & i \in M_2 \\ & p_i \text{ free} & i \in M_3 \\ & \mathbf{p}^\intercal \mathbf{A}_j \leq \mathbf{c}_j & j \in N_1 \\ & \mathbf{p}^\intercal \mathbf{A}_j \geq \mathbf{c}_j & j \in N_2 \\ & \mathbf{p}^\intercal \mathbf{A}_j = \mathbf{c}_j & j \in N_3 \end{array}$$

This will be shown in the example sheets.

5.6. Optimality conditions

If $\mathbf{x}$ is feasible for the primal, $\mathbf{p}$ is feasible for the dual, and complementary slackness holds, then $\mathbf{x}$ is optimal for the primal and $\mathbf{p}$ is optimal for the dual.

Theorem (Fundamental Theorem of Linear Programming)**.** Let $\mathbf{x}, \mathbf{p}$ be feasible solutions to the primal and dual problems respectively. Then $\mathbf{x}, \mathbf{p}$ are optimal for these problems if and only if

- $p_i(\mathbf{a}_i^\intercal \mathbf{x} - b_i) = 0$ for all i, and
- $(c_j - \mathbf{p}^\intercal \mathbf{A}_j)x_j = 0$ for all j.

Proof. First, let us define $u_i = p_i(\mathbf{a}_i^\intercal \mathbf{x} - b_i)$ and $v_j = (c_j - \mathbf{p}^\intercal \mathbf{A}_j)x_j$. Observe that if $\mathbf{x}, \mathbf{p}$ are feasible, then $u_i \geq 0$ for all i, and $v_j \geq 0$ for all j. This can be seen by the signs of the constraints on the primal and dual problems. Now,

$$\sum u_i = \sum p_i(\mathbf{a}_i^\intercal \mathbf{x} - b_i) = \mathbf{p}^\intercal A\mathbf{x} - \mathbf{p}^\intercal \mathbf{b}$$

Similarly,

$$\sum v_j = \sum (c_j - \mathbf{p}^\intercal \mathbf{A}_j)x_j = \mathbf{c}^\intercal \mathbf{x} - \mathbf{p}^\intercal A\mathbf{x}$$

Then,

$$\sum u_i + \sum v_j = \mathbf{c}^\intercal \mathbf{x} - \mathbf{p}^\intercal \mathbf{b}$$

which is the difference between the two objective functions in the primal and the dual. Hence,

$$0 \leq \sum u_i + \sum v_j = \mathbf{c}^\intercal \mathbf{x} - \mathbf{p}^\intercal \mathbf{b}$$

So if complementary slackness holds, then $u_i = 0$ and $v_j = 0$ for all i, j. This then implies that $\mathbf{c}^\intercal \mathbf{x} = \mathbf{p}^\intercal \mathbf{b}$. By weak duality, $\mathbf{x}$ and $\mathbf{p}$ must be optimal. Conversely, suppose $\mathbf{x}, \mathbf{p}$ are optimal. By strong duality, $\mathbf{c}^\intercal \mathbf{x} = \mathbf{p}^\intercal \mathbf{b}$.

$$0 \leq \sum u_i + \sum v_j = \mathbf{c}^\intercal \mathbf{x} - \mathbf{p}^\intercal \mathbf{b} = 0$$

Thus $\sum u_i + \sum v_j = 0$. Since all u_i, v_j are non-negative, $u_i = 0$ and $v_j = 0$ for all i, j. Equivalently, complementary slackness holds. □

6. Simplex method

6.1. Introduction

Consider the problem

$$\begin{array}{ll} \text{minimise} & \mathbf{c}^\intercal \mathbf{x} \\ \text{subject to} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array}$$

The dual problem is

$$\begin{array}{ll} \text{maximise} & \boldsymbol{\lambda}^\intercal \mathbf{b} \\ \text{subject to} & \boldsymbol{\lambda}^\intercal A \leq \mathbf{c}^\intercal \end{array}$$

The optimality conditions are

- (primal feasibility) $A\mathbf{x} = \mathbf{b}; \mathbf{x} \geq 0$
- (dual feasibility) $A^\intercal \boldsymbol{\lambda} \leq \mathbf{c}$
- (complementary slackness) $\mathbf{x}^\intercal(\mathbf{c} - A^\intercal \boldsymbol{\lambda}) = 0$

Suppose $\mathbf{x}$ is a basic feasible solution given by

$$\mathbf{x}_B = (x_{B(1)}, \dots, x_{B(m)})$$

Substituting this $\mathbf{x}$ into the complementary slackness equation gives

$$\mathbf{x}_B^\intercal \mathbf{c}_B - \mathbf{x}_B^\intercal B^\intercal \boldsymbol{\lambda} = 0 \implies \mathbf{x}_B^\intercal (\mathbf{c}_B - B^\intercal \boldsymbol{\lambda}) = 0$$

For a basic feasible solution, $\mathbf{x}_B > 0$. Hence,

$$\mathbf{c}_B - B^\intercal \boldsymbol{\lambda} = 0$$

Hence

$$\boldsymbol{\lambda} = (B^\intercal)^{-1} \mathbf{c}_B$$

So for this $\mathbf{x}$ and this calculated $\boldsymbol{\lambda}$, primal feasibility and complementary slackness both hold. What remains now is to check if dual feasibility holds. Equivalently,

$$A^\intercal \boldsymbol{\lambda} \leq \mathbf{c} \implies A^\intercal (B^\intercal)^{-1} \mathbf{c}_B \leq \mathbf{c}$$

If this holds, then the optimality conditions are met. This means that we do not even need to explicitly find $\boldsymbol{\lambda}$ in order to check optimality; it suffices to check whether this single inequality holds. We define

$$\overline{\mathbf{c}} = \mathbf{c} - A^\intercal (B^\intercal)^{-1} \mathbf{c}_B$$

This is called the *vector of reduced costs*. Then the inequality $\overline{\mathbf{c}} \geq 0$ implies $\mathbf{x}$ is optimal.

6.2. Feasibility of basic directions

Definition. Let $P = \{\mathbf{x} : A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$ be the feasible set of a problem in standard form. Further, let $\mathbf{x} \in P$. A vector $\mathbf{d} \in \mathbb{R}^n$ is called a *feasible direction* if there exists $\theta > 0$ such that $\mathbf{x} + \theta\mathbf{d} \in P$.

Let $\mathbf{x}$ be a basic feasible solution. Let $B(1), \dots, B(m)$ be the indices of the basic variables, and let B be the basis matrix $\left[A_{B(1)}, \dots, A_{B(m)}\right]$. Let $\mathbf{x}_B = \left(x_{B(1)}, \dots, x_{B(m)}\right)^\intercal$. Suppose we move in a direction $\mathbf{d}$ such that $d_j = 1$, and $d_i = 0$ for all non-basic $i \neq j$, or more explicitly $i \in \{1, 2, \dots, n\} \setminus \{B(1), \dots, B(m), j\}$. This direction $\mathbf{d}$ is called the jth *basic direction*, since it moves in the direction of the jth basic variable. Note that we can write

$$\mathbf{d} = (d_{B(1)}, \dots, d_{B(m)}, 0, 0, \dots, \underbrace{1}_{j\text{th entry}}, \dots, 0, 0)$$

When we move in this direction, we want to move to a feasible point. This means that we require

$$\begin{aligned} A(\mathbf{x} + \theta\mathbf{d}) &= \mathbf{b} \\ A\mathbf{d} &= 0 \\ B\mathbf{d}_B + A_j &= 0 \\ \mathbf{d}_B &= -B^{-1}A_j \end{aligned}$$

For the positivity condition, note that

$$\mathbf{x} + \theta\mathbf{d} = (x_{B(1)} + \theta d_{B(1)}, \dots, x_{B(m)} + \theta d_{B(m)}, 0, 0, \dots, \underbrace{\theta}_{j\text{th entry}}, \dots, 0, 0)$$

For this $\mathbf{x}$ to be feasible, all x_i must be non-negative. Since $x_{B(i)} > 0$, there exists a small enough θ such that $\mathbf{x} + \theta\mathbf{d} \geq 0$. Hence, the jth basic direction is feasible.

6.3. Cost of basic directions

How does the cost change when $\mathbf{x} \mapsto \mathbf{x} + \theta\mathbf{d}$ where $\mathbf{d}$ is the (feasible) jth basic direction? The new cost is

$$\begin{aligned} \mathbf{c}^\intercal(\mathbf{x} + \theta\mathbf{d}) &= \mathbf{c}^\intercal(\mathbf{x} + \theta(-B^{-1}A_j)) \\ &= \mathbf{c}^\intercal\mathbf{x} + \theta(c_j - \mathbf{c}_B^\intercal B^{-1}A_j) \\ &= \mathbf{c}^\intercal\mathbf{x} + \theta\overline{c_j} \end{aligned}$$

Theorem. Let $\mathbf{x}$ be a basic feasible solution associated with a basis matrix B, and let $\overline{\mathbf{c}}$ be the vector of reduced costs. Then $\mathbf{x}$ is optimal if and only if $\overline{\mathbf{c}} \geq 0$.

Proof. This follows from the optimality conditions given previously. □

Now, if $\overline{c_j} \geq 0$ for all j, then this is an optimal solution. However, if any $\overline{c_j} < 0$, then we can move in the jth direction and decrease the cost.

6.4. Moving to basic feasible solutions

Suppose $\mathbf{x}$ is a basic feasible solution. If $\overline{\mathbf{c}} \geq 0$, then this is the optimum and we can stop. If $c_j < 0$ for some j, then moving in the jth feasible direction will reduce the cost by $\theta\overline{c_j}$. The amount by which the cost decreases is proportional to θ, so we should choose the largest possible value of θ while retaining feasibility. We denote this largest θ with $\theta^\star$. There are two cases:

- If $\mathbf{d} \geq 0$, then θ is unbounded since $\mathbf{x} + \theta\mathbf{d} \geq 0$ for all $\theta > 0$. Therefore the optimal cost of this problem is $-\infty$.
- If $d_i < 0$ for some i, then we need $x_i + \theta d_i \geq 0$, so $\theta^\star \leq -\frac{x_i}{d_i}$. This then gives

$$\theta^\star = \min_{\{i\,:\,d_i<0\}} -\frac{x_i}{d_i}$$

  or equivalently,

$$\theta^\star = \min_{\left\{i\in\{1,\ldots,m\}\,:\,d_{B(i)<0}\right\}} -\frac{x_{B(i)}}{d_{B(i)}}$$

Suppose the optimal cost is bounded. Let ℓ be the index minimising $\theta^\star$, so

$$\theta^\star = -\frac{x_{B(\ell)}}{d_{B(\ell)}}$$

Now, let us move in this direction by this amount.

Theorem. Let $\mathbf{y} = \mathbf{x} + \theta^\star\mathbf{d}$. $\mathbf{y}$ is feasible, and $\mathbf{c}^\intercal\mathbf{y} < \mathbf{c}^\intercal\mathbf{x}$. Then, $\mathbf{y}$ is a basic feasible solution with basis matrix

$$\overline{B} = \begin{pmatrix} \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{B(1)} & \cdots & A_{B(\ell-1)} & A_j & A_{B(\ell+1)} & \cdots & A_{B(m)} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \end{pmatrix}$$

Proof. We know that $\mathbf{y}$ has exactly m nonzero entries, since $y_{B(\ell)} = 0$. We know it is feasible, hence $\mathbf{y}$ is a basic feasible solution. j becomes a basic variable and $B(\ell)$ is no longer. □

6.5. Simplex method

Algorithm 3: Simplex Method

Result: Global minimum of $\mathbf{c}^\intercal \mathbf{x}$
start at a basic feasible solution $\mathbf{x}$ with basis matrix $B = [A_{B(1)}, \dots, A_{B(m)}]$;
repeat
  choose j such that $\overline{c_j} < 0$;
  $\mathbf{u} \leftarrow -B^{-1}A_j$;
  if $\mathbf{u} \leq 0$ **then** cost is $-\infty$ so terminate algorithm;
  $\theta^\star \leftarrow \min \frac{x_{B(i)}}{u_i}$ where $i \in \{1, \dots, m\}$ and $u_i > 0$;
  $\ell \leftarrow$ an index i from the step above that gives the minimal value of $\theta^\star$;
  $\mathbf{x} \leftarrow \mathbf{x} - \theta^\star \mathbf{u}$;
until $\overline{\mathbf{c}} \geq 0$;
since $\overline{\mathbf{c}} \geq 0$, $\mathbf{x}$ is optimal

6.6. Tableau implementation

The full tableau implementation of the simplex method is a convenient way of executing the simplex algorithm without excessive computation. A simplex tableau contains four values of information:

$-\mathbf{c}_B^\intercal$	$\overline{\mathbf{c}}$
$B^{-1}\mathbf{b}$	$B^{-1}A$

The information is essentially

$-$cost	reduced costs
vector to generate current basic feasible solution	matrix to generate basic directions

In a more detailed form,

$-\mathbf{c}_B^\intercal$	$\overline{c_1}$	$\overline{c_2}$	$\cdots$	$\overline{c_n}$
$x_{B(1)}$	$\vdots$	$\vdots$		$\vdots$
$\vdots$	$B^{-1}A_1$	$B^{-1}A_2$	$\cdots$	$B^{-1}A_n$
$x_{B(m)}$	$\vdots$	$\vdots$		$\vdots$

To execute the simplex algorithm using this table, use the following algorithm.

Algorithm 4: Simplex Method (Tableau Implementation)

Result: Global minimum of $\mathbf{c}^\intercal\mathbf{x}$
start at a basic feasible solution $\mathbf{x}$ with basis matrix $B = [A_{B(1)}, \dots, A_{B(m)}]$;
repeat
  choose j such that $\overline{c_j} < 0$;
  $\mathbf{u} \leftarrow -B^{-1}A_j$;
  if $\mathbf{u} \leq 0$ **then** cost is $-\infty$ so terminate algorithm;
  $\theta^\star \leftarrow \min \frac{x_{B(i)}}{u_i}$ where $i \in \{1, \dots, m\}$ and $u_i > 0$;
  $\ell \leftarrow$ an index i from the step above that gives the minimal value of $\theta^\star$;
  $(*)$ add to each row of the tableau a constant multiple of the ℓth row so that u_ℓ becomes 1 and all other entries of the pivot column are 0;
until $\overline{\mathbf{c}} \geq 0$ *(when all entries in the 0th row are non-negative)*;
since $\overline{\mathbf{c}} \geq 0$, $\mathbf{x}$ is optimal

This is just the same as the simplex method discussed before, apart from step $(*)$. No proof will be given for why this step achieves the same result as the full simplex algorithm.

Example. Consider the problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^3}{\text{minimise}} \quad & -x_1 - x_2 - x_3 \\ \text{subject to} \quad & x_1 + 2x_2 + 2x_3 \leq 10 \\ & 2x_1 + x_2 + 2x_3 \leq 10 \\ & 2x_1 + 2x_2 + x_3 \leq 20 \\ & x_1, x_2, x_3 \geq 0 \end{aligned}$$

By introducing slack variables, we can write this in standard form.

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^6}{\text{minimise}} \quad & -x_1 - x_2 - x_3 \\ \text{subject to} \quad & x_1 + 2x_2 + 2x_3 + x_4 = 10 \\ & 2x_1 + x_2 + 2x_3 + x_5 = 10 \\ & 2x_1 + 2x_2 + x_3 + x_6 = 20 \\ & x_1, x_2, x_3, x_4, x_5, x_6 \geq 0 \end{aligned}$$

Observe that $(0, 0, 0, 10, 10, 20)$ is a basic feasible solution. We will use this to initiate the simplex algorithm. The corresponding basis matrix is the 3×3 identity matrix. We construct the simplex tableau by first constructing the 0th row:

- $\mathbf{c}_B = 0$ hence $\mathbf{c}_B^\intercal \mathbf{x}_B = 0$.
- $\overline{\mathbf{c}} = \mathbf{c}$.

We construct the tableau as follows.

0	-1	-1	-1	0	0	0
10	1	2	2	1	0	0
10	2	1	2	0	1	0
20	2	2	1	0	0	1

$\overline{c_1} < 0$, so we will descend in the 1st basic direction. Consider $\frac{10}{1}, \frac{10}{2}, \frac{20}{2}$. The smallest is $\frac{10}{2} = 5$, so the *favourite element* is the number 2 in the 1st column and 2nd row. We want to change this column to $(0, 0, 1, 0)^\intercal$ by using row operations. Denoting the rows as $R_0, \dots, R_3$, we want to perform the operations

$$\begin{aligned} R_0 &\mapsto R_0 + \frac{1}{2}R_2 \\ R_1 &\mapsto R_1 - \frac{1}{2}R_2 \\ R_2 &\mapsto \frac{1}{2}R_2 \\ R_3 &\mapsto R_3 - R_2 \end{aligned}$$

The tableau now looks like this.

5	0	−0.5	0	0	0.5	0
5	0	1.5	1	1	−0.5	0
5	1	0.5	1	0	0.5	0
10	0	1	−1	0	−1	1

Now, $\overline{c_2} < 0$, so we will descend in the 2nd basic direction. Consider $\frac{5}{1.5}, \frac{5}{0.5}, \frac{10}{1}$. The smallest is $\frac{5}{1.5}$, so the favourite element is the 1.5 in the 1st row and 2nd column. To make the column a one-hot vector, we perform

$$\begin{aligned} R_0 &\mapsto R_0 + \frac{1}{3}R_1 \\ R_1 &\mapsto \frac{2}{3}R_1 \\ R_2 &\mapsto R_2 - \frac{1}{3}R_1 \\ R_3 &\mapsto R_3 - \frac{2}{3}R_1 \end{aligned}$$

This yields

$\frac{20}{3}$	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
$\frac{10}{3}$	0	1	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{3}{4}$	0
$\frac{10}{3}$	1	0	$\frac{2}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	0
$\frac{20}{3}$	0	0	$-\frac{5}{3}$	$-\frac{2}{3}$	$-\frac{2}{3}$	1

Now, the 0th row has no negative values, so we are at the optimum. The optimal cost therefore is $-\frac{20}{3}$. The solution is at $\left(\frac{10}{3}, \frac{10}{3}, 0, 0, 0, \frac{20}{3}\right)$.

7. Game theory

7.1. Zero-sum games

Definition. A *zero-sum two-person game* is a scenario in which two players (denoted P1 and P2) have different actions they can take:

- P1 has m possible actions $\{1, 2, \dots, m\}$, and
- P2 has n possible actions $\{1, 2, \dots, n\}$; such that

if P1 plays move i and P2 plays move j, then we say P1 'wins' an amount a_{ij} and P2 'loses' the same amount a_{ij}. The matrix of results A is called the *payoff matrix*. P1 chooses a row of the matrix, and P2 chooses a column, and the intersection is the outcome of the game.

Suppose P1 plays first, and chooses row i. P1 knows that P2 will choose the column j such that a_{ij} is minimised, since that will maximise P2's winnings. In particular, if P1 picks row i then they can expect to win $\min_{j\in\{1,\dots,m\}} a_{ij}$. So P1 will try to solve the problem

$$\begin{aligned} &\text{maximise} && \min_{j\in\{1,\dots,m\}} a_{ij} \\ &\text{subject to} && i \in \{1, \dots, n\} \end{aligned}$$

If P2 plays first, they will try to solve the problem

$$\begin{aligned} &\text{minimise} && \max_{i\in\{1,\dots,n\}} a_{ij} \\ &\text{subject to} && j \in \{1, \dots, m\} \end{aligned}$$

Example. Suppose the payoff matrix is

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

P1 chooses a row, and P2 chooses a column. If P1 plays first, they choose row 2, then P2 chooses row 1, and the payoff is 3. If P2 plays first, they choose column 1, then P1 chooses row 2, and the payoff is again 3. Since the solution is the same for both problems, this point $(2, 1)$ is called a *saddle point*. The value $a_{21} = 3$ is called the *value* of the game.

Example. Consider the payoff matrix

$$A = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$$

If P1 plays first, they choose row 1, then P2 chooses column 2, and the payoff is 2. If P2 plays first, they choose column 2, then P1 chooses row 2, and the payoff is 3. Here, both players cannot play optimally simultaneously since different outcomes will occur depending on what they think their opponent will do.

7.2. Mixed strategies

In a mixed strategy, the players are allowed to choose their action randomly. Such mixed strategies are employed when we do not know what our opponent will pick; for example, when both players choose their option at the same time. P1 picks action i with probability p_i, and P2 picks action j with probability q_j, such that $\sum p_i = \sum q_j = 1$. Now, a player's strategy is encoded as a probability vector. If P1 picks the mixed strategy $(p_1, \dots, p_m)$, then the expected reward of P1 (if P2 picks a pure strategy j) is

$$\sum_i a_{ij} p_i$$

The optimisation problem for P1 is

$$\begin{aligned} \text{maximise} \quad & \min_{j \in \{1,\dots,n\}} \sum_i a_{ij} p_i \\ \text{subject to} \quad & \sum p_i = 1 \\ & \mathbf{p} \geq 0 \end{aligned}$$

Equivalently, where $\mathbf{e} = (1, 1, \dots, 1)^\intercal$,

$$\begin{aligned} \text{maximise} \quad & v \\ \text{subject to} \quad & A^\intercal \mathbf{p} \geq v\mathbf{e} \\ & \mathbf{e}^\intercal \mathbf{p} = 1 \\ & \mathbf{p} \geq 0 \end{aligned}$$

This v is the minimum value of $A^\intercal \mathbf{p}$. P2's optimisation problem is

$$\begin{aligned} \text{minimise} \quad & \max_{i \in \{1,\dots,m\}} \sum_i a_{ij} q_j \\ \text{subject to} \quad & \sum q_j = 1 \\ & \mathbf{q} \geq 0 \end{aligned}$$

or equivalently,

$$\begin{aligned} \text{minimise} \quad & w \\ \text{subject to} \quad & A\mathbf{q} \leq w\mathbf{e} \\ & \mathbf{e}^\intercal \mathbf{q} = 1 \\ & \mathbf{q} \geq 0 \end{aligned}$$

7.3. Duality of mixed strategy problems

The two problems above are duals of each other. Adding slack variables, P2's problem is

$$\begin{array}{ll} \text{minimise} & w \\ \text{subject to} & A\mathbf{q} + \mathbf{s} = w\mathbf{e} \\ & \mathbf{e}^\intercal \mathbf{q} = 1 \\ & \mathbf{q} \geq 0 \\ & \mathbf{s} \geq 0 \end{array}$$

The Lagrangian of this problem is

$$\begin{aligned} L(w, \mathbf{q}, \mathbf{s}, \boldsymbol{\lambda}_1, \lambda_2) &= w + \boldsymbol{\lambda}_1^\intercal (A\mathbf{q} + \mathbf{s} - w\mathbf{e}) - \lambda_2(\mathbf{e}^\intercal \mathbf{q} - 1) \\ &= w(1 - \boldsymbol{\lambda}_1^\intercal \mathbf{e}) + (\boldsymbol{\lambda}_1^\intercal A - \lambda_2 \mathbf{e}^\intercal)\mathbf{q} + \boldsymbol{\lambda}_1^\intercal \mathbf{s} + \lambda_2 \end{aligned}$$

Thus,

$$\boldsymbol{\Lambda} = \left\{ \boldsymbol{\lambda} : \ \boldsymbol{\lambda}_1^\intercal \mathbf{e} = 1, \boldsymbol{\lambda}_1^\intercal A - \lambda_2 \mathbf{e}^\intercal \geq 0, \boldsymbol{\lambda}_1 \geq 0 \right\}$$

When $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$,

$$\inf L = \lambda_2$$

Hence the dual is

$$\begin{array}{ll} \text{maximise} & \lambda_2 \\ \text{subject to} & \boldsymbol{\lambda}_1^\intercal \mathbf{e} = 1 \\ & \boldsymbol{\lambda}_1^\intercal A \geq \lambda_2 \mathbf{e}^\intercal \\ & \boldsymbol{\lambda}_1 \geq 0 \end{array}$$

Note that $\boldsymbol{\lambda}_1 = \mathbf{p}$ and $\lambda_2 = v$ in the above formulation of P1's problem.

Theorem. A strategy $\mathbf{p}$ is optimal for P1 if there exist $\mathbf{q}, v$ such that

- (primal feasibility) $A^\intercal \mathbf{p} \geq v\mathbf{e}, \mathbf{e}^\intercal \mathbf{p} = 1, \mathbf{p} \geq 0$;
- (dual feasibility) $A\mathbf{q} \leq v\mathbf{e}, \mathbf{e}^\intercal \mathbf{q} = 1, \mathbf{q} \geq 0$; and
- (complementary slackness) $v = \mathbf{p}^\intercal A \mathbf{q}$

Proof. $(\mathbf{p}, v)$ and $(\mathbf{q}, w)$ are optimal if

$$(A\mathbf{q} - w\mathbf{e})^\intercal \mathbf{p} = 0; \mathbf{q}^\intercal (A^\intercal \mathbf{p} - v\mathbf{e}) = 0$$

which gives

$$v = w = \mathbf{p}^\intercal A \mathbf{q}$$

□

7.4. Finding optimal strategies

There are a number of strategies for finding optimal strategies.

(i) We can search for saddle points in the payoff matrix. If such a saddle point is found, a pure strategy aiming for this saddle point is optimal for both players.

(ii) We can search for *dominating actions*. Suppose there exist i, i' such that $a_{ij} \geq a_{i'j}$ for all j. Then i dominates i', so P1 will never play i' and we can simply drop this row in the matrix. A similar technique can be used to drop columns.

(iii) If these simplification techniques are not sufficient, we can simply solve the linear program using (for instance) the simplex method.

Example. Suppose we have a payoff matrix

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 3 & 1 & \frac{1}{2} \\ 1 & 3 & 2 \end{pmatrix}$$

First, observe that there is no saddle point. Note that the first row dominates the last row, so we can simplify the payoff matrix to

$$\widetilde{A} = \begin{pmatrix} 2 & 3 & 4 \\ 3 & 1 & \frac{1}{2} \end{pmatrix}$$

P1's strategy is $\mathbf{p} = (p, 1-p, 0)$, and the optimisation problem is

$$\begin{array}{ll} \text{maximise} & v \\ \text{subject to} & A^\intercal \mathbf{p} \geq v\mathbf{e} \\ & \mathbf{e}^\intercal \mathbf{p} = 1 \\ & \mathbf{p} \geq 0 \end{array}$$

which is

$$\begin{array}{ll} \text{maximise} & v \\ \text{subject to} & 2p + 3(1-p) \geq v \\ & 3p + (1-p) \geq v \\ & 4p + \frac{1}{2}(1-p) \geq v \\ & 0 \leq p \leq 1 \end{array}$$

and by simplifying,

$$\begin{aligned} \text{maximise} \quad & v \\ \text{subject to} \quad & v \leq 3 - p \\ & v \leq 1 + 2p \\ & v \leq \frac{1}{2} + \frac{7}{2}p \\ & 0 \leq p \leq 1 \end{aligned}$$

We can solve this graphically since it is a one-dimensional problem, or use the simplex method. We arrive at the solution $\mathbf{p} = \left(\frac{2}{3}, \frac{1}{3}, 0\right)$, i.e. $p = \frac{2}{3}$. The payoff is $\frac{7}{3}$. Player 2 has the dual optimisation problem, so we can use complementary slackness to compute P2's strategy. The first two constraints are tight, but the final constraint may not be (since it is zero in P1's strategy). Therefore $q_3 = 0$, and P2's strategy is $\mathbf{q} = (q, 1 - q, 0)$. Since the value of the game is $\frac{7}{3}$, we have

$$\frac{7}{3} = \mathbf{p}^\intercal A \mathbf{q}$$

which lets us find q. Alternatively, we can use complementary slackness. Since $p_1, p_2 > 0$, the first two constraints in the dual problem must be tight.

$$2q + 3(1 - q) = \frac{7}{3} \implies q = \frac{2}{3}$$

8. Network flows

8.1. Minimum cost flow

Definition. A *directed graph* (also known as a *digraph*) G consists of a set of vertices and a set of edges; $G = (V, E)$. The edges are such that $E \subseteq V \times V$. Each edge (i, j) can be thought of as an edge pointing from vertex i to vertex j. When E is symmetric (that is, $(i, j) \in E \iff (j, i) \in E$), we call G an *undirected graph*.

Definition. Given a graph $G = (V, E)$ on n vertices, we associate to every $(i, j) \in E$ the number x_{ij}. This represents the flow of a quantity from vertex i to vertex j. The collection x of x_{ij} is called the *flow*. The flow x is affected by

(i) A vector $\mathbf{b} \in \mathbb{R}^n$, where b_i is the amount of flow entering vertex i from outside the graph. If $b_i > 0$, then vertex i is called a *source*. If $b_i < 0$, then vertex i is called a *sink*.

(ii) The cost matrix $c \in \mathbb{R}^{n \times n}$, which gives the cost c_{ij} per unit of flow on $(i, j) \in E$. If the flow along (i, j) is x_{ij}, the cost for this flow is $c_{ij}x_{ij}$ (without the summation convention).

(iii) The lower bound matrix $\underline{M}$ and the upper bound matrix $\overline{M}$, which give lower and upper bounds on x_{ij}. In particular, for all $(i, j) \in E$, we require $\underline{m}_{ij} \leq x_{ij} \leq \overline{m}_{ij}$.

Definition. The *minimum cost flow* is the linear program

$$\begin{aligned} \text{minimise} \quad & \sum_{(i,j)\in E} c_{ij}x_{ij} \\ \text{subject to} \quad & \underline{m}_{ij} \leq x_{ij} \leq \overline{m}_{ij} \quad \forall (i, j) \in E \\ & b_i + \sum_{(j,i)\in E} x_{ji} = \sum_{(i,j)\in E} x_{ij} \quad \forall i \in V \end{aligned}$$

The second constraint is a conservation of flow equation. The amount of flow entering and leaving the vertex must be equal. Note that in order for the problem to be feasible, $\sum b_i = 0$; since the graph has no storage capacity at any vertex, the amount of flow that enters the graph must be the amount of flow that exits. Alternatively, we could prove this by finding the sum of the conservation of flow equations for all i.

Definition. We can define the *incidence matrix* $A : \mathbb{R}^{|V|\times|E|}$. Each column of A is associated with an edge (i, j). We define that this column is filled with zeroes, except for $+1$ at position i and -1 at position j. We can now rewrite the conservation of flow equation as

$$A\mathbf{x} = \mathbf{b}$$

8.2. Transport problem

The transport problem is a special case of the minimum cost flow problem. Consider n suppliers, and m consumers. Each supplier i has some capacity s_i for how much of this good

they can satisfy, and each consumer j has some demand d_j that they want to be fulfilled. We will assume that there is exactly as much supply as demand; that is, $\sum s_i = \sum d_j$. The cost of transporting one unit of this good from supplier i to consumer j is c_{ij}. For this problem, the graph G is a *bipartite graph*; it can be separated into a set of sources and a set of sinks, and the edges are only from the sources to the sinks. The optimisation problem is

$$\begin{aligned} \text{minimise} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{m} x_{ij} = s_i \quad \forall i \in \{1,\dots,n\} \\ & \sum_{i=1}^{m} x_{ij} = d_j \quad \forall j \in \{1,\dots,m\} \end{aligned}$$

which is a special case of the minimum flow problem.

8.3. Sufficiency of transport problem

Theorem. Every minimum cost flow problem with either finite capacities or non-negative capacities can be translated into an equivalent transport problem.

Proof. Consider the minimum cost flow problem on a graph $G = (V, E)$. We may assume without loss of generality that $\underline{m}_{ij} = 0$ for all $(i, j) \in E$, because we may write $x_{ij} = \underline{m}_{ij} + \tilde{x}_{ij}$ where $\tilde{x}_{ij} > 0$. Then the conservation equation becomes

$$\tilde{b}_i + \sum_{(j,i)\in E} \tilde{x}_{ji} = \sum_{(i,j)\in E} \tilde{x}_{ij}$$

where $\tilde{b}_i = \sum_{(j,i)\in E} \underline{m}_{ji} - \sum_{(i,j)\in E} \underline{m}_{ij}$. The regional constraints are now

$$0 \le \tilde{x}_{ij} \le \overline{m}_{ij} - \underline{m}_{ij}$$

We assume that $\underline{m}_{ij} \equiv 0$ from now. If all the costs are non-negative and a particular capacity is infinite, then we can replace that capacity by a large number e.g. $\sum |b_i|$, which is the maximum amount of flow that could possibly travel along this edge. This transformation does not change the optimal solution. We have now reduced to the case where all capacities are finite.

Now, for each such minimum cost flow problem, we will construct an equivalent transport problem that has the same feasible solutions and the same costs. For each vertex i, we create a consumer with demand $\sum_{(i,j)\in E} \overline{m}_{ik} - b_i$. For every edge (i, j), we create a supplier with supply $\overline{m}_{ij}$. The total supply and the total demand are equal, since $\sum_i b_i = 0$. We now

define the cost of moving from $(i, j) \to i$ is zero. We further define the cost of moving from $(i, j) \to j$ is c_{ij}.

Now, suppose x_{ij} flows from $(i, j) \to j$. Then $\overline{m}_{ij} - x_{ij}$ flows from $(i, j) \to i$, since the total incoming and outgoing flow from (i, j) must balance. Then, since the demand at i is $\sum_{(i,j)\in E} \overline{m}_{ik} - b_i$, the total flow into i satisfies

$$\sum_{(i,k)\in E} (\overline{m}_{ik} - x_{ik}) + \sum_{(k,i)\in E} x_{ki} = \sum_{(i,j)\in E} \overline{m}_{ik} - b_i$$

which simplifies to the conservation equation for the minimum cost flow problem. We can easily check that $0 \leq x_{ij} \leq \overline{m}_{ij}$. So this mapping between the minimum cost flow problem and the transport problem preserves feasibility of solutions.

It now suffices to show that the costs of the two feasible solutions for the two problems are the same; since then we will have demonstrated a mapping between the two problems. The cost in the transport problem is $\sum_{(i,j)\in E} x_{ij}c_{ij}$ since the edge from (i, j) to i has zero cost. This is identical to the cost in the minimum cost flow problem. □

8.4. Optimality conditions for transport problem

Recall that for a linear program, there are three optimality conditions: primal feasibility, dual feasibility, and complementary slackness. These have various interpretations in the context of a transport problem.

Theorem. If for some feasible x we have dual variables $\boldsymbol{\lambda} \in \mathbb{R}^n$ (for suppliers) and $\boldsymbol{\mu} \in \mathbb{R}^m$ (for consumers), such that:

(i) $\lambda_i + \mu_j \leq c_{ij} \quad \forall (i, j) \in E$; and

(ii) $(c_{ij} - (\lambda_i + \mu_j))x_{ij} = 0 \quad \forall (i, j) \in E$

then x is an optimal solution.

Proof. The Lagrangian of the transport problem is

$$\begin{aligned} L(x, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} - \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{m} x_{ij} - s_i \right) - \sum_{j=1}^{m} \mu_j \left(\sum_{i=1}^{n} x_{ij} - d_j \right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} (c_{ij} - \lambda_i - \mu_j)x_{ij} + \sum_{i=1}^{n} \lambda_i s_i + \sum_{j=1}^{m} \mu_j d_j \end{aligned}$$

$(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is dual feasible if $\lambda_i + \mu_j \leq c_{ij}$ for all i, j. We have primal feasibility, dual feasibility, and complementary slackness, so optimality holds. □

Note that if $\boldsymbol{\lambda}, \boldsymbol{\mu}$ are optimal, then $\boldsymbol{\lambda}+k, \boldsymbol{\mu}-k$ are also optimal, since $(\lambda_i+k)+(\mu_j-k) = \lambda_i+\mu_j$. So for simplicity, we can always choose $\lambda_1 = 0$. This gives $m + n - 1$ remaining Lagrange multipliers.

9. The transport algorithm

9.1. Transportation tableaux

Analogously to the simplex tableaux, for the transport problem we can create transportation tableaux. This is a convenient format for storing all relevant information for the transport problem while solving it. The transportation tableau is as follows:

	μ_1	μ_2	$\cdots$	μ_m	
λ_1	$\lambda_1+\mu_1$ x_{11} c_{11}	$\lambda_1+\mu_2$ x_{12} c_{12}	$\cdots$	$\lambda_1+\mu_m$ x_{1m} c_{1m}	s_1
λ_2	$\lambda_2+\mu_1$ x_{21} c_{21}	$\lambda_2+\mu_2$ x_{22} c_{22}	$\cdots$	$\lambda_2+\mu_m$ x_{2m} c_{2m}	s_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
λ_n	$\lambda_n+\mu_1$ x_{n1} c_{n1}	$\lambda_n+\mu_2$ x_{n2} c_{n2}	$\cdots$	$\lambda_n+\mu_m$ x_{nm} c_{nm}	s_n
	d_1	d_2	$\cdots$	d_m	

Like with the simplex method, we must begin with a basic feasible solution to construct the initial tableau. We can construct such a basic feasible solution by using the first supplier to satisfy the first consumer, then gradually using the next suppliers and consumers as we run out of supply or demand. $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ can be deduced by considering complementary slackness. That is, if $x_{ij} > 0$ then $\lambda_i + \mu_j = c_{ij}$. For instance, consider this problem with three suppliers and four consumers. The general transportation tableau would look like this:

	μ_1	μ_2	μ_3	μ_4	
λ_1	$\lambda_1+\mu_1$ x_{11} c_{11}	$\lambda_1+\mu_2$ x_{12} c_{12}	$\lambda_1+\mu_3$ x_{13} c_{13}	$\lambda_1+\mu_4$ x_{14} c_{14}	s_1
λ_2	$\lambda_2+\mu_1$ x_{21} c_{21}	$\lambda_2+\mu_2$ x_{22} c_{22}	$\lambda_2+\mu_3$ x_{23} c_{23}	$\lambda_2+\mu_4$ x_{24} c_{24}	s_2
λ_3	$\lambda_3+\mu_1$ x_{31} c_{31}	$\lambda_3+\mu_2$ x_{32} c_{32}	$\lambda_3+\mu_3$ x_{33} c_{33}	$\lambda_3+\mu_4$ x_{34} c_{34}	s_3
	d_1	d_2	d_3	d_4	

We will consider the problem given by

$$\mathbf{s} = \begin{pmatrix} 14 \\ 10 \\ 9 \end{pmatrix}; \quad \mathbf{d} = \begin{pmatrix} 12 \\ 5 \\ 8 \\ 8 \end{pmatrix}; \quad C = \begin{pmatrix} 5 & 3 & 4 & 6 \\ 2 & 7 & 4 & 1 \\ 5 & 6 & 2 & 4 \end{pmatrix}$$

A basic feasible solution is given by

$$X = \begin{pmatrix} 12 & 2 & 0 & 0 \\ 0 & 3 & 7 & 0 \\ 0 & 0 & 1 & 8 \end{pmatrix}$$

Complementary slackness gives

$$\begin{aligned} \lambda_1 + \mu_1 &= 5 \\ \lambda_1 + \mu_2 &= 3 \\ \lambda_2 + \mu_2 &= 7 \\ \lambda_2 + \mu_3 &= 4 \\ \lambda_3 + \mu_3 &= 2 \\ \lambda_3 + \mu_4 &= 4 \end{aligned}$$

This is a system of seven equations for six unknowns. However, since we can always set $\lambda_1 = 0$, we can reduce this to a system of six equations for six unknowns.

$$\begin{aligned} \mu_1 &= 5 \\ \mu_2 &= 3 \\ \lambda_2 + \mu_2 &= 7 \\ \lambda_2 + \mu_3 &= 4 \\ \lambda_3 + \mu_3 &= 2 \\ \lambda_3 + \mu_4 &= 4 \end{aligned}$$

Hence,

$$\boldsymbol{\lambda} = \begin{pmatrix} 0 \\ 4 \\ 2 \end{pmatrix}; \quad \boldsymbol{\mu} = \begin{pmatrix} 5 \\ 3 \\ 0 \\ 2 \end{pmatrix}$$

Theorem. When constructing a basic feasible solution in this way, the set of edges with strictly positive flow form a connected graph with no cycles. In particular, this graph is a *spanning tree T* with exactly $m + n - 1$ edges. This allows us to always construct a system of equations as above.

No proof is given.

	5	3	0	2	
0	5 / 12 | 5	3 / 2 | 3	0 / 0 | 4	2 / 0 | 6	14
4	9 / 0 | 2	7 / 3 | 7	4 / 7 | 4	6 / 0 | 1	10
2	7 / 0 | 5	5 / 0 | 6	2 / 1 | 2	4 / 8 | 4	9
	12	5	8	8	

9.2. Updating the transportation tableau

First, we check if $c_{ij} \geq \lambda_i + \mu_j$ for all i, j. If this is true, then our solution is optimal. In our example $c_{21} \geq \lambda_2 + \mu_1$, so we are not at an optimal solution. If $(i, j) \notin T$ (where T is the spanning tree above, i.e. $x_{ij} = 0$) and $c_{ij} < \lambda_i + \mu_j$, then (i, j) and the edges of T form a loop. We then increase x_{ij} as much as possible until another flow $x_{i'j'}$ is forced to be zero. Then we update the dual variables $\boldsymbol{\lambda}, \boldsymbol{\mu}$ and repeat.

In our example, we will introduce a flow of $x_{21} = \theta$. This will change the amount of flow along some nonzero edges. Doing this will force an update $x_{11} \mapsto x_{11} - \theta$ due to constrained demand, $x_{12} \mapsto x_{12} + \theta$ due to supply, and $x_{22} \mapsto x_{22} - \theta$ due to demand. We can then increase θ to a maximum value of 3. Now,

$$x = \begin{pmatrix} 9 & 5 & 0 & 0 \\ 3 & 0 & 7 & 0 \\ 0 & 0 & 1 & 8 \end{pmatrix}$$

We now recalculate $\boldsymbol{\lambda}, \boldsymbol{\mu}$ in the same way as above, which will give

$$\boldsymbol{\lambda} = \begin{pmatrix} 0 \\ -3 \\ -5 \end{pmatrix}; \quad \boldsymbol{\mu} = \begin{pmatrix} 5 \\ 3 \\ 7 \\ 9 \end{pmatrix}$$

Reconstructing the tableau gives

	5	3	7	9	
0	5 / 9 | 5	3 / 3 | 3	7 / 0 | 4	9 / 0 | 6	14
−3	2 / 3 | 2	0 / 0 | 7	4 / 7 | 4	6 / 0 | 1	10
5	0 / 0 | 5	−2 / 0 | 6	2 / 1 | 2	4 / 8 | 4	9
	12	5	8	8	

Once again there is an edge where $c_{ij} < \lambda_i + \mu_j$, notably $(i, j) = (2, 4)$, with zero flow. If $x_{ij} = \theta$, then $x_{23} \mapsto x_{23} - \theta$, $x_{34} \mapsto x_{34} - \theta$, $x_{33} \mapsto x_{33} + \theta$. We can increase θ only to 7. Once again, updating the tableau gives

	5	3	2	4	
0	5 | 9	3 | 5	2 | 4	4 | 6	14
−3	2 | 3	0 | 7	−1 | 4	1 | 1	10
0	5 | 0	3 | 6	2 | 2	4 | 4	9
	12	5	8	8	

In this current table, all optimality conditions are satisfied. So the solution is

$$x = \begin{pmatrix} 9 & 5 & 0 & 0 \\ 3 & 0 & 0 & 7 \\ 0 & 0 & 8 & 1 \end{pmatrix}$$

10. Maximum flow, minimum cut

10.1. Introduction

Consider the problem

$$\begin{aligned}
&\text{maximise} && \delta \\
&\text{subject to} && \sum_{\{j\,:\,(i,j)\in E\}} x_{ij} - \sum_{\{j\,:\,(j,i)\in E\}} x_{ji} = 0 \text{ for all } i \neq 1, i \neq n \\
& && \sum_{\{j\,:\,(1,j)\in E\}} x_{1j} - \sum_{\{j\,:\,(j,1)\in E\}} x_{j1} = \delta \\
& && \sum_{\{j\,:\,(n,j)\in E\}} x_{nj} - \sum_{\{j\,:\,(j,n)\in E\}} x_{jn} = -\delta \\
& && 0 \leq x_{ij} \leq c_{ij} \text{ for all } (i,j) \in E
\end{aligned}$$

This is a graph where vertex 1 is a source and vertex n is a sink, and δ is the flow from vertex 1 to vertex n. We want to maximise the total amount of flow on the graph, constrained by a certain maximum flow c_{ij} on each edge.

10.2. Cuts and flows

Definition. A *cut* of a graph $G = (V, E)$ is a partition of its vertices into two sets $(S, V \setminus S)$. The *capacity* of a cut is given by

$$C(S) = \sum_{\{(i,j)\in E\,:\,i\in S, j\in V\setminus S\}} c_{ij}$$

Theorem. For any feasible flow x with value δ, then for any cut $(S, V \setminus S)$ such that $1 \in S, n \in V \setminus S$, we have

$$\delta \leq C(S)$$

Proof. For any sets $X, Y \subseteq V$, we define the function

$$f_x(X, Y) = \sum_{\{(i,j)\in E\,:\,i\in X, j\in Y\}} x_{ij}$$

Note that X, Y need not be disjoint. Let $(S, V \setminus S)$ be a cut such that $1 \in S, n \in V \setminus S$. We have

$$\delta = \sum_{i\in S} \left(\sum_{\{j\,:\,(i,j)\in E\}} x_{ij} - \sum_{\{j\,:\,(j,i)\in E\}} x_{ji} \right)$$

since for $i = 1$ the bracket is δ and for all others it is zero. Therefore,

$$\begin{aligned}\delta &= f_x(S, V) - f_x(V, S)\\ &= f_x(S, S) + f_x(S, V \setminus S) - f_x(S, S) - f_x(V \setminus S, S)\\ &= f_x(S, V \setminus S) - \underbrace{f_x(V \setminus S, S)}_{\geq 0}\\ &\leq f_x(S, V \setminus S)\\ &\leq C(S)\end{aligned}$$

□

10.3. Max-flow min-cut theorem

Theorem. Let $\delta^\star$ be the value of the maximum flow. Then we have

$$\delta^\star = \min\{C(S) : 1 \in S, n \in V \setminus S\}$$

So the value of the maximum flow is equal to the cut of smallest capacity.

Proof. A *path* $v_0, v_1, \dots, v_k$ is a sequence of vertices such that every pair of adjacent vertices is connected by an edge, either in the forward direction or in the reverse direction. A path is called an *augmenting* path if

$$\begin{aligned} x_{v_i v_{i+1}} &< c_{v_i v_{i+1}} && \text{for all forward edges;}\\ x_{v_i v_{i+1}} &> 0 && \text{for all backward edges}\end{aligned}$$

So each forward edge must have remaining capacity, and reverse edges must have some flow. This definition allows us to state that augmenting paths are actually all paths such that altering the flow on all edges in the path can increase the total flow from 1 to n, while keeping the amount of flow into each vertex the same (excluding the first and last vertices in the path). Therefore, an optimal flow x cannot have an augmenting path from vertex 1 to vertex n. Now, suppose x is optimal. We define a cut:

$$S = \{1\} \cup \{i : \exists \text{ an augmenting path } 1 \to i\}$$

Therefore $n \in V \setminus S$, since there is no augmenting path from 1 to n. Then,

$$\delta^\star = f_x(S, V \setminus S) - f_x(V \setminus S, S)$$

But we can show that $f_x(V \setminus S, S) = 0$, so

$$\delta^\star = f_x(S, V \setminus S) = C(S)$$

as required. □

10.4. Ford–Fulkerson algorithm

The above proof provides a convenient method for finding an optimal flow.

Algorithm 5: Ford–Fulkerson Algorithm

Result: Optimal flow x
start with a feasible flow, such as $x = 0$;
repeat
  choose an augmenting path from 1 to n, and increase the flow along this path as much as possible;
until *no augmenting paths from* 1 *to* n;

Example. *Note that typically such graphs are represented pictorially, but due to difficulty of typesetting abstract diagrams, a matrix is substituted here.* Consider a graph given by the capacity matrix

$C =$

c_{ij}	1	a	b	c	d	n
1		5		5		
a			1		4	
b						5
c					2	
d						5
n						

First consider the feasible flow of $x = 0$. There exists an augmenting path $1, a, b, n$. We increase the flow by 1 in all edges, saturating edge (a, b), giving the flow matrix

$x =$

x_{ij}	1	a	b	c	d	n
1		1				
a			1			
b						1
c						
d						
n						

The path $1, a, d, n$ is now augmenting. We can increase the flow by 4 to saturate the edge (a, d):

$x =$

x_{ij}	1	a	b	c	d	n
1		5				
a			1		4	
b						1
c						
d						4
n						

The path $1, c, d, n$ is augmenting. Increasing by 1,

$$x = \begin{array}{c|cccccc} x_{ij} & 1 & a & b & c & d & n \\ \hline 1 & & 5 & & 1 & & \\ a & & & 1 & & 4 & \\ b & & & & & & 1 \\ c & & & & & 1 & \\ d & & & & & & 5 \\ n & & & & & & \end{array}$$

There are no augmenting paths. We can also check that the cut $(\{1\}, \{a, b, c, d, n\})$ gives the capacity 6, equivalent to the value at n so this must be optimal. We now have $\delta^\star = 6$.

Example. Consider a graph given by the capacity matrix

$$C = \begin{array}{c|cccccc} c_{ij} & 1 & a & b & c & d & n \\ \hline 1 & & 10 & & 10 & & \\ a & & & 4 & 2 & 8 & \\ b & & & & & & 10 \\ c & & & & & 9 & \\ d & & & 6 & & & 10 \\ n & & & & & & \end{array}$$

The path $1, a, d, n$ is augmenting. We can increase the (currently zero) flow by 8.

$$x = \begin{array}{c|cccccc} x_{ij} & 1 & a & b & c & d & n \\ \hline 1 & & 8 & & & & \\ a & & & & & 8 & \\ b & & & & & & \\ c & & & & & & \\ d & & & & & & 8 \\ n & & & & & & \end{array}$$

The path $1, c, d, n$ is also augmenting. We increase the flow by 2.

$$x = \begin{array}{c|cccccc} x_{ij} & 1 & a & b & c & d & n \\ \hline 1 & & 8 & & 2 & & \\ a & & & & & 8 & \\ b & & & & & & \\ c & & & & & 2 & \\ d & & & & & & 10 \\ n & & & & & & \end{array}$$

Now, the path $1, c, d, a, b, n$ is augmenting. (b, a) here is a reverse edge. Here, we can in-

crease the flow by 4. This will decrease the (a, b) by 4.

$x =$

x_{ij}	1	a	b	c	d	n
1		8		6		
a			4		4	
b						4
c					6	
d						10
n						

The path $1, a, d, b, n$ is augmenting, with all forward edges. Increasing by 2,

$x =$

x_{ij}	1	a	b	c	d	n
1		10		6		
a			4		6	
b						6
c					6	
d			2			10
n						

Finally, $1, c, d, b, n$ is augmenting, with all forward edges. Increasing by 3,

$x =$

x_{ij}	1	a	b	c	d	n
1		10		9		
a			4		6	
b						9
c					9	
d			5			10
n						

The flow δ is now 19. The cut given by $\{1, c\}$ has capacity 19, so we are at the optimum.

10.5. Termination of Ford–Fulkerson

If all capacities are integers, then the algorithm will always find the optimal flow. The same argument can be used for rational numbers. At each step, the flow increases by a positive integer value, so after a finite amount of steps it will stop, as the maximum flow is bounded.

10.6. Bipartite matching problem

A k-regular bipartite graph is a graph with $\frac{n}{2}$ vertices on the left and $\frac{n}{2}$ vertices on the right, where each vertex on both the left and right has exactly k edges. Suppose we want to match up the vertices on the left and right, such that each pair (a, b) is joined with an edge that already exists in this graph.

Theorem. Every k-regular bipartite graph has a perfect matching.

Proof. First, we construct a new graph with extra vertices $1, n$. We construct edges from vertex 1 to all vertices a on the left, with capacity 1. We then construct edges from all vertices b on the right to vertex n, also with capacity 1. The original edges in the graph will be given capacity ∞. Then by using the cut given by 1, $\delta^\star < \frac{n}{2}$. We can easily achieve the value $\delta^\star$ by attaching a flow $\frac{1}{k}$ to every edge from the left to the right, and of course sending a flow of 1 along each new edge. So the maximum flow for this new graph is $\frac{n}{2}$.

Now, we know that the Ford–Fulkerson algorithm will terminate, when given the initial flow $x = 0$. But with this algorithm, all edge weights are always integers, since all capacities are integral or infinite. The only way for all edge weights to be integer values are when we have a perfect matching. So this algorithm will generate a perfect matching. □

II. Variational Principles

Lectured in Easter 2021 by Dr. M. Dunajski

In this course, we solve problems of the form 'find the optimal function such that…'. Examples include 'find the shortest path between points A and B on surface Σ', or 'find the shape of a wire under the influence of gravity between points A and B in the plane'. The latter is called the brachistochrone problem, and is of central importance in motivating the subject.

In the same way that turning points of functions can often be located by setting the derivative to zero, optimal functions can be located by setting the functional derivative to zero. This is called the Euler–Lagrange equation, and is a main tool that we use to find solutions to such problems. An application of the Euler–Lagrange equation is Noether's theorem, which roughly states that any symmetry of a physical system gives rise to a conserved quantity. For example, uniformity of space in the laws of physics shows that momentum is conserved, and uniformity of time shows that energy is conserved.

Contents

1. History and motivation 66
1.1. The brachistochrone problem 66
1.2. Geodesics 66
1.3. Calculus of variations 66
1.4. Variational principles 67
2. Calculus for functions on $\mathbb{R}^n$ 68
2.1. Introduction 68
2.2. Constraints and Lagrange multipliers 69
2.3. Geometric justification of Lagrange multipliers 70
3. Euler–Lagrange equation 71
3.1. Fundamental lemma of calculus of variations 71
3.2. Euler–Lagrange equation 72
3.3. First integral of Euler–Lagrange equation (eliminating y) 73
3.4. Geodesics on a sphere 74
3.5. First integral of Euler–Lagrange equation (eliminating x) 75
3.6. Solving the brachistochrone problem 75
3.7. Fermat's principle 77
4. Extensions to the Euler–Lagrange equation 78
4.1. Euler–Lagrange equation with constraints 78
4.2. Dido's isoparametric problem 78
4.3. The Sturm–Liouville problem 79
4.4. Multiple dependent variables 80
4.5. Geodesics on surfaces 80
4.6. Multiple independent variables 81
4.7. Potential energy and the Laplace equation 82
4.8. Minimal surfaces 82
4.9. Higher derivatives 84
4.10. First integral for $n = 2$ 84
4.11. Principle of least action 85
4.12. Central forces 86
4.13. Configuration space and generalised coordinates 87
5. Noether's theorem 88
5.1. Statement and proof 88
5.2. Conservation of momentum 89
5.3. Conservation of angular momentum under central force 89
6. Convexity and the Legendre transform 90
6.1. Convex functions 90

6.2. Conditions for convexity 90
6.3. Legendre transform 91
6.4. Applications to thermodynamics 92
6.5. Legendre transform of the Lagrangian 93
6.6. Hamilton's equations from Euler–Lagrange equation 94
6.7. Hamilton's equations from extremising a functional 95
7. Second variations 96
7.1. Conditions for local minimisers 96
7.2. Legendre condition for minimisers 97
7.3. Associated eigenvalue problem 98
7.4. Jacobi accessory condition 99
7.5. Solving the Jacobi condition 100

1. History and motivation

1.1. The brachistochrone problem

Consider a particle sliding on a wire under the influence of gravity between two fixed points in the plane. What is the shape of the wire that produces the shortest travel time between the end points, given that the particle starts at rest? This problem is known as the brachistochrone problem, an archetypical variational problem. Suppose the end points are labelled A and B, where A is the origin, i.e. $(x_1, y_1) = (0, 0)$, and where B has coordinates (x_2, y_2). Note that $y_2 < 0$ in order that the particle has sufficient energy to reach the destination. The travel time T is given by

$$T = \int \mathrm{d}t = \int_A^B \frac{\mathrm{d}\ell}{v(x, y)}$$

Note that the kinetic energy and the potential energy sum to a constant.

$$\frac{1}{2}mv^2 + mgy = mgy_1 = 0 \implies v = \sqrt{2g}\sqrt{-y}$$

So we must find the function y that minimises

$$T[y] = \frac{1}{\sqrt{2g}} \int_0^{x_2} \frac{\sqrt{1 + y'^2}}{\sqrt{-y}}\,\mathrm{d}x$$

subject to $y_0 = 0$, $y(x_2) = y_2$. This problem's solution will be explored in a later lecture.

1.2. Geodesics

A geodesic is the shortest path γ between two points on a surface Σ, assuming such a path exists. Initially, let $\Sigma = \mathbb{R}^2$. On this plane, the Pythagorean theorem for measuring distances holds. Using a Cartesian coordinate system, we can say that a point A has coordinates (x_1, y_1), and a point B has coordinates (x_2, y_2). The distance from A to B along any path γ can be computed using a line integral.

$$D[y] = \int_A^B \mathrm{d}\ell = \int_{x_1}^{x_2} \sqrt{1 + y'^2}\,\mathrm{d}x$$

In this case, we have defined y as a function of x, and we seek to minimise D by varying the path γ on which we are moving.

1.3. Calculus of variations

A variational problem involves minimising an object of the form

$$F[y] = \int_{x_1}^{x_2} f(x, y(x), y'(x))\,\mathrm{d}x$$

subject to fixed values of y at the end points. We call such an F a *functional*; it is a function on the space of functions. Calculus applied to functionals is called the calculus of variations; we would like to find minima and maxima of functionals. In order to talk about functionals rigorously, we must define first the space of functions we are operating on; analogously to how we must define the domain of a function we are analysing when dealing with real or complex analysis. We write $C(\mathbb{R})$ for the space of continuous functions on $\mathbb{R}$, and $C^k(\mathbb{R})$ for the space of functions with continuous kth derivatives on $\mathbb{R}$. Sometimes, the notation $C^k_{(\alpha,\beta)}(\mathbb{R})$ is used to denote $C^k(\mathbb{R})$ such that $f(\alpha)$ and $f(\beta)$ are fixed, typically fixed to zero.

1.4. Variational principles

We can now define what variational principles are: they are such principles where laws follow from finding the minima or maxima of functionals. An introductory example is Fermat's principle, which states that light that travels between two points takes the path which requires the least travel time. There is also the principle of least action. Consider a particle moving under some potential $V(\mathbf{x})$, and let $T = \frac{1}{2}m|\dot{\mathbf{x}}|^2$ be its kinetic energy. We can define

$$S[\gamma] = \int_{t_1}^{t_2} (T - V)\,\mathrm{d}t$$

where γ represents the path along which the particle travels. The left hand side $S[\gamma]$ is called the *action*, and the principle of least action states that the action is minimised along paths of motion. Then, Newton's laws of motion should follow from this principle by minimising action.

2. Calculus for functions on $\mathbb{R}^n$

2.1. Introduction

Let $f \in C^2(\mathbb{R}^n)$, so $f : \mathbb{R}^n \to \mathbb{R}$ with all continuous second partial derivatives. We say that the point $\mathbf{a} \in \mathbb{R}^n$ is stationary if

$$\nabla f(\mathbf{a}) = \mathbf{0}$$

Consider a Taylor series expansion near a stationary point.

$$f(\mathbf{x}) = f(\mathbf{a}) + \frac{1}{2}(x_i - a_i)(x_j - a_j)\, \partial_{ij}^2 f\Big|_{\mathbf{a}} + O(\|\mathbf{x} - \mathbf{a}\|^2)$$

The Hessian matrix is defined as $H_{ij} = \partial_i \partial_j f = H_{ji}$, where $\partial_i \equiv \frac{\partial}{\partial x_i}$. For convenience, we will shift the origin to let $\mathbf{a} = \mathbf{0}$. The Hessian, evaluated at $\mathbf{0}$, written $H(\mathbf{0})$, is a real symmetric matrix and hence can be diagonalised using an orthogonal transformation.

$$H' = R^\intercal H(\mathbf{0}) R = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Then

$$f(\mathbf{x}') - f(\mathbf{0}) = \frac{1}{2} \sum \lambda_i (x_i')^2 + O(\|\mathbf{x}\|^2)$$

We can characterise the stationary point using the eigenvalues of the Hessian.

(i) If all $\lambda_i > 0$, then $f(\mathbf{x}') > f(\mathbf{0})$ so $f(\mathbf{x}')$ is a local minimum.

(ii) If all $\lambda_i < 0$, then $f(\mathbf{x}') < f(\mathbf{0})$ so $f(\mathbf{x}')$ is a local maxmimum.

(iii) If the eigenvalues have mixed signs, this is a saddle point. $f(\mathbf{x}')$ increases in some directions, but decreases in other directions.

(iv) If some eigenvalues are zero, we must consider higher-order terms of the Taylor expansion.

When $n = 2$, this is a special case. We can compute properties of the eigenvalues using the trace and determinant of the matrix.

$$\det H = \lambda_1 \lambda_2;\ \operatorname{tr} H = \lambda_1 + \lambda_2$$

(i) If $\det H > 0,\ \operatorname{tr} H > 0$ then we have a local minimum.

(ii) If $\det H > 0,\ \operatorname{tr} H < 0$ then we have a local maximum.

(iii) If $\det H < 0$ then we have a saddle point.

(iv) If $\det H = 0$ we need to consider higher-order terms.

Note that if $f : D \to \mathbb{R}$ where $D \subset \mathbb{R}^n$, it is possible that we have a local maximum which is not the global maximum, if such a global maximum actually lies on the boundary and is not a stationary point.

Now, let us suppose that f is harmonic, i.e. $\nabla^2 f(\mathbf{x}) = 0$ on $D \subset \mathbb{R}^2$. Hence, $\operatorname{tr} H = 0$ which implies that if there exists a turning point it is a saddle point. The minimum or maximum of a harmonic function must therefore occur on the boundary.

Example. Let

$$f(x, y) = x^3 + y^3 - 3xy$$

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 3x^2 - 3y \\ 3y^2 - 3x \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ or } \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The Hessian is

$$H = \begin{pmatrix} 6x & -3 \\ -3 & 6y \end{pmatrix} \implies H(\mathbf{0}) = \begin{pmatrix} 0 & -3 \\ -3 & 0 \end{pmatrix}; \quad H\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ -3 & 6 \end{pmatrix}$$

The determinant is negative at zero, giving us a saddle point. At the other point, the determinant is positive and the trace is positive, giving a local minimum.

2.2. Constraints and Lagrange multipliers

Example. Find the circle centered at $(0, 0)$ with smallest radius that intersects the parabola $y = x^2 - 1$. There are essentially two approaches.

- First, we consider the 'direct' method. We solve the constraints directly, which in this case means solving the equations

  $$f = x^2 + y^2$$
  $$y = x^2 - 1$$

  for minimal f. This gives

  $$f = x^2 + (x^2 - 1)^2 = x^4 - x^2 + 1$$

  Then by setting $\partial_x f = 0$ we have

  $$4x^3 - 2x = 0 \implies x \in \left\{0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right\}$$

  which gives

  $$x = \frac{\pm 1}{\sqrt{2}} \implies y = \frac{-1}{2};\ r = \frac{\sqrt{3}}{2}$$

  The other solution for x yields a larger radius. This method works fine for simple problems like this where the constraints are solvable. Therefore, we present an alternative method that works in the more general case.

- This method uses 'Lagrange multipliers'. We define a new function

$$h(x, y, \lambda) = f(x, y) - \lambda g(x, y)$$

where $g(x, y)$ is defined such that $g = 0$ is the constraint. λ is called the Lagrange multiplier. In this example,

$$h(x, y, \lambda) = x^2 + y^2 - \lambda(y - x^2 + 1)$$

We now extremise h over all free variables without constraints.

$$\nabla h = \begin{pmatrix} \partial h/\partial x \\ \partial h/\partial y \\ \partial h/\partial \lambda \end{pmatrix} = \begin{pmatrix} 2x + 2\lambda x \\ 2y - \lambda \\ y - x^2 + 1 \end{pmatrix}$$

Solving $\nabla h = 0$, we have

$$2x + 4xy = 0 \implies x = 0 \text{ or } y = \frac{-1}{2}$$

and the same results follow as before by substitution.

2.3. Geometric justification of Lagrange multipliers

Consider a curve given by $g = 0$. At each point on this curve, there is a normal to the curve of gradient ∇g. In particular, ∇g is perpendicular to $g = 0$. The function f has gradient perpendicular to the function $f = c$ for some constant c. So at the extremum, $\nabla f \propto \nabla g$, so $\nabla f - \lambda g = 0$ for some λ. This guides the creation of the new function h, for which we can optimise without constraints. This same reasoning generalises to functions in higher dimensions and with multiple constraints.

3. Euler–Lagrange equation

3.1. Fundamental lemma of calculus of variations

Consider again the functional

$$F[y] = \int_\alpha^\beta f(x, y, y')\,\mathrm{d}x$$

where f is given, and $f(\alpha, \cdot, \cdot)$ and $f(\beta, \cdot, \cdot)$ are fixed. Consider a small perturbation

$$y \mapsto y + \varepsilon\eta(x); \quad \eta(\alpha) = \eta(\beta) = 0$$

In order to compute the functional for this new function, we first need an additional lemma.

Lemma (Fundamental lemma of calculus of variations)**.** If $g : [\alpha, \beta] \to \mathbb{R}$ is continuous on this interval, and is such that

$$\forall \eta \text{ continuous}, \eta(\alpha) = \eta(\beta) = 0, \int_\alpha^\beta g(x)\eta(x)\,\mathrm{d}x = 0$$

Then

$$\forall x \in (\alpha, \beta),\ g(x) \equiv 0$$

Proof. Suppose that there exists a value $\overline{x} \in (\alpha, \beta)$ such that $g(\overline{x}) \neq 0$. Without loss of generality suppose that this value is positive. Then, by continuity, there exists a sub-interval $[x_1, x_2] \subset (\alpha, \beta)$ where $g(x) > c$ for some positive real c in this sub-interval. So we will construct an η such that $\eta > 0$ in $[x_1, x_2]$ and $\eta = 0$ outside this interval, for example

$$\eta(x) = \begin{cases} (x - x_1)(x_2 - x) & x \in [x_1, x_2] \\ 0 & \text{otherwise} \end{cases}$$

Then the integrand is non-negative everywhere, and is lower bounded by a positive number:

$$\int_\alpha^\beta g(x)\eta(x) > c \int_{x_1}^{x_2} (x - x_1)(x_2 - x)\,\mathrm{d}x > 0$$

So this leads to a contradiction. □

Remark. We call such an η function a 'bump function'. In general it is possible to construct a C^k bump function, e.g.

$$\eta = \begin{cases} [(x - x_1)(x_2 - x)]^{k+1} & x \in [x_1, x_2] \\ 0 & \text{otherwise} \end{cases}$$

3.2. Euler–Lagrange equation

Now, we can evaluate the original functional. Using a Taylor expansion,

$$\begin{aligned} F[y+\varepsilon\eta] &= \int_\alpha^\beta f(x, y+\varepsilon\eta, y'+\varepsilon\eta') \\ &= F[y] + \varepsilon\int_\alpha^\beta \left(\frac{\partial f}{\partial y}\eta + \frac{\partial f}{\partial y'}\eta'\right)\mathrm{d}x + O(\varepsilon^2) \end{aligned}$$

For an extremum,

$$\left.\frac{\mathrm{d}F}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} = 0$$

So we want the first order term to vanish, so

$$\varepsilon\int_\alpha^\beta \left(\frac{\partial f}{\partial y}\eta + \frac{\partial f}{\partial y'}\eta'\right)\mathrm{d}x = 0$$

Integrating by parts, we have

$$\begin{aligned} 0 &= \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\eta - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\eta\right)\right)\mathrm{d}x + \left[\frac{\partial f}{\partial y'}\eta\right]_\alpha^\beta \\ &= \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\eta - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\eta\right)\right)\mathrm{d}x \\ &= \int_\alpha^\beta \underbrace{\left(\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\right)}_{g(x)}\eta\,\mathrm{d}x \end{aligned}$$

We can apply the lemma above, showing that a necessary condition for the optimum is

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) - \frac{\partial f}{\partial y} = 0$$

This is the Euler–Lagrange equation.

Remark. Note that

- This can be seen as a second-order differential equation for $y(x)$ with boundary conditions at α and β.
- The left hand side of the Euler–Lagrange equation is called a 'functional derivative' of y, and is written

  $$\frac{\delta F[y]}{\delta y(x)}$$

  Sometimes, the notation

  $$\delta y = \varepsilon\eta(x)$$

is used, but is not used in this course. Note that in this notation,

$$F[y+\delta y] = F[y] + \delta F[y]; \quad \delta F[y] = \int_\alpha^\beta \left[\frac{\delta F[y]}{\delta y(x)}\delta y(x)\right] \mathrm{d}x$$

- Other boundary conditions, such as $\left.\frac{\partial f}{\partial y'}\right|_{\alpha,\beta}$ can be used.
- Note that when computing the derivatives, we regard x, y, y' as independent;

$$\frac{\partial f}{\partial y} = \left.\frac{\partial f}{\partial y}\right|_{x,y'}$$

We can also compute a total derivative, for instance

$$\frac{\mathrm{d}}{\mathrm{d}x} = \frac{\partial}{\partial x} + \frac{\partial}{\partial y}y' + \frac{\partial}{\partial y'}y''$$

Note that these give different results. As an example, let $f(x, y, y') = x[(y')^2 - y^2]$. Then

$$\frac{\partial f}{\partial x} = (y')^2 - y^2; \quad \frac{\partial f}{\partial y} = -2xy; \quad \frac{\partial f}{\partial y'} = 2xy'$$

Hence

$$\frac{\mathrm{d}f}{\mathrm{d}x} = (y')^2 - y^2 - 2xyy' + 2xy'y''$$

3.3. First integral of Euler–Lagrange equation (eliminating y)

In some cases, we can integrate the Euler–Lagrange equation to give a first-order ordinary differential equation. Suppose f does not explicitly depend on y. Then

$$\frac{\partial f}{\partial y} = 0 \implies \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) = 0$$

Hence,

$$\frac{\partial f}{\partial y'} = c; \quad c \in \mathbb{R}$$

Example. Consider geodesics on $\mathbb{R}^2$; we want to find curves on which the length is minimised.

$$F[y] = \int_\alpha^\beta \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2} = \int_\alpha^\beta \underbrace{\sqrt{1 + \frac{\mathrm{d}y}{\mathrm{d}x}^2}}_{f(y')} \mathrm{d}x$$

We can apply this 'first integral' form of the Euler–Lagrange equation to get

$$\frac{y'}{\sqrt{1+(y')^2}} = c$$

Hence y' is a constant, so let $y' = m$ for $m \in \mathbb{R}$. Hence $y = mx + c$.

3.4. Geodesics on a sphere

Consider the unit sphere $S^2 \subset \mathbb{R}^3$, and two points $A, B \in S^2$ which we wish to connect by a path of minimal length, where the path is constrained to the sphere. We will parametrise the sphere with spherical polar coordinates:

$$\begin{aligned} x &= \sin\theta \sin\phi \\ y &= \sin\theta \cos\phi \\ z &= \cos\theta \end{aligned}$$

where $\theta \in [0, \pi]; \phi \in [0, 2\pi]$. We can calculate the length of a path using the Pythagorean theorem:

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$$

We will parametrise the path by thinking of ϕ as a function of θ. This gives

$$\mathrm{d}s = \sqrt{1 + \sin^2\theta(\phi')^2}\,\mathrm{d}\theta$$

We wish to extremise the functional F, given by

$$F[\phi] = \int_{\theta_1=\alpha}^{\theta_2=\beta} \mathrm{d}s = \int_{\theta_1}^{\theta_2} \mathrm{d}s = \sqrt{1 + \sin^2\theta(\phi')^2}\,\mathrm{d}\theta$$

The integrand does not depend on ϕ but only on its derivative; so $\frac{\mathrm{d}f}{\mathrm{d}\phi} = 0$. Using the first integral form of the Euler–Lagrange equation, we have

$$\frac{\partial f}{\partial \phi'} = k$$

Now, we have

$$\begin{aligned} \frac{\sin^2\theta\phi'}{\sqrt{1 + \sin^2\theta(\phi')^2}} &= k \\ \sin^4\theta(\phi')^2 &= k^2(1 + \sin^2\theta(\phi')^2) \\ (\phi')^2 &= \frac{k^2}{\sin^2\theta(\sin^2\theta - k^2)} \\ \frac{\mathrm{d}\phi}{\mathrm{d}\theta} &= \pm\sqrt{\frac{k^2}{\sin^2\theta(\sin^2\theta - k^2)}} \\ \phi &= \pm\int \frac{k\,\mathrm{d}\theta}{\sin\theta\sqrt{\sin^2\theta - k^2}} \end{aligned}$$

The two solutions correspond to the two directions in which we can trace the path. We then can arrive at

$$\pm\frac{\sqrt{1-k^2}}{k}\cos(\phi-\phi_0)=\cot\theta$$

We will be able to see that this corresponds to a great circle; that is, the intersection of a plane through the origin with the sphere. We will show later that geodesics on a sphere are *only* segments of a great circle.

3.5. First integral of Euler–Lagrange equation (eliminating x)

For any $f(x,y,y')$, consider the quantity

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(f-y'\frac{\partial f}{\partial y'}\right)$$

This is exactly

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(f-y'\frac{\partial f}{\partial y'}\right)&=\frac{\partial f}{\partial x}+y'\frac{\partial f}{\partial y}+y''\frac{\partial f}{\partial y'}-y''\frac{\partial f}{\partial y'}-y'\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\\&=\frac{\partial f}{\partial x}+y'\frac{\partial f}{\partial y}-y'\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\\&=y'\underbrace{\left(\frac{\partial f}{\partial y}-\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y'}\right)}_{\text{zero by Euler–Lagrange}}+\frac{\partial f}{\partial x}\\&=\frac{\partial f}{\partial x}\end{aligned}$$

So, in the case that f does not depend explicitly on x (that is, $\frac{\partial f}{\partial x}\equiv 0$), then we have another first integral condition from the Euler–Lagrange equation:

$$f-y'\frac{\partial f}{\partial y'}=\text{constant}$$

3.6. Solving the brachistochrone problem

Consider a curve in the plane with a fixed endpoint at the origin and another fixed endpoint at $x=\beta$. We want to find a path such that the time taken for a particle to travel along this curve is minimised. We previously computed that the travel time is given by

$$F[y]=\frac{1}{\sqrt{2g}}\int_0^\beta\frac{\sqrt{1+(y')^2}}{\sqrt{-y}}\,\mathrm{d}x$$

This does not depend on x, so we can write (ignoring the $\frac{1}{\sqrt{2g}}$ factor)

$$f - y' \frac{\partial f}{\partial y'} = \frac{\sqrt{1 + (y')^2}}{\sqrt{-y}} - y' \frac{y'}{\sqrt{1 + (y')^2}\sqrt{-y}} = k$$

This gives

$$\begin{aligned} \frac{1}{\sqrt{1 + (y')^2}} &= k\sqrt{-y} \\ y' &= \pm \frac{\sqrt{1 + k^2 y^2}}{k\sqrt{-y}} \\ x &= \pm k \int \frac{\sqrt{-y}}{\sqrt{1 + k^2 y}} \, \mathrm{d}y \end{aligned}$$

We will parametrise further:

$$y = \frac{-1}{k^2} \sin^2 \frac{\theta}{2} \implies \mathrm{d}y = \frac{-1}{k^2} \sin \frac{\theta}{2} \cos \frac{\theta}{2}$$

Hence,

$$\begin{aligned} x &= \pm k \int \frac{-1}{k^2} \frac{\sin^2 \frac{\theta}{2} \cos \frac{\theta}{2}}{\sqrt{1 - \sin^2 \frac{\theta}{2}}} \, \mathrm{d}\theta \\ &= \mp \frac{1}{2k^2} \int (1 - \cos\theta) \, \mathrm{d}\theta \\ &= \mp \frac{1}{2k^2} (\theta - \sin\theta) + c \end{aligned}$$

The initial condition at $(0, 0)$ gives

$$\theta_0 = 0 \implies c = 0$$

Taking the positive solution, we have

$$\begin{aligned} x &= \frac{\theta - \sin\theta}{2k^2} \\ y &= \frac{-1}{k^2} \sin^2 \frac{\theta}{2} \end{aligned}$$

This can be shown to be a parametrised equation of a cycloid.

3.7. Fermat's principle

Fermat's principle states that as light travels between two points, it takes the path of least time. Let a ray of light be represented by a path $y(x)$. The speed of light is given by a function $c(x, y)$ since it depends on the material it is in. Then the time taken is

$$F[y] = \int \frac{\mathrm{d}\ell}{c} = \int_\alpha^\beta \frac{\sqrt{1 + (y')^2}}{c(x, y)}\,\mathrm{d}x$$

In this general form, f depends on x, y, y'. Now, let us assume c depends only on x and not on y. Then we can use a first integral form to get

$$\frac{\partial f}{\partial y'} = \text{constant}$$

This gives

$$\frac{y'}{c(x)\sqrt{1 + (y')^2}} = \text{constant}$$

Suppose that at α, the light ray's path has an angle θ_1 with the x-axis, and at β the angle is θ_2. Note that $\theta_1 = \arctan y'|_\alpha$ and the corresponding result for β. Then,

$$\frac{\sin\theta_1}{c(x_1)} = \frac{\sin\theta}{c(x)}$$

This is known as Snell's law.

Suppose we have a material in which c increases with x. In such a material, we then have that θ increases with x. In a material in which c decreases as x increases, θ naturally decreases.

Now, suppose we have a slow material with $c = c_S$ and a fast material with $c = c_F$ adjacent to each other. We might like to find the path that light takes in its path between points that cross the material boundary. Snell's law can be used to determine that the ratio between the sine of the angle and the speed of light remains constant along the light ray's path.

4. Extensions to the Euler–Lagrange equation

4.1. Euler–Lagrange equation with constraints

Given a functional $F[y] = \int_\alpha^\beta f(x, y, y')\,\mathrm{d}x$, we would like to extremise F subject to $G[y] = \int_\alpha^\beta g(x, y, y')\,\mathrm{d}x = k$ for some constant k. We can use the method of Lagrange multipliers. Instead of extremising F, we will extremise

$$\Phi[y; \lambda] = F[y] - \lambda G[y]$$

Thus, we replace f in the Euler–Lagrange equation with $f - \lambda g$, giving

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial}{\partial y'}(f - \lambda g)\right) - \frac{\partial}{\partial y}(f - \lambda g) = 0$$

4.2. Dido's isoparametric problem

Given a fixed perimeter, we wish to find the simple and closed plane curve which maximises the enclosed area. We can restrict ourselves to convex curves. This is because any concave curve can be transformed into a convex curve with greater area and equal perimeter, by reflecting the non-convex region. We will parametrise the curve in $\mathbb{R}^2$ by letting the minimal and maximal values of x be α, β. Then, as we trace out the curve, x monotonically increases from α to β, and then monotonically decreases as we return from β to α. This induces two functions y_1, y_2 on (α, β) where $y_2 > y_1$. The infinitesimal area is given by $\mathrm{d}A = (y_2 - y_1)\,\mathrm{d}x$. Thus, the area functional is given by

$$A[y] = \int_\alpha^\beta (y_2(x) - y_1(x))\,\mathrm{d}x = \oint_C y(x)\,\mathrm{d}y$$

The constraint functional is

$$L[y] = \oint_C \mathrm{d}\ell = \oint_C \sqrt{1 + (y')^2}\,\mathrm{d}x = L$$

where L is the fixed perimeter. Using Lagrange multipliers, we can define

$$h = y - \lambda\sqrt{1 + (y')^2}$$

Note that we do not need to consider a boundary term in the derivation of the Euler–Lagrange equation, since the curve has no boundary. Using a first integral form of the Euler–Lagrange equation on h, we have

$$k = h - y'\frac{\mathrm{d}h}{\mathrm{d}y'} = y - \lambda\sqrt{1 + (y')^2} + y'\lambda\frac{y'}{\sqrt{1 + (y')^2}} = y - \frac{\lambda}{\sqrt{1 + (y')^2}}$$

for some constant k. Hence,

$$(y')^2 = \frac{\lambda^2}{(y - k)^2} - 1$$

A solution here is the circle of radius λ:

$$(x - x_0)^2 + (y - y_0)^2 = \lambda^2$$

Here, $L = 2\pi\lambda$ so we can write the solution in terms of L instead, giving

$$(x - x_0)^2 + (y - y_0)^2 = \frac{L^2}{4\pi^2}$$

4.3. The Sturm–Liouville problem

Let $\rho(x), \sigma(x)$ be defined for $x \in [\alpha, \beta]$, and let $\rho(x) > 0$ on this interval. Consider the functional

$$F[y] = \int_\alpha^\beta \left[\rho(y')^2 + \sigma y^2\right] \mathrm{d}x$$

Let us extremise F subject to the constraint

$$G[y] = \int_\alpha^\beta y^2 \, \mathrm{d}x = 1$$

We have

$$\Phi[y; \lambda]F[y] - \lambda(G[y] - 1)$$

This induces the integrand

$$h = \rho(y')^2 + \sigma y^2 - \lambda(y^2 - \frac{1}{\beta - \alpha})$$

We consider the derivatives for the Euler–Lagrange equation:

$$\frac{\partial h}{\partial y'} = 2\rho y'; \quad \frac{\partial h}{\partial y} = 2\sigma y - 2\lambda y$$

Hence,

$$-\frac{\mathrm{d}}{\mathrm{d}x}(\rho y') + \sigma y = \lambda y$$

We can write this as $\mathcal{L}(y) = \lambda y$, where the $\mathcal{L}$ is known as the Sturm–Liouville operator. This is essentially an eigenvalue problem, since $\mathcal{L}$ is a linear operator. For example, if $\rho = 1$, this eigenvalue problem is exactly the time-independent Schrödinger equation where σ is the quantum-mechanical potential.

Suppose $\sigma > 0$. Then the functional $F[y]$ is also greater than zero. Then, the positive minimum of F (if it exists) is the lowest eigenvalue.

Proof. Using the result from the Euler–Lagrange equation, we can multiply by y and integrate by parts giving

$$-y\frac{\mathrm{d}}{\mathrm{d}x}(\rho y') + \sigma y^2 = \lambda y^2$$
$$F[y] - \underbrace{[yy'\rho]_\alpha^\beta}_{\text{zero}} = \lambda \underbrace{G[y]}_{\text{one}}$$

Thus, the lowest eigenvalue is the minimum of $F[y]/G[y]$. □

4.4. Multiple dependent variables

Suppose we have some vector

$$\mathbf{y}(x) = (y_1(x), y_2(x), \dots, y_n(x))$$

Suppose we want to extremise the functional

$$F[\mathbf{y}] = \int_\alpha^\beta f(x, y_1, \dots, y_n, y_1', \dots, y_n')\,\mathrm{d}x$$

If there is some critical point $\mathbf{y}$, we perturb by a small amount $\varepsilon\boldsymbol{\eta} = \varepsilon(\eta_1(x), \dots, \eta_n(x))$, where $\boldsymbol{\eta}(\alpha) = \boldsymbol{\eta}(\beta) = \mathbf{0}$. Following the derivation of the one-dimensional Euler–Lagrange equation, we can deduce that

$$F[\mathbf{y} + \varepsilon\boldsymbol{\eta}] - F[\mathbf{y}] = \int_\alpha^\beta \sum_{i=1}^n \eta_i \left(\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y_i'} - \frac{\partial f}{\partial y_i} \right) \mathrm{d}x + \text{boundary term} + O(\varepsilon^2)$$

We can apply the fundamental lemma, choosing η_i in a useful way, we can show that a necessary condition for a critical point is

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y_i'} - \frac{\partial f}{\partial y_i} = 0$$

for all i. This is a second-order *system* of n ODEs that we can solve. If f does not depend on one of the y_i, then we have a first integral form for this particular equation. In particular, if $\frac{\partial f}{\partial y_j} \equiv 0$ then $\frac{\partial f}{\partial y_j'} =$ constant. If f does not depend on x, then we have $f - \sum_i y_i' \frac{\partial f}{\partial y_i'} =$ constant.

4.5. Geodesics on surfaces

Consider a surface Σ in $\mathbb{R}^3$, given by

$$\Sigma = \{\mathbf{x} : \, g(\mathbf{x}) = 0\}$$

Consider two points A, B on Σ. What are the geodesics (the shortest paths on the surface) between the two points, if one exists at all? Consider a parametrisation of such a path given by $t \in [0, 1]$ where $A = \mathbf{x}(0), B = \mathbf{x}(1)$. We wish to extremise

$$\Phi[\mathbf{x}, \lambda] = \int_0^1 \left\{ \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} - \lambda(t) g(\mathbf{x}) \right\} \mathrm{d}t$$

The Lagrange multiplier, a function of t, since we want the entire curve (for all t) to lie on Σ. We substitute the integrand h in the Euler–Lagrange equation. Considering the variation with respect to λ, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial h}{\partial \dot{\lambda}} - \frac{\partial h}{\partial \lambda} = 0$$

But h does not depend on $\dot{\lambda}$, hence $\frac{\partial h}{\partial \lambda} = 0$, giving $g(\mathbf{x}) = 0$ for all $\mathbf{x}$. Considering the variation with respect to x_i, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial h}{\partial \dot{x}_i} - \frac{\partial h}{\partial x_i} = 0$$

Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{x}_i}{\sqrt{\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2}}\right) + \lambda\frac{\partial g}{\partial x_i} = 0$$

We could alternatively solve the constraint $g = 0$, and parametrise the surface according to this solution.

4.6. Multiple independent variables

In the most general case, we may have multiple independent variables in a variational problem. This converts the Euler–Lagrange equation into a partial differential equation. Suppose $\phi : \mathbb{R}^n \to \mathbb{R}^m$. If $n = 3$, for example, we have

$$F[\phi] = \iiint_{\mathcal{D}} f(\underbrace{x, y, z}_{\text{independent}}, \phi, \phi_x, \phi_y, \phi_z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

where $\mathcal{D} \subset \mathbb{R}^3$, and $\phi_{x_i} \coloneqq \partial\phi/\partial x_i$. Suppose there exists some extremum ϕ, and consider a small variation $\phi \mapsto \phi(x, y, z) + \varepsilon\eta(x, y, z)$ where $\eta = 0$ on $\partial\mathcal{D}$. Evaluating the functional on this perturbed ϕ gives

$$F[\phi + \varepsilon\eta] - F[\phi] = \varepsilon \iiint_{\mathcal{D}} \left\{\eta\frac{\partial f}{\partial \phi} + \eta_x\frac{\partial f}{\partial \phi_x} + \eta_y\frac{\partial f}{\partial \phi_y} + \eta_z\frac{\partial f}{\partial \phi_z}\right\}\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + O(\varepsilon^2)$$

$$= \varepsilon \iiint_{\mathcal{D}} \left\{\eta\frac{\partial f}{\partial \phi} + \underbrace{\nabla\cdot\left(\eta\left(\frac{\partial f}{\partial \phi_x}, \frac{\partial f}{\partial \phi_y}, \frac{\partial f}{\partial \phi_z}\right)\right)}_{\text{apply divergence theorem since } \eta \text{ vanishes on } \partial\mathcal{D}} - \eta\nabla\cdot\left(\frac{\partial f}{\partial \phi_x}, \frac{\partial f}{\partial \phi_y}, \frac{\partial f}{\partial \phi_z}\right)\right\}\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + O(\varepsilon^2)$$

$$= \varepsilon \iiint_{\mathcal{D}} \eta\left\{\frac{\partial f}{\partial \phi} - \nabla\cdot\left(\frac{\partial f}{\partial \phi_x}, \frac{\partial f}{\partial \phi_y}, \frac{\partial f}{\partial \phi_z}\right)\right\}\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + O(\varepsilon^2)$$

Now, we can apply the fundamental lemma to give the Euler–Lagrange equation for multiple independent variables.

$$\frac{\partial f}{\partial \phi} - \nabla\cdot\left(\frac{\partial f}{\partial \phi_x}, \frac{\partial f}{\partial \phi_y}, \frac{\partial f}{\partial \phi_z}\right) = 0$$

Or, in suffix notation (with the summation convention),

$$\frac{\partial f}{\partial \phi} - \partial_i\frac{\partial f}{\partial(\partial_i\phi)} = 0$$

This result applies for any n. Note that this is now a partial differential equation for ϕ, instead of an ordinary differential equation.

4.7. Potential energy and the Laplace equation

Consider the functional

$$F[\phi] = \iint_{\mathcal{D}\subset\mathbb{R}^2} \frac{1}{2}\left[\phi_x^2 + \phi_y^2\right] \mathrm{d}x\,\mathrm{d}y$$

Note that $\frac{\partial f}{\partial \phi} = 0$ and $\frac{\partial f}{\partial \phi_x} = \phi_x$; $\frac{\partial f}{\partial \phi_y} = \phi_y$. The Euler–Lagrange equation becomes

$$\frac{\partial}{\partial x}\phi_x + \frac{\partial}{\partial y}\phi_y = 0 \implies \phi_{xx} + \phi_{yy} = 0$$

This produces the Laplace equation.

4.8. Minimal surfaces

Consider minimising the area of a surface $\Sigma \subset \mathbb{R}^3$, where we want the surface to have two boundaries defined by fixed closed curves. This is sometimes known as Plateau's problem. We will let $\Sigma = \left\{\mathbf{x} = \mathbb{R}^3 : k(x,y,z) = 0\right\}$, and assume there exists a parametrisation of Σ given by $z = \phi(x,y)$. The line element is given by

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$$

We have $\mathrm{d}z = \phi_x\,\mathrm{d}x + \phi_y\,\mathrm{d}y$ hence

$$\mathrm{d}s^2 = (1+\phi_x^2)\,\mathrm{d}x^2 + (1+\phi_y^2)\,\mathrm{d}y^2 + 2\phi_x\phi_y\,\mathrm{d}x\,\mathrm{d}y$$

This is a quadratic form in the differentials $\mathrm{d}x\,,\mathrm{d}y$, known as the first fundamental form (also the Riemannian metric). Alternatively,

$$\mathrm{d}s^2 = g_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j$$

where

$$g = \begin{pmatrix} 1+\phi_x^2 & \phi_x\phi_y \\ \phi_x\phi_y & 1+\phi_y^2 \end{pmatrix}$$

From this, we can compute the area element, which is defined as

$$\mathrm{d}A = \sqrt{\det g}\,\mathrm{d}x\,\mathrm{d}y$$

We will extremise the area functional

$$A[\phi] = \int_{\mathcal{D}} \sqrt{1+\phi_x^2+\phi_y^2}\,\mathrm{d}x\,\mathrm{d}y$$

Let the integrand be h, and apply the Euler–Lagrange equation.

$$\frac{\partial h}{\partial \phi_x} = \frac{\phi_x}{\sqrt{1+\phi_x^2+\phi_y^2}}; \quad \frac{\partial h}{\partial \phi_y} = \frac{\phi_y}{\sqrt{1+\phi_x^2+\phi_y^2}}$$

Hence

$$\partial_x\left(\frac{\phi_x}{\sqrt{1+\phi_x^2+\phi_y^2}}\right)+\partial_y\left(\frac{\phi_x}{\sqrt{1+\phi_x^2+\phi_y^2}}\right)=0$$

which can be expanded to give

$$(1+\phi_y^2)\phi_{xx}+(1+\phi_x^2)\phi_{yy}-2\phi_x\phi_y\phi_{xy}=0$$

This is known as the minimal surface equation. We will solve a special case, where there is circular (cylindrical) symmetry, so $z=\phi(r)$. Since $r=\sqrt{x^2+y^2}$, we can find that

$$\phi_x=z'\frac{x}{r};\quad \phi_y=z'\frac{y}{r}$$

and we can analogously compute $\phi_{xx},\phi_{yy},\phi_{xy}$. This gives

$$rz''+z'+(z')^3=0$$

We can integrate this by first setting $z'=w$ and multiplying through by w.

$$\frac{1}{2}r\frac{\mathrm{d}}{\mathrm{d}r}w^2+w^2+w^4=0$$

Now let $w^2=u$ to make this a separable equation for u. Solving this, we can find that the solution surface is given by

$$r=r_0\cosh\left(\frac{z-z_0}{r_0}\right)$$

This is known as the *catenoid*. At the maximal and minimal values of z, we have the circular boundaries with radii R. At $z=z_0$, the radius is minimal, and the circle here has radius r_0. Supposing $z_0=0$ and that the maximal value of z is L, we have

$$\frac{R}{L}=\frac{r_0}{L}\cosh\left(\frac{L}{r_0}\right)$$

Let $L=1$ without loss of generality. This essentially chooses a scale for the coordinate system. This gives

$$R=r_0\cosh\frac{1}{r_0}$$

Plotting R as a function of r_0, there exists a minimum point $r_0=\mu\approx 0.833$ which gives $R\approx 1.5$. So if $R>1.5$, there exist two distinct minimal surfaces, one with $r_0>\mu$ and one with $r_0<\mu$. The 'tighter' minimal surface (with $r_0<\mu$) is unstable, but the 'looser' surface is stable (however this cannot be shown from our current understanding of variational principles).

4.9. Higher derivatives

Consider the functional

$$F[y] = \int_\alpha^\beta f(x, y, y', \dots, y^{(n)})\,\mathrm{d}x$$

We can find an analogous Euler–Lagrange equation to extremise this functional. Let η be a variation where $\eta^{(k)} = 0$ for $k \in \{1, \dots, n-1\}$ at the endpoints α, β. Now,

$$F[y + \varepsilon\eta] - F[y] = \varepsilon \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\eta + \frac{\partial f}{\partial y'}\eta' + \cdots + \frac{\partial f}{\partial y^{(n)}}\eta^{(n)} \right) \mathrm{d}x + O(\varepsilon^2)$$

We can repeatedly integrate each term by parts, integrating the $\eta^{(k)}$ term k times. Many of these terms will vanish due to the boundary conditions we specified for η. This then gives

$$F[y + \varepsilon\eta] - F[y] = \varepsilon \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\eta - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y'}\eta + \cdots + (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n}\frac{\partial f}{\partial y^{(n)}}\eta \right) \mathrm{d}x + O(\varepsilon^2)$$

Applying the fundamental lemma of calculus of variations, we have

$$\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y'} + \cdots + (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n}\frac{\partial f}{\partial y^{(n)}} = 0$$

This is the Euler–Lagrange equation in the context of a function with higher derivatives. The alternating signs come from the negative signs produced in the iterated integration by parts.

4.10. First integral for $n = 2$

Suppose $n = 2$. If $\frac{\partial f}{\partial y} = 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y'} - \frac{\mathrm{d}^2}{\mathrm{d}x^2}\frac{\partial f}{\partial y''} = 0$$

Hence

$$\frac{\partial f}{\partial y'} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y''} = \text{constant}$$

Example. Extremise the functional

$$F[y] = \int_0^1 (y'')^2\,\mathrm{d}x$$

subject to the conditions

$$y(0) = y'(0) = 0; \quad y(1) = 0; \quad y'(1) = 1$$

Using the above first integral form, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}(2y'') = \text{constant} \implies y''' = k$$

for some $k \in \mathbb{R}$. Imposing the boundary conditions on this cubic gives

$$y = x^3 - x^2$$

Now, we are going to show that this is an absolute *minimum* of the functional, not just a stationary point. Let $y_0 = x^2 - x^2$. Consider a variation η of y_0, where all relevant endpoints of η are zero. In this case, we are *not* going to assume that η is small; we will simply look at all possible variations.

$$F[y_0 + \eta] - F[y_0] = \underbrace{\int_0^1 (\eta'')^2 \,\mathrm{d}x}_{>0} + 2\int_0^1 y_0''\eta'' \,\mathrm{d}x$$

Substituting for y_0, given that $\eta \not\equiv 0$,

$$\begin{aligned} F[y_0 + \eta] - F[y_0] &> 4\int_0^1 (3x-1)\eta'' \,\mathrm{d}x \\ &= 4\left\{[-\eta']_0^1 + \int_0^1 \left[\frac{\mathrm{d}}{\mathrm{d}x}(3x\eta') - \eta'\right]\mathrm{d}x\right\} \\ &= 4\left\{\int_0^1 \left[\frac{\mathrm{d}}{\mathrm{d}x}(3x\eta') - \eta'\right]\mathrm{d}x\right\} \\ &= 4\left\{[3x\eta']_0^1 - [3\eta]_0^1\right\} \\ &= 0 \end{aligned}$$

Hence y_0 is an absolute minimum of F. This method of showing y_0 is an absolute minimum is easier than calculating second variations, where we know the solution y_0.

4.11. Principle of least action

Consider a particle moving in $\mathbb{R}^3$ with kinetic energy T and potential energy V. We define the *Lagrangian* to be

$$L(\mathbf{x}, \dot{\mathbf{x}}, t) = T - V$$

We now define the *action* to be

$$S[\mathbf{x}] = \int_{t_1}^{t_2} L \,\mathrm{d}t$$

We can now formulate the principle of least (or stationary) action: on the path of motion of a particle,

$$\frac{\delta S}{\delta \mathbf{x}} = 0$$

Equivalently, L satisfies the Euler–Lagrange equations:

$$\frac{\partial L}{\partial x_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}_i} = 0$$

Consider

$$T = \frac{1}{2}m|\dot{\mathbf{x}}|^2; \quad V = V(\mathbf{x})$$

The Euler–Lagrange equations are now

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}_i} &= \frac{\partial L}{\partial x_i} \\ m\ddot{x}_i &= -\frac{\partial V}{\partial x_i} \\ \implies m\ddot{\mathbf{x}} &= -\nabla V \end{aligned}$$

This is exactly Newton's second law, derived from the principle of stationary action.

4.12. Central forces

Example. Consider a central force in the plane. The Lagrangian is

$$L = T - V = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) - V(r)$$

The Euler–Lagrange equation gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{r}} - \frac{\partial L}{\partial r} &= 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}} - \frac{\partial L}{\partial \theta} &= 0 \end{aligned}$$

Since $\frac{\partial L}{\partial \theta} = 0$, we have a first integral form:

$$\frac{\partial L}{\partial \dot{\theta}} = mr^2\dot{\theta} = \text{constant}$$

This can be interpreted physically as the law of conservation of angular momentum. Further, we have $\frac{\partial L}{\partial t} = 0$ so we have another first integral:

$$\begin{aligned} \dot{r}\frac{\partial L}{\partial \dot{r}} + \dot{\theta}\frac{\partial L}{\partial \dot{\theta}} - L &= \text{constant} \\ m\dot{r}^2 + mr^2\dot{\theta}^2 - \frac{1}{2}m\dot{r}^2 - \frac{1}{2}mr^2\dot{\theta}^2 + V(r) &= \text{constant} \\ \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + V(r) &= \text{constant} \end{aligned}$$

The left hand side is the total energy of the system, denoted E. This is the law of conservation of energy.

4.13. Configuration space and generalised coordinates

Example. Consider N particles moving in $\mathbb{R}^3$. Typically we represent each point as a distinct vector in $\mathbb{R}^3$ that changes over time. We can alternatively consider a point in $\mathbb{R}^{3N}$, which contains the information about every point. This is called the configuration space. The Lagrangian in configuration space is

$$L = L(q_i, \dot{q}_i, t)$$

where $\mathbf{q}$ is the combined position vector of all N points, and likewise $\dot{\mathbf{q}}$ is the combined velocity.

5. Noether's theorem

5.1. Statement and proof

Consider a functional

$$F[\mathbf{y}] = \int_\alpha^\beta f(y_i, y_i', x)\,\mathrm{d}x\,; \quad i = 1, \dots, n$$

Suppose there exists a one-parameter family of transformations

$$y_i(x) \mapsto Y_i(x, s); \quad Y_i(x, 0) = y_i(x)$$

This can be thought of as a change of variables parametrised by $s \in \mathbb{R}$, where $s = 0$ implies no change of variables. This family is called a *continuous symmetry* of the Lagrangian f if

$$\frac{\mathrm{d}}{\mathrm{d}s} f(Y_i(x, s), Y_i'(x, s), x) = 0$$

In this course, we only consider continuous symmetries, so they may be abbreviated as just 'symmetries'.

Theorem (Noether's Theorem)**.** Given a continuous symmetry $Y_i(x, s)$ of f,

$$\left.\frac{\partial f}{\partial y_i'}\frac{\partial Y_i}{\partial s}\right|_{s=0}$$

is a first integral of the Euler–Lagrange equation (where the summation convention applies).

Proof.

$$\begin{aligned}
0 &= \left.\frac{\mathrm{d}}{\mathrm{d}s} f\right|_{s=0} \\
&= \left.\frac{\partial f}{\partial y_i}\frac{\mathrm{d}Y_i}{\mathrm{d}s}\right|_{s=0} + \left.\frac{\partial f}{\partial y_i'}\frac{\partial Y_i'}{\partial s}\right|_{s=0} \\
&= \left.\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y_i'}\right)\frac{\mathrm{d}Y_i}{\mathrm{d}s} + \frac{\partial f}{\partial y_i'}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}Y_i}{\mathrm{d}s}\right)\right]\right|_{s=0} \\
&= \left.\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\partial f}{\partial y_i'}\frac{\partial Y_i}{\partial s}\right]\right|_{s=0} \\
\therefore \text{constant} &= \frac{\partial f}{\partial y_i'}\frac{\partial Y_i}{\partial s}
\end{aligned}$$

□

5.2. Conservation of momentum

Example. Consider a vector $\mathbf{y} = (y, z)$ and the function

$$f = \frac{1}{2}y'^2 + \frac{1}{2}z'^2 - V(y - z)$$

Consider the symmetry

$$\begin{aligned} Y = y + s &\implies Y' = y' \\ Z = z + s &\implies Z = z' \\ \therefore V(Y - Z) = V(y - z) &\implies \frac{\mathrm{d}}{\mathrm{d}s} f = 0 \end{aligned}$$

Then from Noether's theorem,

$$\text{constant} = \left[\frac{\partial f}{\partial y'}\frac{\mathrm{d}Y}{\mathrm{d}s} + \frac{\partial f}{\partial z'}\frac{\mathrm{d}Z}{\mathrm{d}s}\right]\Bigg|_{s=0} = y' + z'$$

This can be thought of as a conserved momentum in the $y + z$ direction.

5.3. Conservation of angular momentum under central force

Example. Suppose $\Theta = \theta + s, R = r$. Our space is isotropic, so $\frac{\mathrm{d}L}{\mathrm{d}s} = 0$, hence

$$\left[\frac{\partial L}{\partial \dot{\theta}}\frac{\partial \Theta}{\partial s} + \frac{\partial L}{\partial \dot{r}}\frac{\partial R}{\partial s}\right]\Bigg|_{s=0} = mr^2\dot{\theta}$$

which shows that angular momentum is conserved.

6. Convexity and the Legendre transform

6.1. Convex functions

This subsection is covered by Lecture 1 of the IB Optimisation course.

Definition. A set $S \subset \mathbb{R}^n$ is convex if $\forall \mathbf{x}, \mathbf{y} \in S, \forall t \in [0,1], (1-t)\mathbf{x} + t\mathbf{y} \in S$.

Definition. The graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the surface $\{(\mathbf{x}, z) \in \mathbb{R}^{n+1} : z - f(\mathbf{x}) = 0\}$.

Definition. A chord of a function $f : \mathbb{R}^n \to \mathbb{R}$ is a line segment connecting two points on the graph of f.

Definition. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

(i) the domain of f is a convex set; and

(ii) $\forall \mathbf{x}, \mathbf{y} \in S, \forall t \in (0,1), f((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y})$

Equivalently, f is convex if the graph of f lies below (or on) all of its chords. We say that f is concave if f lies above (or on) all of its chords. Clearly, f is convex if and only if $-f$ is concave. We say f is *strictly* convex (or concave) if the inequality in (ii) becomes strict.

Example. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^2$$

The domain is clearly convex. To show convexity, we need

$$f((1-t)x + ty) - (1-t)f(x) - tf(y) \leq 0$$

We have

$$[(1-t)x+ty]^2-(1-t)x^2-ty^2 = x^2(1-t)(-t)+ty^2(1-t)+2(1-t)txy = -(1-t)t(x-y)^2 < 0$$

as required. Hence $f(x) = x^2$ is a strictly convex function.

Example. Consider

$$f(x) = \frac{1}{x}$$

where the domain is $\mathbb{R} \setminus \{0\}$. This domain is not convex, so f is not convex. However, restricted to the domain $\{x \in \mathbb{R} : x > 0\}$, f can be shown to be convex.

6.2. Conditions for convexity

Proofs for these conditions, where appropriate, are given in Lecture 1 of the IB Optimisation course.

Theorem. If f is a once-differentiable function, then f is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x}) \cdot \nabla f(\mathbf{x})$$

Corollary. If f is convex, and has a stationary point, then it is a global minimum.

Proof. Suppose the stationary point is at $\mathbf{x}_0$, so $\nabla f(\mathbf{x}_0) = \mathbf{0}$. We then have

$$f(\mathbf{y}) \geq f(\mathbf{x}_0) + (\mathbf{y} - \mathbf{x}_0) \cdot \mathbf{0}$$

which is larger than $f(\mathbf{x}_0)$ as required. □

Theorem. If f is a once-differentiable function, then f is convex if

$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})) \cdot (\mathbf{y} - \mathbf{x}) \geq 0$$

This can be thought of as stating that f' is monotonically increasing.

Theorem. If f is a twice-differentiable function, then f is convex if and only if

$$\nabla^2 f \succeq 0$$

i.e. all eigenvalues of the Hessian matrix are non-negative. Note that $\nabla^2 f \succ 0$ implies strict convexity.

Example. Consider the function

$$f(x, y) = \frac{1}{xy}$$

for $x > 0, y > 0$. Then the Hessian is

$$H = \frac{1}{xy} \begin{pmatrix} \frac{2}{x^2} & \frac{1}{xy} \\ \frac{1}{xy} & \frac{2}{y^2} \end{pmatrix}$$

Then,

$$\det H = \frac{3}{x^3 y^3} > 0$$

$$\operatorname{tr} H > 0$$

Hence the eigenvalues are both positive. So f is strictly convex.

6.3. Legendre transform

Definition. The Legendre transform of a function $f : \mathbb{R}^n \to \mathbb{R}$ is a function $f^\star$ given by

$$f^\star(\mathbf{p}) = \sup_{\mathbf{x}} (\mathbf{p} \cdot \mathbf{x} - f(\mathbf{x}))$$

The domain of $f^\star$ is such that the supremum provided is finite. In one dimension, we can consider $f^\star(p)$ to be the maximum vertical distance between the graphs of $y = f(x)$ and $y = px$.

Example. Consider the function $f(x) = ax^2$, which is convex where $a > 0$. Computing the derivative of the right hand side and setting it to zero,

$$\begin{aligned} f^\star(p) &= \sup_x (px - ax^2) \\ &= p\Big(\frac{p}{2a}\Big) - a\Big(\frac{p}{2a}\Big)^2 \\ &= \frac{p^2}{4a} \end{aligned}$$

We can apply the Legendre transform twice:

$$f^{\star\star}(s) = \sup_p (sp - f^\star(p)) = as^2 = f(s)$$

In fact, if f is convex, then we always have $f^{\star\star} = f$. If $a < 0$, the supremum does not exist so $f^\star$ has an empty domain, and thus $f^{\star\star} \neq f$.

Proposition. If the domain of $f^\star$ is non-empty, it is a convex set, and $f^\star$ is convex.

Proof. Given $\mathbf{p}, \mathbf{q}$ in the domain of $f^\star$,

$$\begin{aligned} f^\star((1-t)\mathbf{p} + t\mathbf{q}) &= \sup_{\mathbf{x}} [(1-t)\mathbf{p}\cdot\mathbf{x} + t\mathbf{q}\cdot\mathbf{x} - f(\mathbf{x})] \\ &= \sup_{\mathbf{x}} [(1-t)(\mathbf{p}\cdot\mathbf{x} - f(\mathbf{x})) + t(\mathbf{q}\cdot\mathbf{x} - f(\mathbf{x}))] \\ &\leq \sup_{\mathbf{x}} [(1-t)(\mathbf{p}\cdot\mathbf{x} - f(\mathbf{x}))] + \sup_{\mathbf{x}} [t(\mathbf{q}\cdot\mathbf{x} - f(\mathbf{x}))] \\ &< \infty \end{aligned}$$

as required. □

In practice, if f is convex and differentiable, we compute $f^\star(\mathbf{p})$ by considering the derivative:

$$\nabla(\mathbf{p}\cdot\mathbf{x} - f(\mathbf{x})) = 0 \implies \mathbf{p} = \nabla f$$

If f is strictly convex, the condition $\mathbf{p} = \nabla f$ has a unique inverse to give $\mathbf{x}$ as a function of $\mathbf{p}$, so $f^\star(\mathbf{p}) = \mathbf{p}\cdot\mathbf{x}(\mathbf{p}) - f(\mathbf{x}(\mathbf{p}))$. This eliminates the supremum condition.

6.4. Applications to thermodynamics

If we consider the particles in a gas, we could theoretically solve the Euler–Lagrange equations for a system of around 10^{23} particles. However, solving such a complicated system is difficult. Instead of solving for each particle, we instead consider macroscopic quantities

such as pressure P, volume V, temperature T, and entropy S. A system has *internal energy* $U(S,V)$. The *Helmholtz free energy* is

$$\begin{aligned} F(T,V) &= \min_S(U(S,V) - TS) \\ &= -\max_S(TS - U(S,V)) \\ &= -U^\star(T,V) \end{aligned}$$

where $U^\star$ is the Legendre transform of U with respect to S, fixing V constant. Assuming U is convex,

$$\frac{\partial}{\partial S}(TS - U(S,V))\bigg|_{T,V} = 0 \implies T = \frac{\partial U}{\partial S}\bigg|_V$$

There are other thermodynamical quantities that can be represented using a Legendre transform, for instance enthalpy $H(S,P)$.

$$\begin{aligned} H(S,P) &= \min_V(U(S,V) + PV) \\ &= -U^\star(-P,S) \end{aligned}$$

At this minimum, $P = -\frac{\partial U}{\partial V}\big|_S$. We can think of the Legendre transform in this context as a way of swapping from dependence on entropy and volume to dependence on other variables.

6.5. Legendre transform of the Lagrangian

Recall that the Lagrangian in mechanics was defined as

$$L = T - V = L(\mathbf{q}, \dot{\mathbf{q}}, t)$$

This is a function on the configuration space. We define the *Hamiltonian* to be the Legendre transform of L with respect to $\dot{\mathbf{q}}$. We find, assuming that L is convex,

$$\begin{aligned} H(\mathbf{q},\mathbf{p},\mathbf{t}) &= \sup_{\mathbf{v}}(\mathbf{p}\cdot\mathbf{v} - L) \\ &= \mathbf{p}\cdot\mathbf{v}(\mathbf{p}) - L(\mathbf{q},\mathbf{v}(\mathbf{p}),t) \end{aligned}$$

where $\mathbf{v}(\mathbf{p})$ is the solution to $p_i = \frac{\partial L}{\partial \dot{q}_i}$. The $\mathbf{p}$ are referred to as *generalised momenta* or *conjugate momenta*. Consider

$$T = \frac{1}{2}m|\dot{\mathbf{q}}|^2; \quad V = V(\mathbf{q})$$

Then,

$$\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{q}}} = m\dot{\mathbf{q}} \implies \dot{\mathbf{q}} = \frac{1}{m}\mathbf{p}$$

The Hamiltonian is therefore

$$\begin{aligned} H(\mathbf{q},\mathbf{p},t) &= \mathbf{p}\cdot\frac{1}{m}\mathbf{p} - L \\ &= \mathbf{p}\cdot\frac{1}{m}\mathbf{p} - \left(\frac{1}{2}m\frac{|\mathbf{p}|^2}{m^2} - V(\mathbf{q})\right) \\ &= \frac{1}{2m}|\mathbf{p}|^2 + V(\mathbf{q}) \\ &= T + V \end{aligned}$$

6.6. Hamilton's equations from Euler–Lagrange equation

Given that the Lagrangian satisfies the Euler–Lagrange equation, we can deduce analogous equations for the Hamiltonian. We often write the indices of the generalised coordinates in superscript, as follows, where the summation convention applies:

$$H = H(\mathbf{q},\mathbf{p},t) = p_i\dot{q}^i - L(q^i,\dot{q}^i,t)$$

Using this equation, we can compute two expressions for the differential of the Hamiltonian:

$$\begin{aligned} \mathrm{d}H &= \frac{\partial H}{\partial q^i}\,\mathrm{d}q^i + \frac{\partial H}{\partial p_i}\,\mathrm{d}p_i + \frac{\partial H}{\partial t}\,\mathrm{d}t \\ &= p_i\,\mathrm{d}\dot{q}^i + \dot{q}^i\,\mathrm{d}p_i - \frac{\partial L}{\partial q^i}\,\mathrm{d}q^i - \frac{\partial L}{\partial \dot{q}^i}\,\mathrm{d}\dot{q}^i - \frac{\partial L}{\partial t}\,\mathrm{d}t \end{aligned}$$

Now, note that $\frac{\partial L}{\partial \dot{q}^i} = p_i$. This cancels some terms. Making use of the Euler–Lagrange equation,

$$\frac{\partial L}{\partial q^i} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}^i} = \frac{\mathrm{d}}{\mathrm{d}t}p_i = \dot{p}_i$$

This gives

$$\mathrm{d}H = \frac{\partial H}{\partial q^i}\,\mathrm{d}q^i + \frac{\partial H}{\partial p_i}\,\mathrm{d}p_i + \frac{\partial H}{\partial t}\,\mathrm{d}t = \dot{q}^i\,\mathrm{d}p_i - \dot{p}_i\,\mathrm{d}q^i - \frac{\partial L}{\partial t}\,\mathrm{d}t$$

Comparing the differentials, we can see that

$$\dot{q}^i = \frac{\partial H}{\partial p_i};\quad \dot{p}_i = -\frac{\partial H}{\partial q^i};\quad \frac{\partial L}{\partial t} = -\frac{\partial H}{\partial t}$$

This system of equations is known as Hamilton's equations. Note that in the last equation, $\frac{\partial}{\partial t}\big|_{q,\dot{q}} \neq \frac{\partial}{\partial t}\big|_{p,q}$. For now, we will assume that there is no explicit t dependence in the Lagrangian. Then, Hamilton's equations are a system of $2n$ first-order ordinary differential equations. (Note, for comparison, that the Euler–Lagrange equations were a system of n second-order differential equations, which gives the same amount of initial conditions.) The initial conditions are typically a configuration of $\mathbf{p},\mathbf{q}$ at some fixed t_0. The solutions to Hamilton's equations are called the *trajectories* in $2n$-dimensional phase space.

6.7. Hamilton's equations from extremising a functional

Note that we can also arrive at Hamilton's equations by extremising a functional in phase space.

$$S[\mathbf{q},\mathbf{p}] = \int_{t_1}^{t_2} \left(\dot{q}^i p_i - H(\mathbf{q},\mathbf{p},t)\right) \mathrm{d}t$$

The integrand, denoted f, is a function of $\mathbf{q},\mathbf{p},\dot{\mathbf{q}},t$. Writing the Euler–Lagrange equations for S, varying first with respect to p_i,

$$\frac{\partial f}{\partial p_i} - \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial f}{\partial \dot{p}_i}}_{0} = 0 \implies \dot{q}^i = \frac{\partial H}{\partial p_i}$$

Now varying with respect to q^i,

$$\frac{\partial f}{\partial q^i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial f}{\partial \dot{q}^i} = 0 \implies \dot{p}_i = -\frac{\partial H}{\partial q^i}$$

These results are exactly Hamilton's equations.

7. Second variations

7.1. Conditions for local minimisers

The Euler–Lagrange equation gives a necessary condition for a stationary point. We cannot tell whether this leads to a minimum, a maximum, or a saddle point, just from the Euler–Lagrange equation. We can analyse the nature of the stationary points by considering the second variation. Consider the functional

$$F[y] = \int_\alpha^\beta f(x, y, y')\,\mathrm{d}x$$

where y is perturbed by a perturbation $\varepsilon\eta$. Let us assume that y is a solution to the Euler–Lagrange equation, so has no first variation. We will then expand $F[y + \varepsilon\eta]$ to second order.

$$\begin{aligned} F[y + \varepsilon\eta] &= \int_\alpha^\beta [f(x, y + \varepsilon\eta, y' + \varepsilon\eta')]\,\mathrm{d}x \\ F[y + \varepsilon\eta] - F[y] &= \int_\alpha^\beta [f(x, y + \varepsilon\eta, y' + \varepsilon\eta') - f(x, y, y')]\,\mathrm{d}x \\ &= 0 + \varepsilon \underbrace{\int_\alpha^\beta \eta\left(\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial f}{\partial y'}\right)\mathrm{d}x}_{\text{zero by Euler–Lagrange equation}} \\ &\quad + \frac{1}{2}\varepsilon^2 \int_\alpha^\beta \left(\eta^2 \frac{\partial^2 f}{\partial y^2} + \eta'^2 \frac{\partial^2 f}{\partial (y')^2} + 2\eta\eta' \frac{\partial^2 f}{\partial y \partial y'}\right)\mathrm{d}x + O(\varepsilon^3) \end{aligned}$$

The last term (excluding the ε^2 component) is called the second variation. We write

$$\delta^2 F[y] \equiv \frac{1}{2}\int_\alpha^\beta \left(\eta^2 \frac{\partial^2 f}{\partial y^2} + \eta'^2 \frac{\partial^2 f}{\partial (y')^2} + \frac{\mathrm{d}}{\mathrm{d}x}(\eta^2)\frac{\partial^2 f}{\partial y \partial y'}\right)\mathrm{d}x$$

Integrating the last term by parts, using $\eta = 0$ at α, β, we have

$$\delta^2 F[y] = \frac{1}{2}\int_\alpha^\beta \left(Q\eta^2 + P(\eta')^2\right)\mathrm{d}x$$

where

$$P = \frac{\partial^2 f}{\partial (y')^2}; \quad Q = \frac{\partial^2 f}{\partial y^2} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial^2 f}{\partial y \partial y'}\right)$$

Thus, if y is a solution to the Euler–Lagrange equation, and also $Q\eta^2 + P(\eta')^2 > 0$ for all η vanishing at α, β, then y is a local minimiser of F.

Example. We will prove that the geodesic on a plane is a local minimiser of path length. The functional we will analyse is given by

$$f = \sqrt{1 + (y')^2}$$

Hence,

$$P = \frac{\partial^2 f}{\partial (y')^2} = \frac{\partial}{\partial y'}\left(\frac{y'}{\sqrt{1+(y')^2}}\right) = \frac{1}{(1+(y')^2)^{\frac{3}{2}}} > 0$$

$$Q = 0$$

Therefore the second variation is positive, so any y that satisfies the Euler–Lagrange equation minimises path length. In particular, straight lines minimise path length on the plane.

7.2. Legendre condition for minimisers

Proposition (Legendre condition)**.** If $y_0(x)$ is a local minimiser, then $P|_{y=y_0} \geq 0$.

We can say that the Legendre condition is a necessary condition for a minimiser. In less formal terms, P is 'more important' than Q when determining if a stationary point is a minimiser.

Proof. This condition is not proven rigorously. However, the general idea of the proof is to construct a function η which is small everywhere (giving a small Q contribution), but oscillates very rapidly near some point x_0, at which $P < 0$. This gives a large P contribution which can overpower the Q contribution. Then this gives $Q\eta^2 + P(\eta')^2 < 0$ if there exists some x_0 where $P|_{y=y_0} < 0$. □

Note that the Legendre condition is not a sufficient condition for local minima, but $P > 0$ and $Q \geq 0$ is sufficient.

Example. Consider again the brachistochrone problem.

$$f = \sqrt{\frac{1+(y')^2}{-y}}$$

We have

$$\frac{\partial f}{\partial y} = -\frac{1}{2y} f$$

$$\frac{\partial f}{\partial y'} = \frac{y'}{\sqrt{1+(y')^2}\sqrt{-y}}$$

Hence

$$P = \frac{1}{(1+(y')^2)^{\frac{3}{2}}\sqrt{-y}} > 0$$

$$Q = \frac{1}{2\sqrt{1+(y^2)^2}y^2\sqrt{-y}} > 0$$

Hence the cycloid is a local minimiser of the time taken to travel between the two points.

7.3. Associated eigenvalue problem

When deriving the minimiser condition, we had the integrand

$$Q\eta^2 + P(\eta')^2$$

We can integrate this by parts:

$$Q\eta^2 + \frac{\mathrm{d}}{\mathrm{d}x}(P\eta\eta') - \eta\frac{\mathrm{d}}{\mathrm{d}x}(P\eta')$$

giving

$$\delta^2 F[y] = \frac{1}{2}\int_\alpha^\beta \eta[-(P\eta')' + Q\eta]\,\mathrm{d}x$$

The bracketed term $-(P\eta')' + Q\eta$ is known as the Sturm–Liouville operator acting on η, denoted $\mathcal{L}(\eta)$. If there exists η such that $\mathcal{L}(\eta) = -\omega^2\eta$, $\omega \in \mathbb{R}$, and $\eta(\alpha) = \eta(\beta) = 0$, then y is not a minimiser, since the integrand will be $-\omega^2\eta^2 < 0$.

Example. Consider

$$F[y] = \int_0^\beta \left((y')^2 - y^2\right)\mathrm{d}x$$

such that

$$y(0) = y(\beta) = 0; \quad \beta \neq k\pi, k \in \mathbb{N}$$

The Euler–Lagrange equation gives

$$y'' + y = 0$$

Thus, constrained to the boundary conditions, the only stationary point of F is

$$y \equiv 0$$

Analysing the second variation,

$$\delta^2 F[0] = \frac{1}{2}\int_0^\beta \left[\eta'^2 - \eta^2\right]\mathrm{d}x$$

giving

$$P = 1 > 0; \quad Q < 0$$

Let us now examine the eigenvalue problem, since we cannot find whether $y \equiv 0$ is a minimiser from what we know already. Consider the eigenvalue problem

$$-\eta'' - \eta = -\omega^2\eta; \quad \eta(0) = \eta(\beta) = 0$$

Let us take

$$\eta = A\sin\left(\frac{\pi x}{\beta}\right)$$

to give

$$\left(\frac{\pi}{\beta}\right)^2 = 1 - \omega^2$$

So this has a solution $\omega > 0$ if and only if $\beta > \pi$. If $P > 0$, a problem may arise if the interval of integration is 'too large' (in this case $\beta > \pi$). Next lecture we will make this notion precise.

7.4. Jacobi accessory condition

Legendre tried to prove that $P > 0$ implied local minimality; obviously this was impossible due to the counterexample shown above. However, the method he used is still useful to analyse, since we can find an actual sufficient condition using the same idea. Let $\phi(x)$ be any differentiable function of x on $[\alpha, \beta]$. Then note that

$$\int_\alpha^\beta \frac{\mathrm{d}}{\mathrm{d}x}(\phi\eta^2)\,\mathrm{d}x = 0$$

since $\eta(\alpha) = \eta(\beta) = 0$. We can expand the integrand to give

$$\int_\alpha^\beta \left(\phi'\eta^2 + 2\eta\eta'\phi\right)\mathrm{d}x = 0$$

We can add this new zero to both sides of the second variation equation.

$$\delta^2 F[y] = \frac{1}{2}\int_\alpha^\beta \left(P(\eta')^2 + 2\eta\eta'\phi + (Q+\phi')\eta^2\right)\mathrm{d}x$$

Now, suppose that $P > 0$ at a particular y. Then, we can complete the square on the integrand, giving

$$\delta^2 F[y] = \frac{1}{2}\int_\alpha^\beta \left(P\left(\eta' + \frac{\phi}{P}\eta\right)^2 + \left(Q + \phi' - \frac{\phi^2}{P}\right)\eta^2\right)\mathrm{d}x$$

If we could choose a ϕ such that the second bracket vanishes, then the integrand would be $P\left(\eta' + \frac{\phi}{P}\eta\right)^2$. The only way the integral can be zero is if $\eta' + \frac{\phi}{P}\eta \equiv 0$. Since $\eta = 0$ at α, we have $\eta'(\alpha) = 0$. Hence, $\eta \equiv 0$ by the uniqueness of solutions to first order differential equations. Therefore, by contradiction, the integrand is not identically zero, and the second variation is positive. Now, such a ϕ function is given by

$$\phi^2 = P(Q + \phi')$$

If a solution to this differential equation exists, then $\delta^2 F[y] > 0$. We can transform this non-linear equation into a second order equation by the substitution $\phi = -P\frac{u'}{u}$ for some function $u \neq 0$. We have

$$P\left(\frac{u'}{u}\right)^2 = Q - \left(\frac{Pu'}{u}\right)' = Q - \frac{(Pu')'}{u} + P\left(\frac{u'}{u}\right)^2$$

Hence,

$$-(Pu')' + Qu = 0$$

This is known as the Jacobi accessory condition. Note that the left hand side is just $\mathcal{L}(u)$, where $\mathcal{L}$ is the Sturm–Liouville operator.

7.5. Solving the Jacobi condition

We need to find a solution to $\mathcal{L}(u) = 0$, where $u \neq 0$ on $[\alpha, \beta]$. The solution we find may not be nonzero on a large enough interval, in which case we would not have a local minimum.

Example. Consider

$$F[y] = \frac{1}{2}\int_\alpha^\beta \left((y')^2 - y^2\right) \mathrm{d}x$$

The second variation is

$$\delta^2 F[y] = \frac{1}{2}\int_\alpha^\beta \left((\eta')^2 - \eta^2\right) \mathrm{d}x$$

In this case, $P = 1, Q = -1$. The Jacobi accessory equation is

$$u'' + u = 0$$

We can solve this to find

$$u = A\sin x - B\cos x; \quad A, B \in \mathbb{R}$$

We want this to be nonzero on the interval $[\alpha, \beta]$. In particular,

$$\tan x \neq \frac{B}{A}; \quad \forall x \in [\alpha, \beta]$$

Note that $\tan x$ repeats every π, so if $|\beta - \alpha| < \pi$ we have a positive second variation for any stationary y.

Example. Consider again the geodesic on a sphere.

$$F[\theta] = \int \sqrt{\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2} = \int \sqrt{(\theta')^2 + \sin^2\theta}\,\mathrm{d}\phi$$

We have already proven that critical points of this functional are segments of great circles. Considering an equatorial great circle (since all great circles are equatorial under a change of perspective),

$$\theta = \frac{\pi}{2}$$

Consider ϕ_1, ϕ_2 on this great circle. The minor arc is clearly the shortest path, but the major arc is also a stationary point and must still be analysed.

$$P = 1; \quad Q = -1$$

Thus,

$$\delta^2 F\left[\theta_0 = \frac{\pi}{2}\right] = \frac{1}{2}\int_{\phi_1}^{\phi_2} \left((\eta')^2 - \eta^2\right) \mathrm{d}\phi$$

which is exactly the example from above. This is a minimiser if $|\phi_2 - \phi_1| < \pi$, which is exactly the condition of being a minor arc. If $\phi_2 - \phi_1 = \pi$, we have an infinite amount of geodesics, since these represent antipodal points. The set of geodesics exhibit rotational symmetry.

III. Markov Chains

Lectured in Michaelmas 2021 by Dr. P. Sousi

A Markov chain is a common type of random process, where each state in the process depends only on the previous one. Due to their simplicity, Markov processes show up in many areas of probability theory and have lots of real-world applications, for example in computer science.

One example of a Markov chain is a simple random walk, where a particle moves around an infinite lattice of points, choosing its next direction to move at random. It turns out that if the lattice is one- or two-dimensional, the particle will return to its starting point infinitely many times, with probability 1. However, if the lattice is three-dimensional or higher, the particle has probability 0 of ever returning to its starting point.

Contents

1. Introduction 103
1.1. Definition 103
1.2. Sequence definition 103
1.3. Point masses 104
1.4. Independence of sequences 104
1.5. Simple Markov property 104
1.6. Powers of the transition matrix 106
1.7. Calculating powers 106
2. Elementary properties 108
2.1. Communicating classes 108
2.2. Hitting times 109
2.3. Birth and death chain 113
2.4. Mean hitting times 113
2.5. Strong Markov property 115
3. Transience and recurrence 117
3.1. Definitions 117
3.2. Probability of visits 117
3.3. Duality of transience and recurrence 118
3.4. Recurrent communicating classes 119
4. Pólya's recurrence theorem 121
4.1. Statement of theorem 121
4.2. One-dimensional proof 121
4.3. Two-dimensional proof 122
4.4. Three-dimensional proof 123
5. Invariant distributions 124
5.1. Invariant distributions 124
5.2. Conditions for unique invariant distribution 125
5.3. Uniqueness of invariant distributions 126
5.4. Positive and null recurrence 128
5.5. Time reversibility 130
5.6. Aperiodicity 132
5.7. Positive recurrent limiting behaviour 133
5.8. Null recurrent limiting behaviour 135

1. Introduction

1.1. Definition

Let I be a finite or countable set. All of our random variables will be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition. A stochastic process $(X_n)_{n\geq 0}$ is called a *Markov chain* if for all $n \geq 0$ and for all $x_1 \dots x_{n+1} \in I$,

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, \dots, X_1 = x_1) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

We can think of n as a discrete measure of time. If $\mathbb{P}(X_{n+1} = y \mid X_n = x)$ for all x, y is independent of n, then X is called a time-homogeneous Markov chain. Otherwise, X is called time-inhomogeneous. In this course, we only study time-homogeneous Markov chains. If we consider only time-homogeneous chains, we may as well take $n = 0$ and we can write

$$P(x, y) = \mathbb{P}(X_1 = y \mid X_0 = x); \quad \forall x, y \in I$$

Definition. A *stochastic matrix* is a matrix where the sum of each row is equal to 1.

We call P the *transition matrix*. It is a stochastic matrix:

$$\sum_{y \in I} P(x, y) = 1$$

Remark. The index set does not need to be $\mathbb{N}$; it could alternatively be the set $\{0, 1, \dots, N\}$ for $N \in \mathbb{N}$.

We say that X is Markov (λ, P) if X_0 has distribution λ, and P is the transition matrix. Hence,

(i) $\mathbb{P}(X_0 = x_0) = \lambda_{x_0}$

(ii) $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, \dots, X_0 = x_0) = P(x_n, x_{n+1}) =: P_{x_n x_{n+1}}$

We usually draw a diagram of the transition matrix using a graph. Directed edges between nodes are labelled with their transition probabilities.

1.2. Sequence definition

Theorem. The process X is Markov (λ, P) if and only if $\forall n \geq 0$ and all $x_0, \dots, x_n \in I$, we have

$$\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \lambda_{x_0} P(x_0, x_1) P(x_1, x_2) \dots P(x_{n-1}, x_n)$$

Proof. If X is Markov, then we have

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) &= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &\quad \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &= P(x_{n-1}, x_n)\mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &= P(x_{n-1}, x_n) \dots P(x_0, x_1)\lambda_{x_0}\end{aligned}$$

as required. Conversely, $\mathbb{P}(X_0 = x_0) = \lambda_{x_0}$ satisfies (i). The transition matrix is given by

$$\mathbb{P}(X_n = x_n \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1}) = \frac{\lambda_{x_0} P(x_0, x_1) \dots P(x_{n-1}, x_n)}{\lambda_{x_0} P(x_0, x_1) \dots P(x_{n-2}, x_{n-1})} = P(x_{n-1}, x_n)$$

which is exactly the Markov property as required. □

1.3. Point masses

Definition. For $i \in I$, the δ_i-mass at i is defined by

$$\delta_{ij} = \mathbb{1}(i = j)$$

This is a probability measure that has probability 1 at i only.

1.4. Independence of sequences

Recall that discrete random variables (X_n) are considered independent if for all $x_1, \dots, x_n \in I$, we have

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \dots \mathbb{P}(X_n = x_n)$$

A sequence (X_n) is independent if for all k, $i_1 < i_2 < \cdots < i_n$ and for all $x_1, \dots, x_k$, we have

$$\mathbb{P}\left(X_{i_1} = x_1, \dots, X_{i_k} = x_k\right) = \prod_{j=1}^{n} \mathbb{P}\left(X_{i_j} = x_j\right)$$

Let $X = (X_n), Y = (Y_n)$ be sequences of discrete random variables. They are independent if for all $k, m, i_1 < \cdots < i_k, j_1 < \cdots < j_m$,

$$\begin{aligned}&prob X_1 = x_1, \dots, X_{i_k} = x_{i_k}, Y_{j_1} = y_{j_1}, \dots, Y_{j_m} \\ &= \mathbb{P}\left(X_1 = x_1, \dots, X_{i_k} = x_{i_k}\right)\mathbb{P}\left(Y_{j_1} = y_{j_1}, \dots, Y_{j_m}\right)\end{aligned}$$

1.5. Simple Markov property

Theorem. Suppose X is Markov(λ, P). Let $m \in \mathbb{N}$ and $i \in I$. Given that $X_m = i$, we have that the process after time m, written $(X_{m+n})_{n \geq 0}$, is Markov(δ_i, P), and it is independent of $X_0, \dots, X_m$.

Informally, the past and the future are independent given the present.

Proof. We must show that

$$\mathbb{P}(X_m = x_0, \dots, X_{m+n} = x_n \mid X_m = i) = \delta_{ix_0} P(x_0, x_1) \dots P(x_{n-1}, x_n)$$

We have

$$\mathbb{P}(X_{m+n} = x_{m+n}, \dots, X_m = x_m \mid X_m = i) = \frac{\mathbb{P}(X_{m+n} = x_{m+n}, \dots, X_m = x_m)\, \delta_{ix_m}}{\mathbb{P}(X_m = i)}$$

The numerator is

$$\begin{aligned}
&\mathbb{P}(X_{m+n}, \dots, X_m = x_m) \\
&= \sum_{x_0, \dots, x_{m-1} \in I} \mathbb{P}(X_{m+n} = x_{m+n}, \dots, X_m = x_m, X_{m-1} = x_{m-1}, \dots, X_0 = x_0) \\
&= \sum_{x_0, \dots, x_{m-1}} \lambda_{x_0} P(x_0, x_1) \dots P(x_{m-1}, x_m) P(x_m, x_{m+1}) \dots P(x_{m+n-1}, x_{m+n}) \\
&= P(x_m, x_{m+1}) \dots P(x_{m+n-1}, x_{m+n}) \sum_{x_0, \dots, x_{m-1}} \lambda_{x_0} P(x_0, x_1) \dots P(x_{m-1}, x_m) \\
&= P(x_m, x_{m+1}) \dots P(x_{m+n-1}, x_{m+n}) \mathbb{P}(X_m = x_m)
\end{aligned}$$

Thus we have

$$\mathbb{P}(X_{m+n} = x_{m+n}, \dots, X_m = x_m \mid X_m = i) = P(x_m, x_{m+1}) \dots P(x_{m+n-1}, x_{m+n}) \delta_{ix_m}$$

Hence $(X_{m+n})_{n \geq 0} \sim \text{Markov}(\delta_i, P)$ conditional on $X_m = i$. Now it suffices to show independence between the past and future variables. In particular, we need to show $m \leq i_1 < \dots < i_k$ for some $k \in \mathbb{N}$ implies that

$$\begin{aligned}
&\mathbb{P}\left(X_{i_1} = x_{m+1}, \dots, X_{i_k} = x_{m+k}, X_0 = x_0, \dots, X_m = x_m \mid X_m = i\right) \\
&= \mathbb{P}\left(X_{i_1} = x_{m+1}, \dots, X_{i_k} = x_{m+k} \mid X_m = i\right) \mathbb{P}(X_0 = x_0, \dots, X_m = x_m \mid X_m = i)
\end{aligned}$$

So let $i = x_m$, and then

$$\begin{aligned}
&= \frac{\mathbb{P}\left(X_{i_1} = x_{m+1}, \dots, X_{i_k} = x_{m+k}, X_0 = x_0, \dots, X_m = x_m\right)}{\mathbb{P}(X_m = i)} \\
&= \frac{\lambda_{x_0} P(x_0, x_1) \dots P(x_{m-1}, x_m) \mathbb{P}\left(X_{i_1} = x_{m+1}, \dots, X_{i_k} = x_{m+k} \mid X_m = x_m\right)}{\mathbb{P}(x_m = i)} \\
&= \frac{\mathbb{P}(X_0 = x_0, \dots, X_m = x_m)}{\mathbb{P}(X_m = x_m)} \mathbb{P}\left(X_{i_1} = x_{m+1}, \dots, X_{i_k} = x_{m+k} \mid X_m = x_m\right)
\end{aligned}$$

which gives the result as required. □

1.6. Powers of the transition matrix

Suppose $X \sim \text{Markov}(\lambda, P)$ with values in I. If I is finite, then P is an $|I| \times |I|$ square matrix. In this case, we can label the states as $1, \dots, |I|$. If I is infinite, then we label the states using the natural numbers $\mathbb{N}$. Let $x \in I$ and $n \in \mathbb{N}$. Then,

$$\begin{aligned}\mathbb{P}(X_n = x) &= \sum_{x_0,\dots,x_{n-1} \in I} \mathbb{P}(X_n = x, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &= \sum_{x_0,\dots,x_{n-1} \in I} \lambda_{x_0} P(x_0, x_1) \dots P(x_{n-1}, x)\end{aligned}$$

We can think of λ as a row vector. So we can write this as

$$= (\lambda P^n)_x$$

By convention, we take $P^0 = I$, the identity matrix. Now, suppose $m, n \in \mathbb{N}$. By the simple Markov property,

$$\mathbb{P}(X_{m+n} = y \mid X_m = x) = \mathbb{P}(X_n = y \mid X_0 = x) = (\delta_x P^n)_y$$

We will write $\mathbb{P}_x(A) \coloneqq \mathbb{P}(A \mid X_0 = x)$ as an abbreviation. Further, we write $p_{ij}(n)$ for the (i, j) element of P^n. We have therefore proven the following theorem.

Theorem.

$$\mathbb{P}(X_n = x) = (\lambda P^n)_x;$$

$$\mathbb{P}(X_{n+m} = y \mid X_m = x) = \mathbb{P}_x(X_n = y) = p_{xy}(n)$$

1.7. Calculating powers

Example. Consider

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}; \quad \alpha, \beta \in [0, 1]$$

Note that for any stochastic matrix P, P^n is a stochastic matrix. First, we have $P^{n+1} = P^n P$. Let us begin by finding $p_{11}(n+1)$.

$$p_{11}(n+1) = p_{11}(n)(1-\alpha) + p_{12}(n)\beta$$

Note that $p_{11}(n) + p_{12}(n) = 1$ since P^n is stochastic. Therefore,

$$p_{11}(n+1) = p_{11}(n)(1-\alpha-\beta) + \beta$$

We can solve this recursion relation to find

$$p_{11}(n) = \begin{cases} \frac{\alpha}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n & \alpha+\beta > 0 \\ 1 & \alpha+\beta = 0 \end{cases}$$

The general procedure for finding P^n is as follows. Suppose that P is a $k \times k$ matrix. Then let $\lambda_1, \dots, \lambda_k$ be its eigenvalues (which may not be all distinct).

(1) All λ_i distinct. In this case, P is diagonalisable, and hence we can write $P = UDU^{-1}$ where U is a diagonal matrix, whose diagonal entries are the λ_i. Then, $P^n = UD^nU^{-1}$. Calculating D^n may be done termwise since D is diagonal. In this case, we have terms such as

$$p_{11}(n) = a_1\lambda_1^n + \cdots + a_k\lambda_k^n; \quad a_i \in \mathbb{R}$$

First, note $P^0 = I$ hence $p_{11}(0) = 1$. We can substitute small values of n and then solve the system of equations. Now, suppose λ_k is complex for some k. In this case, $\overline{\lambda_k}$ is also an eigenvalue. Then, up to reordering,

$$\lambda_k = re^{i\theta} = r(\cos\theta + i\sin\theta); \lambda_{k-1} = \overline{\lambda_k} = re^{i\theta} = r(\cos\theta - i\sin\theta)$$

We can instead write $p_{11}(n)$ as

$$p_{11}(n) = a_1\lambda_1^n + \cdots + a_{k-1}r^n\cos(n\theta) + a_kr^n\sin(n\theta)$$

Since $p_{11}(n)$ is real, all the imaginary parts disappear, so we can simply ignore them.

(2) Not all λ_i distinct. In this case, λ appears with multiplicity 2, then we include also the term $(an + b)\lambda^n$ as well as $b\lambda^n$. This can be shown by considering the Jordan normal form of P.

Example. Let

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

The eigenvalues are $1, \frac{1}{2}i, -\frac{1}{2}i$. Then, writing $\frac{i}{2} = \frac{1}{2}(\cos\frac{\pi}{2} + i\sin\frac{\pi}{2})$, we can write

$$p_{11}(n) = \alpha + \beta\left(\frac{1}{2}\right)^n\cos\frac{n\pi}{2} + \gamma\left(\frac{1}{2}\right)^n\sin\frac{n\pi}{2}$$

For $n = 0$ we have $p_{11}(0) = 1$, and for $n = 1$ we have $p_{11}(1) = 0$, and for $n = 2$ we can calculate P^2 and find $p_{11}(2) = 0$. Solving this system of equations for α, β, γ, we can find

$$p_{11}(n) = \frac{1}{5} + \left(\frac{1}{2}\right)^n\left(\frac{4}{5}\cos\frac{n\pi}{2} - \frac{2}{5}\sin\frac{n\pi}{2}\right)$$

2. Elementary properties

2.1. Communicating classes

Definition. Let X be a Markov chain with transition matrix P and values in I. For $x, y \in I$, we say that x *leads to* y, written $x \to y$, if

$$\mathbb{P}_x(\exists n \geq 0, X_n = y) > 0$$

We say that x *communicates with* y and write $x \leftrightarrow y$ if $x \to y$ and $y \to x$.

Theorem. The following are equivalent:

(i) $x \to y$

(ii) There exists a sequence of states $x = x_0, x_1, \dots, x_k = y$ such that

$$P(x_0, x_1)P(x_1, x_2) \dots P(x_{k-1}, x_k) > 0$$

(iii) There exists $n \geq 0$ such that $p_{xy}(n) > 0$.

Proof. First, we show (i) and (iii) are equivalent. If $x \to y$, then $\mathbb{P}_x(\exists n \geq 0, X_n = y) > 0$. Then if $\mathbb{P}_x(\exists n \geq 0, X_n = y) > 0$ we must have some $n \geq 0$ such that $\mathbb{P}_x(X_n = y) = p_{xy}(n) > 0$. Note that we can write (i) as $\mathbb{P}_x\left(\bigcup_{n=0}^{\infty} X_n = y\right) > 0$. If there exists $n \geq 0$ such that $p_{xy}(n) > 0$, then certainly the probability of the union is also positive.

Now we show (ii) and (iii) are equivalent. We can write

$$p_{xy}(n) = \sum_{x_1, \dots, x_{n-1}} P(x, x_1) \dots P(x_{n-1}, y)$$

which leads directly to the equivalence of (ii) with (iii). □

Corollary. Communication is an equivalence relation on I.

Proof. $x \leftrightarrow x$ since $p_{xx}(0) = 1$. If $x \to y$ and $y \to z$ then by (ii) above, $x \to z$. □

Definition. The equivalence classes induced on I by the communication equivalence relation are called *communicating classes*. A communicating class C is *closed* if $x \in C, x \to y \implies y \in C$.

Definition. A transition matrix P is called *irreducible* if it has a single communicating class. In other words, $\forall x, y \in I, x \leftrightarrow y$.

Definition. A state x is called *absorbing* if $\{x\}$ is a closed (communicating) class.

2.2. Hitting times

Definition. For $A \subseteq I$, we define the *hitting time* of A to be a random variable $T_A : \Omega \to \{0, 1, 2 \dots\} \cup \{\infty\}$, defined by

$$T_A(\omega) = \inf\{n \geq 0 : X_n(\omega) \in A\}$$

with the convention that $\inf \varnothing = \infty$. The *hitting probability* of A is $h^A : I \to [0, 1]$, defined by

$$h_i^A = \mathbb{P}_i(T_A < \infty)$$

The *mean hitting time* of A is $k^A : I \to [0, \infty]$, defined by

$$k_i^A = \mathbb{E}_i[T_A] = \sum_{n=0}^{\infty} n\mathbb{P}_i(T_A = n) + \infty\mathbb{P}_i(T_A = \infty)$$

Example. Consider

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Consider $A = \{4\}$.

$$h_1^A = 0$$

$$h_2^A = \mathbb{P}_2(T_A < \infty) = \frac{1}{2}h_1^A + \frac{1}{2}h_3^A$$

$$h_3^A = \frac{1}{2} \cdot 1 + \frac{1}{2}h_2^A$$

Hence $h_2^A = \frac{1}{3}$. Now, consider $B = \{1, 4\}$.

$$k_1^B = k_4^B = 0$$

$$k_2^B = 1 + \frac{1}{2}k_1^B + \frac{1}{2}k_3^B$$

$$k_3^B = 1 + \frac{1}{2}k_4^B + \frac{1}{2}k_2^B$$

Hence $k_2^B = 2$.

Theorem. Let $A \subset I$. Then the vector $(h_i^A)_{i \in A}$ is the minimal non-negative solution to the system

$$h_i^A = \begin{cases} 1 & i \in A \\ \sum_j P(i, j)h_j^A & i \notin A \end{cases}$$

Minimality here means that if $(x_i)_{i \in I}$ is another non-negative solution, then $\forall i, h_i^A \leq x_i$.

Note. The vector $h_i^A = 1$ always satisfies the equation, since P is stochastic, but is typically not minimal.

Proof. First, we will show that $(h_i)_{i\in A}$ solves the system of equations. Certainly if $i \in A$ then $h_i^A = 1$. Suppose $i \notin A$. Consider the event $\{T_A < \infty\}$. We can write this event as a disjoint union of the following events:

$$\{T_A < \infty\} = \{X_0 \in A\} \cup \bigcup_{n=1}^{\infty} \{X_0 \notin A, \dots, X_{n-1} \notin A, X_n \in A\}$$

By countable additivity,

$$\begin{aligned}
\mathbb{P}_i(T_A < \infty) &= \underbrace{\mathbb{P}_i(X_0 \in A)}_{=0} + \sum_{n=1}^{\infty} \mathbb{P}_i(X_0 \notin A, \dots, X_{n-1} \notin A, X_n \in A) \\
&= \sum_{n=1}^{\infty} \sum_j \mathbb{P}(X_0 \notin A, \dots, X_{n-1} \notin A, X_n \in A, X_1 \in j \mid X_0 = i) \\
&= \sum_j \mathbb{P}(X_1 \in A, X_1 = j \mid X_0 = i) \\
&+ \sum_{n=2}^{\infty} \sum_j \mathbb{P}(X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A, X_1 \in j \mid X_0 = i) \\
&= \sum_j P(i,j) \mathbb{P}(X_1 \in A \mid X_1 = j, X_0 = i) \\
&+ \sum_j P(i,j) \sum_{n=2}^{\infty} \mathbb{P}(X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A \mid X_1 \in j, X_0 = i)
\end{aligned}$$

By the definition of the Markov chain, we can drop the condition on X_0, and subtract one from all indices.

$$
\begin{aligned}
&= \sum_j P(i,j)\mathbb{P}\left(X_0 \in A \mid X_0 = j\right) \\
&+ \sum_j P(i,j) \sum_{n=2}^{\infty} \mathbb{P}\left(X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A \mid X_1 \in j\right) \\
&= \sum_j P(i,j)\mathbb{P}\left(X_0 \in A \mid X_0 = j\right) \\
&+ \sum_j P(i,j) \sum_{n=2}^{\infty} \mathbb{P}_j\left(X_0 \notin A, \dots, X_{n-2} \notin A, X_{n-1} \in A\right) \\
&= \sum_j P(i,j)\left(\mathbb{P}_j\left(X_0 \in A\right) + \sum_2^{\infty} \mathbb{P}_j\left(X_0 \notin A, \dots, X_{n-1} \notin A, X_n \in A\right)\right) \\
&= \sum_j P(i,j)\left(\mathbb{P}_j\left(T_A = 0\right) + \sum_{n=1}^{\infty} \mathbb{P}_j\left(T_A = n\right)\right) \\
&= \sum_j P(i,j)\mathbb{P}_j\left(T_A < \infty\right) \\
&= \sum_j P(i,j) h_j^A
\end{aligned}
$$

Now we must show minimality. If (x_i) is another non-negative solution, we must show that $h_i^A \leq x_i$. We have

$$x_i = \sum_j P(i,j)x_j = \sum_{j \in A} P(i,j) + \sum_{j \notin A} P(i,j)x_j$$

Substituting again,

$$x_i = \sum_{j \in A} P(i,j)x_j + \sum_{j \notin A} P(i,j)\left(\sum_{k \in A} P(j,k) + \sum k \notin A P(j,k)x_k\right)$$

Then

$$
\begin{aligned}
x_i = &\sum_{j_1 \in A} P(i,j_1) + \sum_{j_1 \notin A} \sum_{j_2 \in A} P(i,j_1)P(j_1,j_2) + \cdots \\
&+ \sum_{j_1 \notin A, \dots, j_{n-1} \notin A, j_n \in A} P(i,j_1) \dots P(j_{n-1},j_n) \\
&+ \sum_{j_1 \notin A \dots, j_n \notin A} P(i,j_1) \dots P(j_{n-1},j_n)x_{j_n}
\end{aligned}
$$

The last term is non-negative since x is non-negative. So

$$x_i \geq \mathbb{P}_i\left(T_A = 1\right) + \mathbb{P}_i\left(T_A = 2\right) + \cdots + \mathbb{P}_i\left(T_A = n\right) \geq \mathbb{P}_i\left(T_A \leq n\right), \ \forall n \in \mathbb{N}$$

Now, note $\{T_A \leq n\}$ are a set of increasing functions of n, so by continuity of the probability measure, the probability increases to that of the union, $\{T_A < \infty\} = h_i^A$. □

Example. Consider the Markov chain previously explored:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Let $A = \{4\}$. Then $h_1^A = 0$ since there is no route from 1 to 4. From the theorem above, the system of linear equations is

$$h_2 = \frac{1}{2}h_1 + \frac{1}{2}h_3$$

$$h_3 = \frac{1}{2}h_4 + \frac{1}{2}h_2$$

$$h_4 = 1$$

Hence,

$$h_2 = \frac{2}{3}h_1 + \frac{1}{3}$$

$$h_3 = \frac{1}{3}h_1 + \frac{2}{3}$$

So the minimal solution arises at $h_1 = 0$.

Example. Consider $I = \mathbb{N}$, and

$$P(i, i+1) = p \in (0,1); \quad P(i, i-1) = 1 - p = q$$

Then $h_i = \mathbb{P}_i(T_0 < \infty)$ hence $h_0 = 1$. The linear equations are

$$p \neq q \implies h_i = ph_{i+1} + qh_{i-1}$$

$$p(h_{i+1} - h_i) = q(h_i - h_{i-1})$$

Let $u_i = h_i - h_{i-1}$. Then,

$$\frac{q}{p}u_i = \cdots = \left(\frac{q}{p}\right)^i u_1$$

Hence

$$h_i = \sum_{j=1}^{i}(h_j - h_{j-1}) + 1 = 1 - (1 - h_1)\sum_{j=1}^{i}\left(\frac{q}{p}\right)^j$$

The general solution is therefore

$$h_i = a + b\left(\frac{q}{p}\right)^i$$

If $q > p$, then minimality of h_i implies $b = 0$, $a = 1$. Hence,

$$h_i = 1$$

Otherwise, if $p > q$, minimality of h_i implies $a = 0, b = 1$. Hence,

$$h_i = \left(\frac{q}{p}\right)^i$$

If $p = q = \frac{1}{2}$, then

$$h_i = \frac{1}{2}h_{i+1} + \frac{1}{2}h_{i-1}$$

Hence, $h_i = a + bi$. Minimality implies $a = 1$ and $b = 0$.

$$h_i = 1$$

2.3. Birth and death chain

Consider a Markov chain on $\mathbb{N}$ with

$$P(i, i+1) = p_i; \quad P(i, i-1) = q_i; \quad \forall i,\ p_i + q_i = 1$$

Now, consider $h_i = \mathbb{P}_i(T_0 < \infty)$. $h_0 = 1$, and $h_i = p_i h_{i+1} + q_i h_{i-1}$.

$$p_i(h_{i+1} - h_i) = q_i(h_i - h_{i-1})$$

Let $u_i = h_i - h_{i-1}$ to give

$$u_{i+1} = \frac{q_i}{p_i}u_i = \underbrace{\prod j = 1^i \frac{q_i}{p_i}}_{\gamma_i} u_i$$

Then

$$h_i = 1 - (1 - h_1)(\gamma_0 + \gamma_1 + \cdots + \gamma_{i-1})$$

where we let $\gamma_0 = 1$. Since h_i is the minimal non-negative solution,

$$h_i \geq 0 \implies 1 - h_1 \leq \frac{1}{\sum_{j=0}^{i-1}\gamma_j} \leq \frac{1}{\sum_{j=0}^{\infty}\gamma_j}$$

By minimality, we must have exactly this bound. If $\sum_{j=0}^{\infty}\gamma_j = \infty$ then $1-h_1 = 0 \implies h_i = 1$ for all i. If $\sum_{j=0}^{\infty}\gamma_j < \infty$ then

$$h_i = \frac{\sum_{j=i}^{\infty}\gamma_j}{\sum_{j=0}^{\infty}\gamma_j}$$

2.4. Mean hitting times

Recall that

$$k_i^A = \mathbb{E}_i[T_A] = \sum_n n\mathbb{P}_i(T_A = n) + \infty\mathbb{P}_i(T_A = \infty)$$

Theorem. The vector $(k_i^A)_{i\in I}$ is the minimal non-negative solution to the system of equations

$$k_i^A = \begin{cases} 0 & \text{if } i \in A \\ 1 + \sum_{j\notin A} P(i,j)k_j^A & \text{if } i \notin A \end{cases}$$

Proof. Suppose $i \in A$. Then $k_i = 0$. Now suppose $i \notin A$. Further, we may assume that $\mathbb{P}_i(T_A = \infty) = 0$, since if that probability is positive then the claim is trivial. Indeed, if $\mathbb{P}_i(T_A = \infty) > 0$, then there must exist j such that $P(i,j) > 0$ and $\mathbb{P}_j(T_A = \infty) > 0$ since

$$\mathbb{P}_i(T_A < \infty) = \sum_j P(i,j)h_j^A \implies 1 - \mathbb{P}_i(T_A = \infty) = \sum_j P(i,j)\big(1 - \mathbb{P}_j(T_A = \infty)\big)$$

Because P is stochastic,

$$\mathbb{P}_i(T_A = \infty) = \sum_j P(i,j)\mathbb{P}_j(T_A = \infty)$$

so since the left hand side is positive, there must exist j with $P(i,j) > 0$ and $\mathbb{P}_j(T_A = \infty > 0)$. For this j, we also have $k_j^A = \infty$. Now we only need to compute $\sum_n n\mathbb{P}_i(T_A = n)$.

$$\mathbb{P}_i(T_A = n) = \mathbb{P}_i(X_0 \notin A, \dots, X_{n-1} \notin A, X_n \in A)$$

Then, using the same method as the previous theorem,

$$k_i^A = \sum_n n\mathbb{P}_i(T_A = n) = 1 + \sum_{j\notin A} P(i,j)k_j^A$$

It now suffices to prove minimality. Suppose (x_i) is another solution to this system of equations. We need to show that $x_i \geq k_i^A$ for all i. Suppose $i \notin A$. Then

$$x_i = 1 + \sum_{j\notin A} P(i,j)x_j = 1 + \sum_{j\notin A} P(i,j)\left(1 + \sum_{k\notin A} P(j,k)x_k\right)$$

Expanding inductively,

$$\begin{aligned} x_i = 1 &+ \sum_{j_1\notin A} P(i,j_1) + \sum_{j_1\notin A, j_2\notin A} P(i,j_1)P(j_1,j_2) + \cdots \\ &+ \sum_{j_1\notin A,\dots,j_n\notin A} P(i,j_1)\dots P(j_{n-1},j_n) + \sum_{j_1\notin A,\dots,j_{n+1}\notin A} P(i,j)\dots P(j_n,j_{n+1})x_{j_{n+1}} \end{aligned}$$

Since x is non-negative, we can remove the last term and reach an inequality.

$$x_i \geq 1 + \sum_{j_1\notin A} P(i,j_1) + \sum_{j_1\notin A, j_2\notin A} P(i,j_1)P(j_1,j_2) + \cdots + \sum_{j_1\notin A,\dots,j_n\notin A} P(i,j_1)\dots P(j_{n-1},j_n)$$

Hence

$$\begin{aligned} x_i &\geq 1 + \mathbb{P}_i(T_A > 1) + \mathbb{P}_i(T_A > 2) + \cdots + \mathbb{P}_i(T_A > n) \\ &= \mathbb{P}_i(T_A > 0) + \mathbb{P}_i(T_A > 1) + \mathbb{P}_i(T_A > 2) + \cdots + \mathbb{P}_i(T_A > n) \\ &= \sum_{k=0}^{n} \mathbb{P}_i(T_A > k) \end{aligned}$$

for all n. Hence, the limit of this sum is

$$x_i \geq \sum_{k=0}^{\infty} \mathbb{P}_i(T_A > k) = \mathbb{E}_i[T_A]$$

which gives minimality as required. □

2.5. Strong Markov property

The simple Markov property shows that, if $X_m = i$,

$$X_{m+n} \sim \text{Markov}(\delta_i, P)$$

and this is independent of $X_0, \ldots, X_m$. The strong Markov property will show that the same property holds when we replace m with a finite random 'time' variable. It is not the case that *any* random variable will work; indeed, an m very dependent on the Markov chain itself might not satisfy this property.

Definition. A random time $T : \Omega \to \{0, 1, \ldots\} \cup \{\infty\}$ is called a *stopping time* if, for all $n \in \mathbb{N}$, $\{T = n\}$ depends only on $X_0, \ldots, X_n$.

Example. The hitting time $T_A = \inf\{n \geq 0 : X_n \in A\}$ is a stopping time. This is because we can write

$$\{T_A = n\} = \{X_0 \notin A, \ldots, X_{n-1} \notin A, X_n \in A\}$$

Example. The time $L_A = \sup\{n \geq 0 : X_n \in A\}$ is not a stopping time. This is because we need to know information about the future behaviour of X_n in order to guarantee that we are at the supremum of such events.

Theorem (Strong Markov Property)**.** Let $X \sim \text{Markov}(\lambda, P)$ and T be a stopping time. Conditional on $T < \infty$ and $X_T = i$,

$$(X_{n+T})_{n \geq 0} \sim \text{Markov}(\delta_i, P)$$

and this distribution is independent of $X_0, \ldots, X_T$.

Proof. We need to show that, for all $x_0, \ldots, x_n$ and for all vectors w of any length,

$$\begin{aligned} &\mathbb{P}(X_T = x_0, \ldots, X_{T+n} = x_n, (X_0, \ldots, X_T) = w \mid T < \infty, X_T = i) \\ &= \delta_{ix_0} P(x_0, x_1) \ldots P(x_{n-1}, x_n) \mathbb{P}((X_0, \ldots, X_T) = w : T < \infty, X_T = i) \end{aligned}$$

Suppose that w is of the form $w = (w_0, \dots, w_k)$. Then,

$$\begin{aligned}&\mathbb{P}(X_T = X_0, \dots, X_{T+n} = x_n, (X_0, \dots, X_T) = w \mid T < \infty, X_T = i)\\ &= \frac{\mathbb{P}(X_k = x_0, \dots, X_{k+n} = x_n, (X_0, \dots, X_k) = w, T = k, X_k = i)}{\mathbb{P}(T < \infty, X_T = i)}\end{aligned}$$

Now, since $\{T = k\}$ depends only on $X_0, \dots, X_k$, by the simple Markov property we have

$$\begin{aligned}&\mathbb{P}(X_k = x_0, \dots, X_{k+n} = x_n \mid (X_0, \dots, X_k) = w, T = k, X_k = i)\\ &= \mathbb{P}(X_k = x_0, \dots, X_{k+n} = x_n \mid X_k = i) = \delta_{ix_0} P(x_0, x_1) \dots P(x_{n-1}, x_n)\end{aligned}$$

Now,

$$\begin{aligned}&\mathbb{P}(X_T = x_0, \dots, X_{T+n} = x_n, (X_0, \dots, X_T) = w \mid T < \infty, X_T = i)\\ &= \frac{\delta_{ix_0} P(x_0, x_1) \dots P(x_{n-1}, x_n) \mathbb{P}((X_0, \dots, X_k) = w : T = k, X_k = i)}{\mathbb{P}(T < \infty, X_T = i)}\\ &= \delta_{ix_0} P(x_0, x_1) \dots P(x_{n-1}, x_n) \mathbb{P}((X_0, \dots, X_T) = w : T < \infty, X_T = i)\end{aligned}$$

as required. □

Example. Consider a simple random walk on $I = \mathbb{N}$, where $P(x, x \pm 1) = \frac{1}{2}$ for $x \neq 0$, and $P(0, 1) = 1$. Now, let $h_i = \mathbb{P}_i(T_0 < \infty)$. We want to calculate h_1. We can write

$$h_1 = \frac{1}{2} + \frac{1}{2} h_2$$

but the system of recursion relations this generates is difficult to solve. Instead, we will write

$$h_2 = \mathbb{P}_2(T_0 < \infty)$$

Note that in order to hit 0, we must first hit 1. So conditioning on the first hitting time of 1 being finite, after this time the process starts again from 1. We can write $T_0 = T_1 + \widetilde{T}_0$, where $\widetilde{T}_0$ is independent of T_1, with the same distribution as T_0 under $\mathbb{P}_1$. Now,

$$h_2 = \mathbb{P}_2(T_0 < \infty, T_1 < \infty) = \mathbb{P}_2(T_0 < \infty \mid T_1 < \infty)\, \mathbb{P}_2(T_2 < \infty)$$

Note that

$$\begin{aligned}\mathbb{P}_2(T_0 < \infty \mid T_1 < \infty) &= \mathbb{P}_2\left(T_1 + \widetilde{T}_0 < \infty \mid T_1 < \infty\right)\\ &= \mathbb{P}_2\left(\widetilde{T}_0 < \infty \mid T_1 < \infty\right)\\ &= \mathbb{P}_1(T_0 < \infty)\end{aligned}$$

But $\mathbb{P}_2(T_1 < \infty) = \mathbb{P}_1(T_0 < \infty)$, so

$$h_2 = \mathbb{P}_2(T_1 < \infty)\, \mathbb{P}_1(T_0 < \infty)$$

By translation invariance,

$$h_2 = h_1^2$$

In general, therefore, for any $n \in \mathbb{N}$,

$$h_n = h_1^n$$

3. Transience and recurrence

3.1. Definitions

Definition. Let X be a Markov chain, and let $i \in I$. i is called *recurrent* if

$$\mathbb{P}_i\left(X_n = i \text{ for infinitely many } n\right) = 1$$

i is called *transient* if

$$\mathbb{P}_i\left(X_n = i \text{ for infinitely many } n\right) = 0$$

We will prove that any i is either recurrent or transient.

3.2. Probability of visits

Definition. Let $T_i^{(0)} = 0$ and inductively define

$$T_i^{(r+1)} = \inf\left\{n \geq T_i^{(r)} + 1 : X_n = i\right\}$$

We write $T_i^{(1)} = T_i$, called the first return time (or first passage time) to i. Further, let

$$f_i = \mathbb{P}_i\left(T_i < \infty\right)$$

and let the number of visits to i be defined by

$$V_i = \sum_{n=0}^{\infty} 1(X_n = i)$$

Lemma. For all $r \in \mathbb{N}, i \in I$, $\mathbb{P}_i\left(V_i > r\right) = f_i^r$.

Proof. For $r = 0$, this is trivially true. Now, suppose that the statement is true for r, and we will show that it is true for $r + 1$.

$$\begin{aligned}
\mathbb{P}_i\left(V_i > r+1\right) &= \mathbb{P}_i\left(T_i^{(r+1)} < \infty\right) \\
&= \mathbb{P}_i\left(T_i^{(r+1)} < \infty, T_i^{(r)} < \infty\right) \\
&= \mathbb{P}_i\left(T_i^{(r+1)} < \infty \mid T_i^{(r)} < \infty\right) \mathbb{P}_i\left(T_i^{(r)} < \infty\right) \\
&= \mathbb{P}_i\left(T_i^{(r+1)} < \infty \mid T_i^{(r)} < \infty\right) \mathbb{P}_i\left(V_i > r\right) \\
&= \mathbb{P}_i\left(T_i^{(r+1)} < \infty \mid T_i^{(r)} < \infty\right) f_i^r
\end{aligned}$$

By the strong Markov property applied to the stopping time $T_i^{(r)}$,

$$\begin{aligned}
&= \mathbb{P}_i\left(T_i < \infty\right) f_i^r \\
&= f_i f_i^r \\
&= f_i^{r+1}
\end{aligned}$$

□

3.3. Duality of transience and recurrence

Theorem. Let X be a Markov chain with transition matrix P, and let $i \in I$. Then, exactly one of the following is true.

(i) If $\mathbb{P}_i(T_i < \infty) = 1$, then i is recurrent, and

$$\sum_{n=0}^{\infty} p_{ii}(n) = \infty$$

(ii) If $\mathbb{P}_i(T_i < \infty) < 1$, then i is transient, and

$$\sum_{n=0}^{\infty} p_{ii}(n) < \infty$$

Proof.

$$\begin{aligned}
\mathbb{E}_i[V_i] &= \mathbb{E}_i\left[\sum_{n=0}^{\infty} 1(X_n = i)\right] \\
&= \sum_{n=0}^{\infty} \mathbb{E}_i[1(X_n = i)] \\
&= \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = i) \\
&= \sum_{n=0}^{\infty} p_{ii}(n)
\end{aligned}$$

First, suppose $\mathbb{P}_i(T_i < \infty) = 1$. Then, for all r, $\mathbb{P}_i(V_i > r) = 1$, so $\mathbb{P}_i(V_i = \infty) = 1$. Hence, i is recurrent. Further, $\mathbb{E}_i[V_i] = \infty$ so $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$.

Now, if $f_i < 1$, by the previous lemma we see that $\mathbb{E}_i[V_i] = \frac{1}{1-f_i} < \infty$ hence $\mathbb{P}_i(V_i < \infty) = 1$. Thus, i is transient. Further, $\mathbb{E}_i[V_i] < \infty$ so $\sum_{n=0}^{\infty} p_{ii}(n) < \infty$. □

Theorem. Let x, y be states that communicate. Then, either x and y are both recurrent, or they are both transient.

Proof. Suppose x is recurrent. Then, since x and y communicate, $\exists m, \ell \in \mathbb{N}$ such that

$$p_{xy}(m) > 0; \quad p_{yx}(\ell) > 0$$

Note, $\sum_n p_{xx}(n) = \infty$. Then,

$$p_{yy}(n) \geq \sum_n p_{yy}(n + m + \ell) \geq \sum_n p_{yx}(\ell) p_{xx}(n) p_{xy}(m) \geq p_{yx}(\ell) p_{xy}(m) p_{xx}(n) = \infty$$

□

Corollary. Either all states in a communicating class are recurrent or they are all transient.

3.4. Recurrent communicating classes

Theorem. Any recurrent communicating class is closed.

Proof. Suppose a communicating class C is not closed. Then there exists $x \in C$ and $y \notin C$ such that $x \to y$. Let m be such that $p_{xy}(m) > 0$. If, starting from x, we hit y which is outside the communicating class, then we can never return to the communicating class (including x) again. In particular,

$$\mathbb{P}_x(V_x < \infty) \geq \mathbb{P}_x(X_m = y) = p_{xy}(m) > 0$$

Hence x is not recurrent, which is a contradiction. □

Theorem. Any finite closed communicating class is recurrent.

Proof. Let C be a finite closed communicating class. Let $x \in C$. Then, by the pigeonhole principle, there must exist $y \in C$ such that

$$\mathbb{P}_x(X_n = y \text{ for infinitely many } n) > 0$$

Since x and y communicate, there exists $m \in \mathbb{N}$ such that $p_{yx}(m) > 0$. Now,

$$\begin{aligned}\mathbb{P}_y(X_m = y \text{ for infinitely many } n) &\geq \mathbb{P}_x(X_m = x, X_n = y \text{ for infinitely many } n \geq m)\\ &= \mathbb{P}_x(X_n = y \text{ for infinitely many } n \geq m \mid X_m = x)\,\mathbb{P}_y(X_m = x)\\ &= \mathbb{P}_x(X_n = y \text{ for infinitely many } n)\,\mathbb{P}_y(X_m = x) > 0\end{aligned}$$

Thus y is recurrent. Since recurrence is a class property, C is recurrent. □

Theorem. Let P be irreducible and recurrent. Then, for all x, y,

$$\mathbb{P}_x\left(T_y < \infty\right) = 1$$

Proof. Since y is recurrent,

$$1 = \mathbb{P}_y(X_n = y \text{ for infinitely many } n)$$

Let m such that $p_{yx}(m) > 0$. Now,

$$\begin{aligned}1 &= \mathbb{P}_y(X_n = y \text{ infinitely often})\\ &= \sum_z \mathbb{P}_y(X_m = z, X_n = y \text{ for infinitely many } n \geq m)\\ &= \sum_z \mathbb{P}_y(X_n = y \text{ for infinitely many } n \geq m \mid X_m = z)\,\mathbb{P}_y(X_m = z)\\ &= \sum_z \mathbb{P}_z(X_n = y \text{ for infinitely many } n)\,\mathbb{P}_y(X_m = z)\end{aligned}$$

By the strong Markov property,

$$= \sum_z \mathbb{P}_z\left(T_y < \infty\right) \mathbb{P}_y\left(X_n = y \text{ for infinitely many } n\right) \mathbb{P}_y\left(X_m = z\right)$$

Since y is recurrent,

$$= \sum_z \mathbb{P}_z\left(T_y < \infty\right) \mathbb{P}_y\left(X_m = z\right)$$
$$= \sum_z \mathbb{P}_z\left(T_y < \infty\right) p_{yz}(m)$$

Since $p_{yz}(m) > 0$ and $\sum_z p_{yz}(m) = 1$, $\mathbb{P}_x\left(T_y < \infty\right) = 1$. □

4. Pólya's recurrence theorem

4.1. Statement of theorem

Definition. The simple random walk in $\mathbb{Z}^d$ is the Markov chain defined by

$$P(x, x+e_i) = P(x, x-e_i) = \frac{1}{2d}$$

where e_i is the standard basis.

Theorem. The simple random walk in $\mathbb{Z}^d$ is recurrent for $d = 1, d = 2$ and transient for $d \geq 3$.

4.2. One-dimensional proof

Consider $d = 1$. In this case, $P(x, x+1) = P(x, x-1) = \frac{1}{2}$. We will show that $\sum_n p_{00}(n) = \infty$, then recurrence will hold. We have $p_{00}(n) = \mathbb{P}_0(X_n = 0)$. Note that if n is odd, X_n is odd, so $\mathbb{P}_0(X_{2k+1} = 0) = 0$. So we will consider only even numbers. In order to be back at zero after $2n$ steps, we must make n steps 'up' away from the origin and make n steps 'down'. There are $\binom{2n}{n}$ ways of choosing which steps are 'up' steps. The probability of a specific choice of n 'up' and n 'down' is $\left(\frac{1}{2}\right)^{2n}$. Hence,

$$p_{00}(2n) = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{(n!)^2} \cdot \frac{1}{2^{2n}}$$

Recall Stirling's formula:

$$n! \sim n^n e^{-n}\sqrt{2\pi n}$$

Substituting in,

$$\frac{(2n)!}{(n!)^2} \cdot \frac{1}{2^{2n}} \sim \frac{1}{\sqrt{\pi n}} = \frac{A}{\sqrt{n}}$$

for $A > 0$; the precise value of A is unnecessary. Hence, for some large n_0, $\forall n \geq n_0$, $p_{00}(2n) \geq \frac{A}{2\sqrt{n}}$. So

$$\sum_n p_{00}(2n) \geq \sum_{n \geq n_0} \frac{A}{2\sqrt{n}} = \infty$$

Now, let us consider the asymmetric random walk for $d = 1$, defined by $P(x, x+1) = p$ and $P(x, x-1) = q$. We can compute $p_{00}(2n) = \binom{2n}{n}(pq)^n \sim A\frac{(4pq)^n}{\sqrt{n}}$. If $p \neq q$, then $4pq < 1$ so by the geometric series we have

$$\sum_{n \geq n_0} p_{00}(2n) \leq \sum_{n \geq n_0} 2A(4pq)^n < \infty$$

So the asymmetric random walk is transient.

4.3. Two-dimensional proof

Now, let us consider the simple random walk on $\mathbb{Z}^2$. For each point $(x, y) \in \mathbb{Z}^2$, we will project this coordinate onto the lines $y = x$ and $y = -x$. In particular, we define

$$f(x, y) = \left(\frac{x+y}{\sqrt{2}}, \frac{x-y}{\sqrt{2}} \right)$$

If X_n is the simple random walk on $\mathbb{Z}^2$, we consider $f(X_n) = (X_n^+, X_n^-)$.

Lemma. $(X_n^+), (X_n^-)$ are independent simple random walks on $\frac{1}{\sqrt{2}}\mathbb{Z}$.

Proof. We can write X_n as

$$X_n = \sum_{i=1}^{n} \xi_i$$

where ξ_i are independent and identically distributed by

$$\mathbb{P}(\xi_1 = (1,0)) = \mathbb{P}(\xi_1 = (-1,0)) = \mathbb{P}(\xi_1 = (0,1)) = \mathbb{P}(\xi_1 = (0,-1)) = \frac{1}{4}$$

and we write $\xi_i = (\xi_i^1, \xi_i^2)$. We can then see that

$$X_n^+ = \sum_{i=1}^{n} \frac{\xi_i^1 + \xi_i^2}{\sqrt{2}}; \quad X_n^- = \sum_{i=1}^{n} \frac{\xi_i^1 - \xi_i^2}{\sqrt{2}}$$

We can check that $(X_n^+), (X_n^-)$ are simple random walks on $\frac{1}{\sqrt{2}}\mathbb{Z}$. It now suffices to prove the independence property. Note that it suffices to show that $\xi_i^1 + \xi_1^2$ and $\xi_i^1 - \xi_i^2$ are independent, since the X_n^+, X_n^- are sums of independent and identically distributed copies of these random variables. We can simply enumerate all possible values of ξ_i^1, ξ_i^2 and the result follows. □

We know that $p_{00}(n) = 0$ if n is odd. We want to find $p_{00}(2n) = \mathbb{P}_0(X_{2n} = 0)$. Note, $X_n = 0 \iff X_n^+ = X_n^- = 0$. Using the lemma above,

$$\mathbb{P}_0(X_{2n} = 0) = \mathbb{P}_0(X_n^+ = 0, X_n^- = 0) = \mathbb{P}_0(X_n^+ = 0)\,\mathbb{P}_0(X_n^- = 0) \sim \frac{A}{\sqrt{n}} \frac{A}{\sqrt{n}} = \frac{A^2}{n}$$

Hence,

$$\sum_{n \geq n_0} \mathbb{P}_0(X_{2n} = 0) \geq \sum_{n \geq n_0} = \frac{A^2}{2n} = \infty$$

which gives recurrence as required.

4.4. Three-dimensional proof

Consider $d = 3$. Again, $p_{00}(n) = 0$ if n odd. In order to return to zero after $2n$ steps, we must make i steps both up and down, j steps north and south, and k steps east and west, with $i + j + k = n$. There are $\binom{2n}{i,i,j,j,k,k}$ ways of choosing which steps in each direction we take. Each combination has probability $\left(\frac{1}{6}\right)^{2n}$ of happening. Hence,

$$p_{00}(2n) = \sum_{i,j,k\geq 0, i+j+k=n} \binom{2n}{i,i,j,j,k,k} \left(\frac{1}{6}\right)^{2n} = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} \sum_{i,j,k\geq 0, i+j+k=n} \binom{n}{i,j,k}^2 \left(\frac{1}{3}\right)^{2n}$$

The sum on the right hand side is the total probability of the number of ways of placing n balls in three boxes uniformly at random, so equals one. Suppose $n = 3m$. Then we can show that $\binom{n}{i,j,k} \leq \binom{n}{m,m,m}$.

$$p_{00}(6m) \geq \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} \binom{n}{m,m,m}\left(\frac{1}{3}\right)^{n}$$

Applying Stirling's formula again, we have

$$p_{00}(6m) \sim \frac{A}{n^{3/2}}$$

It is sufficient to consider $n = 3m$:

$$p_{00}(6m) \geq \frac{1}{6^2} p_{00}(6m-2); \quad p_{00}(6m) \geq \frac{1}{6^4} p_{00}(6m-4)$$

Hence

$$\sum_n p_{00}(n) < \infty$$

So the Markov chain is transient.

5. Invariant distributions

5.1. Invariant distributions

Let I be a countable set. (λ_i) is a probability distribution if $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.

Example. Consider a Markov chain with two elements, and $P(1,1) = P(1,2) = P(2,1) = P(2,2) = \frac{1}{2}$. As $n \to \infty$, it is easy to see here that both states should be equally likely to occur. In fact, $p_{11}(n) = p_{12}(n) = p_{21}(n) = p_{22}(n) = \frac{1}{2}$. In this case, the row vector $\left(\frac{1}{2}, \frac{1}{2}\right)$ is an equilibrium probability distribution.

In general, we want to find a distribution π such that if $X_0 \sim \pi$, we have $X_n \sim \pi$ for all n. Suppose $X_0 \sim \pi$. Then,

$$\begin{aligned}\mathbb{P}(X_1 = j) &= \sum_{i \in I} \mathbb{P}(X_0 = i, X_1 = j) \\ &= \sum_{i \in I} \mathbb{P}(X_1 = j \mid X_0 = i)\,\mathbb{P}(X_0 = i) \\ &= \sum_{i \in I} \pi(i) P(i,j)\end{aligned}$$

Since we want $X_1 \sim \pi$, we must have $\pi(j) = \sum_{i \in I} \pi(i) P(i,j)$ for all j. In matrix form, $\pi = \pi P$.

Definition. An *invariant* (or *equilibrium*, or *stationary*) distribution for P is a probability distribution π such that $\pi = \pi P$.

Theorem. Let π be invariant. Then, if $X_0 \sim \pi$, for all n we have $X_n \sim \pi$.

Proof. If $X_0 \sim \pi$, then $X_n \sim \pi P^n = \pi$. □

Theorem. Suppose I is finite, and there exists $i \in I$ such that $p_{ij}(n) \to \pi_j$ as $n \to \infty$ for all j. Then $\pi = (\pi_j)$ is an invariant distribution.

Proof. First, we check that the sum of π_j is one. Since I is finite, we can interchange the sum and limit.

$$\sum_{j \in I} \pi_j = \sum_{j \in I} \lim_{n \to \infty} p_{ij}(n) = \lim_{n \to \infty} \sum_{j \in I} p_{ij}(n) = \lim_{n \to \infty} 1 = 1$$

So π_j is a probability distribution. We now must show $\pi = \pi P$.

$$\pi_j = \lim_{n \to \infty} p_{ij}(n) = \lim_{n \to \infty} \sum_{k \in I} p_{ik}(n-1) P(k,j) = \sum_{k \in I} \lim_{n \to \infty} p_{ik}(n-1) P(k,j) = \sum_{k \in I} \pi_k P(k,j)$$

as required. □

Remark. If I is infinite, the theorem does not necessarily hold. For example, let $I = \mathbb{Z}$, X be a simple symmetric random walk. We know that $p_{00}(n) \sim \frac{c}{\sqrt{n}}$, and $p_{0x}(n) \to 0$ as $n \to \infty$ for all $x \in \mathbb{Z}$. So zero is given by the limit but this is not a distribution.

5.2. Conditions for unique invariant distribution

In this section, we restrict our analysis to irreducible chains. If P is finite and irreducible, then 1 is an eigenvalue, since P is stochastic. The corresponding right eigenvector is $(1, \dots, 1)^\intercal$. We know that 1 is an eigenvalue of $P^\intercal$, so $P^\intercal$ has a right eigenvector corresponding to the eigenvalue of 1, which can be transposed to find a left eigenvector for P. It is possible to show using the Perron–Frobenius theorem that the eigenvector has non-negative components since P is irreducible. Since I is finite, we can normalise the left eigenvector such that its components sum to 1, giving an invariant distribution.

Definition. Let $k \in I$. Recall that T_k is the first return time to k. For every $i \in I$, we define

$$\nu_k(i) = \mathbb{E}_k\left[\sum_{n=0}^{T_k-1} 1(X_n = i)\right]$$

which is the expected number of times that we hit i while on an excursion from k (returning back to k).

Theorem. If P is irreducible and recurrent, then ν_k is an invariant measure: $\nu_k = \nu_k P$. Further, ν_k satisfies $\nu_k(k) = 1$ and in general $\nu_k(i) \in (0, \infty)$ for all i.

Proof. It is clear from the definition that $\nu_k(k) = 1$, since we must hit k exactly once on the outset, and we do not count the return. We will now prove that $\nu_k = \nu_k P$. $T_k < \infty$ with probability 1 by recurrence, and $X_{T_k} = k$. Then,

$$\begin{aligned}
\nu_k(i) &= \mathbb{E}_k\left[\sum_{n=0}^{T_k-1} 1(X_n = i)\right] \\
&= \mathbb{E}_k\left[\sum_{n=1}^{T_k} 1(X_n = i)\right] \\
&= \mathbb{E}_k\left[\sum_{n=1}^{\infty} 1(X_n = i, T_k \geq n)\right] \\
&= \sum_{n=1}^{\infty} \mathbb{E}_k\left[1(X_n = i, T_k \geq n)\right] \\
&= \sum_{n=1}^{\infty} \mathbb{P}_k\left(X_n = i, T_k \geq n\right) \\
&= \sum_{n=1}^{\infty} \sum_{j \in I} \mathbb{P}_k\left(X_n = i, X_{n-1} = j, T_k \geq n\right) \\
&= \sum_{n=1}^{\infty} \sum_{j \in I} \mathbb{P}_k\left(X_n = i \mid X_{n-1} = j, T_k \geq n\right) \mathbb{P}_k\left(X_{n-1} = j, T_k \geq n\right)
\end{aligned}$$

T_k is a stopping time, so the event $\{T_k \geq n\} = \{T_k \leq n-1\}^c$ depends only on values we already know or don't care about. Hence, we can remove it.

$$
\begin{aligned}
&= \sum_{n=1}^{\infty} \sum_{j \in I} \mathbb{P}_k (X_n = i \mid X_{n-1} = j) \, \mathbb{P}_k (X_{n-1} = j, T_k \geq n) \\
&= \sum_{n=1}^{\infty} \sum_{j \in I} P(j,i) \mathbb{P}_k (X_{n-1} = j, T_k \geq n) \\
&= \sum_{j \in I} \sum_{n=1}^{\infty} P(j,i) \mathbb{P}_k (X_{n-1} = j, T_k \geq n) \\
&= \sum_{j \in I} \sum_{n=0}^{\infty} P(j,i) \mathbb{P}_k (X_n = j, T_k \geq n+1) \\
&= \sum_{j \in I} P(j,i) \mathbb{E}_k \left[\sum_{n=0}^{T_k - 1} 1(X_n = j) \right] \\
&= \sum_{j \in I} P(j,i) \nu_k(j)
\end{aligned}
$$

Hence $\nu_k = \nu_k P$. We must show $\nu_k > 0$. P is irreducible, hence there exists n such that $p_{ki}(n) > 0$. Then

$$\nu_k(i) = \sum_{j \in I} \nu_k(j) P^n(j,i) \geq \nu_k(k) p_{ki}(n) > 0$$

To show $\nu_k < \infty$, let m such that $p_{ik}(m) > 0$.

$$1 = \nu_k(k) = \sum_{j \in I} \nu_k(j) P^m(j,k) \geq \nu_k(i) P^m(i,k) \implies \nu_k(i) \leq \frac{1}{P^m(i,k)} < \infty$$

□

5.3. Uniqueness of invariant distributions

Theorem. Let P be irreducible. Let λ be an invariant measure ($\lambda = \lambda P$) with $\lambda_k = 1$. Then $\lambda \geq \nu_k$. If P is recurrent, then $\lambda = \nu_k$.

Proof. Let λ be an invariant measure with $\lambda_k = 1$. Then,

$$\begin{aligned}
\lambda_i &= \sum_{j_1} \lambda_{j_1} P(j_1, i) \\
&= P(k,i) + \sum_{j_1 \neq k} \lambda_{j_1} P(j_1, i) \\
&= P(k,i) + \sum_{j_1 \neq k} P(k, j_1) P(j_1, i) + \sum_{j_1, j_2 \neq k} P(j_2, j_1) P(j_1, i) \lambda_{j_2} \\
&= P(k,i) + \sum_{j_1 \neq k} P(k, j_1) P(j_1, i) + \dots \\
&+ \sum_{j_1, \dots j_{n-1} \neq k} P(k, j_{n-1}) P(j_{n-1}, j_{n-2}) \dots P(j_2, j_1) P(j_1 i) + \underbrace{\sum_{j_1, \dots, j_n \neq k} P(j_n, j_{n-1}) \dots P(j_n, i) \lambda_{j_n}}_{\geq 0} \\
&\geq \mathbb{P}_k (X_1 = i, T_k \geq 1) + \mathbb{P}_k (X_2 = i, T_k \geq 2) + \dots + \mathbb{P}_k (X_n = i, T_k \geq n) \\
&\geq \sum_{i=1}^{n} \mathbb{P}_k (X_n = i, T_k \geq n) \\
&\to \nu_k(i)
\end{aligned}$$

as $n \to \infty$. Now, suppose P is recurrent, so ν_k is invariant. We define $\mu = \lambda - \nu_k$. Then $\mu \geq 0$ is an invariant measure satisfying $\mu_k = 0$. We need to show $\mu_i = 0$ for all i. By invariance, for all n,

$$\mu_k = \sum_j \mu_j P^n(j,k)$$

By irreducibility, we can choose n such that $P^n(i,k) > 0$.

$$\mu_k \geq P^n(i,k) \mu_i \implies \mu_i = 0$$

□

Remark. In the irreducible and recurrent case, all invariant measures are equal up to a scaling factor.

Let k be fixed. Then, ν_k is invariant, and unique in the above sense. If $\sum_i \nu_k(i)$ is finite, we can take

$$\pi_i = \frac{\nu_k(i)}{\sum_j \nu_k(j)}$$

which is an invariant distribution. The sum as required is

$$\begin{aligned}\sum_{i\in I}\nu_k(i) &= \sum_{i\in I}\mathbb{E}_k\left[\sum_{n=0}^{T_k-1}1(X_n=i)\right]\\ &= \mathbb{E}_k\left[\sum_{n=0}^{T_k-1}\sum_{i\in I}1(X_n=i)\right]\\ &= \mathbb{E}_k\left[\sum_{n=0}^{T_k-1}1\right]\\ &= \mathbb{E}_k\left[T_k\right]\end{aligned}$$

So we require that the expectation of the first return time is finite. If $\mathbb{E}_k[T_k]$ is finite, we can normalise ν_k into a (unique) invariant distribution.

5.4. Positive and null recurrence

Definition. Let $k \in I$ be a recurrent state (so $\mathbb{P}_k(T_k < \infty) = 1$). k is *positive recurrent* if $\mathbb{E}_k[T_k] < \infty$. k is called *null recurrent* otherwise; so if $\mathbb{E}_k[T_k] = \infty$.

Theorem. Let P be irreducible. Then the following are equivalent.

(i) every state is positive recurrent;

(ii) some state is positive recurrent;

(iii) P has an invariant distribution π.

If any of these conditions hold, we have

$$\pi_i = \frac{1}{\mathbb{E}_i[T_i]}$$

for all i.

Proof. First, (i) clearly implies (ii). We now show (ii) implies (iii). Let k be the a positive recurrent state, and consider ν_k. Since k is recurrent, we know that ν_k is an invariant measure. Then,

$$\sum_{i\in I}\nu_k(i) = \mathbb{E}_k[T_k] < \infty$$

since k is positive recurrent. If we define

$$\pi_i = \frac{\nu_k(i)}{\mathbb{E}_k[T_k]}$$

we have that π is an invariant distribution.

Now we show that (iii) implies (i). Let k be a state, which we will prove is positive recurrent. First, we show that $\pi_k > 0$. There exists i such that $\pi_i > 0$, and we will choose n such that $P^n(i,k) > 0$ by irreducibility. Then,

$$\pi_k = \sum_j \pi_j P^n(j,k) \geq \pi_i P^n(i,k) > 0$$

Now, we define $\lambda_i = \frac{\pi_i}{\pi_k}$. This is an invariant measure with $\lambda_k = 1$. So from the above theorem, $\lambda \geq \nu_k$. Now, since π is a distribution,

$$\mathbb{E}_k[T_k] = \sum_i \nu_k(i) \leq \sum_i \lambda_i = \sum_i \frac{\pi_i}{\pi_k} = \frac{1}{\pi_k} \sum_i \pi_i = \frac{1}{\pi_k}$$

Hence $\mathbb{E}_k[T_k] < \infty$, so k is positive recurrent.

For the last part, we know that P is recurrent and $\lambda_i = \frac{\pi_i}{\pi_k}$ is an invariant measure with $\lambda_k = 1$. From the previous theorem, $\lambda_i = \nu_k(i)$. Hence, $\frac{\pi_i}{\pi_k} = \nu_k(i)$. Taking the sum over all i,

$$\frac{1}{\pi_k} = \mathbb{E}_k[T_k]$$

which proves the last part. □

Corollary. If P is irreducible and π is an invariant distribution, then for all x, y, the expected number of visits to y starting from x is given by

$$\nu_x(y) = \frac{\pi(y)}{\pi(x)}$$

Example. Consider the simple symmetric random walk on $\mathbb{Z}$. We have proven that this is recurrent. Suppose π is an invariant measure. So $\pi = \pi P$, giving

$$\pi_i = \frac{1}{2}\pi_{i-1} + \frac{1}{2}\pi_{i+1}$$

So $\pi_i = 1$ is an invariant measure. So all invariant measures are multiples of this. But since this is not normalisable, there exists no invariant distribution. So this walk is null recurrent.

Remark. If I is finite, we can always normalise the distribution, since we have only a finite sum.

Remark. Consider a simple random walk on $\mathbb{Z}^3$. This is transient. However, $\lambda_i = 1$ for all $i \in \mathbb{Z}^3$, this is clearly an invariant measure, so existence of an invariant measure does not imply recurrence.

Example. Consider a random walk on $\mathbb{Z}$ with transition probabilities $P(i, i+1) = p, P(i, i-1) = q$ such that $1 > p > q > 0$ and $p + q = 1$. This random walk is transient. Suppose there is an invariant distribution π, so $\pi = \pi P$. Then

$$\pi_i = \pi_{i-1} q + \pi_{i+1} p$$

Solving the recursion gives

$$\pi_i = a + b\left(\frac{p}{q}\right)^i$$

This is not unique up to a multiplicative constant, due to the constant a.

Example. Consider a random walk on $\mathbb{Z}^+$ with transition probabilities $P(i, i+1) = p, P(i, i-1) = q, P(0,0) = q$, and $p < q$ so there is a drift towards zero. We can check that this is recurrent. We will look for a solution to $\pi = \pi P$.

$$\pi_0 = q\pi_0 + q\pi_1; \quad \pi_i = p\pi_{i-1} + q\pi_{i+1}$$

Solving this system yields

$$\pi_1 = \frac{p}{q}\pi_0; \quad \pi_i = \left(\frac{p}{q}\right)^i \pi_0$$

This is unique up to a multiplicative constant. Since $p < q$, we can normalise this to reach an invariant distribution. Let $\pi_0 = 1 - \frac{p}{q}$. Then,

$$\pi_i = \left(\frac{p}{q}\right)^i \left(1 - \frac{p}{q}\right)$$

Hence the walk is positive recurrent.

5.5. Time reversibility

Theorem. Let P be irreducible, and π be an invariant distribution. Let $N \in \mathbb{N}$ and let $Y_n = X_{N-n}$ for $0 \leq n \leq N$. If $X_0 \sim \pi$, then $(Y_n)_{0 \leq n \leq N}$ is a Markov chain with transition matrix

$$\hat{P}(x, y) = \frac{\pi(y)}{\pi(x)} P(y, x)$$

and has invariant distribution π, so $\pi\hat{P} = \pi$. Further, $\hat{P}$ is also irreducible.

Proof. First, note that $\hat{P}$ is stochastic. Since $\pi = \pi P$,

$$\sum_y \hat{P}(x, y) = \sum_y \frac{\pi(y)P(y, x)}{\pi(x)} = \frac{\pi(x)}{\pi(x)} = 1$$

Now we show Y is a Markov chain.

$$\begin{aligned}
\mathbb{P}(Y_0 = y_0, \dots, Y_N = y_N) &= \mathbb{P}(X_N = y_0, \dots, X_0 = y_n) \\
&= \pi(y_N)P(y_N, y_{N-1}) \dots P(y_1, y_0) \\
&= \hat{P}(y_{N-1}, y_N)\pi(y_{N-1})P(y_{N-1}, y_{N-2}) \dots P(y_1, y_0) \\
&= \dots \\
&= \pi(y_0)\hat{P}(y_0, y_1) \dots P(y_{N-1}, y_N)
\end{aligned}$$

Hence $Y \sim \text{Markov}(\pi, \hat{P})$. Now, we must show $\pi = \pi\hat{P}$.

$$\sum_x \pi(x)\hat{P}(x,y) = \sum_x \pi(x)\frac{P(y,x)\pi(y)}{\pi(x)} = \pi(y)\sum_x P(y,x) = \pi(y)$$

Hence π is invariant for $\hat{P}$. Now we show $\hat{P}$ is irreducible. Let $x, y \in I$. Then there exists $x = x_0, x_1, \dots, x_k = y$ such that

$$P(x_0, x_1) \dots P(x_{k-1}, x_k) > 0$$

Hence

$$\hat{P}(x_k, x_{k-1}) \dots \hat{P}(x_1, x_0) = \pi(x_0)P(x_0, x_1) \dots \frac{P(x_{k-1}, x_k)}{\pi(x_k)} > 0$$

So $\hat{P}$ is irreducible. □

Definition. A Markov chain X with transition matrix P and invariant distribution π is called *reversible* or time reversible if $\hat{P} = P$. Equivalently, for all x, y,

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

These equations are called the *detailed balance equations*. Equivalently, X is reversible if, for any fixed $N \in \mathbb{N}$, $X_0 \sim \pi$ implies

$$(X_0, \dots, X_N) \overset{d}{=} (X_N, \dots, X_0)$$

which means that they are equal in distribution.

Remark. Intuitively, X is reversible if, starting from π, we cannot tell if we are watching X evolve forwards in time or backwards in time.

Lemma. Let P be a transition matrix, and μ a distribution satisfying the detailed balance equations.

$$\mu(x)P(x,y) = \mu(y)P(y,x)$$

Then μ is invariant for P.

Proof.

$$\sum_x \mu(x)P(x,y) = \sum_x \mu(y)P(y,x) = \mu(y)$$

□

Remark. If we can find a solution to the detailed balance equations which is a distribution, it must be an invariant distribution. It is simpler to solve this set of equations than to solve $\pi = \pi P$. If there is no solution to the detailed balance equations, then even if there exists an invariant distribution, the Markov chain is not reversible.

Example. Consider a random walk on the integers modulo n, with $P(i, i+1) = \frac{2}{3}$ and $P(i, i-1) = \frac{1}{3}$. We can check $\pi_i = \frac{1}{n}$ is an invariant distribution. This does not satisfy the detailed balance equations. Hence the Markov chain is not reversible.

Example. Consider a random walk on $\{0, \dots, n-1\}$ with $P(i, i+1) = \frac{2}{3}, P(i, i-1) = \frac{1}{3}$ and $P(0,0) = \frac{1}{3}, P(n-1, n-1) = \frac{2}{3}$. This is an 'opened up' version of the previous example; the circle is 'cut' open into a line at zero. The detailed balance equations give

$$\pi_i P(i, i+1) = \pi_{i+1} P(i+1, i) \implies \pi_i = k2^i$$

We can normalise this by setting k such that π is a distribution. Hence the chain is reversible.

Example. Consider a random walk on a graph. Let $G = (V, E)$ be a finite connected graph, where V is a set of vertices and E is a set of edges. The simple random walk on G has the transition matrix

$$P(x, y) = \begin{cases} \frac{1}{d(x)} & (x, y) \in E \\ 0 & (x, y) \notin E \end{cases}$$

where $d(x) = \sum_y 1((x, y) \in E)$ is the degree of x. The detailed balance equations give, for $(x, y) \in E$,

$$\pi(x)P(x, y) = \pi(y)P(y, x) \implies \frac{\pi(x)}{d(x)} = \frac{\pi(y)}{d(y)}$$

Let $\pi(x) \propto d(x)$. Then this is an invariant distribution with normalising constant $\frac{1}{\sum_y d(y)} = \frac{1}{2|E|}$. So the simple random walk on a finite connected graph is always reversible.

5.6. Aperiodicity

Definition. Let P be a transition matrix. For all i, we write

$$d_i = \gcd\{n \geq 1 : P^n(i, i) > 0\}$$

This is called the *period* of i. If $d_i = 1$, we say that i is aperiodic.

Lemma. $d_i = 1$ if and only if $P^n(i, i) > 0$ for all n sufficiently large. More rigorously, there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$, $P^n(i, i) > 0$.

Proof. First, if $P^n(i, i) > 0$ for all n sufficiently large, the greatest common divisor of all sufficiently large numbers is one so this direction is trivial. Conversely, let

$$D(i) = \{n \geq 1 : P^n(i, i) > 0\}$$

Observe that if $a, b \in D(i)$ then $a + b \in D(i)$.

We claim that $D(i)$ contains two consecutive integers. Suppose that it does not, so for all $a, b \in D(i)$ we must have $|a - b| > 1$. Let r be the minimal distance between two integers

in $D(i)$, so $r \geq 2$. Let n, m be numbers in $D(i)$ separated by r, so $n = m + r$. Then we can show there exists $k \in D(i)$ which can be written as $\ell r + s$ with $0 < s < r$. Indeed, if there were not such a k, we would have $d_i = 1$, since all elements would be multiples of r. Now, let $a = (\ell + 1)n$ and $b = (\ell + 1)m + k$. Then $a, b \in D(i)$, and $a - b = r - s < r$. This is a contradiction, since we have found two points in $D(i)$ with a distance smaller than the minimal distance.

Now, let $n_1, n_1 + 1$ be elements of $D(i)$. Then

$$\{xn_1 + y(n_1 + 1) : x, y \in \mathbb{N}\} \subseteq D(i)$$

It is then easy to check that $D(i) \supseteq \{n : n \geq n_1^2\}$. □

Lemma. Suppose P is irreducible and i is aperiodic. Then for all $j \in I$, j is aperiodic. Hence, aperiodicity is a class property.

Proof. There exist n, m such that $P^n(i, j) > 0, P^m(i, j) > 0$. Hence,

$$P^{n+m+r}(j, j) \geq P^n(j, i)P^r(i, i)P^n(i, j)$$

The first and last terms are positive, and the middle term is positive for sufficiently large r. □

5.7. Positive recurrent limiting behaviour

Theorem. Let P be irreducible and aperiodic with invariant distribution π, and further let $X \sim \text{Markov}(\lambda, P)$. Then for all $y \in I$, $\mathbb{P}(X_n = y) \to \pi_y$ as $n \to \infty$. Taking $\lambda = \delta_x$, we get $p_{xy}(n) \to \pi(y)$ as $n \to \infty$.

Proof. This proof will use the idea of 'coupling' of Markov chains. Let $Y \sim \text{Markov}(\pi, P)$ be independent of X. Consider the pair $((X_n, Y_n))_{n \geq 0}$. This is a Markov chain on the state space $I \times I$, because X and Y are independent. The initial distribution is $\lambda \times \pi$. We have $\mathbb{P}((X_0, Y_0) = (x, y)) = \lambda(x)\pi(y)$ and transition matrix $\widetilde{P}$ given by

$$\widetilde{P}((x, y), (x', y')) = P(x, x')P(y, y')$$

This product chain has invariant distribution $\widetilde{\pi}$ given by

$$\widetilde{\pi}(x, y) = \pi(x)\pi(y)$$

Let $a \in I$, and let $T = \inf n \geq 1 : (X_n, Y_n) = (a, a)$ be the hitting time of (a, a).

First, we want to show that $\mathbb{P}(T < \infty) = 1$. We show that $\widetilde{P}$ is irreducible. Let $(x, y), (x', y') \in I \times I$. By irreducibility of P, there exist ℓ, m such that $P^\ell(x, x') > 0$ and $P^m(y, y') > 0$. Now,

$$\widetilde{P}^{\ell+m+n}((x, y), (x', y')) = P^{\ell+m+n}(x, x')P^{\ell+m+n}(y, y')$$

Note that

$$P^{\ell+m+n}(x,x') \geq P^{\ell}(x,x')P^{m+n}(x',x')$$

By taking n large, by aperiodicity the product is positive. Therefore, for sufficiently large n, $P^n(x,x') > 0$. So $\widetilde{P}$ is irreducible, and there exists an invariant distribution $\widetilde{\pi}$. Hence $\widetilde{P}$ is positive recurrent. So $\mathbb{P}(T < \infty) = 1$.

Now, we define

$$Z_n = \begin{cases} X_n & n < T \\ Y_n & n \geq T \end{cases}$$

We wish to show $Z = (Z_n)n \geq 0$ has the same distribution as X, that is, $Z \sim \text{Markov}(\lambda, P)$. Now,

$$\mathbb{P}(Z_0 = x) = \mathbb{P}(X_0 = x) = \lambda(x)$$

so the initial distribution is the same. Now, we will check that Z evolves with transition matrix P. Let $A = \{Z_{n-1} = z_{n-1}, \dots, Z_0 = z_0\}$. We need to show $\mathbb{P}(Z_{n+1} = y \mid Z_n = x, A) = P(x,y)$.

$$\begin{aligned}\mathbb{P}(Z_{n+1} = y \mid Z_n = x, A) &= \mathbb{P}(Z_{n+1} = y, T > n \mid Z_n = x, A) \\ &\quad + \mathbb{P}(Z_{n+1} = y, T \leq n \mid Z_n = x, A) \\ &= \mathbb{P}(X_{n+1} = y \mid T > n, Z_n = x, A)\,\mathbb{P}(T > n \mid Z_n = x, A) \\ &\quad + \mathbb{P}(Y_n + 1 = y \mid T \leq n, Z_n = x, A)\,\mathbb{P}(T \leq n \mid Z_n = x, A)\end{aligned}$$

Now,

$$\begin{aligned}&\mathbb{P}(X_{n+1} = y \mid T > n, Z_n = x, A) \\ &= \sum_z \mathbb{P}(X_{n+1} = y \mid T > n, Z_n = x, Y_n = z, A)\,\mathbb{P}(Y_n = z \mid T > n, Z - n = x, A)\end{aligned}$$

Note, $\{T > n\}$ depends only on $(X_0, Y_0), \dots, (X_n, Y_n)$ since it is the complement of $\{T \leq n\}$, so it is a stopping time. Hence,

$$\mathbb{P}(X_{n+1} = y \mid T > n, Z_n = x, A) = \sum_z P(x,y)\mathbb{P}(Y_n = z \mid T > n, Z - n = x, A) = P(x,y)$$

Similarly,

$$\mathbb{P}(Y_{n+1} = y \mid T > n, Z_n = x, A) = P(x,y)$$

Hence,

$$\begin{aligned}\mathbb{P}(Z_{n+1} = y \mid Z_n = x, A) &= P(x,y)\mathbb{P}(T > n \mid Z_n = x, A) + P(x,y)\mathbb{P}(T \leq n \mid Z_n = x, A) \\ &= P(x,y)[\mathbb{P}(T > n \mid Z_n = x, A) + \mathbb{P}(T \leq n \mid Z_n = x, A)] \\ &= P(x,y)\end{aligned}$$

as required. Hence $Z \sim \text{Markov}(\lambda, P)$. Thus,

$$\begin{aligned}|\mathbb{P}(X_n = y) - \pi(y)| &= |\mathbb{P}(Z_n = y) - \mathbb{P}(Y_n = y)| \\ &= |\mathbb{P}(X_n = y, n < T) + \mathbb{P}(Y_n = y, n \geq T) \\ &- Y_n = y, n < T - \mathbb{P}(Y_n = y, n \geq T)| \\ &= |\mathbb{P}(X_n = y, n < T) - \mathbb{P}(Y_n = y, n < T)| \\ &\leq \mathbb{P}(n < T)\end{aligned}$$

As $n \to \infty$, this upper bound becomes zero, since $\mathbb{P}(T < \infty) = 1$. □

5.8. Null recurrent limiting behaviour

Theorem. Let P be irreducible, aperiodic, and null recurrent. Then, for all x, y,

$$\lim_{n\to\infty} P^n(x, y) = 0$$

Proof. Let $\widetilde{P}((x, y), (x', y')) = P(x, x')P(y, y')$ as before. We have shown previously that $\widetilde{P}$ is also irreducible. Suppose first that $\widetilde{P}$ is transient. Then,

$$\sum_n \widetilde{P}^n((x, y), (x, y)) < \infty$$

This sum is equal to

$$\sum_n (P^n(x, y))^2 < \infty$$

Hence,

$$P^n(x, y) \to 0$$

Now, conversely suppose that $\widetilde{P}$ is recurrent. Let $y \in I$. Define as before

$$\nu_y(x) = \mathbb{E}_y\left[\sum_{i=0}^{T_y - 1} 1(X_i = x)\right]$$

This measure is invariant for P since P is recurrent. Since P is null recurrent in particular, $\mathbb{E}_y\left[T_y\right] = \infty$. Hence,

$$\nu_y(I) = \sum_{x\in I} \nu_y(x) = \mathbb{E}_y\left[\sum_{i=0}^{T_y - 1} 1\right] = \mathbb{E}_y\left[T_y\right] = \infty$$

Because $\nu_y(I)$ is infinite, for all $M > 0$ there exists a finite set $A \subset I$ with $\nu_y(A) > M$. Now, we define a probability measure

$$\mu(z) = \frac{\nu_y(z)}{\nu_y(A)} 1(z \in A)$$

Now, for all $z \in I$,

$$\mu P^n(z) = \sum_x \mu(x) P^n(x,z) = \sum_x \frac{\nu_y(x)}{\nu_y(A)} 1(z \in A) P^n(x,z) \leq \frac{1}{\nu_y(A)} \sum_x \nu_y(x) P^n(x,z)$$

Since ν_y is invariant,

$$\mu P^n(z) \leq \frac{1}{\nu_y(A)} \nu_y(z) = \frac{\nu_y(z)}{\nu_y(A)}$$

Let (X, Y) be a Markov chain with matrix $\widetilde{P}$, started according to $\mu \times \delta_x$, so

$$\mathbb{P}(X_0 = z, Y_0 = w) = \mu(z)\delta_x(w)$$

Now, let

$$T = \inf\{n \geq 1 : (X_n, Y_n) = (x, x)\}$$

Since $\widetilde{P}$ is recurrent, T is finite with probability 1. Let

$$Z_n = \begin{cases} X_n & n < T \\ Y_n & n \geq T \end{cases}$$

We have already proven that Z is a Markov chain with transition matrix P, started according to μ; it has the same distribution as X. Hence,

$$\mathbb{P}(Z_n = y) = \mu P^n(y) \leq \frac{\nu_y(y)}{\nu_y(A)} = \frac{1}{\nu_y(A)}$$

Note,

$$\mathbb{P}_x(Y_n = y) \leq \mathbb{P}_x(Y_n = y, n \geq T) + \mathbb{P}_x(T > n) = \mathbb{P}_x(Z_n = y) + \mathbb{P}_x(T > n)$$

Hence,

$$\limsup_{n \to \infty} \mathbb{P}_x(Y_n = y) \leq \frac{1}{M} + 0 = \frac{1}{M}$$

Since this is true for all M, $P^n(x,y) \to 0$ as $n \to \infty$. □

IV. Analysis and Topology

Lectured in Michaelmas 2021 by Dr. V. Zsák

In the analysis part of the course, we continue the study of convergence from Analysis I. We define a stronger version of convergence, called uniform convergence, and show that it has some very desirable properties. For example, if integrable functions f_n converge uniformly to the integrable function f, then the integrals of the f_n converge to the integral of f. The same cannot be said in general about non-uniform convergence. We also extend our study of differentiation to functions with multiple input and output variables, and rigorously define the derivative in this higher-dimensional context.

In the topology part of the course, we consider familiar spaces such as $[a, b], \mathbb{C}, \mathbb{R}^n$, and generalise their properties. We arrive at the definition of a metric space, which encapsulates all of the information about how near or far points are from others. From here, we can define notions such as continuous functions between metric spaces in such a way that does not depend on the underlying space.

We then generalise even further to define topological spaces. The only information a topological space contains is the neighbourhoods of each point, but it turns out that this is still enough to define continuous functions and similar things. We study topological spaces in an abstract setting, and prove important facts that are used in many later courses.

Contents

1. Uniform convergence 141
1.1. Definition 141
1.2. Pointwise convergence 141
1.3. Uniform limit of bounded functions 143
1.4. Integrability 143
1.5. Differentiability 145
1.6. Conditions for uniform convergence 146
1.7. General principle of uniform convergence 146
1.8. Weierstrass M-test 146
1.9. Power series 147
2. Uniform continuity 149
2.1. Definition 149
2.2. Properties of continuous functions 149
3. Metric spaces 151
3.1. Definition 151
3.2. Subspaces 153
3.3. Product spaces 153
3.4. Convergence 153
3.5. Continuity 155
3.6. Isometric, Lipschitz, and uniformly continuous functions 157
3.7. Generalised triangle inequality 158
4. Topology of metric spaces 159
4.1. Open balls 159
4.2. Neighbourhoods and openness 159
4.3. Continuity and convergence using topology 160
4.4. Properties of topology of metric space 162
4.5. Homeomorphisms 163
4.6. Equivalence of metrics 163
5. Completeness 164
5.1. Cauchy sequences 164
5.2. Definition of completeness 164
5.3. Completeness of product spaces 165
5.4. Completeness of subspaces and function spaces 165
6. Contraction mapping theorem 169
6.1. Contraction mappings 169
6.2. Contraction mapping theorem 169
6.3. Application of contraction mapping theorem 170

6.4. Lindelöf–Picard theorem . . . 170
7. Topology . . . 173
7.1. Definitions . . . 173
7.2. Closed subsets . . . 174
7.3. Neighbourhoods . . . 174
7.4. Convergence . . . 174
7.5. Interiors and closures . . . 175
7.6. Dense subsets . . . 176
7.7. Subspaces . . . 176
7.8. Continuity . . . 177
7.9. Homeomorphisms and topological invariance . . . 178
7.10. Products . . . 178
7.11. Continuity in product topology . . . 179
7.12. Quotients . . . 180
7.13. Continuity of functions in quotient spaces . . . 181
8. Connectedness . . . 183
8.1. Definition . . . 183
8.2. Consequences of definition . . . 184
8.3. Partitioning into connected components . . . 186
8.4. Path-connectedness . . . 187
8.5. Gluing lemma . . . 187
9. Compactness . . . 190
9.1. Motivation and definition . . . 190
9.2. Subspaces . . . 191
9.3. Continuous images of compact spaces . . . 192
9.4. Topological inverse function theorem . . . 192
9.5. Tychonov's theorem . . . 193
9.6. Heine–Borel theorem . . . 193
9.7. Sequential compactness . . . 194
9.8. Compactness and sequential compactness in metric spaces . . . 194
10. Differentiation . . . 197
10.1. Linear maps . . . 197
10.2. Differentiation . . . 198
10.3. Derivatives on open subsets . . . 200
10.4. Properties of derivative . . . 201
10.5. Linearity and product rule . . . 203
11. Partial derivatives . . . 204
11.1. Directional and partial derivatives . . . 204
11.2. Jacobian matrix . . . 205
11.3. Constructing total derivative from partial derivatives . . . 205

11.4. Mean value inequality 206
11.5. Zero derivatives 207
11.6. Inverse function theorem 207
12. Second derivatives 210
12.1. Definition 210
12.2. Second derivatives and partial derivatives 212
12.3. Symmetry of mixed directional derivatives 212

1. Uniform convergence

1.1. Definition

Recall that $x_n \to x$ as $n \to \infty$ (for $x \in \mathbb{R}$ or $\mathbb{C}$) if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |x_n - x| < \varepsilon$$

This is essentially considering the ε-neighbourhood of x. We aim to define the same notion of convergence for functions, by defining an analogous concept of an ε-neighbourhood. In particular, each value on the domain should converge in its own ε-neighbourhood.

Definition. Let S be a set, and $f, f_n : S \to \mathbb{R}$, be functions. We say that (f_n) converges to f uniformly on S if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall x \in S, |f_n(x) - f(x)| < \varepsilon$$

Note. N depends only on ε, *not* on any x. Each x converges therefore at a 'similar speed', hence the name 'uniform convergence'.

Equivalently, we can write

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \sup_{x \in S} |f_n(x) - f(x)| < \varepsilon$$

The supremum condition is equivalent overall because the inequality on the right is weakened to a possible equality, but we can always decrease ε to retain the inequality. Alternatively, we could write

$$\lim_{n \to \infty} \sup_{x \in S} |f_n - f| = 0$$

For each $x \in S$, $(f_n(x))_{n=1}^{\infty} \to f(x)$. Hence, f is unique given (f_n), since limits are unique. We call f the *uniform limit* of (f_n) on S.

1.2. Pointwise convergence

Definition. (f_n) converges *pointwise* to f on S if $(f_n(x))_{n=1}^{\infty}$ converges to $f(x)$ for every $x \in S$. In other words,

$$\underbrace{\forall x \in S,}_{\text{order rearranged}} \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon$$

Now, N depends both on ε and on x. Note that the pointwise limit of (f_n) on S is also unique since limits are unique.

Remark. Uniform convergence implies pointwise convergence, and the uniform limit is the pointwise limit.

Example. Let $f_n(x) = x^2 e^{-nx}$ on $[0, \infty), n \in \mathbb{N}$. Does (f_n) converge uniformly on the domain? First let us check pointwise convergence. We have $x^2 e^{-nx} \to 0$ hence pointwise convergence to $f(x) = 0$ is satisfied. Now, we need only check uniform convergence to the function $f(x) = 0$.

$$\sup_{x \in [0,\infty)} |f_n(x) - 0| = \sup_{x \in [0,\infty)} f_n(x)$$

We could differentiate f_n and find the maximum if it exists, but we might not find the maximum if it is (for example) on the endpoints. A much better method is to find an upper bound on $|f_n(x) - f(x)|$ (which, in this example, is $f_n(x)$) that does not depend on x. In this case, we can expand e^{nx} on the denominator and isolate a single term to get

$$x^2 e^{-nx} = \frac{x^2}{e^{nx}} \leq \frac{2}{n^2}; \quad \forall x$$

Hence,

$$\sup_{x \in [0,\infty)} |f_n(x) - 0| \to 0$$

and uniform convergence is satisfied.

Example. Consider $f_n(x) = x^n$ on $[0, 1], n \in \mathbb{N}$. A pointwise limit is reached by

$$f(x) = \begin{cases} 1 & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Consider $\sup |f_n(x) - f(x)|$ excluding 1 (since at 1 the supremum is zero). Note $f_n(x) \to 1$ as $x \to 1$ from below, for all n. Hence the supremum is always 1 by choosing an x sufficiently close to 1. So $f_n \nrightarrow f$ uniformly on $[0, 1]$, hence (f_n) does not converge at all uniformly on this domain. Or,

$$\sup f_n(x) \geq f_n\left(\left(\frac{1}{2}\right)^{1/n}\right) = \frac{1}{2}$$

Remark. If $f_n \nrightarrow f$ uniformly on S,

$$\exists \varepsilon > 0, \forall N \in \mathbb{N}, \exists n \geq N, \exists x \in S, |f_n(x) - f(x)| \geq \varepsilon$$

In the above example, we proved something stronger:

$$\forall n, \exists x \in S, f_n(x) \geq \frac{1}{2}$$

We could have alternatively stated, for example, $f_n(x)$ is continuous so there exists some subset of $[0, 1]$ greater than $\frac{1}{2}$ always.

Theorem. Let $S \subseteq \mathbb{R}, \mathbb{C}$. Let $(f_n), f : S \to \mathbb{R}(\text{or } \mathbb{C})$, where f_n is continuous and $(f_n) \to f$ uniformly on S. Then f is continuous.

Informally, the uniform limit of continuous functions is continuous.

Proof. Fix some point $a \in S$, $\varepsilon > 0$. We seek $\delta > 0$ such that $\forall x \in S, |x-a| < \delta \implies |f(x)-f(a)| < \varepsilon$. We fix an $n \in \mathbb{N}$ such that $\forall x \in S, |f_n(x)-f(x)| < \varepsilon$. Since f_n is continuous, there exists $\delta > 0$ such that $\forall x \in S, |x-a| < \delta \implies |f_n(x)-f_n(a)| < \varepsilon$. So, $\forall x \in S$,

$$|x-a| < \delta \implies |f(x)-f(a)| \leq |f(x)-f_n(x)| + |f_n(x)-f_n(a)| + |f_n(a)-f(a)| < 3\varepsilon$$

□

Remark. The above proof is often called a 3ε-proof. Note, the proof is not true for pointwise convergence; if $f_n \to f$ pointwise and f_n continuous, f is not necessarily continuous. Further, it is not true for differentiability; f_n differentiable does not imply f differentiable (see example sheet). Another way to interpret the result of the above theorem is to swap limits:

$$\lim_{x\to a}\lim_{n\to\infty} f_n(x) = \lim_{x\to a} f(x) = f(a) = \lim_{n\to\infty} f_n(a) = \lim_{n\to\infty}\lim_{x\to a} f_n(x)$$

1.3. Uniform limit of bounded functions

Lemma. Let $f_n \to f$ uniformly on S. If f_n is bounded for every n, then so is f.

In other words, the uniform limit of bounded functions is bounded.

Proof. Fix some $n \in \mathbb{N}$ such that $\forall x \in S, |f_n(x)-f(x)| < 1$. Since f_n is bounded, $\exists M \in \mathbb{R}$ such that $\forall x \in S, |f_n(x)| < M$. Hence, $\forall x \in S, |f(x)| \leq |f(x)-f_n(x)| + |f_n(x)| \leq 1 + M$. So f is bounded. □

1.4. Integrability

Let $f : [a,b] \to \mathbb{R}$ be a bounded function. Recall that for a dissection $\mathcal{D}$ of $[a,b]$, we define the upper and lower sums of f with respect to $\mathcal{D}$ by

$$U_{\mathcal{D}}(f) = \sum_{k=1}^{n} (x_k - x_{k-1}) \sup_{[x_{k-1},x_k]} f(x)$$

$$L_{\mathcal{D}}(f) = \sum_{k=1}^{n} (x_k - x_{k-1}) \inf_{[x_{k-1},x_k]} f(x)$$

Riemann's integrability criterion states that f is integrable if and only if

$$\forall \varepsilon, \exists \mathcal{D}, U_{\mathcal{D}}(f) - L_{\mathcal{D}}(f) < \varepsilon$$

Equivalently, for any $I \subset [a,b]$, we have

$$\sup_I f - \inf_I f = \sup_{x,y\in I} (f(x)-f(y)) = \sup_{x,y\in I} |f(x)-f(y)|$$

This is called the oscillation of f on I. So an integrable function 'doesn't oscillate too much'.

Theorem. Let $f_n : [a,b] \to \mathbb{R}$ be integrable for all n. If $f_n \to f$ uniformly on $[a,b]$, then f is integrable and

$$\int_a^b f_n \to \int_a^b f$$

Proof. First, we prove f to be bounded, then we will check Riemann's criterion. We know f is bounded because each f_n is bounded, hence by the lemma above f is bounded. Now fix $\varepsilon > 0$, and choose $n \in \mathbb{N}$ such that $\forall x \in [a,b], |f_n(x) - f(x)| < \varepsilon$. Since f_n is integrable, $\exists \mathcal{D} : a = x_0 < x_1 < \cdots < x_N = b$ of $[a,b]$ such that $U_{\mathcal{D}} - L_{\mathcal{D}} < \varepsilon$. Now, we fix $k \in \{1, \dots, N\}$ and then for any $x, y \in [x_{k-1}, x_k]$ we have

$$|f(x) - f(y)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| < 2\varepsilon + |f_n(x) - f_n(y)|$$

Taking the supremum,

$$\sup_{x,y \in [x_{k-1}, x_k]} (f(x) - f(y)) \leq \sup_{x,y \in [x_{k-1}, x_k]} |f_n(x) - f_n(y)| + 2\varepsilon$$

Multiplying by $(x_k - x_{k-1})$ and taking the sum over all k,

$$U(f) - L(f) \leq U(f_n) - L(f_n) + 2\varepsilon(b-a) \leq \varepsilon(2(b-a) + 1)$$

Hence f is integrable. We can now show that

$$\left| \int_a^b f_n - \int_a^b f \right| \leq \int_a^b |f_n - f| \leq (b-a) \sup_{[a,b]} |f_n - f| \to 0$$

□

Remark. We can interpret this as

$$\int_a^b \lim_{n \to \infty} f_n(x)\,\mathrm{d}x = \lim_{n \to \infty} \int_a^b f_n(x)\,\mathrm{d}x$$

This is another 'allowed' way to swap limits.

Corollary. Let $f_n : [a,b] \to \mathbb{R}$ be integrable for all n. If $\sum_{n=1}^{\infty} f_n(x)$ converges uniformly on $[a,b]$, then

$$F(x) = \sum_{n=1}^{\infty} f_n(x)$$

is integrable, and

$$\int_a^b \sum_{n=1}^{\infty} f_n(x)\,\mathrm{d}x = \sum_{n=1}^{\infty} \int_a^b f_n(x)\,\mathrm{d}x$$

Proof. Let $F_n(x) = \sum_{k=1}^{n} f_k(x)$. By assumption, $F_n \to F$ uniformly on $[a, b]$. F_n is integrable where the integral of F_n is the sum of the integrals:

$$\int_a^b F_n = \sum_{k=1}^{n} \int_a^b f_k$$

Then the result follows from the theorem above. □

1.5. Differentiability

Theorem. Let $f_n : [a, b] \to \mathbb{R}$ be continuously differentiable for all n. Suppose $\sum_{k=1}^{\infty} f_k'(x)$ converges uniformly on $[a, b]$, and that $\forall c \in [a, b]$, $\sum_{n-1}^{\infty} f_n(c)$ converges. Then, $\sum_{k=1}^{\infty} f_k(x)$ converges uniformly on $[a, b]$ to a continuously differentiable function f, and

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{k=1}^{\infty} f_k\right) = \sum_{k=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x} f_k(x)$$

Proof. Let $g(x) = \sum_{k=1}^{\infty} f_k'(x)$, for $x \in [a, b]$. The general idea is that we want to solve the differential equation $f' = g$ subject to the initial condition $f(c) = \sum_{n=1}^{\infty} f_n(c)$. Let $\lambda = \sum_{n=1}^{\infty} f_n(c)$ and define $f : [a, b] \to \mathbb{R}$ by

$$f(x) = \lambda + \int_c^x g(t)\,\mathrm{d}t$$

Note that g is integrable; $\sum_{k=1}^{\infty} f_k'(x) \to g$ uniformly implies that g is continuous and hence integrable. By the fundamental theorem of calculus, $f' = g$ and $f(c) = \lambda$. So we have found such an f that satisfies the conditions set out. All that remains is to prove uniform convergence of $\sum_{k=1}^{\infty} f_k \to f$. Also by the fundamental theorem, $f_k(x) = f_k(c) + \int_c^x f_k'(t)\,\mathrm{d}t$. Let $\varepsilon > 0$. There exists $N \in \mathbb{N}$ such that $\left|\lambda - \sum_{k=1}^{N} f_k(c)\right| < \varepsilon$ and $\left|g(t) - \sum_{k=1}^{N} f_k'(t)\right| < \varepsilon$. Now, for $n \geq N$ we have

$$\begin{aligned}
\left|f(x) - \sum_{k=1}^{n} f_k(x)\right| &= \left|\lambda + \int_c^x g(t)\,\mathrm{d}t - \sum_{k=1}^{n}\left(f_k(c) + \int_c^x f_k'(t)\,\mathrm{d}t\right)\right| \\
&\leq \left|\lambda - \sum_{k=1}^{n} f_k(c)\right| + \left|\int_c^x \left(g(t) - \sum_{k-1}^{n} f_k'(t)\right)\mathrm{d}t\right| \\
&\leq \varepsilon + |x - c|\varepsilon \\
&\leq \varepsilon(b - a + 1)
\end{aligned}$$

□

1.6. Conditions for uniform convergence

Recall that a scalar sequence x_n is Cauchy if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, |x_m - x_n| < \varepsilon$$

and that the general principle of convergence shows that any Cauchy sequence converges.

1.7. General principle of uniform convergence

Definition. A sequence (f_n) of scalar functions on a set S is called *uniformly Cauchy* if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, \forall x \in S, |f_m(x) - f_n(x)| < \varepsilon$$

Theorem. A uniformly Cauchy sequence of functions is uniformly convergent.

Proof. Let $x \in S$ and we will show that $(f_n(x))_{n=1}^{\infty}$ converges. Given $\varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, \forall t \in S, |f_m(t) - f_n(t)| < \varepsilon$. In particular, $\forall m, n \geq N, |f_m(x) - f_n(x)| < \varepsilon$. So certainly $(f_n(x))_{n=1}^{\infty}$ is Cauchy and hence convergent by the general principle of convergence. Therefore f_n converges pointwise. Now, let $f(x)$ be the limit $f(x) = \lim_{n\to\infty} f_n(x)$. Then $f_n \to f$ pointwise on S. Now we must extend this to show $f_n \to f$ uniformly on S. Given $\varepsilon > 0$, we know that $\exists N \in \mathbb{N}, \forall m, n \geq N, \forall x \in S, |f_m(x) - f_n(x)| < \varepsilon$. Now, we must show $\forall n \geq N, \forall x \in S, |f_n(x) - f(x)| < 2\varepsilon$, then we are done. We will fix $x \in S, n \geq N$. Since $f_n(x) \to f(x)$, we can choose $m \in \mathbb{N}$ such that $|f_m(x) - f(x)| < \varepsilon$, and $m \geq N$. Note however that m depends on x in this statement, but this doesn't matter—we have shown that

$$|f_n(x) - f(x)| \leq |f_n(x) - f_m(x)| + |f_m(x) - f(x)| \leq \varepsilon + \varepsilon = 2\varepsilon$$

which is a result that, in itself, does *not* depend on x. □

Note. Alternatively, we could end the proof as the following. Fix $x \in S, n \geq N$. Then

$$\forall m \geq N, |f_n(x) - f_m(x)| < \varepsilon$$

Then let $m \to \infty$, and

$$|f_n(x) - f(x)| \leq \varepsilon$$

1.8. Weierstrass M-test

Theorem. Let (f_n) be a sequence of scalar functions on S. Assume that $\forall n \in \mathbb{N}, \exists M_n \in \mathbb{R}^+, \forall x \in S, |f_n(x)| \leq M_n$. In other words, (f_n) is a sequence of bounded scalar functions. Then,

$$\sum_{n=1}^{\infty} M_n < \infty \implies \sum_{n=1}^{\infty} f_n(x) \text{ is uniformly convergent on } S$$

Proof. Let $F_n(x) = \sum_{k=1}^{n} f_k(x)$ for $x \in S, n \in \mathbb{N}$. Then

$$|F_n(x) - F_m(x)| \leq \sum_{k=m+1}^{n} |f_k(x)| \leq \sum_{k=m+1}^{n} M_k$$

Hence, given $\varepsilon > 0$, we can choose $N \in \mathbb{N}$ such that $\sum_{k=N+1}^{n} M_k < \varepsilon$. Thus, $\forall x \in S, \forall n \geq m \geq N$, we have

$$|F_n(x) - F_m(x)| \leq \sum_{k=m+1}^{n} M_k < \varepsilon$$

We have shown (F_n) is uniformly Cauchy on S and hence uniformly convergent on S. □

1.9. Power series

Consider the power series

$$\sum_{n=0}^{\infty} c_n (z-a)^n$$

where $c_n \in \mathbb{C}, a \in \mathbb{C}$ are constants, and $z \in \mathbb{C}$. Let $R \in [0, \infty]$ be the radius of convergence. Recall that

$$|z-a| < R \implies \sum_{n=0}^{\infty} c_n (z-a)^n \text{ converges absolutely;}$$

$$|z-a| > R \implies \sum_{n=0}^{\infty} c_n (z-a)^n \text{ diverges}$$

Let $D(a, R) \coloneqq \{z \in \mathbb{C} \mid |z-a| < R\}$ be the open disc centred on a with radius R. Then we can create $f : D(a, \mathbb{R}) \to \mathbb{C}$ to be defined by the power series, which is well-defined. f is the pointwise limit of the power series on D. In general, the convergence of the power series is not uniformly convergent.

Example. $\sum_{n=1}^{\infty} \frac{z^n}{n^2}$ has $R = 1$. Let $f_n : D(0,1) \to \mathcal{C}$ be defined by $f_n(z) = \frac{z^n}{n^2}$. Then for every $z \in D(0,1), |z| \leq \frac{1}{n^2}$. Since $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, by the Weierstrass M-test, the power series converges uniformly on the disc.

Example. Consider $\sum_{n=0}^{\infty} z^n = \frac{1}{1-z}$ with $R = 1$. Now,

$$\forall z \in D(0,1), \left| \sum_{n=0}^{\infty} z^n \right| \leq N + 1$$

Therefore, the series does not converge uniformly on the disc since $\frac{1}{1-z}$ is unbounded on the disc. Alternatively, consider

$$\sup_{|z|<1} \left| \frac{1}{1-z} - \sum_{k=0}^{n} z_k \right| = \sup_{|z|<1} \left| \frac{z^{n+1}}{1-z} \right| = \infty$$

In some sense, the problem with uniform convergence here is that we are allowed to go too close too the boundary.

Theorem. Suppose the power series $\sum_{n=0}^{\infty} c_n(z-a)^n$ has radius of convergence R. Then for all $0 < r < R$, the power series converges uniformly on $D(a, r)$.

Proof. Let $w \in \mathbb{C}$ such that $r < |w-a| < R$, for instance $w = a + \frac{r+R}{2}$. Now, let $\rho = \frac{r}{|w-a|} \in (0,1)$. Since $\sum_{n=0}^{\infty} c_n(w-a)^n$ converges, we have that $c_n(w-a)^n \to 0$ as $n \to \infty$. Therefore, $\exists M \in \mathbb{R}+$ such that $|c_n(w-a)^n| \leq M$ for all $n \in \mathbb{N}$, since convergence implies boundedness. Now, for $z \in D(a,r), n \in \mathbb{N}$ we have

$$|c_n(z-a)^n| = |c_n(w-a)^n|\left(\frac{|z-a|}{|w-a|}\right)^n \leq M\left(\frac{r}{|w-a|}\right)^n = M\rho^n$$

Since the sum $\sum_{n=0}^{\infty} M\rho^n$ converges, the Weierstrass M-test shows us that $\sum_{n=0}^{\infty} c_n(z-a)^n$ converges uniformly on $D(a,r)$. □

Remark. $f : D(a,R) \to \mathbb{C}$ defined by $f(z) = \sum_{n=0}^{\infty} c_n(z-a)^n$ is the uniform limit on $D(a,r)$ of polynomials for any r such that $0 < r < R$. Hence f is continuous on $D(a,r)$. Since $D(a,R) = \bigcup_{0<r<R} D(a,r)$, it follows that f is continuous everywhere inside the radius of convergence.

Recall that the termwise derivative $\sum_{n=1}^{\infty} c_n n(z-a)^{n-1}$ has the same radius of convergence. This sequence therefore also converges uniformly on $D(a,r)$ if $0 < r < R$. Analogously to the previous result about interchanging derivatives and sums, we can show that $\sum c_n(z-a)^n$ is complex differentiable on $D(a,R)$ with derivative $\sum_{n=1}^{\infty} c_n n(z-a)^{n-1}$. This is seen in the IB Complex Analysis course.

Now, fix $w \in D(a,R)$. Then fix r such that $|w-a| < r < R$, and fix $\delta > 0$ such that $|w-a| + \delta < r$. If $|z-w| < \delta$, then $|z-a| \leq |z-w| + |w-a| < \delta + |w-a| < r$. Therefore, geometrically, $D(w,\delta) \subset D(a,r)$. Hence, $\sum_{n=0}^{\infty} c_n(z-a)^n$ converges uniformly on $D(w,\delta)$. This is known as local uniform convergence.

Definition. $U \subset \mathbb{C}$ is called open if $\forall w \in U, \exists \delta > 0, D(w,\delta) \subset U$.

Definition. Let U be an open subset of $\mathbb{C}$, and f_n be a sequence of scalar functions on U. Then f_n converges locally uniformly on U if

$$\forall w \in U, \exists \delta > 0, f_n \text{ converges uniformly on } D(w,\delta) \subset U$$

Remark. Above, we showed that power series always converge locally uniformly inside the radius of convergence, or equivalently inside the disc $D(a,R)$. We will return to this point about local uniform convergence when discussing compactness.

2. Uniform continuity

2.1. Definition

Let $U \subset \mathbb{R}, \mathbb{C}$. Let f be a scalar function on U. Then for $x \in U$, we say f is continuous at x if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall y \in U, |y - x| < \delta \implies |f(y) - f(x)| < \varepsilon$$

We say f is continuous on U if f is continuous at x for all $x \in U$:

$$\forall x \in U, \forall \varepsilon > 0, \exists \delta > 0, \forall y \in U, |y - x| < \delta \implies |f(y) - f(x)| < \varepsilon$$

Note here that δ depends on ε and x.

Definition. Let U, f be as in the previous definition. We say f is *uniformly continuous* if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x, y \in U, |y - x| < \delta \implies |f(y) - f(x)| < \varepsilon$$

Now, δ works for all $x \in U$ simultaneously; δ depends on ε only. Certainly, uniform continuity implies continuity.

Example. Let $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = 2x + 17$. Then f is uniformly continuous; given $\varepsilon > 0$, we can find $\delta = \frac{1}{2}\varepsilon$. Then $\forall x, y \in \mathbb{R}, |y - x| < \delta \implies |f(y) - f(x)| = |2y - 2x| = 2y - x < 2\delta = \varepsilon$.

Example. Let $f : \mathbb{R} \to \mathbb{R}$, defined by $f(x) = x^2$. This is not uniformly continuous, since no δ works for all x given some 'bad' ε. Let us take $\varepsilon = 1$, and we wish to show that no δ exists. Suppose some δ does exist. Then, let $x > 0$ and $y = x + \frac{\delta}{2}$. We should have $|f(y) - f(x)| < 1$.

$$\left(x + \frac{\delta}{2}\right)^2 - x^2 = \delta x + \frac{\delta^2}{4}$$

So for $x = \frac{1}{\delta}$, this condition $|f(y) - f(x)| < 1$ is not satisfied. Hence f is not uniformly continuous.

Note. For U, f as in the above definition, f is not uniformly continuous on U if

$$\exists \varepsilon > 0, \forall \delta > 0, \exists x, y \in U, |y - x| < \delta, |f(y) - f(x)| \geq \varepsilon$$

So there are points arbitrarily close together whose difference of function values exceed some fixed ε.

2.2. Properties of continuous functions

Theorem. Let f be a scalar function on a closed bounded interval $[a, b]$. If f is continuous on $[a, b]$, then f is uniformly continuous on $[a, b]$.

Proof. Suppose there exists $\varepsilon > 0$ such that $\forall \delta > 0, \exists x, y \in [a, b], |y - x| < \delta, |f(y) - f(x)| \geq \varepsilon$. In particular, we can construct a sequence (δ_n) defined by $\delta_n = \frac{1}{n}$, and we can construct sequences $x_n, y_n \in [a, b]$ such that $|y_n - x_n| < \frac{1}{n}$ but $|f(y_n) - f(x_n)| \geq \varepsilon$. By the Bolzano–Weierstrass theorem, there exists a subsequence (x_{k_n}) that converges. Now, let x be the limit of the subsequence, $\lim_{n\to\infty} x_{k_n}$. Then $x \in [a, b]$ since the interval is closed. Then, $|y_{k_n} - x| \leq |y_{k_n} - x_{k_n}| + |x_{k_n} - x| < \frac{1}{n} + |x_{k_n} - x| \to 0$. Hence $y_{k_n} \to x$. Now, since f is continuous $f(x_{k_n}), f(y_{k_n}) \to f(x)$. Now, $\varepsilon \leq |f(x_{k_n}) - f(y_{k_n})| \to |f(x) - f(x)| = 0$, which is a contradiction. □

Corollary. A continuous function $f : [a, b] \to \mathbb{R}$ is Riemann integrable.

Proof. Since a continuous function on a closed bounded interval is bounded, we have that f is bounded. Now, fix $\varepsilon > 0$, and we want to find a dissection $\mathcal{D}$ such that the difference between upper and lower sums is less than ε. By the above theorem, f is uniformly continuous. Hence,

$$\exists \delta > 0, \forall x, y \in [a, b], |y - x| < \delta \implies |f(y) - f(x)| < \varepsilon$$

So we must simply choose a dissection such that all intervals have size smaller than δ. For instance, choose some $n \in \mathbb{N}$ such that $\frac{b-a}{N} < \delta$, and then divide the interval equally into n subintervals. If I is an interval in this dissection, then $\forall x, y \in I$ we have $|y - x| < \delta$ and hence $|f(y) - f(x)| < \varepsilon$. Hence,

$$\sup_{x,y\in I} |f(y) - f(x)| \leq \varepsilon$$

Multiplying by the length of I and summing over all subintervals I,

$$U_{\mathcal{D}}(f) - L_{\mathcal{D}}(f) \leq (b - a)\varepsilon$$

Hence f is Riemann integrable. □

3. Metric spaces

3.1. Definition

Definition. Let M be a set. Then a *metric* on M is a function $d : M \times M \to \mathbb{R}$ such that

(i) (positivity) $\forall x, y \in M, d(x, y) \geq 0$, and in particular, $x = y \iff d(x, y) = 0$

(ii) (symmetric) $\forall x, y \in M, d(x, y) = d(y, x)$

(iii) (triangle inequality) $\forall x, y, z \in M, d(x, z) \leq d(x, y) + d(y, z)$.

A metric space is a set M together with a metric d on M, written as the pair (M, d).

Example. Let $M = \mathbb{R}, \mathbb{C}$ and $d(x, y) = |x - y|$. This is known as the 'standard metric' on M. If a metric is not specified, the standard metric is taken as implied.

Example. Let $M = \mathbb{R}^n, \mathbb{C}^n$, and we define the Euclidean norm (or Euclidean length) to be

$$\|x\| = \|x\|_2 = \left(\sum_{k=1}^{n} |x_k|^2 \right)^{\frac{1}{2}}$$

This satisfies

$$\|x + y\| \leq \|x\| + \|y\|$$

and it then follows that we can define the metric as

$$d_2(x, y) = \|x - y\|_2$$

called the Euclidean metric. We can check that this is indeed a metric easily. This is the standard metric on $\mathbb{R}^n, \mathbb{C}^n$. The metric space (M, d) in this case is called n-dimensional real (or complex) Euclidean space, sometimes denoted ℓ_2^n. The Euclidean norm is sometimes called the ℓ_2 norm, and the Euclidean metric is the ℓ_2 metric.

Example. Let $M = \mathbb{R}^n, \mathbb{C}^n$, and we define the ℓ_1 norm to be

$$|x|_1 = \sum_{k=1}^{n} |x_k|$$

which defines the ℓ_1 metric given by

$$d_1(x, y) = \|x - y\|_1$$

(M, d_1) is denoted ℓ_1^n. We can generalise and form the metric space ℓ_p^n for all $p \in [1, \infty]$.

Example. Again, let $M = \mathbb{R}^n, \mathbb{C}^n$. We can define the ℓ_∞ norm by

$$\|x\|_\infty = \max_{1 \leq k \leq n} |x_k|$$

This defines the ℓ_∞ metric:

$$d_\infty(x, y) = \|x - y\|_\infty = \max_{1 \leq k \leq n} |x_k - y_k|$$

We denote (M, d) by ℓ_∞^n.

In this course, we will only work with $p = 1, 2, \infty$, although the calculations can be made to work for other p.

Example. Let S be a set. Let $\ell_\infty(S)$ be the set of all bounded scalar functions on S. We then define the ℓ_∞ norm of $f \in \ell_\infty(S)$ by

$$\|f\| = \|f\|_\infty = \sup_{x \in S} |f(x)|$$

The supremum exists since the function is always bounded. This is also known as the 'sup norm' or the 'uniform norm'. Note that, for $f, g \in \ell_\infty(S)$, and $x \in S$,

$$\|f + g\| \leq \sup_{x \in S} |f(x) + g(x)| \leq |f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\| + \|g\|$$

Hence $d(f, g) = \|f - g\|$ defines a metric on $\ell_\infty(S)$. This is the standard metric on this space $\ell_\infty(S)$, also called the 'uniform metric'. For example, $\ell_\infty(\{1, \dots, n\}) = \mathbb{R}^n$ with the metric ℓ_∞. Also, for $\ell_\infty(\mathbb{N})$, we typically omit the $\mathbb{N}$ and instead write ℓ_∞ for the space of scalar sequences with the uniform metric.

Example. Consider $C[a, b]$, the set of all continuous functions on $[a, b]$. For $p = 1, 2$, we define the L_p norm of $f \in C[a, b]$ by

$$\|f\|_p = \left(\int_a^b |f(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}}$$

which induces the L_p metric on $C[a, b]$.

Example. Let M be a set. Then

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}$$

is a metric, called the discrete metric on M. In particular, (M, d) is called a discrete metric space.

Example. Let G be a group generated by $S \subset G$. We assume $e \notin S$ and $x \in S \implies x^{-1} \in S$. Then

$$d(x, y) = \min \{n \geq 0 : \exists s_1, \dots, s_n, y = x s_1 \dots s_n\}$$

defines a metric called the word metric.

Example. Let p be prime. Then

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ p^{-n} & \text{otherwise, where } x - y = p^n m, n \geq 0, m \in Z, p \nmid m \end{cases}$$

defines a metric on $\mathbb{Z}$. This is known as the p-adic metric.

3.2. Subspaces

Let (M, d) be a metric space, and $N \subset M$. Then naturally we can restrict d to $N \times N$, giving a metric on N. (N, d) is called a subspace of M.

Example. Consider $\mathbb{Q}$ with the metric $d(x, y) = |x - y|$. This is clearly a subspace of $\mathbb{R}$ (implicitly with the standard metric on $\mathbb{R}$).

Example. Since every continuous function on a closed bounded interval is bounded, $C[a, b]$ is a subset of $\ell_\infty[a, b]$. Hence $C[a, b]$ with the uniform metric is a subspace of $\ell_\infty[a, b]$.

3.3. Product spaces

Let $(M, d), (M', d')$ be metric spaces. Then any of the following defines a metric on the Cartesian product $M \times M'$.

(i) $d_1((x, x'), (y, y')) = d(x, y) + d(x', y')$

(ii) $d_2((x, x'), (y, y')) = \left(d(x, y)^2 + d(x', y')^2\right)^{\frac{1}{2}}$

(iii) $d_\infty((x, x'), (y, y')) = \max\{d(x, y), d(x', y')\}$

We commonly write $(M \times M', p)$ as $M \oplus_p M'$. Note that we always have

$$d_\infty \leq d_2 \leq d_1 \leq 2d_\infty$$

We can generalise for $n \in \mathbb{N}$ and metric spaces (M_k, d_k) for $k \in \{1, \dots, n\}$, by defining

$$\left(\bigoplus_{k=1}^{n} M_k\right)_p = M_1 \oplus_p \cdots \oplus_p M_n = \left(M_1 \times \cdots \times M_n, d_p\right)$$

Example. $\mathbb{R} \oplus_1 \mathbb{R} = \ell_1^2$. Further, $\mathbb{R} \oplus_2 \mathbb{R} \oplus_2 R = \ell_2^3$, and other analogous results hold.

Remark. $\mathbb{R} \oplus_1 \mathbb{R} \oplus_2 \mathbb{R}$ does not make sense since we have not defined the associativity of the $\oplus$ operator. The two choices yield different metric spaces.

3.4. Convergence

Let M be a metric space, and (x_n) a sequence in M. Given $x \in M$, we say that (x_n) converges to x in M if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, d(x_n, x) < \varepsilon$$

We say that (x_n) is convergent in M if $\exists x \in M$ such that $x_n \to x$. Otherwise, we say that (x_n) is divergent. Note that $x_n \to x$ in M if and only if $d(x_n, x) \to 0$ in $\mathbb{R}$.

Lemma. Suppose we have a sequence $x_n \to x$ and $x_n \to y$ in a metric space M. Then $x = y$.

Proof. Suppose $x \neq y$. Then let $\varepsilon = \frac{d(x,y)}{3} > 0$. So, by the definition of convergence,

$$\exists N_1 \in \mathbb{N}, \forall n \geq N_1, d(x_n, x) < \varepsilon;$$

$$\exists N_2 \in \mathbb{N}, \forall n \geq N_2, d(x_n, y) < \varepsilon$$

Now, fix $N \in \mathbb{N}$ such that $n \geq N_1, n \geq N_2$, for instance $N = \max\{N_1, N_2\}$. Then

$$d(x, y) \leq d(x, x_n) + d(x_n, y) < 2\varepsilon = \frac{2}{3} d(x, y)$$

which is a contradiction. □

Definition. Given a convergent sequence (x_n) in a metric space M, we say the *limit* of (x_n) is the unique $x \in M$ such that $x_n \to x$ as $n \to \infty$. This is denoted

$$\lim_{n \to \infty} x_n$$

Example. This definition has the usual meaning when $M = \mathbb{R}, \mathbb{C}$.

Example. The constant sequence defined by $x_n = x$ converges to x. In particular, 'eventually constant' sequences converge; let (x_n) be a sequence in M such that $\exists x \in M, \exists N \in \mathbb{N}, \forall n \geq N, x_n = x$, then $x_n \to x$. It is not necessarily true that sequences only converge if they are eventually constant. However, in a discrete metric space, the converse is true, since we can choose ε smaller than all distances.

Example. Consider the 3-adic metric. Then, $3^n \to 0$ as $n \to \infty$ since $d(3^n, 0) = 3^{-n} \to 0$.

Example. Let S be a set. Then, $f_n \to f$ in $\ell_\infty(S)$ in the uniform metric if and only if $d(f_n, f) = \|f_n - f\|_\infty = \sup_S |f_n - f| \to 0$, which is precisely the condition that $f_n \to f$ uniformly on S. Note, however, that $f_n(x) = x + \frac{1}{n}$ for $x \in \mathbb{R}, n \in \mathbb{N}$ and $f(x) = x$, then certainly $f_n \to x$ uniformly on $\mathbb{R}$. However, $f_n, f \notin \ell_\infty(\mathbb{R})$, so the uniform metric is not defined on these functions. So the notion of uniform convergence visited before is slightly more general than the idea of convergence in this metric space.

Example. Consider Euclidean space $M = \mathbb{R}^n, \mathbb{C}^n$ with the ℓ_2 metric. Then, consider

$$x^{(k)} = \left(x_1^{(k)}, \ldots, x_n^{(k)}\right) \in M$$

for $k \in \mathbb{N}$, and $x = (x_1, \ldots, x_n) \in M$. Then,

$$\left|x_i^{(k)} - x_i\right| \leq \left\|x^{(k)} - x\right\|_2 \leq \sum_{i=1}^{n} \left|x_i^{(k)} - x_i\right|$$

So $x^{(k)} \to x$ if and only if all i satisfy $x_i^{(k)} \to x_i$. This can be thought of as convergence being equivalent to coordinate-wise (or pointwise) convergence.

Example. Consider $f_n(x) = x^n$ for $x \in [0,1]$, and $n \in \mathbb{N}$. Then (f_n) is a sequence in $C[0,1]$, which converges pointwise but not uniformly. So (f_n) is not convergent in the uniform metric. However, using the L_1 metric, we have

$$d_1(f_n, 0) = \|f_n\|_1 = \int_0^1 f_n = \frac{1}{n+1} \to 0$$

So, $f_n \to 0$ in $(C[0,1], L_1)$.

Example. Let N be a subspace of a metric space M, and (x_n) be a convergent sequence in N. Then (x_n) converges in M. The converse is not necessarily true; consider $M = \mathbb{R}$ and $N = (0, \infty)$ with $(x_n) = \frac{1}{n}$. This is divergent in N but convergent in M.

Example. Let $(M, d), (M', d')$ be metric spaces. Let $N = M \oplus_p M'$. Let $a_n = (x_n, y_n) \in N$ for all $n \in \mathbb{N}$, and $a = (x, y) \in N$. Then

$$a_n \to a \text{ in } N \iff x_n \to x \text{ in } M, y_n \to y \text{ in } M'$$

Indeed,

$$\max\{d(x_n, x), d'(y_n, y)\} = d_\infty(a_n, a) \leq d_p(a_n, a) \leq 2d_1(a_n, a) = 2d(x_n, x) + 2d'(y_n, y)$$

3.5. Continuity

Definition. Let $f : M \to M'$ be a function between metric spaces $(M, d), (M', d')$. Then for $a \in M$, we say f is continuous at a if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x \in M, d(x, a) < \delta \implies d'(f(x), f(a)) < \varepsilon$$

We say f is continuous if f is continuous at a for all $a \in M$. In other words,

$$\forall a \in M, \forall \varepsilon > 0, \exists \delta > 0, \forall x \in M, d(x, a) < \delta \implies d'(f(x), f(a)) < \varepsilon$$

Note that δ depends both on ε and a.

Proposition. Let $f : M \to M'$ be as above. Let $a \in M$. Then the following are equivalent:

(i) f is continuous at a;

(ii) $x_n \to a$ in M implies $f(x_n) \to f(a)$ in M

Proof. First we show (i) implies (ii). Suppose $x_n \to a$ in M. Then fix $\varepsilon > 0$, and seek $N \in \mathbb{N}$ such that $\forall n \geq N, d'(f(x_n), f(a)) < \varepsilon$. By continuity, there exists $\delta > 0$ such that $\forall x \in M, d(x, a) < \delta \implies d'(f(x_n), f(a)) < \varepsilon$ as required. So we want N such that $\forall n \geq N, d(x, a) < \delta$, which must exist since $x_n \to a$.

Now, we show (ii) implies (i). Suppose that f is not continuous at a. Then,

$$\exists \varepsilon > 0, \forall \delta > 0, \exists x \in M, d(x, a) < \delta, d'(f(x), f(a)) \geq \varepsilon$$

So fix such an ε for which no suitable δ exists. Choose the sequence $\delta_n = \frac{1}{n}$, so

$$d(x_n, a) < \frac{1}{n}; \quad d'(f(x_n), f(a)) \geq \varepsilon$$

Then $x_n \to a$ in M but $f(x_n) \nrightarrow f(a)$ in M, which is a contradiction. □

Proposition. Let f, g be scalar functions on a metric space M. Let $a \in M$. Then if f, g are continuous at a, so are $f + g$ and $f \cdot g$. Moreover, letting $N = \{x \in M : g(x) \neq 0\}$ and assuming $a \in N$, $\frac{f}{g}$ is continuous at a. Hence if f, g are continuous, then so are $f+g, f \cdot g, \frac{f}{g}$ where they are defined.

Proof. Suppose $x_n \to a$. Then by the above proposition, $(f \cdot g)(x_n) = f(x_n) \cdot g(x_n) \to f(a) \cdot g(a) = (f \cdot g)(a)$, and similar results hold for the other operators. □

Remark. If $f : M \to M'$ is continuous everywhere,

$$\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right)$$

by the second proposition.

Proposition. Let $f : M \to M', g : M' \to M''$ be functions between metric spaces. If f is continuous at a and g is continuous at $f(a)$, then $g \circ f$ is continuous at a. If f, g are continuous, $g \circ f$ is continuous.

Proof. Let $\varepsilon > 0$. We want to find $\delta > 0$ such that $\forall x \in M$,

$$d(x, a) < \delta \implies d''(g(f(x)), g(f(a))) < \varepsilon$$

Since g is continuous at $f(a)$, there exists $\eta > 0$ such that $\forall y \in M'$,

$$d'(y, f(a)) < \eta \implies d''(g(y), g(f(a))) < \varepsilon$$

Now, since f is continuous at a, for this η there exists δ such that for all $x \in M$,

$$d(x, a) < \delta \implies d'(f(x) - f(a)) < \eta$$

Then

$$d(x, a) < \delta \implies d''(g(f(x)), g(f(a))) < \varepsilon$$

as required. □

Example. Constant functions are continuous. For instance, let $b \in M$ and let $f(x) = b$. Then this is continuous since $d'(f(x) - f(a)) = d'(b, b) = 0$ so any $\delta > 0$ will satisfy the condition.

Example. The identity function $f : M \to M$ defined by $x \mapsto x$ is continuous. Consider $d(f(x) - f(a)) = d(x - a)$. So $\delta = \varepsilon$ will suffice.

Example. All real and complex polynomials and rational functions are continuous wherever they are defined by the propositions and examples above. In fact, using uniform convergence, the uniform limits of such functions are also continuous. For example, exponential and trigonometric functions are continuous.

Example. Let (M, d) be a metric space. Then $d : M \oplus_p M \to \mathbb{R}$, which can be viewed as a function between metric spaces $M \oplus_p M$ and $\mathbb{R}$. Then, given $v = (x, x'), w = (y, y') \in M \oplus_p M$,

$$|d(v) - d(w)| = |d(x, x') - d(y, y')| \leq d(x, y) + d(x', y') = d_1(v, w) \leq 2d_p(v, w)$$

Hence $\delta = \frac{\varepsilon}{2}$ will suffice.

3.6. Isometric, Lipschitz, and uniformly continuous functions

Definition. Let $f : M \to M'$ be a function between metric spaces. Then, f is

(i) *isometric*, if
$$\forall x, y \in M, d'(f(x), f(y)) = d(x, y)$$

(ii) *Lipschitz*, or c-Lipschitz, if
$$\exists c \in \mathbb{R}^+, \forall x, y \in M, d'(f(x), f(y)) \leq c \cdot d(x, y)$$

(iii) *uniformly continuous*, if
$$\forall \varepsilon > 0, \exists \delta > 0, \forall x, y \in M, d(x, y) < \delta \implies d'(f(x), f(y)) < \varepsilon$$

Remark. Any isometric function is 1-Lipschitz. Any Lipschitz function is uniformly continuous. Any uniformly continuous function is continuous.

Remark. If a function is isometric, it is injective, since $f(x) = f(y) \implies x = y$. For example, if $N \subset M$, the inclusion map $i : N \to M$ defined by $i(x) = x$ is isometric but not surjective. An isometric and surjective map is called an *isometry*. If there exists an isometry $M \to M'$, we say that M and M' are isometric metric spaces, or M' is an isometric copy of M.

Example. Suppose $(M, d), (M', d')$ be metric spaces. Let $y \in M'$. We define $f : M \to M \oplus_p M'$ by $x \mapsto (x, y)$. Then $d_p(f(x), f(z)) = d_p((x, y), (z, y)) = d(x, z)$. So the function f is isometric. Therefore, $M \times \{y\}$ is an isometric copy of M in $M \oplus_p M'$.

Example. Consider the projections $q : M \oplus_p M' \to M$ defined by $q(x, y) = x$ and $q' : M \oplus_p M' \to M'$ defined by $q'(x, y) = y$. These projections are both 1-Lipschitz. Indeed,

$$d(q(x, y), q(x', y')) = d(x, x') \leq d_p((x, y), (x', y'))$$

In particular, polynomials in any finite number of variables are continuous since we can multiply continuous functions together.

3.7. Generalised triangle inequality

Suppose $u, x, y, z \in M$. Then, $|d(u,x) - d(y,z)| \leq d(u,y) + d(x,z)$. First,

$$d(u,x) \leq d(u,y) + d(y,x) \leq d(u,y) + d(y,z) + d(z,x)$$

Rearranging,

$$d(u,x) - d(y,z) \leq d(u,y) + d(x,z)$$

To achieve the negative, satisfying both conditions in the absolute value term,

$$d(y,z) \leq d(y,u) + d(u,x) + d(x,z)$$

which gives

$$d(y,z) - d(u,x) \leq d(u,y) + d(x,z)$$

as required.

4. Topology of metric spaces

4.1. Open balls

Definition. Let M be a metric space, $x \in M$, $r > 0$. Then the *open ball* in M of centre x and radius r is the set

$$\mathcal{D}_r(x) = \{y \in M : \ d(y,x) < r\}$$

The open ball notation is a convenient syntax for denoting closeness in some metric space. Note that, for example, $x_n \to n$ in M is equivalent to saying

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, x_n \in \mathcal{D}_\varepsilon(x)$$

We can also say that $f : M \to M'$ is continuous at x if

$$\forall \varepsilon > 0, \exists \delta > 0, f(\mathcal{D}_\delta(x)) \subset \mathcal{D}_\varepsilon(f(x))$$

Definition. The closed ball of centre x and radius $r \geq 0$ is the set

$$\mathcal{B}_r(x) = \{y \in M : \ d(y,x) \leq r\}$$

Example. In $\mathbb{R}$, $\mathcal{D}_r(x) = (x-r, x+r)$. Further, $\mathcal{B}_r(x) = [x-r, x+r]$. In the plane $(\mathbb{R}^2, d_p)$,

$$\mathcal{B}_1(0) = \left\{x \in \mathbb{R}^2 : \ \|x\|_p \leq 1\right\}$$

Note. $\mathcal{D}_r(x) \subset \mathcal{B}_r(x) \subset \mathcal{D}_s(x)$ for all $r < s$.

Example. Let M be a discrete metric space. Then for $x \in M$,

$$\mathcal{D}_1(x) = \{x\}; \quad \mathcal{B}_1(x) = M$$

4.2. Neighbourhoods and openness

Definition. Let M be a metric space, and $U \subset M$. Then for $x \in M$, we say that U is a *neighbourhood* of x (in M) if

$$\exists r > 0, \mathcal{D}_r(x) \subset U \iff \exists r > 0, \mathcal{B}_r(x) \subset U$$

Definition. We say $U \subset M$ is *open* in M, or that U is an *open subset* of M, if

$$\forall x \in U, \exists r > 0, \mathcal{D}_r(x) \subset U$$

So U is a neighbourhood of all points in U.

Example. $\mathcal{D}_r(x), \mathcal{B}_r(x)$ are neighbourhoods of x.

Example. Let $H = \{z \in \mathbb{C} : \operatorname{Im} z \geq 0\}$. Let $w \in H$ and $\delta = \operatorname{Im} w$. If $\delta > 0$, then $\mathcal{D}_\delta(w) \subset H$. If $\delta = 0$, then for any r, $\mathcal{D}_\delta(w) \not\subset H$. So H is not open.

Lemma. Open balls are open.

Proof. Let $\mathcal{D}_r(x)$ be an open ball in a metric space M. We need to show that

$$\forall y \in \mathcal{D}_r(x), \exists \delta > 0, \mathcal{D}_\delta(y) \subset \mathcal{D}_r(x)$$

So let $y \in \mathcal{D}_r(x)$ and set $\delta = r - d(x, y)$. Note that $d(x, y) > 0$, and by the triangle inequality,

$$d(z, x) \leq d(z, y) + d(y, x) < \delta + (r - \delta) = r$$

as required. □

Corollary. Let M be a metric space, $U \subset M$, $x \in M$. Then U is a neighbourhood of x if and only if there exists an open subset V of M such that $x \in V \subset U$.

Proof. In the forward direction, there exists $r > 0$ such that $\mathcal{D}_r(x) \subset U$, so let $V = \mathcal{D}_r(x)$. Conversely, if V is open we can construct $r > 0$ such that $\mathcal{D}_r(x) \subset V \subset U$. So U is a neighbourhood of x. □

4.3. Continuity and convergence using topology

Proposition. In a metric space M, the following are equivalent.

(i) $x_n \to x$;

(ii) for all neighbourhoods U of x in M, $\exists N \in \mathbb{N}, \forall n \geq N, x_n \in U$;

(iii) for all open neighbourhoods U of x in M, $\exists N \in \mathbb{N}, \forall n \geq N, x_n \in U$.

Proof. First, (i) implies (ii). Let U be a neighbourhood of x. Then by definition $\exists \varepsilon > 0, \mathcal{D}_\varepsilon(x) \subset U$. Since $x_n \to x$,

$$\exists N \in \mathbb{N}, \forall n \geq N, x_n \in \mathcal{D}_\varepsilon(x)$$

hence $\forall n \geq N, x_n \in U$.

Now we show (ii) implies (iii). This is clear since any open set U with $x \in U$ is a neighbourhood of x.

Finally, (iii) implies (i). Fix $\varepsilon > 0$. By the above lemma, $U = \mathcal{D}_\varepsilon(x)$ is open, and $x \in U$. Then by (iii),

$$\exists N \in \mathbb{N}, \forall n \geq n, x_n \in U$$

hence $d(x_n, x) < \varepsilon$. □

Proposition. Let $f : M \to M'$ be a function between metric spaces.

(a) The following are equivalent for all $x \in M$.

(i) f is continuous at x;

(ii) for all neighbourhoods V of $f(x)$ in M', there exists a neighbourhood U of x in M such that $f(U) \subset V$;

(iii) for all neighbourhoods V of $f(x)$ in M', $f^{-1}(V)$ is a neighbourhood of x in M.

(b) The following are equivalent.

(i) f is continuous;

(ii) $f^{-1}(V)$ is open in M for all open subsets V of M'.

Proof. First, we show (a)(i) implies (a)(ii). Let V be a neighbourhood of $f(x)$ in M'. By definition, $\exists \varepsilon > 0$ such that $\mathcal{D}_\varepsilon(f(x)) \subset V$. Since f is continuous at x, there exists $\delta > 0$ such that $f(\mathcal{D}_\delta(x)) \subset \mathcal{D}_\varepsilon(f(x))$. Then, $U = \mathcal{D}_\delta(x)$ is a neighbourhood of x in M, and $f(U) \subset V$.

Now, (a)(ii) implies (a)(iii). Let V be a neighbourhood of $f(x)$ in M'. By (ii), there exists a neighbourhood of x in M such that $f(U) \subset V$. Then $U \subset f^{-1}(V)$ and since U is a neighbourhood of x in M, ther exists $r > 0$ such that $\mathcal{D}_r(x) \subset U \subset f^{-1}(V)$ Thus, $f^{-1}(V)$ is a neighbourhood of x in M.

Finally, (a)(iii) implies (a)(i). Given $\varepsilon > 0$, $V = \mathcal{D}_\varepsilon(f(x))$ is a neighbourhood of $f(x)$ in V. By (iii), $f^{-1}(V)$ is a neighbourhood of x in M. So $\exists \delta > 0$ such that $\mathcal{D}_\delta(x) \subset f^{-1}(V)$. Thus, $f(\mathcal{D}_\delta(x)) \subset V = \mathcal{D}_\varepsilon(f(x))$.

Now, (b)(i) implies (b)(ii). Let V be open in M'. So pick $x \in f^{-1}(V)$. Then, $f(x) \in V$. Since V is open, $\exists \varepsilon > 0, \mathcal{D}_\varepsilon(f(x)) \subset V$. Since f is continuous at x, $\exists \delta > 0, f(\mathcal{D}_\delta(x)) \subset \mathcal{D}_\varepsilon(f(x))$. Then, $\mathcal{D}_\delta(x) \subset f^{-1}(\mathcal{D}_\varepsilon(f(x))) \subset f^{-1}(V)$.

Finally, (b)(ii) implies (b)(i). Consider $x \in M$. We must show f is continuous at x. Let $\varepsilon > 0$. Consider the ball $V = \mathcal{D}_\varepsilon(f(x))$. This is open in M' by the above lemma. By (ii), $f^{-1}(V)$ is open in M. Further, $x \in f^{-1}(V)$. So by definition, $\exists \delta > 0, \mathcal{D}_\delta(x) \subset V$, which is exactly continuity as required. □

Definition. The *topology of a metric space M* is the family of all open subsets of M.

Proposition. The topology of a metric space satisfies

(i) $\varnothing$ and M are open;

(ii) if U_i are open in M for $i \in I$ (I may be countable or uncountable), then $\bigcup_{i \in I} U_i$ is open in M;

(iii) if U, V are open then $U \cap V$ is open.

Proof. (ii): Let $x \in \bigcup_{i \in I} U_i$, then $\exists i_a \in I, x \in U_{i_a}$. Then since U_{i_a} is open, $\exists \delta > 0, \mathcal{D}_r(x) \subset U_{i_a} \subset \bigcup_{i \in I} U_i$

(iii) Given $x \in U \cap V$, since U is open then $\exists r > 0, \mathcal{D}_r(x) \subset U$ and $\exists s > 0, \mathcal{D}_s(x) \subset V$. Then let $t = \min(r, s)$, and $\mathcal{D}_t(x) = \mathcal{D}_r(x) \cap \mathcal{D}_s(x) \subset U \cap V$. □

4.4. Properties of topology of metric space

Definition. A subspace A of a metric space M is *closed in M* if for every sequence $(x_n) \in A$ that is convergent in M,

$$\lim_{n\to\infty} x_n \in A$$

Lemma. Closed balls are closed.

Proof. Consider $\mathcal{B}_r(x)$ in M. Consider further $(x_n) \in \mathcal{B}_r(x)$ such that $x_n \to z$ in M.

$$d(z, x) \leq d(z, x_n) + d(x_n, x) \leq d(z, x_n) + r \to r$$

Hence $d(z, x) \leq r$, so $z \in \mathcal{B}_r(x)$. □

Example. $[0, 1] = \mathcal{B}_{1/2}(1/2)$ is closed in $\mathbb{R}$. This is not open, for instance consider $D_r(0) \not\subset [0, 1]$.

Example. $(0, 1) = \mathcal{D}_{1/2}(1/2)$ is open in $\mathbb{R}$. This is not closed, for instance the sequence $\frac{1}{n+1}$ tends to zero in $\mathbb{R}$.

Example. $\mathbb{R}$ and $\varnothing$ are open and closed in $\mathbb{R}$.

Example. $(0, 1]$ in $\mathbb{R}$ is neither open nor closed. Consider $\mathcal{D}_r(1) \not\subset (0, 1]$ and $\frac{1}{n} \to 0 \notin (0, 1]$.

Lemma. Let $A \subset M$. Then A is closed in M if and only if $M \setminus A$ is open in M.

Proof. Let A be closed. Suppose $M \setminus A$ is not open. Then $\exists x \in M \setminus A, \forall r > 0, \mathcal{D}_r(x) \not\subset M \setminus A$, so $\mathcal{D}_r(x) \cap A \neq \varnothing$. In particular, for every n we can choose a point in $\mathcal{D}_{1/n}(x) \cap A$. Then, $d(x_n, x) < \frac{1}{n} \to 0$ and $x_n \in A$ which contradicts the fact that A is closed.

Conversely, let us assume $M \setminus A$ is open, but suppose A is not closed. Then there exists a sequence $(x_n) \in A$ such that $x_n \to x$ in M but $x \notin A$. Since $x \in M \setminus A$ and $M \setminus A$ is open, there exists $\varepsilon > 0, \mathcal{D}_\varepsilon(x) \subset M \setminus A$. Since $x_n \to x$, we must have $\exists N \in \mathbb{N}, \forall n \geq N, x_n \in \mathcal{D}_\varepsilon(x)$ and hence $x_n \in M \setminus A$, which is a contradiction. □

Example. Let M be a discrete metric space. Let $A \subset M$. Then for all $x \in A, \mathcal{D}_1(x) = \{x\} \subset A$. Hence A is open. So in a discrete metric space, all subsets are open. Hence every subset is closed.

4.5. Homeomorphisms

Definition. A map $f : M \to M'$ between metric spaces is called a *homeomorphism* if f is a bijection and f, f^{-1} are continuous. Equivalently, f is a bijection, and for all open sets V in M', $f^{-1}(V)$ is open in M, and for all open sets U in M, $f(U)$ is open in M'. If there exists a homeomorphism between M, M', we say that M, M' are homeomorphic.

Example. Consider $(0, \infty)$ and $(0, 1)$. Consider the map $x \mapsto \frac{1}{x+1}$ with inverse $x \mapsto \frac{1}{x} - 1$. These are continuous, so the metric spaces are homeomorphic.

Remark. Every isometry is a homeomorphism, since it is bijective by definition. It is not true that every homeomorphism is an isometry.

Consider the identity on $\mathbb{R}$ with the discrete metric to $\mathbb{R}$ with the Euclidean metric. This is a continuous bijection whose inverse is not continuous. So it is not true that a continuous bijection always has a continuous inverse.

4.6. Equivalence of metrics

Definition. Let d, d' be metrics on a set M. We say that d, d' are *equivalent*, written $d \sim d'$, if they define the same topology. In particular, $U \subset M$ is open in (M, d) if and only if U is open in (M, d'). So $d \sim d'$ if and only if $\mathrm{id} : (M, d) \to (M, d')$ is a homeomorphism.

Remark. If $d \sim d'$, then (M, d) and (M, d') have the same convergent sequences and continuous maps.

Definition. Let d, d' be metrics on M. Then we say d, d' are *uniformly equivalent*, written $d \sim_u d'$ if

$$\mathrm{id} : (M, d) \to (M, d'); \quad \mathrm{id} : (M, d') \to (M, d)$$

are uniformly continuous. We say d, d' are *Lipschitz equivalent*, written $d \sim_{\mathrm{Lip}} d'$, if the identity maps above are Lipschitz. Equivalently, $d \sim_{\mathrm{Lip}} d'$ if $\exists a > 0, b > 0, ad(x, y) \leq d'(x, y) \leq bd(x, y)$. Note, $d \sim_{\mathrm{Lip}} d' \implies d \sim_u d' \implies d \sim d'$.

Example. Given a metric space (M, d), we define $d'(x, y) = \min(1, d(x, y))$. This defines a metric on M, and $d' \sim_u d$.

Example. On $M \times M'$, d_1, d_2, d_∞ are pairwise Lipschitz equivalent.

Example. Consider $C[0, 1]$. The L_1 metric and the uniform metric are not equivalent. Consider $f_n(x) = x^n$. This is convergent to zero in the L_1 metric but is not convergent in the uniform metric.

Example. The discrete metric and Euclidean metric on $\mathbb{R}$ are not equivalent. This is because in the discrete metric all sets are open, but in the Euclidean metric there are some non-open sets.

5. Completeness

5.1. Cauchy sequences

In $\mathbb{R}, \mathbb{C}$, every Cauchy sequence is convergent. We wish to generalise this notion to an arbitrary metric space. Recall that a sequence (x_n) in $\mathbb{R}$ or $\mathbb{C}$ is bounded if there exists $c \in \mathbb{R}^+$ such that $\forall n \in \mathbb{N}, |x_n| \leq c$.

Definition. A sequence (x_n) in a metric space M is said to be *Cauchy* if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, d(x_m, x_n) < \varepsilon$$

The sequence is bounded if

$$\exists z \in M, \exists r > 0, \forall n \in \mathbb{N}, x_n \in \mathcal{B}_r(z)$$

This is equivalent to

$$\forall z \in M, \exists r > 0, \forall n \in \mathbb{N}, x_n \in \mathcal{B}_r(z)$$

by considering the triangle inequality around the given z point. In particular, if the metric arises from a norm, (x_n) is bounded if and only if $\|x_n\|$ is bounded.

Lemma. If a sequence is convergent, it is Cauchy. If a sequence is Cauchy, it is bounded.

Proof. Let (x_n) be a sequence in M. First, we assume that (x_n) is convergent in M, so let x be the limit. Given $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $\forall n \geq N$, $d(x_n, x) < \varepsilon$. Then, for all $m, n \geq N$, we have $d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) \leq 2\varepsilon$ as required. So (x_n) is Cauchy.

Now conversely, we assume (x_n) is Cauchy. There exists $n \in \mathbb{N}$ such that $\forall m, n \geq N$, we have $d(x_m, x_n) < 1$. In particular, $d(x_n, x_N) < 1$ for $n \geq N$. In other words, $x_n \in \mathcal{B}_1(x_N)$. Now, let $r = \max\{d(x_1, x_N), \dots, d(x_{N-1}, x_N), 1\}$. This r is a bound for all elements of the sequence; for all $n \in \mathbb{N}, x_n \in \mathcal{B}_r(x_N)$. □

Remark. Boundedness does not imply the sequence is Cauchy. For instance, consider the sequence $0, 1, 0, 1, \dots$ in $\mathbb{R}$. If a sequence is Cauchy, it is not necessarily convergent in an arbitrary metric space (not $\mathbb{R}, \mathbb{C}$). For instance, consider $x_n = \frac{1}{n}$ in $(0, \infty)$. This is certainly not convergent, since the limit cannot be zero.

5.2. Definition of completeness

Definition. A metric space M is called *complete* if every Cauchy sequence in M converges in M.

Example. $\mathbb{R}, \mathbb{C}$ are complete.

5.3. Completeness of product spaces

Proposition. Product spaces of complete spaces are complete. More precisely, if M, M' are complete, then so is $M \oplus_p M'$.

Proof. Let (a_n) be a Cauchy sequence in the product space $M \oplus_p M'$. We will write $a_n = (x_n, x'_n)$ for all n. Then, since (a_n) is Cauchy,

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \in N, d_p(a_m, a_n) < \varepsilon$$

Then, for all $m, n \geq N$,

$$d(x_m, x_n) \leq \max\{d(x_m, x_n), d(x'_m, x'_n)\} \leq d_p(a_m, a_n) < \varepsilon$$

Hence (x_n) is Cauchy in M, and similarly (x'_m) is Cauchy in M'. Since M, M' are complete, $(x_n), (x'_n)$ are convergent in M, M' to x, x'. Now, let $a = (x, x')$. Then,

$$d_p(a_n, a) \leq d_1(a_n, a) = d(x_n, x) + d(x'_n, x') \to 0$$

So the product space is complete. □

Remark. (a_n) is Cauchy in $M \oplus_p M'$ if and only if (x_n) is Cauchy in M and (x'_n) is Cauchy in M'.

Corollary. $\mathbb{R}^n, \mathbb{C}^n$ are complete in the ℓ_p metric. In particular, n-dimensional real or complex Euclidean space is complete.

5.4. Completeness of subspaces and function spaces

Theorem. Let S be any set. Then, $\ell_\infty(S)$, the set of bounded scalar functions on S, is complete in the uniform metric D.

Proof. Let (f_n) be a Cauchy sequence in $\ell_\infty(S)$. Then,

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, D(x_m, x_n) = \sup_{x \in S} |f_m(x) - f_n(x)| < \varepsilon$$

In other words, $\forall m, n \geq N, \forall x \in S, |f_m(x) - f_n(x)| < \varepsilon$. So (f_n) is uniformly Cauchy as defined previously. As shown previously, (f_n) is uniformly convergent. Hence, there is a scalar function f on S such that $f_n \to f$ uniformly on S. We have also shown previously that the uniform limit f of bounded functions (f_n) is bounded. In other words, $f \in \ell_\infty(S)$. Now it remains to show that $f_n \to f$ in the uniform metric.

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq M, \forall x \in S, |f_n(x) - f(x)| < \varepsilon$$

Hence,

$$\forall n \geq N, \sup_{x \in S} |f_n(x) - f(x)| = D(f_n, f) \leq \varepsilon$$

which is convergence in the metric as required. □

Proposition. Let N be a subspace of a metric space M. Then,

(i) If N is complete, N is closed in M.

(ii) If M is complete and N is closed in M, then N is complete.

In other words, in a complete metric space, a subspace is complete if and only if it is closed.

Proof. To prove (i), we let (x_n) be a sequence in N and assume that $x_n \to x$ in M. We want to show that $x \in N$. We know (x_n) is convergent in M, so it is Cauchy in M. So (x_n) is Cauchy in N. Since N is complete, $x_n \to y$ in N. So $x_n \to y$ in M. By uniqueness of limits, $x = y$ as required.

Now we want to prove (ii) is complete. Let (x_n) be a Cauchy sequence in N. Then (x_n) is Cauchy in M. Since M is complete, $x_n \to x$ in M for some $x \in M$. Since N is closed in M, $x \in N$. So $x_n \to x$ in N. □

Theorem. Let (M, d) be a metric space, and define $C_b(M)$ to be the set of functions f in $\ell_\infty(M)$ such that f is continuous. This is a subspace of $\ell_\infty(M)$ in the uniform metric D. $C_b(M)$ is complete in the uniform metric.

Proof. By the above proposition, it is sufficient to show that $C_b(M)$ is closed in $\ell_\infty(M)$. Let (f_n) be a sequence in $C_b(M)$ and we assume that $f_n \to f$ in $\ell_\infty(M)$. We want to show that $f_n \in C_b(M)$. It is now sufficient to show that f is continuous, or equivalently, continuous at every point in M. Let $a \in M$, and let $\varepsilon > 0$. Since $f_n \to f$ in $\ell_\infty(M)$, we can fix $n \in \mathbb{N}$ such that $F(f_n, f) < \varepsilon$. Since f_n is continuous (at a),

$$\exists \delta > 0, \forall x \in M, d(x, a) < \delta \implies |f_n(x) - f_n(a)| < \varepsilon$$

Hence, $\forall x \in M$, if $d(x, a) < \delta$ we have

$$\begin{aligned} |f(x) - f(a)| &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(a)| + |f_n(a) - f(a)| \\ &\leq 2D(f_n, f) + |f_n(x) - f_n(a)| \\ &< 3\varepsilon \end{aligned}$$

□

Corollary. Consider $C[a, b]$, the space of continuous functions on $[a, b]$. This space is complete in the uniform metric, since $C[a, b] = C_b[a, b]$.

Definition. Let S be a set, and (N, e) be a metric space. Then we generalise $\ell_\infty(S)$ to the following definition.

$$\ell_\infty(S, N) = \{f : S \to N : f \text{ is bounded}\}$$

where f is bounded if there exists $y \in N, r > 0$ such that $\forall x \in S, f(x) \in \mathcal{B}_r(y)$. If $g : S \to N$ is a bounded function, $\forall x \in S, g(x) \in \mathcal{B}_s(z)$, then

$$\forall x \in S, e(f(x), g(x)) \leq e(f(x), y) + e(y, z) + e(z, g(x)) \leq r + e(y, z) + s$$

This is a uniform bound for all x, so we may take the supremum. So $\sup_{x\in S} e(f(x), g(x))$ exists and we denote this by

$$\mathcal{D}(f, g) = \sup_{x\in S} e(f(x), g(x))$$

This can be shown to be a metric, called the uniform metric on $\ell_\infty(S, N)$. Now, let $S = M$, where (M, d) is a metric space. We define

$$C_b(M, N) = \{f : M \to N : f \text{ continuous and bounded}\}$$

Note that $C_b(M, N)$ is a subspace of $\ell_\infty(M, N)$ with the uniform metric.

Theorem. Let S be a set, let (M, d) be a metric space, and let (N, e) be a complete metric space. Then

(i) $\ell_\infty(S, N)$ is complete in the uniform metric D;

(ii) $C_b(M, N)$ is complete in the uniform metric D.

Proof. We first prove part (i). Let (f_n) be a Cauchy sequence in $\ell_\infty(S, N)$. We first show that (f_n) is pointwise Cauchy. Let $x \in S$.

$$\forall\varepsilon > 0, \exists K \in \mathbb{N}, \forall i, j \geq K, D(f_i, f_j) < \varepsilon$$

In particular, $e(f_i(x), f_j(x)) \leq D(f_i, f_j) < \varepsilon$ for $i, j \geq K$. So the sequence $(f_k(x))_k$ is Cauchy in N. Since N is complete, $(f_k(x))_k$ converges. This holds for all $x \in S$, hence we can define $f : S \to N$ by $f(x) = \lim_{k\to\infty} f_k(x)$.

Now, we must show that f is bounded, such that $f \in \ell_\infty(S, N)$. Since f_k is Cauchy in the uniform metric D, there exists $K \in \mathbb{N}$ such that $\forall i, j \geq K$, $D(f_i, f_j) < 1$. In particular, for all $i \geq K$, $D(f_i, f_K) < 1$. Since f_K is bounded, there exists $y \in N, r > 0$ such that $\forall x \in S, f_K(x) \in \mathcal{B}_r(y)$. Then, by the triangle inequality, for a fixed $x \in S$, $\forall i \geq K, e(f_i(x), f_K(x)) \leq D(f_i(x), f_K(x)) < 1$. Let $i \to \infty$, then $e(f_i(x), f_K(x)) \leq 1$. Hence $e(f(x), y) \leq e(f(x), f_K(x)) + e(f_K(x), y) \leq 1 + r$. But since this is true for all x, $1 + r$ is a uniform bound; $\forall x \in S, f(x) \in \mathcal{B}_{r+1}(y)$.

Now we will show that $f_k \to f$ uniformly in D. Again, we use

$$\forall\varepsilon > 0, \exists K \in \mathbb{N}, \forall i, j \geq K, D(f_i, f_j) < \varepsilon$$

So choose $i \geq K$, and $x \in S$. Then for all $j \geq K$, $e(f_i(x), f_j(x)) \leq D(f_i, f_j) < \varepsilon$. As $j \to \infty$, $e(f(x), f_i(x)) \leq \varepsilon$, because metrics are continuous. But since x was arbitrary, we have a uniform distance $D(f, f_i) < \varepsilon$. This holds for all $i \geq K$, so we have uniform convergence.

Now we prove part (ii). By part (i) and an above proposition, it is enough to show that $C_b(M, N)$ is closed in $\ell_\infty(M, N)$. Let (f_k) be a sequence in $C_b(M, N)$ and $f_k \to f$ in $\ell_\infty(M, N)$. We require $f \in C_b(M, N)$, so it is enough to show that f is continuous. This is exactly the proof that the uniform limit of continuous functions is continuous. Let $a \in M, \varepsilon > 0$. Then,

since $f_k \to f$ in $\ell_\infty(M,N)$, we can fix $k \in \mathbb{N}$ such that $D(f_k, f) < \varepsilon$. Since f_k is continuous, $\exists \delta > 0, \forall x \in M, d(x,a) < \delta \implies e(f_k(x), f_k(a)) < \varepsilon$.

$$\begin{aligned} \forall x \in M, f(x,a) < \delta \implies e(f(x), f(a)) &\leq e(f(x), f_k(x)) + e(f_k(x), f_k(a)) + e(f_k(a), f(a)) \\ &\leq 3\varepsilon \end{aligned}$$

□

6. Contraction mapping theorem

6.1. Contraction mappings

Definition. A function $f : M \to M'$ is called a *contraction mapping* if $\exists \lambda, 0 \leq \lambda < 1$ such that

$$\forall x, y \in M, d'(f(x), f(y)) \leq \lambda d(x, y)$$

so f is λ-Lipschitz.

6.2. Contraction mapping theorem

This theorem is also called Banach's fixed point theorem.

Theorem. Let M be a non-empty complete metric space. Let $f : M \to M$ be a contraction mapping. Then f has a unique fixed point:

$$\exists! z \in M, f(z) = z$$

Proof. Let λ such that $0 \leq \lambda < 1$ and $\forall x, y \in M, d'(f(x), f(y)) \leq \lambda d(x, y)$. First we show uniqueness. Suppose there were two fixed points $f(z) = z, f(w) = w$. Then $d(z, w) = d(f(z), f(w)) \leq \lambda d(z, w) < d(z, w)$. Hence $d(z, w) = 0$ so $z = w$.

Now we show existence. Fix a starting point $x_0 \in M$. Let $x_n = f(x_{n-1})$ for all $n \in \mathbb{N}$, so $x_n = f^n(x_0)$. First, observe that for all $n \in \mathbb{N}$,

$$d(x_n, x_{n+1}) = d(f(x_{n-1}), f(x_n)) \leq \lambda d(x_{n-1}, x_n) \leq \cdots \leq \lambda^n d(x_0, x_1)$$

For $m \geq n$, we have

$$d(x_n, x_m) \leq \sum_{k=n}^{m-1} d(x_k, x_{k+1}) \leq \sum_{k=n}^{m-1} \lambda^k d(x_0, x_1) \leq \frac{\lambda^n}{1-\lambda} d(x_0, x_1)$$

Since $\frac{\lambda^n}{1-\lambda} d(x_0, x_1) \to 0$,

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \frac{\lambda^n}{1-\lambda} d(x_0, x_1) < \varepsilon$$

Hence, $\forall m \geq n \geq N, d(x_n, x_m) < \varepsilon$. So the sequence (x_n) is Cauchy. Since M is complete, (x_n) is convergent to some point $z \in M$. f is continuous since it is a contraction, so $f(x_n) \to z$ so $f(z) = z$. So the fixed point exists. □

Remark. Letting $m \to \infty$ in the inequality for $d(x_n, x_m)$, $d(x_n, z) \leq \frac{\lambda^n}{1-\lambda} d(x_0, x_1)$. So $x_n \to z$ exponentially fast. Consider $f : \mathbb{R} \setminus \{0\} \to \mathbb{R} \setminus \{0\}$, and $x \mapsto \frac{x}{2}$. This is a contraction, but there is no fixed point. This is because $\mathbb{R} \setminus \{0\}$ is not complete. Consider instead $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x + 1$. This has no fixed point, since f is an isometry ($\lambda = 1$) and not a contraction. Consider further $f : [1, \infty) \to [1, \infty)$ mapping $x \mapsto x + \frac{1}{x}$. Certainly $|f(x) - f(y)| < |x - y|$. $[1, \infty)$ is closed in $\mathbb{R}$ so it is complete. However this is not a contraction; even though $|f(x) - f(y)| < |x - y|$, there is no upper bound λ. There are no fixed points.

6.3. Application of contraction mapping theorem

Let $y_0 \in \mathbb{R}$. Then the initial value problem $f'(t) = f(t^2)$ and $f(0) = y_0$ has a unique solution on $\left[0, \frac{1}{2}\right]$. In other words, there exists a unique differentiable function $f : \left[0, \frac{1}{2}\right] \to \left[0, \frac{1}{2}\right]$ such that $f(0) = y_0$ and $f'(t) = f(t^2)$ for all t in the domain.

First, observe that if f is a solution then certainly it is continuous, so $f \in C\left[0, \frac{1}{2}\right]$. Further, by the fundamental theorem of calculus, it satisfies

$$f(t) = y_0 + \int_0^t f(s^2)\,\mathrm{d}s$$

Note that $f'(s) = f(s^2)$ is continuous. Conversely, if $f \in C\left[0, \frac{1}{2}\right]$ and $f(t) = y_0 + \int_0^t f(s^2)\,\mathrm{d}s$ then f is a solution to the initial value problem.

Let $M = C\left[0, \frac{1}{2}\right]$ with the uniform metric. This is non-empty and complete. Then we define the map $T : M \to M$ by

$$(Tg)(t) = y_0 + \int_0^t g(s^2)\,\mathrm{d}s$$

Note that Tg is well-defined since $g(s^2)$ is continuous. Moreover, by the fundamental theorem of calculus, Tg is differentiable and $(Tg)'(t) = g(t^2)$. Thus, f is a solution to the initial value problem if and only if $f \in M$ and $Tf = f$.

Now, if T is a contraction, we can use the contraction mapping theorem to assert that there is exactly one fixed point. For $g, h \in M$, $t \in \left[0, \frac{1}{2}\right]$, consider

$$|(Tg)(t) - (Th)(t)| = \left| \int_0^t \left[g(s^2) - h(s^2)\right] \mathrm{d}s \right| \leq t \sup_{s \in \left[0, \frac{1}{2}\right]} |g(s^2) - h(s^2)| \leq \frac{1}{2} D(g, h)$$

Taking the supremum over t gives $D(Tg, Th) \leq \frac{1}{2} D(g, h)$, and so there is exactly one fixed point.

Remark. The above shows that for any $\delta \in (0, 1)$ there is a unique solution to the initial value problem on $[0, \delta]$, called f_δ, since $\delta < 1$ is required for the map to be a contraction. For $0 < \delta < \mu < 1$, $f_\mu|_{[0,\delta]} = f_\delta$ by uniqueness. So we can combine the solutions together to yield a unique solution on $[0, 1)$.

6.4. Lindelöf–Picard theorem

Theorem. Let $n \in \mathbb{N}$, $y_0 \in \mathbb{R}^n$, and $a, b, R \in \mathbb{R}$, such that $a < b$ and $R > 0$. Let $\phi : [a, b] \times \mathcal{B}_R(y_0) \to \mathbb{R}^n$ be a continuous function. Given that there exists $K > 0$ such that $\forall t \in [a, b], \forall x, y \in \mathcal{B}_R(y_0)$, such that

$$\|\phi(t, x) - \phi(t, y)\| \leq K\|x - y\|$$

Then, $\exists \varepsilon > 0$ such that $\forall t, t_0 \in [a, b]$, the initial value problem

$$f'(t) = \phi(t, f(t)); \quad f(t_0) = y_0$$

has a unique solution on $[c, d] = [t_0 - \varepsilon, t_0 + \varepsilon] \cap [a, b]$.

Remark. If f is a solution of the initial value problem, implicitly this includes the assumption that $f(t) \in B_r(y_0)$ for all $t \in [c, d]$. Note that if $f : [c, d] \to \mathbb{R}^n$, we let $f_k : [c, d] \to \mathbb{R}$ be the kth component of f, and $f_k = q_k \circ f$ where q_k is the kth coordinate projection. Then, $f(t) = (f_1(t), \dots, f_n(t))$ and we define f to be differentiable if and only if all of the components are differentiable, with $f'(t) = (f_1'(t), \dots, f_n'(t))$. Note further, if f is continuous, then so are f_k, hence f_k are integrable. So we define

$$\int_c^d f(t)\,\mathrm{d}t = v = \left(\int_c^d f_1(t)\,\mathrm{d}t, \dots, \int_c^d f_n(t)\,\mathrm{d}t \right)$$

Note that we can use the Cauchy–Schwarz inequality to give

$$\begin{aligned}
\|v\|^2 &= \sum_{k=1}^n v_k^2 \\
&= \sum_{k=1}^n v_k \int_c^d f_k(t)\,\mathrm{d}t \\
&= \int_c^d \sum_{k=1}^n v_k f_k(t)\,\mathrm{d}t \\
&= \int_c^d v \cdot f(t)\,\mathrm{d}t \\
&\leq \int_c^d \|v\| \cdot \|f(t)\|\,\mathrm{d}t \\
&= \|v\| \int_c^d \|f(t)\|\,\mathrm{d}t
\end{aligned}$$

Hence,

$$\left\| \int_c^d f(t)\,\mathrm{d}t \right\| \leq \int_c^d \|f(t)\|\,\mathrm{d}t \leq (d - c) \sup_{t \in [c,d]} \|f(t)\|$$

Proof. Recall that closed balls are closed, hence $\mathcal{B}_R(y_0)$ is a closed subset of $\mathbb{R}^n$. So ϕ is a continuous function on the closed and bounded set $[a, b] \times \mathcal{B}_R(y_0)$. It follows that ϕ is bounded. Now, let $c = \sup\{\|\phi(t, x)\| : t \in [a, b], x \in \mathcal{B}_R(y_0)\}$. Let $\varepsilon = \min(\frac{R}{c}, \frac{1}{2K})$. Let $t_0 \in [a, b]$ and let $[c, d] = [t_0 - \varepsilon, t_0 + \varepsilon] \cap [a, b]$. We need to show that there exists a unique differentiable function $f : [c, d] \to \mathbb{R}^n$ such that $f(t_0) = y_0$ and $f'(t) = \phi(t, f(t))$ for all $t \in [c, d]$. Since $\mathcal{B}_R(y_0)$ is closed in $\mathbb{R}^n$, and since $\mathbb{R}^n$ is complete, $\mathcal{B}_R(y_0)$ is complete. Then, $M = C([c, d], \mathcal{B}_R(y_0))$ is complete in the uniform metric D. This is certainly non-empty;

consider the constant function yielding y_0. f is a solution to the initial value problem if $f \in M$ and $f'(t) = y_0 + \int_{t_0}^{t} \phi(s, f(s))\,\mathrm{d}s$, from the fundamental theorem of calculus applied coordinatewise. We define $T : M \to M$ mapping $g \mapsto Tg$ where Tg is given by

$$(Tg)(t) = y_0 + \int_{t_0}^{t} \phi(s, g(s))\,\mathrm{d}s$$

We must show T is well defined. First, note that the integral is well defined; $s \mapsto \phi(s, g(s))$ is continuous so integrable. By the fundamental theorem of calculus, Tg is differentiable and the derivative is $(Tg)'(t) = \phi(t, g(t))$. In particular, $Tg : [c, d] \to \mathbb{R}^n$ is continuous. Finally, for $t \in [c, d]$,

$$\|(Tg)(t) - y_0\| = \left\| \int_{t_0}^{t} \phi(s, g(s))\,\mathrm{d}s \right\| \leq |t - t_0| \sup_{s \in [c,d]} \|\phi(s, g(s))\| \leq \varepsilon c \leq R$$

So $Tg \in M$. Recall that f is a solution of the initial value problem if and only if $f \in M$ and $Tf = f$. Now we must show that T has a unique fixed point, so we will show that T is a contraction. Let $t \in [c, d]$ and $g, h \in M$.

$$\|(Tg)(t) - (Th)(t)\| = \left\| \int_{t_0}^{t} [\phi(s, g(s)) - \phi(s, h(s))]\,\mathrm{d}s \right\|$$

Note that $\|\phi(s, g(s)) - \phi(s, h(s))\| \leq K\|g(s) - h(s)\| \leq K \cdot D(g, h)$.

$$\|(Tg)(t) - (Th)(t)\| = |t - t_0| \cdot K \cdot K(g, h) \leq \varepsilon K D(g, h)$$

Taking the supremum over $t \in (c, d)$,

$$D(Tg, Th) \leq \varepsilon K D(g, h) \leq \frac{1}{2} D(g, h)$$

So T is a contraction. By the contraction mapping theorem, T has a unique fixed point in M. □

Remark. For any $\delta \in (0, 1)$, taking $\varepsilon = \min(\frac{R}{c}, \frac{\delta}{K})$ works. But by the uniqueness of the solution, the choice does not matter for constructing the solution. So we can construct the solution for $\varepsilon = \min(\frac{R}{c}, \frac{1}{K})$, on $(t_0 - \varepsilon, t_0 + \varepsilon) \cap [a, b]$. In general, there is no solution on $[a, b]$. Finally, note that the above theorem can handle any nth order ODE for any $n \in \mathbb{N}$.

7. Topology

7.1. Definitions

Definition. Let X be a set. A *topology* on X is a family τ of subsets of X (so $\tau \subset \mathcal{P}(X)$) such that

(i) $\emptyset, X \in \tau$;

(ii) if $U_i \in \tau$ for all $i \in I$ where I is some index set, then $\bigcup_{i\in I} U_i \in \tau$; and

(iii) if $U, V \in \tau$ then $U \cap V \in \tau$.

A *topological space* is a pair (X, τ) where X is a set and τ is a topology on X. Members of τ are called *open sets* in the topology. So we say that $U \subset X$ is *open in* X, or U is τ*-open*, if $U \in \tau$.

Remark. If $U_i \in \tau$ for $i = 1, \dots, n$, then $\bigcap_{i=1}^{n} U_i \in \tau$.

Example. Let (M, d) be a metric space. Recall that $U \subset M$ is open in the metric sense if $\forall x \in U, \exists r > 0, \mathcal{B}_r(x) \subset U$. We may say that U is d*-open*. We have already proven that the family of d-open sets is a topology on M. This is a metric topology.

Definition. Let (X, τ) be a topological space. Then we say that X is *metrisable* (or sometimes we say τ is metrisable) if there exists a metric d on X such that τ is the metric topology on X induced by d. In other words, $U \subset X$ is τ-open if and only if U is d-open. If $d' \sim d$, then d' also induces the same topology τ on X.

Example. The indiscrete topology on a set X is a topology $\tau = \{\emptyset, X\}$. If $|X| \geq 2$, then this is not metrisable. Let d be a metric on X. Then let $x \neq y \in X$, let $r = d(x, y)$, and finally let $U = \mathcal{D}_r(x)$. We know that U is d-open. But since $x \in U, y \notin U$, $U \notin \tau$.

Definition. If τ_1, τ_2 are topologies on X, we say that τ_1 is *coarser* than τ_2, or that τ_2 is finer than τ_1, if $\tau_1 \subset \tau_2$. For example, the indiscrete topology on X is the coarsest topology on X.

Example. The discrete topology on a set X is $\tau = \mathcal{P}(X)$. This is the finest topology on X. This is metrisable by the discrete metric.

Definition. A topological space X is *Hausdorff* if $\forall x \neq y$ in X, there exist open sets U, V in X such that $x \in U, y \in V, U \cap V = \emptyset$. Informally, x, y are 'separated by open sets'.

Proposition. Metric spaces are Hausdorff.

Proof. Let $x \neq y$ be points in a metric space (M, d). Let $r > 0$ such that $2r < d(x, y)$. Then let $U = \mathcal{D}_r(x)$, let $V = \mathcal{D}_r(y)$. Certainly U, V are open since they are open balls, and they have no intersection by the triangle inequality, so the metric space is Hausdorff as required. □

Example. The cofinite topology on a set X is

$$\tau = \{\emptyset\} \cup \{U \in X : U \text{ is cofinite in } X\}$$

where U is cofinite in X if $X \setminus U$ is finite. When X is finite, this topology τ is simply $\mathcal{P}(X)$. When X is infinite, τ is not metrisable. Let $x \neq y$ in X, and let $x \in U, y \in V$ where U, V are open in X. Then U and V are cofinite, and hence $U \cap V \neq \varnothing$. So this topology on an infinite set is not Hausdorff and hence not metrisable.

7.2. Closed subsets

Definition. A subset A of a topological space (X, τ) is said to be *closed* in X if $X \setminus A$ is open in X.

Remark. In a metric space, this agrees with the earlier definition of a closed subset, as proven before.

Proposition. The collection of closed sets in a topological space X satisfy

(i) $\varnothing, X$ are closed;

(ii) If A_i are closed in X for i in some non-empty index set I, then $\bigcap_{i \in I} A_i$ is closed;

(iii) If A_1, A_2 are closed in X then $A_1 \cup A_2$ is closed.

Example. In a discrete topological space, every set is closed.

Example. In the cofinite topology, a subset is closed if and only if it is finite or the full set.

7.3. Neighbourhoods

Definition. Let X be a topological space, and let $U \subset X$ and $x \in X$. We say that U is a *neighbourhood* of x in X if there exists an open set V in X such that $X \in V \subset U$.

Remark. In a metric space, we defined this in terms of open balls not open sets. However, we have already proven that the definitions agree.

Proposition. Let U be a subset of a topological space X. Then U is open if and only if U is a neighbourhood of x for every $x \in U$.

Proof. If U is open, and $x \in U$, then by letting $V = U$, V is open and $x \in V \subset U$. Conversely, if $x \in U$, there exists V_x in X such that $x \in V_x \subset U$. Then, $U = \bigcup_{x \in U} x = \bigcup_{x \in U} V_x$ is open, since each V_x is open. □

7.4. Convergence

Definition. Let (x_n) be a sequence in a topological space X. Let $x \in X$. We say that (x_n) converges to x if for all neighbourhoods U of x in X, there exists $N \in \mathbb{N}$ such that $\forall n \geq N, x_n \in U$. Equivalently, for all open sets U which contain x, there exists $N \in \mathbb{N}$ such that $\forall n \geq N, x_n \in U$.

Remark. Again, the definition in a metric space agrees with this definition.

Example. Eventually constant sequences converge. If $\exists z \in X, \exists N \in \mathbb{N}, \forall n \geq N, x_n = z$, then $x_n \to z$.

Example. In an indiscrete topological space, every sequence converges to every point.

Example. In the cofinite topology on a set X, let $x_n \to X$. Suppose that $x_n \to x$ in X. Then if $y \neq x$, $X \setminus \{y\}$ is a neighbourhood of x. Then $N_y = \{n \in N : x_n = y\}$ is finite.

Conversely, suppose (x_n) is a sequence such that for some $x \in X$ and for all $y \neq x$, N_y is finite. Then $x_n \to x$.

In particular, if N_y is finite for all $y \in X$, the sequence converges to every point.

Proposition. If $x_n \to x$ and $x_n \to y$ in a Hausdorff space, then $x = y$.

Proof. Suppose $x \neq y$, then we can choose open sets U, V such that $x \in U, y \in V, U \cap V = \varnothing$. Since $x_n \to x$, there exists $N_1 \in \mathbb{N}$ such that $\forall n \geq N_1, x_n \in U$. Similarly there exists an analogous N_2. Hence $\forall n \geq \max(N_1, N_2), x_n \in U, x_n \in V$ which is a contradiction since $U \cap V = \varnothing$. □

Remark. If $x_n \to x$ in a Hausdorff space, we write $x = \lim_{n\to\infty} x_n$ since the limit is unique.

Remark. In a metric space, for a subset A, we say that A is closed if and only if $x_n \to x$ in A implies $x \in A$. In a general topological space, any closed set is closed under limits, but not every subset that is closed under limits is closed.

7.5. Interiors and closures

Definition. Let X be a topological space, and $A \subset X$. We define the *interior* of A in X, denoted A° or $\operatorname{int}(A)$, by

$$A^\circ = \bigcup \{U \subset X : U \text{ is open in } X, U \subset A\}$$

Similarly we define the *closure* of A in X, denoted $\overline{A}$ or $\operatorname{cl}(A)$, by

$$\overline{A} = \bigcap \{F \subset X : F \text{ is closed in } X, F \supset A\}$$

Remark. Note that A° is open in X, and $A^\circ \subset A$. In particular, if U is open in X and $U \subset A$, then $U \subset A^\circ$. Hence, A° is the largest open subset of A.

Similarly, $\overline{A}$ is closed in X, and $\overline{A} \supset A$. The intersection is not empty since X is closed and $X \supset A$, so it is well-defined. We have that $\overline{A}$ is the smallest closed superset of A.

Proposition. Let X be a topological space and let $A \subset X$. Then the interior is exactly those $x \in X$ for which A is a neighbourhood of x. Similarly, the closure is those $x \in X$ such that for all neighbourhoods U of x, $U \cap A \neq \varnothing$.

Proof. If A is a neighbourhood of X, then by definition there exists an open set U such that $x \in U \subset A$, which is true if and only if $x \in A^\circ$.

For the other part, suppose $x \notin \overline{A}$. Then there exists a closed set $F \supset A$ such that $x \notin F$. Let $U = X \setminus F$. Then U is open and $x \in U$. So U is a neighbourhood of x, and $U \cap A = \varnothing$.

Conversely, suppose there exists a neighbourhood U of x such that $U \cap A = \varnothing$. Then there exists an open set V such that $x \in V \subset U$. Since $V \subset U$, $V \cap A = \varnothing$. Let $F = X \setminus V$. Then F is closed, and $A \subset F$. Hence $\overline{A} \subset F$. So $x \notin \overline{A}$. □

Example. In $\mathbb{R}$, let $A = [0,1) \cup \{2\}$. Then $A^\circ = (0,1)$, and $\overline{A} = [0,1] \cup \{2\}$. Further, $\mathbb{Q}^\circ = \varnothing$ and $\overline{\mathbb{Q}} = \mathbb{R}$. Finally, $\mathbb{Z}^\circ = \varnothing$ and $\overline{\mathbb{Z}} = \mathbb{Z}$.

Remark. In a metric space, for a subset A we have that $x \in \overline{A}$ if and only if there exists a sequence (x_n) in A such that $x_n \to x$. In a general topological space, the existence of a sequence implies $x \in \overline{A}$ but the converse is not true.

7.6. Dense subsets

Definition. A subset A of a topological space X is said to be *dense* in X if $\overline{A} = X$. X is *separable* if there exists a countable subset $A \subset X$ such that A is dense in X.

Example. $\mathbb{R}$ is separable as $\mathbb{Q}$ is dense in $\mathbb{R}$. $\mathbb{R}^n$ is separable in the same way as $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$.

Example. An uncountable discrete topological space is not separable, since the closure of any set is itself.

7.7. Subspaces

Definition. Let (X, τ) be a topological space. Let $Y \subset X$. Then the *subspace topology*, or *relative topology* on Y induced by τ is the topology

$$\{V \cap Y : V \in \tau\}$$

on Y. This is the intersection of Y with all open sets in X. We can denote this $\tau|_Y$. So, for $U \subset Y$, U is open in Y if and only if there exists an open set V in X with $U = V \cap Y$.

Example. Let $X = \mathbb{R}, Y = [0,2]$, and $U = (1,2]$. Then certainly $U \subset Y \subset X$. U is open in Y, since $V = (1,3)$ is open in X and $U = V \cap Y$. However, U is not open in X, since no neighbourhood (or ball) around 2 can be constructed in X that is contained within U.

Remark. On a subset of a topological space, this is considered the standard topology. Suppose that (X, τ) is a topological space, and $Z \subset Y \subset X$. There are two natural topologies on Z: $\tau|_Z$ and $\tau|_Y|_Z$. One can easily check that these two topologies are equal.

Let (M, d) be a metric space, and $N \subset M$. Again, there are two natural topologies on N: $\tau(d)|_N$ and $\tau(d|_N)$, where $\tau(e)$ is the metric topology induced by the metric e. These two constructions coincide; indeed, for any $x \in N, r > 0$,

$$\{y \in N : d(y, x) < r\} = \{y \in M : d(y, x) < r\} \cap N$$

Proposition. Let X be a topological space, and let $A \subset Y \subset X$. A is closed in Y if and only if there exists a closed subset $B \subset X$ such that $A = B \cap Y$. Further,

$$\mathrm{cl}_Y(A) = \mathrm{cl}_X(A) \cap Y$$

This is not true for the interior of a subset in general. For instance, consider $X = \mathbb{R}, A = Y = \{0\}$. In this case, $\mathrm{int}_Y(A) = A, \mathrm{int}_X(A) = \varnothing$.

Proof. The first part is true by taking complements: $Y \setminus A$ is open in Y. By definition, $Y \setminus A = V \cap Y$ for some open V in X. So $B = X \setminus V$ is closed in X and $A = B \cap Y$. If $A = B \cap Y$, B is closed in X, then $X \setminus B$ is open in X, and hence $Y \setminus A = (X \setminus B) \cap Y$ is open in Y.

For the second part, we know $\mathrm{cl}_X(A)$ is closed in X, so by the first part, $\mathrm{cl}_X(A) \cap Y$ is closed in Y. Then $A \subset \mathrm{cl}_X(A) \cap Y$. So by definition, $\mathrm{cl}_Y(A) \subset \mathrm{cl}_X(A) \cap Y$. Similarly, since $\mathrm{cl}_Y(A)$ is closed in Y, we can write $\mathrm{cl}_Y(A) = B \cap Y$ for some closed set B in X. But $A \subset B$, and B is closed in X, so $\mathrm{cl}_X(A) \subset B$ and hence $\mathrm{cl}_Y(A) = B \cap Y \supset \mathrm{cl}_X(A) \cap Y$. □

Remark. If $U \subset Y \subset X$, and Y is open in X, then U is open in Y if and only if U is open in X.

7.8. Continuity

Definition. A function $f : X \to Y$ between topological spaces is said to be continuous if for all open sets V in Y, the preimage $f^{-1}(V)$ is open in X.

Remark. We have already proven that this agrees with the definition of continuity of functions between metric spaces.

Example. Constant functions are always continuous. Consider $f : X \to Y$ defined by $f(x) = y_0$ for a fixed $y_0 \in Y$. For any $V \subset Y$, $f^{-1}(V) = \varnothing$ if $y_0 \notin V$, and $f^{-1}(V) = X$ if $y_0 \in V$. So f is continuous.

Example. The identity map is always continuous. If $f : X \to X$ is defined by $x \mapsto x$, $f^{-1}(V) = V$ so if V is open, $f^{-1}(V)$ is trivially open.

Example. Let $Y \subset X$. Let $i : Y \to X$ be the inclusion map. Then for an open set V in X, $i^{-1}(V) = V \cap Y$ which by definition is open in Y. Hence, if $g : X \to Z$ is continuous, then $g|_Y = g \circ i : X \to Y$ is continuous, as we will see below.

Proposition. Let $f : X \to Y$ be a function between topological spaces. Then,

(i) f is continuous if and only if for all closed sets B in Y, $f^{-1}(B)$ is closed in X;

(ii) if f is continuous and $g : Y \to Z$ is continuous, then $g \circ f$ is continuous.

Proof. To prove (i), note that for any subset $D \subset Y$, $f^{-1}(Y \setminus D) = X \setminus f^{-1}(D)$. We can now use the fact that $A \subset X$ is open in X if and only if $X \setminus A$ is closed in X, and vice versa for Y.

To prove (ii), note that if W is an open subset of Z, then $g^{-1}(W)$ is open in Y since g is continuous. Hence $f^{-1}g^{-1}(W)$ is open in X since f is continuous. But then $f^{-1}g^{-1} = (g \circ f)^{-1}$, so $g \circ f$ is continuous. □

Remark. There exists a notion of 'continuity at a point' for topological spaces, but it is not as useful in this course as the global continuity definition.

7.9. Homeomorphisms and topological invariance

Definition. A function $f : X \to Y$ between topological spaces is a homeomorphism if f is a bijection, and both f, f^{-1} are continuous. If such an f exists, we say that X and Y are homeomorphic. This is exactly the definition from metric spaces.

Definition. A property $\mathcal{P}$ of topological spaces is said to be a *topological property* or *topological invariant* if, for all pairs X, Y of homeomorphic spaces, X satisfies $\mathcal{P}$ if and only if Y satisfies $\mathcal{P}$.

Example. Metrisability is a topological invariant. Being Hausdorff is a topological invariant. Being completely metrisable (metrisable into a complete metric space) is *not* a topological invariant. For example, consider metrics d, d' on $\mathbb{R}$ such that $d \sim d'$ but d is complete and d' is not.

Remark. If $f : X \to Y$ is a homeomorphism, for an open set U in X, $f(U) = (f^{-1})^{-1}(U)$ is open in Y since $f^{-1} : Y \to X$ is continuous.

Definition. A function $f : X \to Y$ between topological spaces is an *open map* if for all open sets U in X, $f(U)$ is open in Y.

Remark. $f : X \to Y$ is a homeomorphism if and only if f is a continuous and open bijection.

7.10. Products

Let X, Y be topological spaces. We want to define the topology on $X \times Y$. If U is open in X and V is open in Y, then we would like $U \times V$ to be open in $X \times Y$. Certainly $\varnothing = \varnothing \times \varnothing$ and $X \times Y$ should be open. Further $(U \times V) \cap (U' \times V') = (U \cap U') \times (V \cap V')$, so intersections work. $\bigcup_{i \in I} U_i \times V_i$ must be open for open sets U_i, V_i, but this is not obvious from what we have shown so far, so we must include this in our definition.

Definition. The *product topology* on $X \times Y$ is the topology such that a subset U of $X \times Y$ is open if there exists a set I and open sets U_i, V_i in X, Y for all $i \in I$ such that

$$U = \bigcup_{i \in I} U_i \times V_i$$

Remark. For $W \subset X \times Y$, we know that W is open if and only if for all $z \in W$, there exist open sets $U \subset X, V \subset Y$, such that $z \in U \times V \subset W$. So, thinking of the product as a product of real lines, we might say that W is open if for every point $z \in W$, we can construct a 'box set' (the Cartesian product of open intervals) contained in W that has z as an element. More formally, W is a neighbourhood of z if and only if there exist neighbourhoods U of x in X and V of y in Y such that $U \times V \subset W$.

7.11. Continuity in product topology

Example. Let $(M, d), (M', d')$ be metric spaces. Then, the metric d_∞ on $M \times M'$ is

$$d_\infty((x, x'), (y, y')) = \max(d(x, y), d'(x', y'))$$

This metric is chosen since all d_p metrics induce the same metric topology, but this is easier to work with. Also, M, M' are topological spaces with their metric topologies, which induce the product topology on the product space $M \times M'$. These two constructions create the same topology. For a point $z = (x, x') \in M \times M'$ and $r > 0$, the open ball $\mathcal{D}_r(z)$ is exactly

$$\begin{aligned}\mathcal{D}_r(z) &= \{(y, y') \in M \times M' : d_\infty((y, y'), (x, x')) < r\} \\ &= \{(y, y') \in M \times M' : d(x, y) < r, d(x', y') < r\} \\ &= \mathcal{D}_r(x) \times \mathcal{D}_r(x')\end{aligned}$$

Now, let $W \subset M \times M'$. Then W is open in the product topology if and only if for all $z = (x, x') \in W$, there exist open sets U in M and U' in M' such that $(x, x') \in U \times U' \subset W$. Equivalently, for all $z = (x, x') \in W$, there exists $r > 0$ such that $\mathcal{D}_r(x) \times \mathcal{D}_r(x') \subset W$. But $\mathcal{D}_r(x) \times \mathcal{D}_r(x') = \mathcal{D}_r(z)$, so W is d_∞-open, as required. For instance, the product topology on $\mathbb{R} \times \mathbb{R}$ is the Euclidean topology on $\mathbb{R}^2$.

Proposition. Let X, Y be topological spaces. Let $X \times Y$ be given the product topology. Then, the coordinate projections $q_X : X \times Y \to X$ and $q_Y : X \times Y \to Y$ satisfy

(i) q_X, q_Y are continuous;

(ii) if Z is any topological space, and $g : Z \to X \times Y$ is a function, then g is continuous if and only if $q_X \circ g, q_Y \circ g$ are continuous.

Proof. If U is open in X, then $q_X^{-1}(U) = U \times Y$, which is the product of an open set in X and an open set in Y, so is open in $X \times Y$. Hence q_X is continuous. Similarly, q_Y is continuous.

If g is continuous then certainly $q_X \circ g, q_Y \circ g$ are continuous since the composition of continuous functions are continuous. Conversely, let $h : Z \to X$ and $k : Z \to Y$ be continuous functions with $h = q_X \circ g$ and $k = q_Y \circ g$. Then $g(x) = (h(x), k(x))$ for $x \in Z$. Now, for open sets U in X and V in Y, we have

$$z \in g^{-1}(U \times V) \iff g(z) \in U \times V \iff h(z) \in U, k(z) \in V \iff z \in h^{-1}(U) \cap k^{-1}(V)$$

So $g^{-1}(U \times V) = h^{-1}(U) \cap k^{-1}(V)$ which is open in Z as h, k are continuous. Given an arbitrary open set W in $X \times Y$, we can write $W = \bigcup_{i \in I} U_i \times V_i$, where U_i are open in X and V_i are open in Y. Thus, $g^{-1}(W) = \bigcup_{i \in I} g^{-1}(U_i \times V_i)$ which is open. □

Remark. The product topology may be extended to a finite product $X_1 \times \cdots \times X_n$, consisting of all unions of sets of the form $U_1 \times \cdots \times U_n$ where U_j is open in X_j. Properties of the product topology hold in this more general case. For example, if X_j is metrisable with metric e_j for all j, then the product topology is metrisable with, for instance, the d_∞ metric.

7.12. Quotients

Let X be a set and R an equivalence relation on X. So $R \subset X \times X$, but we will write $x \sim y$ to mean $(x, y) \in R$. For $x \in X$, we define $q(x) = \{y \in X : y \sim x\}$ to be the equivalence class of x, the set of which partition X. Let X/R denote the set of all equivalence classes. The map $q : X \to X/R$ is called the quotient map.

Definition. Let X be a topological space, and R an equivalence relation on X. The *quotient topology* on X/R is given by

$$\tau = \left\{V \subset X/R : q^{-1}(V) \text{ open in } X\right\}$$

This is a topology:

(i) $q^{-1}(\varnothing) = \varnothing$ which is open, and $q^{-1}(X/R) = X$ which is open.

(ii) If V_i are open, then $q^{-1}\left(\bigcup_{i \in I} V_i\right) = \bigcup_{i \in I} q^{-1}(V_i)$ which is a union of open sets which is open.

(iii) If U, V are open, then $q^{-1}(U \cap V) = q^{-1}(U) \cap q^{-1}(V)$ which is open.

Remark. The quotient map $q : X \to X/R$ is continuous. In particular, it is the largest possible topology on X such that q is continuous.

Let $x \in X, t \in X/R$. Then $x \in t$ if and only if $t = q(x)$. For $V \subset X/R$,

$$\begin{aligned} q^{-1}(V) &= \{x \in X : q(x) \in V\} \\ &= \{x \in X : \exists t \in V, t = q(x)\} \\ &= \{x \in X : \exists t \in V, x \in t\} \\ &= \bigcup_{t \in V} t \end{aligned}$$

Example. Consider $\mathbb{R}$, an abelian group under addition, and the subgroup $\mathbb{Z}$. We can form the quotient group $\mathbb{R}/\mathbb{Z}$, which is the set of equivalence classes where $x \sim y \iff x - y \in \mathbb{Z}$. For all $x \in \mathbb{R}$, there exists $y \in [0, 1]$ such that $x \sim y$, and for all $x, y \in [0, 1]$ we have $x \sim y$ if and only if $x = y$ or $\{x, y\} = \{0, 1\}$. So we can think of the quotient topology of $\mathbb{R}/\mathbb{Z}$ as a circle. We can say that $\mathbb{R}/\mathbb{Z}$ is homeomorphic to $S^1 = \left\{(x, y) \in \mathbb{R}^2 : \|(x, y)\| = 1\right\}$, which we will prove later.

Example. Consider the subgroup $\mathbb{Q}$ of $\mathbb{R}$. Let $V \subset \mathbb{R}/\mathbb{Q}$, such that $V \neq \varnothing$ and V is open. Then $q^{-1}(V)$ is open and not empty. Therefore, there exist $a < b \in \mathbb{R}$ such that $(a, b) \subset q^{-1}(V)$. Given $x \in \mathbb{R}$, we can choose a rational r in the interval $(a - x, b - x)$. Then $r + x \in (a, b) \subset q^{-1}(V)$, so $q(x) = q(r + x) \in V$. So $V = \mathbb{R}/\mathbb{Q}$. This is the indiscrete topology, which is not metrisable or Hausdorff. So we cannot (in general) take quotients of metric spaces.

Example. Let $Q = [0, 1] \times [0, 1] \subset \mathbb{R}^2$. We define the equivalence relation R given by

$$(x_1, x_2) \sim (y_1, y_2) \iff \begin{cases} (x_1, x_2) = (y_1, y_2) & \text{or} \\ x_1 = y_1, \{x_2, y_2\} = \{0, 1\} & \text{or} \\ x_2 = y_2, \{x_1, y_1\} = \{0, 1\} & \text{or} \\ x_1, x_2, y_1, y_2 \in \{0, 1\} \end{cases}$$

The space Q/R is homeomorphic to $\mathbb{R}^2/\mathbb{Z}^2$. This is a square where the top and bottom edges are identified as the same, and the left and right edges are also identified as the same. This is homeomorphic to the surface of a torus with the Euclidean topology embedded in Euclidean three-dimensional space.

Proposition. Let X be a set, and let R be an equivalence relation on X. Let $q : X \to X/R$ be the quotient map. Let Y be a set, and $f : X \to Y$ be a function. Suppose that f 'respects' R; that is, $x \sim y \implies f(x) = f(y)$. Then there exists a unique map $\widetilde{f} : X/R \to Y$ such that $f = \widetilde{f} \circ q$. For $z \in X/R$, we write $z = q(x)$ for some $x \in X$, and then define $\widetilde{f}(z) = f(x)$.

Remark. Note that $\operatorname{Im} f = \operatorname{Im} \widetilde{f}$ since q is surjective. $\widetilde{f}$ is injective if for all $x, y \in X$, $\widetilde{f}(q(x)) = \widetilde{f}(q(y))$ implies $q(x) = q(y)$. In other words, for all $x, y \in X$, $f(x) = f(y) \implies x \sim y$. We say that f *fully respects* R if, for all $x, y \in X$,

$$x \sim y \iff f(x) = f(y)$$

In this case, $\widetilde{f}$ is injective.

7.13. Continuity of functions in quotient spaces

Proposition. Let X be a topological space and let R be an equivalence relation on X. Let $q : X \to X/R$ be a quotient map, where X/R has the quotient topology. Let Y be another topological space and $f : X \to Y$ be a function that respects R. Let $\widetilde{f} : X/R \to Y$ be the unique map such that $f = \widetilde{f} \circ q$. Then

(i) if f is continuous then $\widetilde{f}$ is continuous; and

(ii) if f is an open map (the image of an open set is open) then $\widetilde{f}$ is an open map.

In particular, if f is a continuous surjective map that fully respects R, then $\widetilde{f}$ is a continuous bijection. If in addition f is an open map, then $\widetilde{f}$ is a continuous bijective open map, so is a homeomorphism.

Proof. We prove part (i). Let V be an open set in Y.

$$q^{-1}\big(\widetilde{f}^{-1}(V)\big) = (\widetilde{f} \circ q)^{-1}(V) = f^{-1}(V) \text{ is open}$$

So by definition, $\widetilde{f}^{-1}(V)$ is open in X/R. Hence $\widetilde{f}$ is continuous. Now, we prove part (ii). Let V be an open set in X/R. Let $U = q^{-1}(V)$. Then U is open in X by definition of the quotient topology. Since q is surjective, $q(U) = q\big(q^{-1}(V)\big) = V$. Hence,

$$\widetilde{f}(V) = \widetilde{f}(q(U)) = (\widetilde{f} \circ q)(U) = f(U) \text{ is open}$$

since f is an open map. □

Example. $\mathbb{R}/\mathbb{Z}$ is homeomorphic to a circle $S^1 = \big\{x \in \mathbb{R}^2 : \|x\| = 1\big\}$. We define

$$f(t) = (\cos 2\pi t, \sin 2\pi t)$$

Then, $s - t \in \mathbb{Z}$ if and only if $f(s) = f(t)$ so f fully respects the relation, and f is surjective. f is also continuous since each component is continuous. Hence, there exists $\widetilde{f} : \mathbb{R}/\mathbb{Z} \to S^1$ such that $f = \widetilde{f} \circ q$ and $\widetilde{f}$ is a continuous bijection. Now we must show f is an open map, and then $\widetilde{f}$ will be a homeomorphism. Suppose f is not an open map, so there exists an open set U in $\mathbb{R}$ such that $f(U)$ is not open in S^1. So $S^1 \setminus f(U)$ is not closed, so there exists a sequence (z_n) in this complement and $z \in f(U)$ such that $z_n \to z$. f is surjective so for all $n \in N$ we can choose $x_n \in [0, 1]$ such that $f(x_n) = z_n$. This is a bounded sequence, so by the Bolzano–Weierstrass theorem, without loss of generality we can let $x_n \to x \in [0, 1]$. Since f is continuous, $f(x_n) \to f(x)$, so $z_n \to z$. But since $z_n \notin f(U)$, we have $x_n \in \mathbb{R} \setminus U$. Since the complement is closed and $x_n \to x$, we have $x \in \mathbb{R} \setminus U$ so $x \notin U$. Since $z \in f(U)$, there exists $y \in U$ such that $z = f(y)$. Hence $k = y - x \in \mathbb{Z}$. Now, $f(x_n + k) = f(x_n) = z_n \to z$, but also $x_n + k \to x + k = y \in U$. Since $z_n \notin f(U)$, we have $x_n + k \notin U$. Since $\mathbb{R} \setminus U$ is closed and $x_n + k \to y$, we have $y \in \mathbb{R} \setminus U$ which is a contradiction.

Proposition. Let X be a topological space, and R an equivalence relation on X. Then,

(a) If X/R is Hausdorff, then R is closed in $X \times X$.

(b) If R is closed in $X \times X$ and the quotient map $q : X \to X/R$ is an open map, then X/R is Hausdorff.

Proof. Let $W = X \times X \setminus R$. For part (a), we want to show W is open, so is a neighbourhood of all of its points. Given $(x, y) \in W$, we have $x \not\sim y$, so $q(x) \neq q(y)$. Since the quotient is Hausdorff, there exist open sets S, T in X/R such that $S \cap T = \varnothing$ and $q(x) \in S, q(y) \in T$. Let $U = q^{-1}(S), V = q^{-1}(T)$ which are open in X, and $x \in U, y \in V$. For all $(a, b) \in U \times V$, we have $q(a) \in S, q(b) \in T$ hence $a \not\sim b$. So $(x, y) \in U \times V \subset W$. Hence R is closed.

For part (b), let $z \neq w$ be elements of X/R, and we want to separate these points by open sets. Let $x, y \in X$ such that $q(x) = z, q(y) = w$. Then $(x, y) \in W$ since $x \not\sim w$. Since R is closed, W is open, so there exist open sets U, V in X such that $(x, y) \in U \times V \subset W$. Since q is an open map, $q(U)$ and $q(V)$ are open in X/R, and $z = q(x) \in q(U), w = q(y) \in q(V)$. Now it suffices to show $q(U) \cap q(V) = \varnothing$. For $(a, b) \in U \times V \subset W$, $(a, b) \notin R$ hence $q(a) \neq q(b)$ so $q(U) \cap q(V) = \varnothing$. □

8. Connectedness

8.1. Definition

Recall the intermediate value theorem from IA Analysis. If $f : I \to \mathbb{R}$ is continuous, where I is an interval, and $x < y$ in I and $c \in (f(x), f(y))$, then there exists $z \in (x, y)$ such that $f(z) = c$. An interval in this context is a set I such that for all $x < y < z \in \mathbb{R}$, $x, z \in I \implies y \in I$. So the intermediate value theorem essentially states that the continuous image of an interval is an interval.

Example. Consider $[0, 1) \cup (1, 2]$. Let f be a function from this space to $\mathbb{R}$, defined by

$$f(x) = \begin{cases} 0 & x \in [0, 1) \\ 1 & x \in (1, 2] \end{cases}$$

This is continuous, but the image of f is not an interval.

Definition. A topological space X is *disconnected* if there exist open subsets U, V of X such that $U \cap V = \varnothing, U \cup V = X$ and $U, V \neq \varnothing$. We say that U and V *disconnect* X. We say X is *connected* if X is not disconnected.

Theorem. Let X be a topological space. Then the following are equivalent.

(i) X is connected;

(ii) if $f : X \to \mathbb{R}$ is continuous, then $f(X)$ is an interval;

(iii) if $f : X \to \mathbb{Z}$ is continuous, f is constant.

Proof. First we show (i) implies (ii). Suppose X is connected, and $f : X \to \mathbb{R}$ is continuous, but $f(X)$ is not an interval. Then there exist $a < b < c \in \mathbb{R}$ such that $a, c \in f(X)$ and $b \notin f(X)$. Let $x, y \in X$ such that $f(x) = a, f(y) = c$. Let $U = f^{-1}(-\infty, b), V = f^{-1}(b, \infty)$. U, V are open since f is continuous. U, V are non-empty since $x \in U, y \in V$. Their intersection is empty since we are taking the preimage of disjoint sets. Finally, $U \cup V = f^{-1}(\mathbb{R} \setminus b) = X$ since b is not in the image. So U, V disconnect X, which is a contradiction.

Now (ii) implies (iii). This is immediate since an interval containing an integer must only contain one integer.

Finally, (iii) implies (i). Suppose U, V disconnect X. Let $f : X \to \mathbb{Z}$ by

$$f(x) = \begin{cases} 0 & x \in U \\ 1 & x \in V \end{cases}$$

For any $Y \subset \mathbb{R}$,

$$f^{-1}(Y) = \begin{cases} \varnothing & 0, 1 \notin Y \\ U & 0 \in Y, 1 \notin Y \\ V & 0 \notin Y, 1 \in Y \\ X & 0, 1 \in Y \end{cases}$$

which is open. But f is not constant, so this is a contradiction. □

Corollary. Let $X \subset \mathbb{R}$. Then X is connected if and only if X is an interval.

Proof. Suppose X is connected. Then the inclusion map $i : X \to \mathbb{R}$ is continuous. By the theorem above, $i(X) = X$ is an interval. Conversely, suppose X is an interval. Then, for all continuous $f : X \to \mathbb{R}$, $f(X)$ is an interval by the intermediate value theorem. Then X is connected. □

Proof. This is an alternative, direct proof that intervals are connected. Suppose U, V disconnect X. Then let $x \in U, y \in V$ such that $x < y$. Let $z = \sup U \cap [x, y]$. This set is non-empty since it contains x and is bounded above by y. So $z = [x, y] \subset X$. We will show $z \in U \cap V$, which is a contradiction. For all $n \in \mathbb{N}$, we have $z - \frac{1}{n} < n$ so there exists $x_n \in U \cap [x, y]$ which satisfies $z - \frac{1}{n} < x_n \leq z$. Hence $x_n \to z$. Also, $U = X \setminus V$ is closed, so $z \in U$. In particular, $z < y$. Now, choose $N \in \mathbb{N}$ such that $z + \frac{1}{N} < y$. Then for all $n \geq N$ we have $z < z + \frac{1}{n} < y$. Hence $z + \frac{1}{n} \in V$. However, $z + \frac{1}{n} \to z$, and V is closed, so $z \in V$, which is a contradiction. □

8.2. Consequences of definition

Example. Any indiscrete topological space is connected. Any cofinite topological space on an infinite set is connected. The discrete topological space on a set of size at least two is disconnected.

Lemma. Let Y be a subspace of a topological space X. Then, Y is disconnected if and only if there exist open subsets U, V of X such that $U \cap V \cap Y = \varnothing$ and $U \cup V \supset Y$, and $U \cap Y \neq \varnothing, V \cap Y \neq \varnothing$.

Proof. Suppose Y is disconnected. Then there exist open subsets U', V' of Y that disconnect Y. Then there exist open sets U, V in X such that $U' = U \cap Y$ and $V' = V \cap Y$. Then U, V satisfy the requirements from the lemma.

Conversely, suppose U, V are as given. Then, let $U' = U \cap Y, V' = V \cap Y$. They are open in Y by the definition of the subspace topology, and they disconnect Y. □

Remark. In the above lemma, we say subsets U, V of X disconnect Y.

Proposition. Let Y be a subspace of a topological space X. If Y is connected, then so is $\overline{Y}$.

Proof. Suppose $\overline{Y}$ is disconnected. Then there exist open sets U, V in X which disconnect $\overline{Y}$. Then $U \cap V \cap Y \subset U \cap V \cap \overline{Y} = \varnothing$ by definition. Hence $U \cap V \cap Y = \varnothing$. Also, $U \cup V \supset \overline{Y} \supset Y$. So U, V disconnect Y unless $U \cap Y = \varnothing$ or $V \cap Y = \varnothing$. But Y is connected, so without loss of generality let $V \cap Y = \varnothing$. Then $Y \subset X \setminus V$ and $X \setminus V$ is closed, so $\overline{Y} \subset X \setminus V$. Hence $V \cap \overline{Y} = \varnothing$. This is a contradiction since U, V disconnect $\overline{Y}$. □

Remark. More generally, if $Y \subset Z \subset \overline{Y}$, and Y is connected, then Z is connected. This is since $\mathrm{cl}_Z(Y) = \mathrm{cl}_X(Y) \cap Z = Z$.

Theorem. Let $f : X \to Y$ be a continuous function between topological spaces. If X is connected, then so is $f(X)$.

Proof. Let U, V be open subsets of Y which disconnect $f(X)$. For $x \in X$, $f(x) \in f(X) \subset U \cup V$. Hence, $f^{-1}(U) \cup f^{-1}(V) = X$. Also, if $x \in f^{-1}(U) \cap f^{-1}(V)$ then $f(x) \in U \cap V \cap f(X) = \varnothing$. This is a contradiction, so $f^{-1}(U) \cap f^{-1}(V) = \varnothing$. Since f is continuous, $f^{-1}(U), f^{-1}(V)$ are open in X. Since $U \cap f(X) \neq \varnothing$ and $V \cap f(X) \neq \varnothing$, $f^{-1}(U) \neq \varnothing$ and $f^{-1}(V) \neq \varnothing$ So $f^{-1}(U), f^{-1}(V)$ disconnect X. □

Remark. This shows that connectedness is a topological property. If X, Y are homeomorphic spaces, then X is connected if and only if Y is connected. Further, note that if $f : X \to Y$ is continuous and $A \subset X$ and A is connected, then $f(A)$ is connected. This can be shown by restricting f to the domain A.

Corollary. Any quotient of a connected topological space is connected.

Example. Let

$$Y = \left\{ \left(x, \sin \frac{1}{x} \right) : \ x > 0 \right\} \subset \mathbb{R}^2$$

This space is connected; the function $f : (0, \infty) \to \mathbb{R}^2$ defined by $f(x) = \left(x, \sin \frac{1}{x} \right)$ is continuous. So we have that $Y = \mathrm{Im}\, f$ is connected. Hence, $\overline{Y}$ is connected. We claim that

$$Z \equiv Y \cup \{(0, y) : \ y \in [-1, 1]\} = \overline{Y}$$

Indeed, given $y \in [-1, 1]$, for all $n \in \mathbb{N}$ we have that $(0, \frac{1}{n})$ is mapped to (n, ∞) by $x \to \frac{1}{x}$, so by the intermediate value theorem there exists $x_n \in \left(0, \frac{1}{n}\right)$ such that $\sin \frac{1}{x_n} = y$. Hence,

$$\left(x_n, \sin \frac{1}{x_n} \right) = (x_n, y) \to (0, y) \in \overline{Y}$$

So $Y \subset Z \subset \overline{Y}$. If we can show Z is closed, $Z = \overline{Y}$ since $\overline{Y}$ is the smallest closed superset of Y. Suppose $(x_n, y_n) \in Z$ for all $n \in \mathbb{N}$, and $(x_n, y_n) \to (x, y)$ in $\mathbb{R}^2$. Since $y_n \in [-1, 1]$ and $y_n \to y$, we have $y \in [-1, 1]$. If $x = 0$, we have $(x, y) \in Z$. If $x \neq 0$, then $x_n \to x$ implies $x_n \neq 0$ for all sufficiently large n. Hence $y_n = \sin \frac{1}{x_n}$ for all sufficiently large n. Thus

$$(x_n, y_n) \to \left(x, \sin \frac{1}{x} \right) \in Z$$

Lemma. Let X be a topological space and $\mathcal{A}$ be a family of connected subsets of X. Suppose that $A \cap B \neq \varnothing$ for all $A, B \in \mathcal{A}$. Then $\bigcup_{A \in \mathcal{A}} A$ is connected.

Proof. Let $Y = \bigcup_{A \in \mathcal{A}} A$, and let $f : Y \to \mathbb{Z}$ be a continuous function. We must show that f is constant. For all $A \in \mathcal{A}$, $f|_A : A \to \mathbb{Z}$ is continuous and hence constant, since A is connected. For all $A, B \in \mathcal{A}, A \cap B \neq \emptyset$ hence $f|_A$ and $f|_B$ are both constant and have the same value. So f must be constant, and hence Y is connected. □

Theorem. Let X, Y be connected topological spaces. Then $X \times Y$ is connected (in the product topology).

Proof. Without loss of generality, let $X \neq \emptyset, Y \neq \emptyset$. Let $x_0 \in X$. Consider the function $f : Y \to X \times Y$ defined by $f(y) = (x_0, y)$. The components of f are the functions $y \mapsto x_0$ which is continuous as it is constant, and $y \mapsto y$ which is continuous as it is the identity. So f is continuous. Then, the image of f, which is $\{x_0\} \times Y$, is connected. Similarly, for all $y \in Y$, $X \times \{y\}$ is connected. For $y \in Y$, $\{x_0\} \times Y \cap X \times \{y\} = \{(x_0, y)\} \neq \emptyset$. Hence, $A_y = \{x_0\} \times Y \cup X \times \{y\}$ is connected. For all $y, z \in Y$, $A_y \cap A_z \supset \{x_0\} \times Y$ hence $A_y \cap A_z \neq \emptyset$. Hence, $\bigcup_{y \in Y} A_y = X \times Y$ is connected. □

Example. $\mathbb{R}^n$ is connected for all $n \in \mathbb{N}$.

8.3. Partitioning into connected components

Definition. Let X be a topological space. We define a relation $\sim$ on X by $x \sim y$ if and only if there exists a connected subset A of X such that $x, y \in A$. For all $x \in X$, $x \sim x$ since $\{x\}$ is connected. Symmetry is clear from the definition. If $x \sim y$ and $y \sim z$ then by definition there exist connected subsets A, B in X such that $x, y \in A$ and $y, z \in B$. In particular, $A \cap B$ is not empty since $y \in A \cap B$. Hence $A \cup B$ is connected. Since $A \cup B$ contains x, z, we have $x \sim z$ as required for transitivity. Hence $\sim$ is an equivalence relation. For $x \in X$, we write C_x for the equivalence class containing x, called the *connected component* of x. The equivalence classes are called *connected components* of X.

Proposition. The connected components of a topological space X are non-empty, maximal connected subsets of X, they are closed, and they partition X.

Proof. Let C be a connected component of X. So $C = C_x$ for some $x \in X$. Then $x \in C$ hence $C \neq \emptyset$. Suppose $C \subset A \subset X$ and A is connected. Then for all $y \in A$, since $x, y \in A$ we must have $x \sim y$. So $y \in C$. Hence $A \subset C$, giving $A = C$. For all $y \in C$, we have $y \sim x$, so there exists a connected subset $A_y \subset X$ such that $x, y \in A_y$. Let $A = \bigcup_{y \in C} A_y$. A is connected since the union of pairwise intersecting connected sets are connected. Further $A \supset C$ so $A = C$ and C is connected. Since the closure of a connected set is connected, $\overline{C}$ is connected. But $\overline{C} \supset C$, so $C = \overline{C}$ is closed. □

8.4. Path-connectedness

Definition. Let X be a topological space. For points $x, y \in X$, a *path* from x to y in X is a continuous function $\gamma : [0,1] \to X$ such that $\gamma(0) = x, \gamma(1) = y$. We say that X is *path-connected* if for all $x, y \in X$, there exists a path from x to y in X.

Example. In $\mathbb{R}^n$, $\mathcal{D}_r(x)$ is path-connected by a straight line segment between any two points in the ball. In particular, let $\gamma(t) = (1-t)y + tz$. This is continuous and lies entirely inside $\mathcal{D}_r(x)$, since

$$\begin{aligned}\|\gamma(t) = x\| &= \|(1-t)t + tz - x\| \\ &= \|((1-t)y + tz) - ((1-t)x + tx)\| \\ &\leq (1-t)\|y - x\| + t\|z - x\| \\ &< r\end{aligned}$$

In a similar way, any convex subset of $\mathbb{R}^n$ is path-connected.

Theorem. If X is path-connected, X is connected.

Proof. Suppose X is not connected. Let U, V disconnect X. Let $x \in U, y \in V$, and suppose $\gamma : [0,1] \to X$ is continuous with $\gamma(0) = x$ and $\gamma(1) = y$. Then $\gamma^{-1}(U)$ and $\gamma^{-1}(V)$ disconnect $[0,1]$, which contradicts the connectedness of $[0,1]$. □

Example. The converse is false in general. Recall that the space

$$X = \left\{\left(x, \sin\frac{1}{x}\right) : x > 0\right\} \cup \{(0, y) : -1 \leq y \leq 1\}$$

is connected. We will show X is not path-connected. Suppose $\gamma : [0,1] \to X$ is continuous, and $\gamma(0) = (0,0)$ and $\gamma(1) = (1, \sin 1)$. Let $\gamma = (\gamma_1, \gamma_2)$, so γ_1, γ_2 are continuous functions. Suppose $t \in [0,1]$ such that $\gamma_1(t) > 0$. Then $\gamma_1((0,t)) \supset (0, \gamma_1(t))$ by the intermediate value theorem. In particular, there exists $n \in \mathbb{N}$ such that $\frac{1}{2\pi n} \in (0, \gamma_1(t))$. Hence, there exists $s < t$ such that $\gamma_1(s) = \frac{1}{2\pi n}$ so $\gamma_1(s) = 0$. Similarly, $\frac{1}{2\pi n + \frac{\pi}{2}} \in (0, \gamma_1(t))$ so there exists a different $s < t$ such that $\gamma_1(s) = \frac{1}{2\pi n + \frac{\pi}{2}}$ hence $\gamma_2(s) = 1$. In both cases, $\gamma_1(s) > 0$. We can inductively find a sequence $1 > t_1 > t_2 > \cdots > 0$ such that $\gamma_2(t_n)$ alternates between zero and one. But then $t_n \to t$ since it is a decreasing bounded-below sequence, and γ_2 is continuous, so $\gamma_2(t_n) \to \gamma_2(t)$ which is a contradiction.

8.5. Gluing lemma

Lemma. Let X be a topological space. Suppose $X = A \cup B$ where A, B are closed in X. Let $g : A \to Y$ and $h : B \to Y$ be continuous where Y is a topological space, such that for $A \cap B$, we have $g = h$. Then $f : X \to Y$ defined by

$$f(x) = \begin{cases} g(x) & x \in A \\ h(x) & x \in B \end{cases}$$

is well defined and continuous.

Proof. First, observe that if $F \subset A$ and F is closed in A, then there exists a closed set G in X such that $F = A \cap G$. Since A is closed in X, we must have F is closed in X. The same holds for $F \subset B$. Now, let V be a closed set in Y. Then the inverse image of V under f is

$$f^{-1}(V) = (f^{-1}(V) \cap A) \cup (f^{-1}(V) \cap B) = \underbrace{g^{-1}(V)}_{\text{closed in } A} \cup \underbrace{h^{-1}(V)}_{\text{closed in } B}$$

So $f^{-1}(V)$ is closed in X. To prove continuity it suffices to show that the preimage of a closed set is closed, since that implies that the preimage of an open set is open. □

Definition. Let X be a topological space. For $x, y \in X$, we write $x \sim y$ if there exists a path from x to y in X. This is an equivalence relation:

(i) The constant function shows that $x \sim x$ for all x.

(ii) If $\gamma : [0,1] \to X$ is continuous and $\gamma(0) = x$, $\gamma(1) = y$, we define $t \mapsto \gamma(1-t)$, which is a path from y to x.

(iii) Finally, if $x \sim y$ and $y \sim z$, we have continuous functions γ, δ such that $\gamma(0) = x, \gamma(1) = y = \delta(0), \delta(1) = z$. Then let

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

These intervals are closed on $[0,1]$ and their union is $[0,1]$. On the intersection, they are equal. By the gluing lemma, η is continuous, and now since $\eta(0) = x, \eta(1) = z$ we have $x \sim z$.

We call the equivalence classes *path-connected components* of X.

Theorem. Let U be an open subset of $\mathbb{R}^n$. Then U is connected if and only if U is path-connected.

Proof. The converse is trivial. Suppose U is connected. Without loss of generality, suppose $U \neq \varnothing$. Let $x_0 \in U$. Let $P = \{x \in U : x \sim x_0\}$ be the equivalence class of x_0. We want to show $P = U$. To do this, we will show that P is open and closed in U. Then, $P, U \setminus P$ will disconnect U unless $P = \varnothing$ or $P = U$. But we know $x_0 \in P$, hence $P = U$ will be the only possibility.

To show P is open, let $x \in U$. Since U is open, there exists $r > 0$ such that $\mathcal{D}_r(x) \subset U$. Recall that for all $y \in \mathcal{D}_r(x)$, we have $y \sim x$. Now, if $x \in P$, then we have $y \sim x$ and $x \sim x_0$ so $y \sim x_0$. So $\mathcal{D}_r(x) \subset P$. So P is open.

Now, if $x \in U \setminus P$ and $y \in \mathcal{D}_r(x)$ has $y \sim x_0$, then by transitivity $x \sim x_0$. But this is a contradiction since $x \notin P$. Hence $U \setminus P$ is open. So P is open and closed, so $P = U$. □

Theorem. For $n \geq 2$, $\mathbb{R}$ and $\mathbb{R}^n$ are not homeomorphic.

The generalisation $\mathbb{R}^m \not\simeq \mathbb{R}^n$ is true, but significantly harder to prove and outside the scope of this course.

Proof. Suppose $f : \mathbb{R} \to \mathbb{R}^n$ is a homeomorphism. Let $g = f^{-1}$. Then g is continuous. Then, $f|_{\mathbb{R}\setminus\{0\}}$ is a homeomorphism from $\mathbb{R} \setminus \{0\}$ to $\mathbb{R}^n \setminus \{f(0)\}$, with inverse $g|_{\mathbb{R}^n\setminus\{f(0)\}}$. But $\mathbb{R} \setminus \{0\}$ is disconnected, but $\mathbb{R}^n \setminus \{f(0)\}$ is connected since it is path-connected. This is a contradiction. □

9. Compactness

9.1. Motivation and definition

Recall from IA Analysis that a continuous function on a closed bounded interval is bounded and attains its bounds. We wish to generalise this result to more general topological spaces.

Example. (i) If X is finite, any function $X \to \mathbb{R}$ is finite.

(ii) If, for all continuous functions $f : X \to \mathbb{R}$ there exists $n \in \mathbb{N}$ and subsets $A_1, \dots, A_n$ of X such that $X = \bigcup_{j=1}^{n} A_j$ and f is bounded on A_j for all j, then the property holds.

(iii) Note that continuous functions are 'locally bounded'; if $f : X \to \mathbb{R}$ is continuous, then for all $x \in X$ we have $U_x = f^{-1}((f(x)-1, f(x)+1))$ is an open set containing X, and f is bounded on U_x. So each point has an open neighbourhood on which f is bounded. Further, $X = \bigcup_{x \in X} U_x$. If there exists a finite subset $F \subset X$ such that $\bigcup_{x \in F} U_x = X$, then f is bounded on X. This is exactly the definition we will use for compactness.

Definition. Let X be a topological space. An *open cover* for X is a family $\mathcal{U}$ of open subsets of X that cover X; that is, $\bigcup_{U \in \mathcal{U}} U = X$. A *subcover* of $\mathcal{U}$ is a subset $\mathcal{V} \subset \mathcal{U}$ that covers U. This is called a *finite subcover* if $\mathcal{V}$ is finite. We say that X is *compact* if every open cover has a finite subcover.

Remark. Compactness can be thought of as the next best thing to finiteness.

Theorem. Let X be a compact topological space and $f : X \to \mathbb{R}$ be continuous. Then f is bounded, and if X is not empty f attains its bounds.

Proof. For $n \in \mathbb{N}$, let $U_n = \{x \in X : |f(x)| < n\}$. U_n is open since $x \mapsto |f(x)|$ is continuous and $(-n, n)$ is open. It is clear that $X = \bigcup_{n \in \mathbb{N}} U_n$. This is an open cover of X. Hence there exists a finite subcover $F \subset \mathbb{N}$ such that $X = \bigcup_{n \in F} U_n = U_N$ where $N = \max F$. Hence, for all $x \in X$, we have $|f(x)| < N$ so f is bounded.

Let $\alpha = \inf_X f$; this exists since f is bounded. Suppose there exists no $x \in X$ such that $f(x) = \alpha$. Then, for all $x \in X$, $f(x) > \alpha$. Then there exists $n \in \mathbb{N}$ such that $f(x) > \alpha + \frac{1}{n}$. So let

$$V_n = \left\{x \in X : f(x) > \alpha + \frac{1}{n}\right\} = f^{-1}\left(\left(\alpha + \frac{1}{n}, \infty\right)\right)$$

We can see that V_n is open. Now, since $\bigcup_{n \in \mathbb{N}} V_n = X$, there exists a finite subcover $F \subset \mathbb{N}$ such that $\bigcup_{n \in F} V_n = X = V_N$ where N is the maximal F. Then for all $x \in X$, we have $f(x) > \alpha + \frac{1}{N}$. Hence $\inf_X f \geq \alpha + \frac{1}{N}$, which is a contradiction. The same argument applies for the supremum. □

Lemma. Let Y be a subspace of a topological space X. Then Y is compact if and only if whenever $\mathcal{U}$ is a family of open sets in X such that $\bigcup_{U \in \mathcal{U}} \supset Y$, there is a finite subfamily $\mathcal{V} \subset \mathcal{U}$ with $\bigcup_{U \in \mathcal{V}} U \supset Y$.

Theorem. $[0, 1]$ is compact.

Proof. Let $\mathcal{U}$ be a family of open sets in $\mathbb{R}$ that cover $[0,1]$. For a subset $A \subset [0,1]$, we say that $\mathcal{U}$ *finitely covers* A if there exists a finite subcover $\mathcal{V} \subset \mathcal{U}$ of A. Note that if $A = B \cup C$ and $A, B, C \subset [0,1]$ and $\mathcal{U}$ finitely covers B and C, we can take the union of the finite subcovers to find a finite subcover of A, so U finitely covers A. Suppose that $\mathcal{U}$ does not finitely cover $[0,1]$. Then one of the intervals $\left[0, \frac{1}{2}\right]$ and $\left[\frac{1}{2}, 1\right]$ is not finitely coverable by $\mathcal{U}$. Let this interval be $[a_1, b_1]$. Let $c = \frac{1}{2}(a_1 + b_1)$. Then one of the intervals $[a_1, c], [c, b_1]$ is not finitely coverable by $\mathcal{U}$. Inductively, we obtain a nested sequence of intervals $[a_1, b_1] \supset \cdots \supset [a_n, b_n] \supset \cdots$ which are not finitely covered by $\mathcal{U}$ and $b_n - a_n = 2^{-n}$. Now, $a_n \to x$ for some $x \in [0,1]$ and $b_n = a^n + 2^{-n} \to x$. But since $\mathcal{U}$ covers $[0,1]$, there exists $U \in \mathcal{U}$ such that $x \in U$. U is open in $\mathbb{R}$, so for all $\varepsilon > 0$, we have $(x - \varepsilon, x + \varepsilon) \subset U$. Since $a_n, b_n \to x$, we can choose n such that $a_n, b_n \in (x - \varepsilon, x + \varepsilon)$. This is covered by one open set U in $\mathcal{U}$, so this is a finite subcover. This is a contradiction. □

Example. Other examples of compact spaces include the following.

(i) Any finite set is compact.

(ii) On any set X, the cofinite topology is compact. Suppose without loss of generality that X is not empty, and let $\mathcal{U}$ be an open cover for X. Let $U \in \mathcal{U}$ such that $U \neq \varnothing$. Then $F = X \setminus U$ is finite. For all $x \in F$, let $U_x \in \mathcal{U}$ such that $x \in U_x$. Then $\bigcup_{x \in F} U_x \cup U$ is a finite subcover.

(iii) Let $x_n \to x$ in a topological space X. Let $Y = \{x_n : n \in \mathbb{N}\} \cup \{x\}$. Then Y is compact. Indeed, let $\mathcal{U}$ be a family of open sets in X such that $\bigcup_{U \in \mathcal{U}} U \supset Y$. In particular, let $U \in \mathcal{U}$ such that $x \in U$. Since U is open and $x_n \to x$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ we have $x_n \in U$. So we can cover the remaining finitely many elements analogously to the previous example, and this yields a finite subcover.

(iv) The indiscrete topology on any set is compact, since there are only two open sets.

Counterexamples include the following.

(i) An infinite set X in the discrete topology is not compact. Let

$$\mathcal{U} = \{\{x\} : x \in X\}$$

This has no finite subcover.

(ii) $\mathbb{R}$ is not compact. Consider the intervals $(-n, n)$ for all $n \in \mathbb{N}$. This is an open cover with no finite subcover.

9.2. Subspaces

Theorem. Let Y be a subspace of a topological space X. Then,

(i) Let X be compact and Y be closed in X. Then Y is compact.

(ii) Let X be Hausdorff and Y be compact. Then Y is closed in X.

Proof. Let $\mathcal{U}$ be a family of open sets in X such that their union covers Y. Then $\mathcal{U} \cup (X \setminus Y)$ is an open cover for X since Y is closed. This has a finite subcover $\mathcal{V} \subset \mathcal{U}$ such that $\bigcup_{U \in \mathcal{V}} U \cup (X \setminus Y) = X$. Then $\bigcup_{U \in \mathcal{V}} U \supset Y$.

For part (ii), let $x \in X \setminus Y$. For $y \in Y$, since $x \neq y$ there exist open sets U_y, V_y in X such that $x \in U_y, y \in V_y, U_y \cap V_y = \varnothing$. Now, $\{V_y : y \in Y\}$ is an open cover of Y. Hence there exists $F \subset Y$ finite such that $\bigcup_{y \in F} V_y \supset Y$. Now, $U = \bigcap_{y \in F} U_y$ is open, further $x \in U$ and

$$U \cap Y \subset \left(\bigcap_{y \in F} U_y \right) \cap \left(\bigcup_{y \in F} V_y \right) = \varnothing$$

Hence $X \setminus Y$ is a neighbourhood of all of its points, so it is open and Y is closed. □

9.3. Continuous images of compact spaces

Theorem. Let $f : X \to Y$ be a continuous function between topological spaces such that X is compact. Then $f(X)$ is compact.

Proof. Let $\mathcal{U}$ be a family of open sets in Y such that $\bigcup_{U \in \mathcal{U}} U \supset f(X)$. Then $\bigcup_{U \in \mathcal{U}} f^{-1}(U) = X$ and $f^{-1}(U)$ is open in X for all $U \in \mathcal{U}$ since f is continuous. Since X is compact, we have a finite subcover $\mathcal{V} \subset \mathcal{U}$ such that $X = \bigcup_{U \in \mathcal{V}} f^{-1}(V)$. Hence $f(X) \subset \bigcup_{U \in \mathcal{V}} U$. □

Remark. Compactness is a topological property. If $f : X \to Y$ is continuous and $A \subset X$ is compact, then $f(A)$ is compact.

Corollary. Any quotient of a compact space is compact.

Example. Let $a < b \in \mathbb{R}$. Then $[a, b] \simeq [0, 1]$ so is compact.

9.4. Topological inverse function theorem

Theorem. Let $f : X \to Y$ be a continuous bijection from a compact space X to a Hausdorff space Y. Then f^{-1} is continuous, so f is an open map. Hence f is a homeomorphism.

Proof. Let U be an open subset of X. Then $K = X \setminus U$ is closed. Since X is compact, K is compact. Further, $f(K)$ is compact. Hence $f(K)$ is closed in Y. So $f(U) = Y \setminus f(K)$ is open in Y. □

Example. $\mathbb{R}/\mathbb{Z}$ is homeomorphic to $S^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$. Indeed, let $f : \mathbb{R} \to S^1$ by $f(t) = (\cos(2\pi t), \sin(2\pi t))$. For all s, t, we have $f(s) = f(t)$ if and only if $s \sim t$ so f fully respects $\sim$. f is continuous and surjective. Let $\widetilde{f} : \mathbb{R}/\mathbb{Z} \to S^1$ be the unique map such that $\widetilde{f} \circ q = f$. So $\widetilde{f}$ is a continuous bijection. S^1 is Hausdorff, and $\mathbb{R}/\mathbb{Z}$ is the image of $[0, 1]$ under a continuous map, hence is compact. Hence $\widetilde{f}$ is a homeomorphism.

9.5. Tychonov's theorem

Theorem. Let X, Y be compact topological spaces. Then $X \times Y$ is compact in the product topology.

Proof. Let $\mathcal{U}$ be an open cover for $X \times Y$. We want to show that there exists a finite subcover. Without loss of generality, every member of $\mathcal{U}$ can be of the form $U \times V$ where U is open in X and V is open in Y. Indeed, for $z \in X \times Y$ we can choose $W_z \in \mathcal{U}$ such that $z \in W_z$. By definition of the product topology, there exist open sets U_z in X and V_z in Y such that $z \in U_z \times V_z \subset W_z$. So $\{U_z \times V_z : z \in X \times Y\}$ is an open cover for $X \times Y$. If there exists a finite subset $F \subset X \times Y$ such that $\bigcup_{z \in F} U_z \times V_z$ covers $X \times Y$, then $\{W_x : z \in F\}$ is a finite subcover of $\mathcal{U}$.

Let $x \in X$. Recall that $\{x\} \times Y$ is the continuous image of Y under the map $y \mapsto (x, y)$. Hence, $\{x\} \times Y$ is compact, since the continuous image of a compact space is compact. Since $\{x\} \times Y$ is covered by $\bigcup_{W \in \mathcal{U}} W$, $\mathcal{U}$ finitely covers $\{x\} \times Y$. So there exists $n_x \in \mathbb{N}$ such that we can find open sets $U_{x,1}, \dots, U_{x,n_x}$ in X and $V_{x,1}, \dots, V_{x,n_x}$ in Y such that $U_{x,j} \times V_{x,j} \in \mathcal{U}$ and $\{x\} \times Y \subset \bigcup_{j=1}^{n_x} U_{x,j} \times V_{x,j}$.

Without loss of generality, let $x \in U_{x,j}$ for all j, since any other $U_{x,j}$ is not needed in the cover. Now let $U_x = \bigcap_{j=1}^{n_x} U_{x,j}$. We know $x \in U_x$ and U_x is open since it is a finite intersection of open sets. In particular, $U_x \times Y \subset \bigcup_{j=1}^{n_x} U_{x,j} \times V_{x,j}$.

Now, $\{U_x : x \in X\}$ is an open cover for X. So there exists a finite subset $F \subset X$ such that $X = \bigcup_{x \in F} U_x$. Then, $X \times Y = \bigcup_{x \in F} U_x \times Y \subset \bigcup_{x \in F} \bigcup_{j=1}^{n_x} U_{x,j} \times V_{x,j}$. Hence,

$$\left\{U_{x,j} \times V_{x,j} : x \in F, 1 \leq j \leq n_x\right\}$$

is a finite subcover of $\mathcal{U}$. □

Remark. More generally, if $X_1, \dots, X_n$ are compact spaces, then so is $X_1 \times \cdots \times X_n$.

9.6. Heine–Borel theorem

Theorem. A subset K of $\mathbb{R}^n$ is compact if and only if K is closed and bounded.

Proof. Suppose K is compact. $\mathbb{R}^n$ is a metric space and hence Hausdorff. Hence, K is closed in $\mathbb{R}^n$. The function $x \mapsto \|x\|$ is continuous. Therefore, it is bounded on K. So K is bounded.

Conversely, if K is bounded, there exists $M \geq 0$ such that for all $x \in K$ we have $\|x\| \leq M$. Hence, $K \subset [-M, M]^n$. Note that $[-M, M]$ is compact since it is homeomorphic to $[0, 1]$. By Tychonov's theorem, $[-M, M]^n$ is compact in the product topology. Since a closed subset of a compact space is compact, K is compact. □

Example. Closed balls $\mathcal{B}_r(x)$ in $\mathbb{R}^n$ are compact. The start of the proof for the Lindelöf–Picard theorem now makes more sense.

9.7. Sequential compactness

Definition. A topological space X is *sequentially compact* if every sequence in X has a convergent subsequence. Given a sequence (x_n) and an infinite set $M \subset \mathbb{N}$, we will write $(x_m)_{m\in M}$ for the subsequence $(x_{m_n})_{n=1}^{\infty}$ where $m_1 < m_2 < \dots$ are the elements of M. Note that if $L \subset M \subset \mathbb{N}$, then $(x_n)_{n\in L}$ is a subsequence of $(x_n)_{n\in M}$.

Example. Any closed and bounded subset of $\mathbb{R}$ is sequentially compact by the Bolzano–Weierstrass theorem. Similarly, any closed and bounded subset K of $\mathbb{R}^n$ is sequentially compact. Indeed, let (x_m) be a sequence in K. Then, writing $x_m = (x_{m,1}, \dots, x_{m,n})$, since K is bounded we have that $(x_{m,j})$ is bounded for all j. Applying the Bolzano–Weierstrass theorem to the first coordinate, we find $M_1 \subset \mathbb{N}$ such that $(x_{m,1})_{m\in M_1}$ converges in $\mathbb{R}$. Now, $(x_{m,2})_{m\in M_1}$ is bounded in $\mathbb{R}$, so again applying the Bolzano–Weierstrass theorem, we can find $M_2 \subset \mathbb{N}$ such that $(x_{m,2})_{m\in M_2}$ converges. Note that $(x_{m,1})_{m\in M_2}$ converges. So inductively we can find $M_1 \supset \cdots \supset M_n$ such that $(x_{m,j})_{m\in M_n}$ converges for all j. Hence $(x_m)_{m\in M_n}$ converges in $\mathbb{R}^n$. The limit is contained in K since K is closed.

Remark. In $\mathbb{R}^n$, any compact space is sequentially compact. The converse is also true; any sequentially compact subspace must be closed and bounded. We aim to show that compactness and sequential compactness are identical in metric spaces.

9.8. Compactness and sequential compactness in metric spaces

Let (M, d) be a metric space.

Definition. For $\varepsilon > 0$ and $F \subset M$, we say that F is an *ε-net for M* if for all $x \in M$, there exists $y \in F$ such that $d(y, x) \le \varepsilon$. Equivalently, $M = \bigcup_{y\in M} \mathcal{B}_\varepsilon(y)$. This is called a *finite ε-net* if F is finite. We say that M is *totally bounded* if for all $\varepsilon > 0$, there exists a finite ε-net for M.

Example. For $\varepsilon > 0$, let n such that $\frac{1}{n} < \varepsilon$. Then $\left\{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}\right\}$ is an ε-net for $(0, 1)$.

Definition. For a non-empty $A \subset M$, the *diameter* of A is $\operatorname{diam} A = \sup\{d(x, y) : x, y \in A\}$. This is finite if and only if A is a bounded set.

Example. $\operatorname{diam} \mathcal{B}_r(x) \le 2r$.

Lemma. Suppose M is totally bounded. Let A be a non-empty closed subset of M. Let $\varepsilon > 0$. Then there exists $K \in \mathbb{N}$ and non-empty closed sets $B_1, \dots, B_K$ such that $A = \bigcup_{k=1}^{K} B_k$ and $\operatorname{diam} B_k \le \varepsilon$ for all k.

Proof. Let F be a finite $\frac{\varepsilon}{2}$-net for M. So $M = \bigcup_{x\in F} B_{\varepsilon/2}(x)$ and hence $A = \bigcup_{x\in F}(A \cap B_{\varepsilon/2}(x))$. Let $G = \left\{x \in F : A \cap B_{\varepsilon/2}(x) \neq 0\right\}$. Then for $x \in G$ let $B_x = A \cap B_{\varepsilon/2}(x)$. So for $x \in G$, we have $B_x \neq \varnothing$, $B_x \subset B_{\varepsilon/2}(x)$ and so $\operatorname{diam} B_x \le \varepsilon$, and B_x is closed. Then $A = \bigcup_{x\in G} B_x$. □

Theorem. For a metric space (M, d), the following are equivalent.

(i) M is compact;

(ii) M is sequentially compact;

(iii) M is complete and totally bounded.

Proof. We first show (i) implies (ii). Let (x_n) be a sequence in M. Then for $n \in \mathbb{N}$, let $T_n = \{x_k : k > n\}$ be the tail of the sequence. Note that the limit of any convergent subsequence (if it exists) is in the intersection of $\bigcap_{n\in\mathbb{N}} \overline{T}_n$. So first, we prove that this intersection is non-empty. Suppose that it is empty. Then, $\bigcup_{n\in\mathbb{N}} \left(M \setminus \overline{T}_n\right) = M$. But the $M \setminus \overline{T}_n$ are open, and M is compact, there is a finite subcover. So $M \setminus \overline{T}_N = M$ for some N, since the T_n are a decreasing sequence of sets. This is a contradiction since $T_N \neq \varnothing$. Now, let $x \in \bigcap_{n\in\mathbb{N}} \overline{T}_n$, and we want to show the existence of a subsequence converging to x. First, $x \in \overline{T}_1$, so $\mathcal{D}_1(x) \cap T_1 \neq \varnothing$. Hence there exists $k_1 > 1$ such that $d(x_{k_1}, x) < 1$. Now since $x \in \overline{T}_{k_1}$, $\mathcal{D}_{1/2}(x) \cap T_{k_1} \neq \varnothing$. There exists $k_2 > k_1$ such that $d(x_{k_2}, x) < \frac{1}{2}$. Inductively, we can find a strictly increasing sequence $k_1 < k_2 < \ldots$ such that $d(x_{k_n}, x) < \frac{1}{n}$ for all n, so $x_{k_n} \to x$.

Now, we show (ii) implies (iii). To show M is complete, let (x_n) be a Cauchy sequence in M. Let $k_1 < k_2 < \ldots$ such that x_{k_n} converges in M, and let x be the limit. We show $x_n \to x$. Indeed, for $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $\forall m, n \geq N$, we have $d(x_m, x_n) < \varepsilon$. Then $\forall m \geq N$, we have $k_n \geq m \geq N$, so for a fixed $n \geq N$ and $\forall m \geq N$, we have $d(x_n, x) \leq d(x_n, x_{k_m}) + d(x_{k_m}, x) \leq \varepsilon + d(x_{k_m}, x)$. Let $m \to \infty$, so $d(x_n, x) \leq \varepsilon$. So $x_n \to x$. To show M is totally bounded, suppose it is not. There exists $\varepsilon > 0$ such that M has no finite ε-net. Let $x_1 \in M$, and suppose we can find $x_1, \ldots, x_{n-1}$ in M. Then $\bigcup_{j=1}^{n-1} \mathcal{B}_\varepsilon(x_j) \neq M$. So we can pick $x_n \in M \setminus \bigcup_{j-1}^{n-1} \mathcal{B}_\varepsilon(x_j)$. Inductively we obtain (x_n) such that $d(x_m, x_n) > \varepsilon$ for all $n, m \in \mathbb{N}$. So (x_n) has no Cauchy subsequence. There is therefore no convergent subsequence, which is a contradiction.

Finally, we show (iii) implies (i). Let $\mathcal{U}$ be an open cover for M. We must show there exists a finite subcover. Suppose that is not true, so $\mathcal{U}$ does not finitely cover M. We construct non-empty closed subsets $A_0 \supset A_1 \supset \ldots$ of M such that for all $n \geq 0$, $\mathcal{U}$ does not finitely cover A_n, and for all $n \geq 1$ we have $\operatorname{diam} A_n < \frac{1}{n}$. Let $A_0 = M$. Suppose that for some $n \geq 1$ we have already found A_{n-1}. Since M is totally bounded, we can write $A_{n-1} = \bigcup_{k=1}^{K} B_k$ where $K \in \mathbb{N}$ and the B_k are non-empty, closed, and $\operatorname{diam} B_k < \frac{1}{n}$. Since $\mathcal{U}$ does not finitely cover A_{n-1}, there exists $k \leq K$ such that $\mathcal{U}$ does not finitely cover B_k. Let A_n be this B_k. Now, for all n, pick some $x_n \in A_n$. For all N, $\forall m, n \geq N$ we have $x_m, x_n \in A_N$ hence $d(x_m, x_n) \leq \operatorname{diam} A_N \leq \frac{1}{n}$ so the sequence is Cauchy. M is complete, so $x_n \to x$ for some $x \in M$. Let $U \in \mathcal{U}$ such that $x \in U$. U is open, so there exists $r > 0$ such that $\mathcal{D}_r(x) \subset U$. But $x_n \to x$ hence there exists n such that $d(x_n, x) < \frac{r}{2}$ and $\operatorname{diam} A_n < \frac{r}{2}$. For every $y \in A_n$, $d(y, x) \leq d(y, x_n) + d(x_n, x) \leq \operatorname{diam} A_n + \frac{r}{2} < r$. Hence every point in A_n is contained within $\mathcal{D}_r(x) \subset U$. But this contradicts the fact that $\mathcal{U}$ does not finitely cover A_n, but we have constructed a cover using just one open set. □

Remark. We can now deduce the one direction of the Heine–Borel theorem from the Bolzano–Weierstrass theorem; closed and bounded subsets of $\mathbb{R}^n$ are compact. Similarly, we can check that the product of sequentially compact topological spaces is sequentially compact in the product topology. This yields a new proof for Tychonov's theorem for metric spaces. In general, there exist topological spaces that are compact but not sequentially compact, and conversely there exist topological spaces which are sequentially compact but not compact.

10. Differentiation

10.1. Linear maps

Let $m, n \in \mathbb{N}$. Recall that $L(\mathbb{R}^m, \mathbb{R}^n)$ is the vector space of linear maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. This is isomorphic to $M_{n,m}$, the space of $n \times m$ real matrices. There is also an isomorphism to $\mathbb{R}^{mn}$. Let $e_1, \dots, e_m$ be the standard basis of $\mathbb{R}^m$, and similarly let $e'_1, \dots, e'_n$ be the standard basis of $\mathbb{R}^n$. Then $T \in L(\mathbb{R}^m, \mathbb{R}^n)$ is identified with the $n \times m$ matrix (T_{ji}) where $1 \le j \le n$ and $1 \le i \le m$, such that $T_{ji} = \left\langle Te_i, e'_j \right\rangle$. We can therefore view $L(\mathbb{R}^m, \mathbb{R}^n)$ as the mn-dimensional vector space $\mathbb{R}^{mn}$ with the Euclidean norm. So the norm of a linear map T is given by

$$\|T\| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} T_{ji}^2} = \sqrt{\sum_{i=1}^{m} \|Te_i\|^2}$$

where Te_i is the ith column of T. Thus, $L(\mathbb{R}^m, \mathbb{R}^n)$ becomes a metric space together with the Euclidean distance $d(S, T) = \|S - T\|$.

Lemma. For $T \in L(\mathbb{R}^m, \mathbb{R}^n)$ and $x \in \mathbb{R}^m$,

$$\|Tx\| \le \|T\| \cdot \|x\|$$

So T is a Lipschitz map and hence continuous. Further, if $S \in L(\mathbb{R}^n, \mathbb{R}^p)$ then

$$\|ST\| \le \|S\| \cdot \|T\|$$

Proof. We can write

$$x = \sum_{i=1}^{m} x_i e_i$$

Hence,

$$Tx = \sum_{i=1}^{m} x_i Te_i$$

Thus,

$$\|Tx\| \le \sum_{i=1}^{m} |x_i| \|Te_i\| \le \left(\sum_{i=1}^{m} x_i^2 \right)^{1/2} \cdot \left(\sum_{i=1}^{m} \|Te_i\|^2 \right)^{1/2} = \|T\| \cdot \|x\|$$

Further, for $x, y \in \mathbb{R}^m$ we have

$$d(Tx, Ty) = \|Tx - Ty\| = \|T(x - y)\| \le \|T\| \cdot \|x - y\| = \|T\| d(x, y)$$

So T is Lipschitz, and any Lipschitz function is continuous. Now,

$$\|ST\| = \left(\sum_{i=1}^{m} \|STe_i\|^2 \right)^{1/2} \le \left(\sum_{i=1}^{m} \|S\| \|Te_i\|^2 \right)^{1/2} = \|S\| \left(\sum_{i=1}^{m} \|Te_i\|^2 \right)^{1/2} = \|S\| \cdot \|T\|$$

□

10.2. Differentiation

Recall from IA Analysis that a function $f : \mathbb{R} \to \mathbb{R}$ is *differentiable* at a point $a \in \mathbb{R}$ if

$$\lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

exists. The value of this limit is called the *derivative* of f at a, and denoted $f'(a)$. Note that f is differentiable at a if and only if there exists $\lambda \in \mathbb{R}$ and $\varepsilon : \mathbb{R} \to \mathbb{R}$ such that $\varepsilon(0) = 0$ and ε is continuous at 0, and

$$f(a+h) = f(a) + \lambda h + h\varepsilon(h)$$

This is because we can define

$$\varepsilon(h) = \begin{cases} 0 & h = 0 \\ \frac{f(a+h)-f(a)}{h} - \lambda & h \neq 0 \end{cases}$$

Informally, this ε definition states that f is approximated very well (the error $h\varepsilon(h)$ shrinks rapidly since $\varepsilon \to 0$) by a linear function in a small neighbourhood of a. Recall that if f is n times differentiable at a, then

$$f(a+h) = f(a) + \sum_{k=1}^{n} \frac{f^{(k)}(a)}{k!} h^k + o(h^n)$$

Definition. Let $m, n \in \mathbb{N}$. Then $f : \mathbb{R}^m \to \mathbb{R}^n$ and $a \in \mathbb{R}^m$. We say that f is *differentiable* at a if there exists a linear map $T \in L(\mathbb{R}^m, \mathbb{R}^n)$ and a function $\varepsilon : \mathbb{R}^m \to \mathbb{R}^n$ such that $\varepsilon(0) = 0$ and ε is continuous at 0, and

$$f(a+h) = f(a) + T(h) + \|h\|\varepsilon(h)$$

Note that

$$\varepsilon(h) = \begin{cases} 0 & h = 0 \\ \frac{f(a+h)-f(a)-T(h)}{\|h\|} & h \neq 0 \end{cases}$$

So f is differentiable at a if and only if there exists $T \in L(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$\frac{f(a+h) - f(a) - T(h)}{\|h\|} \to 0$$

as $h \to 0$. Such a T is unique. Indeed, suppose S, T satisfy the above limit. Then, by subtracting,

$$\frac{S(h) - T(h)}{\|h\|} \to 0$$

For a fixed $x \in \mathbb{R}^m$, $x \neq 0$, we have $\frac{x}{k} \to 0$ as $k \to \infty$ so

$$\frac{S\left(\frac{x}{k}\right) - T\left(\frac{x}{k}\right)}{\left\|\frac{x}{k}\right\|} \to 0 \implies \frac{S(x) - T(x)}{\|x\|} = 0$$

So $Sx = Tx$. It follows that $S = T$. We say that if a function f is differentiable at a point a, T is the unique *derivative* of f at a. This is denoted $f'(a) = Df(a) = Df|_a$. If $f : \mathbb{R}^m \to \mathbb{R}^n$ is differentiable at $a \in \mathbb{R}^m$ for every a, we say that f is *differentiable on* $\mathbb{R}^m$. The function $f' = D : \mathbb{R}^m \to L(\mathbb{R}^m, \mathbb{R}^n)$ mapping $a \mapsto f'(a)$ is the derivative of f.

Example. Constant functions are differentiable. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ such that $f(x) = b$ for $b \in \mathbb{R}^n$. Then for all $a \in \mathbb{R}^m$, we have

$$f(a + h) = f(a) + 0h + 0$$

so f is differentiable at a and the derivative is zero.

Example. Linear maps are differentiable. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be defined by $f(x) = Tx$ for a linear map $T \in L(\mathbb{R}^m, \mathbb{R}^n)$. Then

$$f(a + h) = f(a) + f(h) + 0$$

so f is differentiable at a with derivative $T = f$. So f' is a constant function.

Example. Consider

$$f(x) = \|x\|^2$$

For $a \in \mathbb{R}^m$, we can find

$$f(a + h) = \|a + h\|^2 = \|a\|^2 + 2\langle a, h\rangle + \|h\|^2 = f(a) + 2\langle a, h\rangle + \|h\|\varepsilon(h)$$

Hence, f is differentiable with derivative

$$f'(a)(h) = 2\langle a, h\rangle$$

Note that $f' : \mathbb{R}^m \to L(\mathbb{R}^m \to \mathbb{R})$ is linear.

Example. Note $M_n \simeq \mathbb{R}^{n^2}$. The function $f : M_n \to M_n$ given by $f(A) = A^2$. For a fixed $A \in M_n$,

$$f(A + H) = (A + H)^2 = A^2 + AH + HA + H^2$$

It suffices to show H^2 is $o(\|H\|)$. We have $\left\|H^2\right\| \leq \|H\|^2$, hence

$$\frac{\left\|H^2\right\|}{\|H\|} \leq \|H\| \to 0$$

So f is differentiable at A and the derivative is given by

$$f'(A)(H) = AH + HA$$

Example. Suppose $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p$ is bilinear. Let $(a, b) \in \mathbb{R}^m \times \mathbb{R}^n$. Then,

$$f((a, b) + (h, k)) = f((a + h, b + k)) = f(a, b) + f(a, k) + f(h, b) + f(h, k)$$

The map $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p$ given by $(h,k) \mapsto f(a,k) + f(h,b)$ is linear as the sum of two linear maps. So it suffices to show $f(h,k)$ is $o(\|(h,k)\|)$.

$$h = \sum_{i=1}^{m} h_i e_i; \quad k = \sum_{j=1}^{n} k_j e'_j$$

Hence,

$$f(h,k) = \sum_{i=1}^{m}\sum_{j=1}^{n} h_i k_j f(e_i, e'_j) \implies \|f(h,k)\| \leq \sum_{i=1}^{m}\sum_{j=1}^{n} |h_i| \cdot |k_j| \cdot \left\|f(e_i, e'_j)\right\| \leq C\|(h,k)\|^2$$

for some constant C, since $|h_i| \leq \|(h,k)\|^2$ and similarly for $|k_j|$. So

$$\frac{\|f(h,k)\|}{\|(h,k)\|} \leq C\|(h,k)\| \to 0$$

Hence f is differentiable with

$$f'(a,b)(h,k) = f(a,k) + f(h,b)$$

10.3. Derivatives on open subsets

We may define the derivative on a subset of $\mathbb{R}^m$. We will use the notion of open subsets since we are typically interested in neigbourhoods of points.

Definition. Let U be an open subset of $\mathbb{R}^m$. Let $f : U \to \mathbb{R}^n$ be a function, and $a \in U$. Then we say f is *differentiable* at a if there exists a linear map $T \in L(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$f(a+h) = f(a) + T(h) + \|h\|\varepsilon(h)$$

where $\varepsilon(0) = 0$ and ε is continuous at zero. Note that ε need only be defined on the set of h such that $a + h \in U$, or more precisely the open set $U - a$. Hence there exists $r > 0$ such that $\mathcal{D}_r(0) \subset U_a$. Then

$$\varepsilon(h) = \begin{cases} 0 & h = 0 \\ \frac{f(a+h)-f(a)-T(h)}{\|h\|} & h \neq 0, a+h \in U \end{cases}$$

So f is differentiable at a if and only if there exists a linear map $T \in L(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$\frac{f(a+h) - f(a) - T(h)}{\|h\|} \to 0$$

Remark. The linear map T is unique, and is called the *derivative* of f at a, denoted $f'(a)$. In particular,

$$f(a+h) = f(a) + f'(a)(h) + o(\|h\|)$$

Remark. If $m = 1$, the space $L(\mathbb{R}, \mathbb{R}^n)$ is isomorphic to $\mathbb{R}^n$. The linear map is defined uniquely by a vector in $\mathbb{R}^n$ which multiplies by the scalar h. Hence, if $U \subset \mathbb{R}$ is open and $f : U \to \mathbb{R}$ be a function and $a \in U$, then f is differentiable at a if there exists a vector $v \in \mathbb{R}^n$ such that

$$\frac{f(a+h) - f(a) - hv}{|v|} \to 0$$

Equivalently, there exists $v \in \mathbb{R}^n$ such that

$$\frac{f(a+h) - f(a)}{h} \to v$$

10.4. Properties of derivative

Proposition. Let $U \subset \mathbb{R}^m$ be open, $f : U \to \mathbb{R}^n$ be a function, and $a \in U$. If f is differentiable at a, f is continuous at a.

Proof. We have

$$f(a+h) = f(a) + f'(a)(h) + \|h\|\varepsilon(h)$$

Hence,

$$f(x) = f(a) + f'(a)(x-a) + \|x-a\|\varepsilon(x-a)$$

The functions $x \mapsto f(a)$, $x \mapsto f'(a)(x-a)$ and $x \mapsto \|x-a\|\varepsilon(x-a)$ are all continuous at a. Hence their sum is continuous. □

Proposition (chain rule)**.** Let $U \subset \mathbb{R}^m$ and $V \subset \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^n$ and $g : V \to \mathbb{R}^p$ be functions, and $a \in U, b \equiv f(a) \in V$. Suppose f is differentiable at a, and g is differentiable at b. Then $g \circ f$ is differentiable at a and

$$(g \circ f)'(a) = g'(b) \circ f'(a)$$

Proof. Let $S = f'(a)$ and $T = g'(b)$. Then by assumption

$$f(a+h) = f(a) + S(h) + \|h\|\varepsilon(h); \quad g(b+k) + g(b) + T(k) + \|k\|\zeta(k)$$

for suitable ε, ζ. Then,

$$\begin{aligned}
(g \circ f)(a+h) &= g(f(a) + S(h) + \|h\|\varepsilon(h)) \\
&= g\left(b + \underbrace{S(h) + \|h\|\varepsilon(h)}_{k}\right) \\
&= g(b) + T(S(h) + \|h\|\varepsilon(h)) + \|S(h) + \|h\|\varepsilon(h)\|\zeta(S(h) + \|h\|\varepsilon(h)) \\
&= (g \circ f)(a) + (T \circ S)(h) + \|h\|T(\varepsilon(h)) + \|k\|\zeta(k)
\end{aligned}$$

It suffices to show that

$$\eta(h) \equiv \|h\|T(\varepsilon(h)) + \|k\|\zeta(k)$$

satisfies $\frac{\eta}{\|h\|} \to 0$. Then the result follows. First,

$$\frac{\|h\| T(\varepsilon(h))}{\|h\|} = T(\varepsilon(h)) \to 0$$

as $\|T(\varepsilon(h))\| \le \|T\| \cdot \|\varepsilon(h)\| \to 0$. Then,

$$\frac{\|k\|}{\|h\|} = \frac{\|S(h)\| + \|h\| \cdot \|\varepsilon(h)\|}{\|h\|} \le \|S\| + \|\varepsilon(h)\|$$

Hence, $k = S(h) + \|h\| \cdot \varepsilon(h) \to 0$ as $h \to 0$. Thus $\zeta(k) \to 0$ as $k \to 0$. So

$$\frac{\eta(h)}{\|h\|} = T(\varepsilon(h)) + \frac{\|k\|}{\|h\|}\zeta(k) \to 0$$

as required. □

Proposition. Let $U \subset \mathbb{R}^m$ be open, $f : U \to \mathbb{R}^n$ be a function, and $a \in U$. Let f_j be the jth component of f, so $f_j = \pi_j \circ f$. Then f is differentiable at a if and only if each f_j is differentiable at a. If this holds,

$$f'(a)(h) = \sum_{j=1}^{n} f_j'(a)(h) e_j'$$

Equivalently,

$$\pi_j[f'(a)(h)] = f_j'(a)(h)$$

Proof. If f is differentiable at a, by the chain rule the composite $\pi_j \circ f$ is differentiable at a. Since the derivative of a linear map is itself, the derivative is given by

$$f_j'(a) = \pi_j'(f(a)) \circ f'(a) = \pi_j \circ f'(a)$$

Hence

$$f'(a)(h) = \sum_{j=1}^{n} \pi_j\Big[f'(a)(h)e_j'\Big] = \sum_{j=1}^{n} f_j'(a)(h)e_j'$$

Conversely suppose each f_j is differentiable. Then

$$f_j(a+h) = f_j(a) + f_j'(a)(h) + \|h\|\varepsilon_j(h)$$

for suitable $\varepsilon(j)$. Now,

$$\begin{aligned} f(a+h) &= \sum_{j=1}^{n} f_j(a+h)e_j' \\ &= \sum_{j=1}^{n} \Big[f_j(a) + f_j'(a)(h) + \|h\|\varepsilon_j(h)\Big]e_j' \\ &= \sum_{j=1}^{n} f_j(a)e_j' + \sum_{j=1}^{n} f_j'(a)(h)e_j' + \|h\| \sum_{j=1}^{n} \varepsilon_j(h)e_j' \end{aligned}$$

Since each ε_j tends to zero as $h \to 0$, so does their sum. □

Remark. This proposition shows that we can prove things for an image $\mathbb{R}^n = \mathbb{R}$ without loss of generality.

10.5. Linearity and product rule

Proposition. Let $U \subset \mathbb{R}^m$ be open and functions $f, g\colon U \to \mathbb{R}^n$, $\phi\colon U \to \mathbb{R}$ which are differentiable at a. Then the functions $f+g$ and $\phi \cdot f$ are also differentiable and their derivatives are

$$(f+g)'(a) = f'(a) + g'(a); \quad (\phi f)'(a)(h) = \phi(a)[f'(a)(h)] + [\phi'(a)(h)]f(a)$$

For $m = n = 1$ this is the usual product rule.

Proof. We have

$$\begin{aligned} f(a+h) &= f(a) + f'(a)(h) + \|h\|\varepsilon(h) \\ g(a+h) &= g(a) + g'(a)(h) + \|h\|\zeta(h) \\ \phi(a+h) &= \phi(a) + \phi'(a)(h) + \|h\|\eta(h) \end{aligned}$$

for suitable ε, ζ, η. The sum gives

$$(f+g)(a+h) = (f+g)(a+h) + (f'(a) + g'(a))(h) + \|h\|(\varepsilon(h) + \zeta(h))$$

It follows that $f+g$ is differentiable at a and its derivative is the sum of the derivatives of its components.

$$\begin{aligned} (\phi \cdot f)(a+h) &= \phi(a+h)f(a+h) \\ &= (\phi \cdot f)(a) + [\phi(a)f'(a)(h) + \phi'(a)(h)f(a)] + f'(a)(h)\phi'(a)(h) \\ &\quad + \|h\| \underbrace{(f'(a)(h)\eta(h) + \phi'(a)(h)\varepsilon(h) + \eta(h)f(a) + \phi(a)\varepsilon(h) + \|h\|\eta(h)\varepsilon(h))}_{\delta(h)} \end{aligned}$$

Now,

$$\frac{\|\phi'(a)(h) \cdot f'(a)(h)\|}{\|h\|} = \frac{|\phi'(a)(h)| \cdot \|f'(a)(h)\|}{\|h\|} \leq \frac{\|\phi'(a)\| \cdot \|h\| \cdot \|f'(a)\| \cdot \|h\|}{\|h\|} \to 0$$

Clearly $\delta \to 0$ since the same is true for all of its components. □

11. Partial derivatives

11.1. Directional and partial derivatives

Definition. Let U, f, a as before. Fix a direction $u \in \mathbb{R}^m$ where $u \neq 0$. If the limit

$$\lim_{t \to 0} \frac{f(a+tu) - f(a)}{t}$$

exists, then the value of this limit is the *directional derivative* of f at a in direction u, denoted $D_u f(a)$.

Remark. Note that $D_u f(a) \in \mathbb{R}^n$. Further, $f(a+tu) = f(a) + tD_u f(a) + o(t)$. Define $\gamma : \mathbb{R} \to \mathbb{R}^m$ by $\gamma(t) = a + tu$. Then $f \circ \gamma$ is defined on $\gamma^{-1}(U)$ which is open as γ is continuous, and $0 \in \gamma^{-1}(U)$. Then,

$$\frac{f(a+tu) - f(a)}{t} = \frac{(f \circ \gamma)(t) - (f \circ \gamma)(0)}{t}$$

Hence $D_u f(a)$ exists if and only if $f \circ \gamma$ is differentiable at zero, and its value is the derivative of $f \circ \gamma$. When $u = e_i$ for a standard basis vector e_i, if $D_{e_i} f(a)$ exists we call it the ith *partial derivative* of f at a, denoted $D_i f(a)$.

Proposition. Let U, f, a as before. If f is differentiable at a, then all directional derivatives $D_u f(a)$ exist. Further,

$$D_u f(a) = f'(a)(u)$$

Further,

$$f'(a)(h) = \sum_{i=1}^{m} h_i D_i f(a)$$

for all $h = \sum_{i=1}^{m} h_i e_i$.

Proof. Since f is differentiable,

$$f(a+h) = f(a) + f'(a)(h) + \|h\| \varepsilon(h)$$

Let $h = tu$. Then,

$$f(a+tu) = f(a) + tf'(a)(u) + |t| \cdot \|u\| \varepsilon(tu)$$

Hence,

$$\frac{f(a+tu) - f(a)}{t} = f'(a)(u) + \frac{|t|}{t} \|u\| \varepsilon(tu)$$

The error term converges to zero, hence the limit becomes $f'(a)(u)$. Moreover, for all h defined as above,

$$f'(a)(h) = \sum_{i=1}^{m} h_i f'(a)(e_i) = \sum_{i=1}^{m} h_i D_i f(a)$$

□

alternative proof. Let $\gamma(t) = a + tu$. Then $f \circ \gamma$ is defined on the open set $\gamma^{-1}(U)$. Note that γ is differentiable and $\gamma'(t) = u$ for all t. By the chain rule, $f \circ \gamma$ is differentiable at zero, and

$$D_u f(a) = (f \circ \gamma)'(0) = f'(\gamma(0))(\gamma'(0)) = f'(a)(u)$$

□

Remark. If $D_u f(a)$ exists, then so does $D_u f_j(a)$ where $f_j = \pi_j \circ f$. Indeed, by linearity and continuity of π,

$$\frac{f_j(a+tu) - f_j(a)}{t} = \pi_j\left(\frac{f(a+tu) - f(t)}{t}\right) \to \pi_j(D_u f(a))$$

The converse of the proposition is false in general.

11.2. Jacobian matrix

Definition. Suppose f is differentiable at a. Then the Jacobian matrix of f at a, denoted $J_f(a)$, is the matrix of $f'(a)$ with respect to the standard bases. For $1 \le i \le m$, the ith column is

$$f'(a)(e_i) = D_i f(a)$$

In particular, for the j, i entry,

$$\big(J_f(a)\big)_{ji} = \left\langle D_i f(a), e_j' \right\rangle = \pi_j(D_i f(a)) = D_i f_j(a) = \frac{\partial f_j}{\partial x_i}$$

11.3. Constructing total derivative from partial derivatives

Theorem. Suppose there exists an open neighbourhood V of a with $V \subset U$ such that $D_i f(x)$ exists for all $x \in V$ and for all $1 \le i \le m$, and the map $x \mapsto D_i f(x)$ from V to $\mathbb{R}^n$ is continuous at a for all i. Then f is differentiable at a.

Proof. By considering components, without loss of generality let $n = 1$. Let $m = 2$ for convenience of notation; this does not change the proof. Let $a = (p, q)$. Let

$$\psi(h,k) = f(p+h, q+k) - f(p,q) - hD_1 f(p,q) - kD_2 f(p,q)$$

We need to show $\psi(h,k) = o(\|(h,k)\|)$, then the derivative of f can be read off from the definition of ψ. Note,

$$\psi(h,k) = [f(p+h,q+k) - f(p+h,q) - kD_2 f(p,q)] + [f(p+h,q) - f(p,q) - hD_1(p,q)]$$

We will show separately that each part is small enough to be an error term. The second term is $o(h)$ and hence $o(\|(h,k)\|)$ by the definition of $D_1 f(p,q)$. For the first term, let $\phi(t) = f(p+h, q+tk)$ for a given fixed h, k. Then ϕ is differentiable and by the chain rule we have

$\phi'(t) = D_2 f(p+h, q+tk) \cdot k$. By the mean value theorem, there exists a point $t(h,k) \in (0,1)$ such that $\phi(1) - \phi(0) = \phi'(t)$. Hence, the first term becomes

$$\phi(1) - \phi(0) - kD_2 f(p,q) = k[D_2 f(p+h, q+tk) - D_2 f(p,q)]$$

As $(h,k) \to (0,0)$, we have $(p+h, q+tk) \to (p,q)$. By continuity of $D_2 f$ at a, the term is $o(k)$ and hence $o(\|(h,k)\|)$. □

11.4. Mean value inequality

The mean value theorem cannot be extended verbatim to higher dimensional spaces, since there can be multiple paths between points.

Theorem. Let $U \subset \mathbb{R}^m$ be open, and $f : U \to \mathbb{R}^n$ be differentiable at every $z \in U$. Let $a, b \in U$ such that the line segment connecting a, b given by

$$[a,b] = \{(1-t)a + tb : 0 \le t \le 1\}$$

is contained inside U. Suppose there exists $M \ge 0$ such that for all $z \in [a,b]$, we have $\|f'(z)\| \le M$. Then

$$\|f(b) - f(a)\| \le M\|b - a\|$$

Proof. Let $u = b - a$ and $v = f(b) - f(a)$. Without loss of generality, let $u \neq 0$. Let $\gamma(t) = a + tu$, so $f \circ \gamma$ is defined on the open set $\gamma^{-1}(U)$, and is differentiable with derivative

$$(f \circ \gamma)'(t) = f'(\gamma(t))(\gamma'(t)) = f'(a+tu)(u)$$

Now,

$$\|f(b) - f(a)\|^2 = \langle f(b) - f(a), v\rangle = \langle (f\circ\gamma)(1) - (f\circ\gamma)(0), v\rangle$$

Let $\phi(t) = \langle (f \circ \gamma)(t), v\rangle$. Note that ϕ is differentiable since the inner product is linear. The derivative is

$$\phi'(t) = \langle (f\circ\gamma)'(t), v\rangle = \langle f'(a+tu)(u), v\rangle$$

By the mean value theorem, there exists $\theta \in (0,1)$ such that $\phi(1) - \phi(0) = \phi'(\theta)$. Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
\|f(b) - f(a)\|^2 &= \phi'(\theta) \\
&= \langle f'(a+\theta u)(u), v\rangle \\
&\le \|f'(a+\theta u)(u)\| \cdot \|v\| \\
&\le \|f'(a+\theta u)\| \cdot \|u\| \cdot \|v\| \\
&\le M\|b-a\| \cdot \|v\|
\end{aligned}$$

Hence,

$$\|f(b) - f(a)\| \le M\|b - a\|$$

as required. □

11.5. Zero derivatives

Corollary. Let U be an open, connected subset of $\mathbb{R}^m$, and $f : U \to \mathbb{R}^n$ be differentiable at every U. If $f'(a) = 0$ for all $a \in U$, then f is constant.

Proof. If $a, b \in U$ satisfy $[a, b] \subset U$, then by the mean value inequality we have

$$\|f(b) - f(a)\| \leq \|b - a\| \sup_{z \in [a,b]} \|f'(z)\| = 0$$

Hence $f(a) = f(b)$. For an arbitrary $x \in U$, there exists $r > 0$ such that $\mathcal{D}_r(x) \subset U$. This open ball is convex, so for all $y \in \mathcal{D}_r(x)$ we have $f(y) = f(x)$. Hence f is locally constant; every point has a neighbourhood on which f is constant. Since U is connected, f is constant (refer to the derivation from the example sheet). □

11.6. Inverse function theorem

Remark. Let $V \subset \mathbb{R}^m$ and $W \subset \mathbb{R}^n$ be open sets. Let $f : V \to W$ be a bijection. Let $a \in V$, and let f be differentiable at a, and the inverse $f^{-1} : W \to V$ is differentiable at $f(a)$. Denoting $S = f'(a), T = \left(f^{-1}\right)'(f(a))$, we can use the chain rule to find

$$TS = \left(f^{-1} \circ f\right)'(a); \quad ST = \left(f \circ f^{-1}\right)'(f(a))$$

The identity function is linear so its derivative is the identity. Hence TS is the identity on $\mathbb{R}^m$ and ST is the identity on $\mathbb{R}^n$. Hence, $m = \operatorname{tr}(TS) = \operatorname{tr}(ST) = n$. So in order for f to be a bijection, the dimensions of the spaces must match. Hence $f'(a)$ is an invertible matrix. This proves that $\mathbb{R}^m, \mathbb{R}^n$ are not homeomorphic in such a way that the maps between them are differentiable. We aim now to prove an inverse; if f is differentiable and f' is invertible, then f is locally a bijection between neighbourhoods.

Definition. Let $U \subset \mathbb{R}^m$ be open, and $f : U \to \mathbb{R}^n$ be a function. We say that f is differentiable on U if f is differentiable at a for all $a \in U$. Then, the *derivative of f on U* is the function $f' : U \to L(\mathbb{R}^m, \mathbb{R}^n)$ mapping points to their derivatives. We say that f is a *C^1-function on U* if f is continuously differentiable on U; f is differentiable on U and $f' : U \to L(\mathbb{R}^m, \mathbb{R}^n)$ is a continuous function.

Theorem. Let $U \subset \mathbb{R}^n$ be open. Let $f : U \to \mathbb{R}^n$ be a C^1-function. Let $a \in U$, and let $f'(a)$ be an invertible linear map $f'(a) : L(\mathbb{R}^n)$. Then there exist open sets V, W such that $a \in V, f(a) \in W, V \subset U$ and $f|_V : V \to W$ is a bijection with inverse function $g : W \to V$. Further, g is a C^1-function, and

$$g'(y) = [f'(g(y))]^{-1}$$

Proof. We first show that without loss of generality we can let $a = f(a) = 0$ and $f'(a) = I$. To see this, let $T = f'(a)$ and define $h(x) = T^{-1}(f(x + a) - f(a))$. Then, h is defined on

$U - a$, which is open. In particular, $U - a$ is an open neighbourhood of zero. By the chain rule, h is differentiable with $h'(x) = T^{-1} \circ f'(x + a)$. For $x, y \in U - a$, we then have

$$\|h'(x) - h'(y)\| = \left\|T^{-1} \circ (f'(a + x) - f'(a + y))\right\| \leq \left\|T^{-1}\right\| \cdot \|f'(a + x) - f'(a + y)\|$$

It then follows that h is a C^1-function, and that $h(0) = 0$, $h'(0) = T^{-1} \circ T = I$. We have transformed into a coordinate system where $a = f(a) = 0$ and $f'(a) = I$. If we can prove the result for this coordinate system, we can translate back using $f(x) = T(h(x - a)) + f(a)$.

Now, let $f(0) = 0$ and $f'(0) = I$. Since f' is continuous, there exists $r > 0$ such that $\mathcal{B}_r(0) \subset U$ and for all $x \in U$, we have

$$\|f'(x) - f'(0)\| = \|f'(x) - I\| \leq \frac{1}{2}$$

We intend to show that for all $x, y \in \mathcal{B}_r(0)$, we have $\|f(x) - f(y)\| \geq \frac{1}{2}\|x - y\|$. Indeed, define $p\colon U \to \mathbb{R}^n$ by $p(x) = f(x) - x$. Then $p'(x) = f'(x) - I$. Then, $\|p'(x)\| \leq \frac{1}{2}$ for all $x \in \mathcal{B}_r(0)$. By the mean value inequality, $\|p(x) - p(y)\| \leq \frac{1}{2}\|x - y\|$ for all $x, y \in \mathcal{B}_r(0)$. Hence,

$$\|f(x) - f(y)\| = \|(p(x) + x) - (p(y) + y)\| \geq \|x - y\| - \|p(x) - p(y)\| \geq \frac{1}{2}\|x - y\|$$

So we have proven the bound as claimed. Now, let $s = \frac{r}{2}$. We will show that $f(\mathcal{D}_r(0)) \subset \mathcal{D}_s(0)$. More precisely, we will show that for all $w \in \mathcal{D}_s(0)$ there exists a unique $x \in \mathcal{D}_r(0)$ such that $f(x) = w$. Let $w \in \mathcal{D}_s(0)$ be fixed. We now define, for all $x \in \mathcal{B}_r(0)$, the function $q(x) = w - f(x) + x = w - p(x)$. Note that $f(x) = w$ if and only if $q(x) = x$. We will show that q is a contraction mapping, and that there exists a fixed point. Since $p(0) = f(0) - 0 = 0$, we have for all $x \in \mathcal{B}_r(0)$ that

$$\|q(x)\| \leq \|w\| + \|p(x)\| = \|w\| + \|p(x) - p(0)\| \leq \|w\| + \frac{1}{2}\|x - 0\| = \frac{1}{2}\|x\| < s + \frac{1}{2}r$$

Hence, $q(\mathcal{B}_r(0)) \subset \mathcal{D}_r(0) \subset \mathcal{B}_r(0)$. We now show q is a contraction mapping. For $x, y \in \mathcal{B}_r(0)$, we have

$$\|q(x) - q(y)\| = \|p(x) - p(y)\| \leq \frac{1}{2}\|x - y\|$$

Hence $q\colon \mathcal{B}_r(0) \to \mathcal{B}_r(0)$ really is a contraction mapping on the non-empty, complete metric space $\mathcal{B}_r(0)$. By the contraction mapping theorem, there exists a unique $x \in \mathcal{B}_r(0)$ such that $q(x) = x$. But since $q(\mathcal{B}_r(0)) \subset \mathcal{D}_r(0)$, we must have $x \in \mathcal{D}_r(0)$. In particular, there exists a unique $x \in \mathcal{D}_r(0)$ such that $f(x) = w$.

Now, let $W = \mathcal{D}_s(0), V = \mathcal{D}_r(0) \cap f^{-1}(W)$. Then, we will now show that $f|_V : V \to W$ is a bijection with inverse $g\colon W \to V$ which is continuous. First, W is open and $f(0) = 0 \in W$. Since f is continuous, $f^{-1}(W)$ is open. Hence V is open, as the intersection of two open sets. We have $0 \in V$. By the previous paragraph, $f|_V : V \to W$ is a bijection since for

every point in W there exists a unique point in V mapping to it. Finally, let $u, v \in W$. Let $x = g(u), y = g(v)$. Then,

$$\|g(u) - g(v)\| = \|x - y\| \leq 2\|f(x) - f(y)\| = 2\|u - v\|$$

Hence g is 2-Lipschitz and hence continuous. Now it suffices to show g is C^1, and for all $y \in W$ we have $g'(y) = [f'(g(y))]^{-1}$. This part of the proof is non-examinable. □

12. Second derivatives

12.1. Definition

Definition. Let $U \subset \mathbb{R}^m$ be an open set, and $f : U \to \mathbb{R}^n$. Let $a \in U$. Suppose that there exists an open neighbourhood V of a contained within U, and f is differentiable on V. We say that f is *twice differentiable* at a if $f' : V \to L(\mathbb{R}^m \to \mathbb{R}^n)$ is differentiable at a. We write $f''(a)$ for the derivative of f' at a, called the *second derivative* of f at a. Note that $f''(a) \in L(\mathbb{R}^m, L(\mathbb{R}^m, \mathbb{R}^n))$.

Remark. We can visualise the second derivative as a bilinear map instead of a nested sequence of linear maps. Note,

$$L(\mathbb{R}^m, L(\mathbb{R}^m, \mathbb{R}^n)) \sim \mathrm{Bil}(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^n)$$

where $\mathrm{Bil}(X \times Y, Z)$ is the vector space of bilinear maps from $X \times Y$ to Z. For $h, k \in \mathbb{R}^m$, and T is the second derivative, we can say $T(h)(k) = \widetilde{T}(h, k)$ where $\widetilde{T}$ is a bilinear map. From now on, this bilinear map notation will be used, and T and $\widetilde{T}$ will be identified as the same.

Proposition. Let $U \subset \mathbb{R}^m$ be open, $f : U \to \mathbb{R}^n$ be a function, and $a \in U$. Let f be differentiable on an open neighbourhood V of A contained in U. Then f is twice differentiable at a if and only if there exists a bilinear map $T \in \mathrm{Bil}(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^n)$ such that for every $k \in \mathbb{R}^m$, we have

$$f'(a + h)(k) = f'(a)(k) + T(h, k) + o(\|h\|)$$

Then $T = f''(a)$.

Proof. Suppose f is twice differentiable at a. Then f' is differentiable at a. So,

$$f'(a + h) = f'(a) + f''(a)(h) + \|h\| \cdot \varepsilon(h)$$

All terms are linear maps $L(\mathbb{R}^m, \mathbb{R}^n)$. In particular, ε is defined on $V - a \to L(\mathbb{R}^m, \mathbb{R}^n)$ such that $\varepsilon(0) = 0$ and ε is continuous at zero. If we evaluate this equation at a fixed $k \in \mathbb{R}^m$,

$$f'(a + h)(k) = f'(a)(k) + f''(a)(h, k) + \|h\| \cdot \varepsilon(h)(k)$$

Here, $f''(a)$ is a bilinear map. Further,

$$\|\varepsilon(h)(k)\| \leq \|\varepsilon(h)\| \cdot \|k\| \to 0$$

Hence, $\|h\| \cdot \varepsilon(h)(k) = o(\|h\|)$. Conversely, suppose T is a bilinear map and

$$\frac{f'(a + h)(k) - f'(a)(k) - T(h, k)}{\|h\|} \to 0$$

for any fixed k, as $h \to 0$. We need to show that

$$\varepsilon(h) = \frac{f'(a + h) - f'(a) - T(h)}{\|h\|} \to 0$$

in the space $L(\mathbb{R}^m, \mathbb{R}^n)$. We know that for a fixed $k \in \mathbb{R}^m$, $\varepsilon(h)(k) \to 0$ in $\mathbb{R}^n$ as $h \to 0$. It then follows that

$$\|\varepsilon(h)\| = \sqrt{\sum_{i=1}^{m} \|\varepsilon(h)(e_i)\|^2} \to 0$$

since we are in a finite-dimensional vector space. □

Example. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be linear. Then f is differentiable on $\mathbb{R}^m$ with $f'(a) = f$ for all a. Hence $f' : \mathbb{R}^m \to L(\mathbb{R}^m, \mathbb{R}^n)$ sends a to f for all a. So this is a constant function, so has derivative $f''(a) = 0$.

Example. Let $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p$ be bilinear. Then f is differentiable on $\mathbb{R}^m \times \mathbb{R}^n$ and for all $(a, b) \in \mathbb{R}^m \times \mathbb{R}^n$, we have

$$f'(a, b)(h, k) = f(a, k) + f(h, b)$$

Note that this is linear in (a, b) for a fixed (h, k). Hence, $f' : \mathbb{R}^m \times \mathbb{R}^n \to L(\mathbb{R}^m, \mathbb{R}^n, \mathbb{R}^p)$ is linear. Hence this is differentiable, and its derivative is

$$f''(a, b) = f' \in L(\mathbb{R}^m, \mathbb{R}^n, L(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R}^p)) \simeq \mathrm{Bil}((\mathbb{R}^m \times \mathbb{R}^n) \times (\mathbb{R}^m \times \mathbb{R}^n), \mathbb{R}^p)$$

Example. Let $f : M_n \to M_n$ be defined by $f(A) = A^3$. Let A be fixed. Then,

$$\begin{aligned} f(A + H) &= (A + H)^3 = A^3 + A^2H + AHA + HA^2 + AH^2 + HAH + H^2A + H^3 \\ &= f(A) + (A^2H + AHA + HA^2) + o(\|H\|) \end{aligned}$$

Hence f is differentiable at A and

$$f'(A)(H) = A^2H + AHA + HA^2$$

Thus, if $n = 1$, we have commutativity and hence $f'(A) = 3A^2$. So f is differentiable on M_n. For a fixed A and fixed K, the second derivative is given by

$$\begin{aligned} f'(A + H)(K) &= (A + H)^2K + (A + H)K(A + H) + K(A + H)^2 \\ &= \underbrace{(A^2K + AKA + KA^2)}_{f'(A)(K)} \\ &\quad + (AHK + HAK + AKH + HKA + KAH + KHA) + (H^2K + HKH + KH^2) \end{aligned}$$

The term $T(H, K) = (AHK + HAK + AKH + HKA + KAH + KHA)$ is bilinear in H and K as required. So the second derivative is T. In one dimension, this is equivalent to saying $f''(A) = 6A$.

12.2. Second derivatives and partial derivatives

Let U be open in $\mathbb{R}^n$, let $f : U \to \mathbb{R}^n$, and let $a \in U$. Let f be twice differentiable at a, so f is differentiable on some open neighbourhood V of a contained within U, and $f' : V \to L(\mathbb{R}^m, \mathbb{R}^n)$ is differentiable at a. Recall that

$$f'(a+h) = f'(a) + f''(a)(h) + o(\|h\|)$$

Evaluating at a fixed k,

$$f'(a+h)(k) = f'(a)(k) + f''(a)(h,k) + o(\|h\|)$$

Let $u, v \in \mathbb{R}^m \setminus \{0\}$ be directions. Let $k = v$. Then,

$$f'(a+h)(v) = D_v f(a+h) = D_v f(a) + f''(a)(h,v) + o(\|h\|)$$

Hence, the map $D_v f : V \to \mathbb{R}^n$ maps $x \mapsto D_v f(x) = f'(x)(v)$. Then this map is differentiable at a and

$$(D_v f)'(a)(h) = f''(a)(h,v)$$

Hence there exist directional derivatives.

$$D_u D_v f(a) \overset{\text{def}}{=} D_u(D_v f)(a) = (D_v f)'(a)(u) = f''(a)(u,v)$$

In particular, we have

$$D_i D_j f(a) = f''(a)(e_i, e_j)$$

for $1 \le i, j \le m$.

12.3. Symmetry of mixed directional derivatives

Theorem. Let U be open in $\mathbb{R}^n$, let $f : U \to \mathbb{R}^n$, and let $a \in U$. Let f be twice differentiable on an open set V with $a \in V \subset U$. Let $f'' : V \to \mathrm{Bil}(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^n)$ be continuous at a. Then, for all directions $u, v \in \mathbb{R}^m \setminus \{0\}$, we have

$$D_u D_v f(a) = D_v D_u f(a)$$

Equivalently,

$$f''(a)(u,v) = f''(a)(v,u)$$

In other words, f'' is a symmetric bilinear map.

Proof. Without loss of generality we can let $n = 1$. Indeed, we have

$$(D_u f)_j(x) = [D_u f(x)]_j = [f'(x)(u)]_j = f_j'(x)(u) = D_u f_j(x)$$

Hence, $(D_u f)_j = D_u f_j$. For v:

$$(D_v D_u f)_j = D_v (D_u f)_j = D_v D_u f_j$$

So it is sufficient to show that $D_v D_u f_j(a) = D_u D_v f_j(a)$. Now, consider

$$\phi(s,t) = f(a+su+tv) - f(a+tv) - f(a+su) + f(a)$$

for $s,t \in \mathbb{R}$. Let s,t be fixed, and consider

$$\psi(y) = f(a+yu+tv) - f(a+yu)$$

Note that $\phi(s,t)$ can be written as

$$\phi(s,t) = \psi(s) - \psi(0)$$

The term $\psi(s) - \psi(0)$ can be interpreted as $(f(a+su+tv) - f(a+tv)) - (f(a+su) - f(a))$, which is the second difference given by the function when traversing the parallelogram with sides su, tv. By the mean value theorem, there exists $\alpha(s,t) \in (0,1)$ such that

$$\phi(s,t) = \psi(s) - \psi(0) = s\psi'(\alpha s) = s[D_u f(a+\alpha su+tv) - D_u f(a+\alpha su)]$$

Now, applying the mean value theorem to the function $y \mapsto D_u f(a+\alpha su+yv)$, we have

$$\phi(s,t) = st D_v D_u f(a+\alpha su+\beta tv)$$

for $\beta(s,t) \in (0,1)$. Now,

$$\frac{\phi(s,t)}{st} = D_v D_u f(a+\alpha su+\beta tv) = f''(a+\alpha su+\beta tv)(u,v)$$

Since f'' is continuous at a, we can let $s,t \to 0$ and find

$$\frac{\phi(s,t)}{st} \to f''(a)(u,v)$$

Now, we can repeat the above using

$$\psi(y) = f(a+su+yv) - f(a+yv)$$

This calculates the second difference from above, but using the other path. We can find

$$\frac{\phi(s,t)}{st} \to f''(a)(v,u)$$

as required. □

V. Methods

Lectured in Michaelmas 2021 by Prof. E. P. Shellard

In this course, we discuss various methods for solving differential equations. Different forms of differential equations need different solution strategies, and we study a wide range of common types of differential equation.

A particularly powerful method for solving differential equations involves the use of Green's functions. For example, physical systems can involve bodies spread over space with constant density. Green's functions allow the equation to be solved for a point mass, and then integrated to find the solution for the larger body.

Fourier transforms are another way to solve differential equations. Sometimes a differential equation is easier to solve after applying the Fourier transform to the relevant function, then the inverse Fourier transform recovers the solution to the original equation.

Contents

1. Fourier series 219
1.1. Periodic functions 219
1.2. Properties of trigonometric functions 219
1.3. Periodic function space 219
1.4. Fourier series 220
1.5. Dirichlet conditions 221
1.6. Integration 222
1.7. Differentiation 223
1.8. Parseval's theorem 223
1.9. Half-range series 224
1.10. Complex representation of Fourier series 224
1.11. Self-adjoint matrices 225
1.12. Solving inhomogeneous ODEs with Fourier series 226
2. Sturm–Liouville theory 228
2.1. Second-order linear ODEs 228
2.2. Sturm–Liouville form 228
2.3. Converting to Sturm–Liouville form 229
2.4. Self-adjoint operators 229
2.5. Self-adjoint compatible boundary conditions 230
2.6. Properties of self-adjoint operators 230
2.7. Real eigenvalues 231
2.8. Orthogonality of eigenfunctions 231
2.9. Eigenfunction expansions 232
2.10. Completeness and Parseval's identity 233
2.11. Legendre's equation 234
2.12. Properties of Legendre polynomials 235
2.13. Legendre polynomials as eigenfunctions 235
2.14. Solving inhomogeneous differential equations 236
2.15. Integral solutions 236
2.16. Waves on an elastic string 237
3. Separation of variables 239
3.1. Separation of variables 239
3.2. Boundary conditions and normal modes 239
3.3. Initial conditions and temporal solutions 240
3.4. Separation of variables methodology 241
3.5. Energy of oscillations 242
3.6. Wave reflection and transmission 243

3.7. Wave equation in plane polar coordinates . . . 244
3.8. Bessel's equation . . . 245
3.9. Asymptotic behaviour of Bessel functions . . . 246
3.10. Zeroes of Bessel functions . . . 246
3.11. Solving the vibrating drum . . . 246
3.12. Diffusion equation derivation with Fourier's law . . . 248
3.13. Diffusion equation derivation with statistical dynamics . . . 248
3.14. Similarity solutions . . . 249
3.15. Heat conduction in a finite bar . . . 250
3.16. Particular solution to diffusion equation . . . 251
3.17. Laplace's equation . . . 252
3.18. Laplace's equation in three-dimensional Cartesian coordinates . . . 252
3.19. Laplace's equation in plane polar coordinates . . . 254
3.20. Laplace's equation in cylindrical polar coordinates . . . 255
3.21. Laplace's equation in spherical polar coordinates . . . 256
3.22. Generating function for Legendre polynomials . . . 257
4. Green's functions . . . 258
4.1. Dirac δ function . . . 258
4.2. Integral and derivative of δ function . . . 259
4.3. Properties of δ function . . . 259
4.4. Fourier series expansion of δ function . . . 260
4.5. Arbitrary eigenfunction expansion of δ function . . . 261
4.6. Motivation for Green's functions . . . 261
4.7. Definition of Green's function . . . 263
4.8. Explicit form for Green's functions . . . 264
4.9. Solving boundary value problems . . . 264
4.10. Higher-order ODEs . . . 266
4.11. Eigenfunction expansions of Green's functions . . . 266
4.12. Constructing Green's function for an initial value problem . . . 266
5. Fourier transforms . . . 268
5.1. Definitions . . . 268
5.2. Converting Fourier series into Fourier transforms . . . 269
5.3. Properties of Fourier series . . . 270
5.4. Convolution theorem . . . 271
5.5. Parseval's theorem . . . 272
5.6. Fourier transforms of generalised functions . . . 273
5.7. Trigonometric functions . . . 274
5.8. Heaviside functions . . . 274
5.9. Dirichlet discontinuous formula . . . 274
5.10. Solving ODEs for boundary value problems . . . 275

5.11. Signal processing . . . 275
5.12. General transfer functions for ODEs . . . 276
5.13. Damped oscillator . . . 277
5.14. Discrete sampling and the Nyquist frequency . . . 278
5.15. Nyquist–Shannon sampling theorem . . . 278
5.16. Discrete Fourier transform . . . 279
5.17. Fast Fourier transform (non-examinable) . . . 281
6. Method of characteristics . . . 282
6.1. Well-posed Cauchy problems . . . 282
6.2. Method of characteristics . . . 282
6.3. Characteristics of a first order PDE . . . 283
6.4. Inhomogeneous first order PDEs . . . 284
6.5. Classification of second order PDEs . . . 285
6.6. Characteristic curves of second order PDEs . . . 286
6.7. Characteristic coordinates . . . 287
6.8. General solution to wave equation . . . 288
7. Solving partial differential equations with Green's functions . . . 289
7.1. Diffusion equation and Fourier transform . . . 289
7.2. Gaussian pulse for heat equation . . . 290
7.3. Forced diffusion equation . . . 290
7.4. Duhamel's principle . . . 291
7.5. Forced wave equation . . . 292
7.6. Poisson's equation . . . 293
7.7. Green's identities . . . 294
7.8. Dirichlet Green's function . . . 295
7.9. Method of images for Laplace's equation . . . 295
7.10. Method of images for wave equation . . . 296

1. Fourier series

1.1. Periodic functions

A function $f(x)$ is *periodic* if $f(x+T) = f(x)$ for all x, where T is the period. For example, simple harmonic motion is periodic. In space, we consider the wavelength $\lambda = \frac{2\pi}{k}$, and the (angular) wave number k is defined conversely by $k = \frac{2\pi}{\lambda}$.

1.2. Properties of trigonometric functions

Consider the set of functions

$$g_n(x) = \cos\frac{n\pi x}{L}; \quad h_n(x) = \sin\frac{n\pi x}{L}$$

where $n \in \mathbb{N}$. These functions are periodic with period $T = 2L$. Recall that

$$\begin{aligned}
\cos A \cos B &= \frac{1}{2}(\cos(A-B) + \cos(A+B)); \\
\sin A \sin B &= \frac{1}{2}(\cos(A-B) - \cos(A+B)); \\
\sin A \cos B &= \frac{1}{2}(\sin(A-B) + \sin(A+B))
\end{aligned}$$

1.3. Periodic function space

We define the inner product

$$\langle f, g\rangle = \int_0^{2L} f(x)g(x)\,\mathrm{d}x$$

The functions g_n and h_n are mutually orthogonal on the interval $[0, 2L)$ with respect to the inner product above.

$$\begin{aligned}
\langle h_n, h_m\rangle &= \int_0^{2L} \sin\frac{n\pi x}{L}\sin\frac{m\pi x}{L}\,\mathrm{d}x \\
&= \frac{1}{2}\int_0^{2L}\left(\cos\frac{(n-m)\pi x}{L} - \cos\frac{(n+m)\pi x}{L}\right)\mathrm{d}x \\
&= \frac{1}{2}\frac{L}{\pi}\left[\frac{1}{n-m}\sin\frac{(n-m)\pi x}{L} - \frac{1}{n+m}\sin\frac{(n+m)\pi x}{L}\right]_0^{2L} \\
&= 0 \text{ when } n \neq m
\end{aligned}$$

If $n = m$, we have

$$\langle h_n, h_n\rangle = \int_0^{2L}\sin^2\frac{n\pi x}{L}\,\mathrm{d}x = \frac{1}{2}\int_0^{2L}\left(1 - \cos\frac{2\pi n x}{L}\right)\mathrm{d}x = L$$

Thus,

$$\langle h_n, h_m \rangle = \begin{cases} L\delta_{nm} & n, m \neq 0 \\ 0 & nm = 0 \end{cases}$$

Similarly, we can show

$$\langle g_n, g_m \rangle = \begin{cases} L\delta_{nm} & n, m \neq 0 \\ 0 & \text{exactly one of } m, n \text{ is zero} \\ 2L & n, m = 0 \end{cases}$$

and

$$\langle h_n, g_m \rangle = 0$$

Now, we assert that $\{g_n, h_n\}$ form a complete orthogonal set; they span the space of all 'well-behaved' periodic functions of period $2L$. Further, the set $\{g_n, h_n\}$ is linearly independent.

1.4. Fourier series

Since g_n, h_n span the space of 'well-behaved' periodic functions of period $2L$, we can express any such function as a sum of such eigenfunctions.

Definition. The Fourier series of f is

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L} + \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}$$

where a_n, b_n are constants such that the right hand side is convergent for all x where f is continuous. At a discontinuity x, the Fourier series approaches the midpoint of the supremum and infimum of the function in a close neighbourhood of x. That is, we replace the left hand side with

$$\frac{1}{2}f(x_+) + \frac{1}{2}f(x_-)$$

Let $m > 0$, and consider taking the inner product $\langle h_m, f \rangle$ and substituting the Fourier series of f.

$$\begin{aligned} \langle h_m, f \rangle &= \int_0^{2L} \sin\frac{m\pi x}{L} f(x)\,\mathrm{d}x \\ &= \langle h_m, b_m h_m \rangle \\ &= Lb_m \end{aligned}$$

Thus,

$$b_n = \frac{1}{L}\langle h_n, f \rangle = \frac{1}{L}\int_0^{2L} \sin\frac{n\pi x}{L} f(x)\,\mathrm{d}x$$

and analogously

$$a_n = \frac{1}{L}\langle g_n, f\rangle = \frac{1}{L}\int_0^{2L} \cos\frac{n\pi x}{L} f(x)\,\mathrm{d}x$$

Note that $\frac{1}{2}a_0$ is the average of the function. Note further that we may integrate over any range as long as the total length is one period, $2L$. Notably, we may integrate over the interval $[-L, L]$.

Example. Consider the *sawtooth wave*; defined by $f(x) = x$ for $x \in [-L, L)$ and periodic elsewhere. Here,

$$a_n = \frac{1}{L}\int_{-L}^{L} x\cos\frac{n\pi x}{L}\,\mathrm{d}x = 0$$

and

$$\begin{aligned} b_n &= \frac{2}{L}\int_0^L x\sin\frac{n\pi x}{L}\,\mathrm{d}x \\ &= \frac{-2}{n\pi}\Big[x\cos\frac{n\pi x}{L}\Big]_0^L + \frac{2}{n\pi}\int_0^L \cos\frac{n\pi x}{L}\,\mathrm{d}x \\ &= \frac{-2L}{n\pi}\cos n\pi + \frac{2L}{(n\pi)^2}\sin n\pi \\ &= \frac{2L}{n\pi}(-1)^{n+1} \end{aligned}$$

1.5. Dirichlet conditions

The Dirichlet conditions are sufficiency conditions for a well-behaved function, that will imply the existence of a unique Fourier series.

Theorem. If $f(x)$ is a bounded periodic function of period $2L$ with a finite number of minima, maxima and discontinuities in $[0, 2L)$, then the Fourier series converges to f at all points at which f is continuous, and at discontinuities the series converges to the midpoint.

Remark. (i) These are some relatively weak conditions for convergence, compared to Taylor series. However, this definition still eliminates pathological functions such as $\frac{1}{x}$, $\sin\frac{1}{x}$, $\mathbb{1}(\mathbb{Q})$ and so on.

(ii) The converse is not true; for example, $\sin\frac{1}{x}$ does in fact have a Fourier series.

(iii) The proof is difficult and will not be given.

The rate of convergence of the Fourier series depends on the smoothness of the function.

Theorem. If $f(x)$ has continuous derivatives up to a pth derivative which is discontinuous, then the Fourier series converges with order $O(n^{-(p+1)})$ as $n \to \infty$.

Example ($p = 0$)**.** Consider the square wave

$$f(x) = \begin{cases} 1 & 0 \le x < 1 \\ -1 & -1 \le x < 0 \end{cases}$$

Then the Fourier series is

$$f(x) = 4 \sum_{m=1}^{\infty} \frac{\sin(2m-1)\pi x}{(2m-1)\pi}$$

Example ($p = 1$)**.** Consider the general 'seesaw' wave, defined by

$$f(x) = \begin{cases} x(1-\xi) & 0 \le x < \xi \\ \xi(1-x) & \xi \le x < 1 \end{cases}$$

and defined as an odd function for $-1 \le x < 0$. The Fourier series is

$$f(x) = 2 \sum_{m=1}^{\infty} \frac{\sin n\pi\xi \sin n\pi x}{(n\pi)^2}$$

For instance, if $\xi = \frac{1}{2}$, we can show that

$$f(x) = 2 \sum_{m=1}^{\infty} (-1)^{m+1} \frac{\sin(2m-1)\pi x}{((2m-1)\pi)^2}$$

Example ($p = 2$)**.** Let

$$f(x) = \frac{1}{2} x(1-x)$$

for $0 \le x < 1$, and defined as an odd function for $-1 \le x < 0$. We can show that

$$f(x) = 4 \sum_{n=1}^{\infty} \frac{\sin(2m-1)\pi x}{((2m-1)\pi)^3}$$

Example ($p = 3$)**.** Consider

$$f(x) = (1-x^2)^2$$

with Fourier series

$$a_n = O\Big(\frac{1}{n^4}\Big)$$

1.6. Integration

It is always valid to take the integral of a Fourier series term by term. Defining $F(x) = \int_{-L}^{x} f(x)\,\mathrm{d}x$, we can show that F satisfies the Dirichlet conditions if f does. For instance, a jump discontinuity becomes continuous in the integral.

1.7. Differentiation

Differentiating term by term is not always valid. For example, consider the square wave above:

$$f(x) \stackrel{?}{=} 4 \sum_{m=1}^{\infty} \cos(2m-1)\pi x$$

which is an unbounded series.

Theorem. If $f(x)$ is continuous and satisfies the Dirichlet conditions, and $f'(x)$ also satisfies the Dirichlet conditions, then $f'(x)$ can be found term by term by differentiating the Fourier series of $f(x)$.

Example. We can differentiate the seesaw function with $\xi = \frac{1}{2}$, even though the derivative is not continuous. The result is an offset square wave, or by mapping $x \mapsto x + \frac{1}{2}$ we recover the original square wave.

1.8. Parseval's theorem

Parseval's theorem relates the integral of the square of a function with the squares of the function's Fourier series coefficients.

Theorem. Suppose f has Fourier coefficients a_i, b_i. Then

$$\int_0^{2L} [f(x)]^2 \,\mathrm{d}x = \int_0^{2L} \left[\frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_k \cos\frac{k\pi x}{L} + \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L} \right]^2 \mathrm{d}x$$

We can remove cross terms, since the basis functions are orthogonal.

$$\begin{aligned} &= \int_0^{2L} \left[\frac{1}{4}a_0^2 + \sum_{n=1}^{\infty} a_n^2 \cos^2\frac{n\pi x}{L} + \sum_{n=1}^{\infty} b_n^2 \sin^2\frac{n\pi x}{L} \right] \mathrm{d}x \\ &= L\left[\frac{1}{2}a_0^2 + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \right] \end{aligned}$$

This is also called the completeness relation: the left hand side is greater than or equal to the right hand side if any of the basis functions are missing.

Example. Let us apply Parseval's theorem to the sawtooth wave.

$$\int_{-L}^{L} [f(x)]^2 \,\mathrm{d}x = \int_{-L}^{L} x^2 \,\mathrm{d}x = \frac{2}{3}L^3$$

The right hand side gives

$$L \sum_{n=1}^{\infty} \frac{4L^2}{n^2\pi^2} = \frac{4L^3}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2}$$

Parseval's theorem then implies

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

Remark. Parseval's theorem for functions is equivalent to Pythagoras' theorem for vectors in $\mathbb{R}^n$: we can find the norm of a linear combination by computing the sum of the norms of the components.

1.9. Half-range series

Consider $f(x)$ defined only on $0 \leq x < L$. We can extend the range of f to be the full range $-L \leq x < L$ in two simple ways:

(i) require f to be odd, so $f(-x) = -f(x)$. Hence, $a_n = 0$ and

$$b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} \, \mathrm{d}x$$

So

$$f(x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}$$

which is called a Fourier sine series.

(ii) require f to be even, so $f(-x) = f(x)$. In this case, $b_n = 0$ and

$$a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{L} \, \mathrm{d}x$$

and

$$Sof(x) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L}$$

which is a Fourier cosine series.

1.10. Complex representation of Fourier series

Recall that

$$\cos \frac{n\pi x}{L} = \frac{1}{2}\left(e^{in\pi x/L} + e^{-in\pi x/L}\right);$$
$$\sin \frac{n\pi x}{L} = \frac{1}{2i}\left(e^{in\pi x/L} - e^{-in\pi x/L}\right)$$

Therefore, a Fourier series can be written as

$$\begin{aligned} f(x) &= \frac{1}{2}a_0 + \frac{1}{2}\sum_{n=1}^{\infty}\left[(a_n - ib_n)e^{in\pi x/L} + (a_n + ib_n)e^{-in\pi x/L}\right] \\ &= \sum_{m=-\infty}^{\infty} c_m e^{im\pi x/L} \end{aligned}$$

where for $m > 0$ we have $m = n, c_m = \frac{1}{2}(a_n - ib_n)$, and for $m < 0$ we have $n = -m, c_m = \frac{1}{2}(a_{-m} + ib_{-m})$, and where $m = 0$ we have $c_0 = \frac{1}{2}a_0$. In particular,

$$c_m = \frac{1}{2L}\int_{-L}^{L} f(x)e^{-im\pi x/L}\,\mathrm{d}x$$

where the negative sign comes from the complex conjugate. This is because, for complex-valued f, g, we have

$$\langle f, g\rangle = \int_{-L}^{L} f^\star g\,\mathrm{d}x$$

The orthogonality conditions are

$$\int_{-L}^{L} e^{-im\pi x/L}e^{in\pi x/L}\,\mathrm{d}x = 2L\delta_{mn}$$

Parseval's theorem now states

$$\int_{-L}^{L} f^\star(x)f(x)\,\mathrm{d}x = \int_{-L}^{L} |f(x)|^2\,\mathrm{d}x = 2L\sum_{m=-\infty}^{\infty}|c_m|^2$$

1.11. Self-adjoint matrices

Much of this section is a recap of IA Vectors and Matrices. Suppose that $u, v \in \mathbb{C}^N$ with inner product

$$\langle u, v\rangle = u^\dagger v$$

The $N \times N$ matrix A is *self-adjoint*, or *Hermitian*, if

$$\forall u, v \in \mathbb{C}^N, \langle Au, v\rangle = \langle u, Av\rangle \iff A^\dagger = A$$

The eigenvalues λ_n and eigenvectors v_n satisfy

$$Av_n = \lambda_n v_n$$

They have the following properties:

(i) $\lambda_n^\star = \lambda_n$;

(ii) $\lambda_n \neq \lambda_m \implies \langle v_n, v_m \rangle = 0$;

(iii) we can create an orthonormal basis from the eigenvectors.

Given $b \in \mathbb{C}^n$, we can solve for x in the general matrix equation $Ax = b$ by expressing b in terms of the eigenvector basis:

$$b = \sum_{n=1}^{N} b_n v_n$$

We seek a solution of the form

$$x = \sum_{n=1}^{N} c_n v_n$$

At this point, the b_n are known and the c_n are our target. Substituting into the matrix equation, orthogonality of basis vectors gives

$$\begin{aligned} A \sum_{n=1}^{N} c_n v_n &= \sum_{n=1}^{N} b_n v_n \\ \sum_{n=1}^{N} c_n \lambda_n v_n &= \sum_{n=1}^{N} b_n v_n \\ c_n \lambda_n &= b_n \\ c_n &= \frac{b_n}{\lambda_n} \end{aligned}$$

Therefore,

$$x = \sum_{n=1}^{N} \frac{b_n}{\lambda_n} v_n$$

provided $\lambda_n \neq 0$, or equivalently, the matrix is invertible.

1.12. Solving inhomogeneous ODEs with Fourier series

We wish to find $y(x)$ given a source term $f(x)$ for the general differential equation

$$\mathcal{L}y \equiv -\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = f(x)$$

with boundary conditions $y(0) = y(L) = 0$. The related eigenvalue problem is

$$\mathcal{L}y_n = \lambda_n y_n, \quad y_n(0) = y_n(L) = 0$$

which has solutions

$$y_n(x) = \sin \frac{n\pi x}{L}, \lambda_n = \left(\frac{n\pi}{L}\right)^2$$

We can show that this is a self-adjoint linear operator with orthogonal eigenfunctions. We seek solutions of the form of a half-range sine series. Consider

$$y(x) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{L}$$

The right hand side is

$$f(x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}$$

We can find b_n by

$$b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} \, \mathrm{d}x$$

Substituting, we have

$$\mathcal{L}y = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(\sum_n c_n \sin \frac{n\pi x}{L}\right) = \sum_n c_n \left(\frac{n\pi}{L}\right)^2 \sin \frac{n\pi x}{L} = \sum_n b_n \sin \frac{n\pi x}{L}$$

By orthogonality,

$$c_n \left(\frac{n\pi}{L}\right)^2 = b_n \implies c_n = \left(\frac{L}{n\pi}\right)^2 b_n$$

Therefore the solution is

$$y(x) = \sum_n \left(\frac{L}{n\pi}\right)^2 b_n \sin \frac{n\pi x}{L} = \sum_n \frac{b_n}{\lambda_n} y_n$$

which is equivalent to the solution we found for self-adjoint matrices for which the eigenvalues and eigenvectors are known.

Example. Consider an odd square wave with $L = 1$, so $f(x) = 1$ from $0 \le x < 1$.

$$f(x) = 4 \sum_m \frac{\sin(2m-1)\pi x}{(2m-1)\pi}$$

Then the solution to $\mathcal{L}y = f$ should be (with odd $n = 2m - 1$)

$$y(x) = \sum_n \frac{b_n}{\lambda_n} y_n = 4 \sum_n \frac{\sin(2m-1)\pi x}{((2m-1)\pi)^3}$$

This is exactly the Fourier series for

$$y(x) = \frac{1}{2} x(1-x)$$

so this y is the solution to the differential equation. We can in fact integrate $\mathcal{L}y = 1$ directly with the boundary conditions to verify the solution. We can also differentiate the Fourier series for y twice to find the square wave.

2. Sturm–Liouville theory

2.1. Second-order linear ODEs

This section is a review of IA Differential Equations.

We wish to solve a general inhomogeneous ODE, written

$$\mathcal{L}y \equiv \alpha(x)y'' + \beta(x)y' + \gamma(x)y = f(x)$$

The homogeneous version has $f(x) = 0$, so $\mathcal{L}y = 0$, which has two independent solutions y_1, y_2. The general solution, also the complementary function for the inhomogeneous ODE, is $y_c(x) = Ay_1(x) + By_2(x)$. The inhomogeneous equation $\mathcal{L}y = f(x)$ has a solution called the particular integral, denoted $y_p(x)$. The general solution to this equation is then $y_p + y_c$.

We need two boundary or initial conditions to find the particular solution to the differential equation. Suppose $x \in [a, b]$. We can create boundary conditions by defining $y(a), y(b)$, often called the Dirichlet conditions. Alternatively, we can consider $y(a), y'(a)$, called the Neumann conditions. We could also used some kind of mixed condition, for instance $y + ky'$. Homogeneous boundary conditions are such that $y(a) = y(b) = 0$. In this part of the course, homogeneous boundary conditions are often assumed. Note that we can add a complementary function y_c to the solution, for instance $\overline{y} = y + Ay_1 + By_2$ such that $\overline{y}(a) = \overline{y}(b) = 0$. This would allow us to construct homogeneous boundary conditions even when they are not present *a priori* in the problem. We could also specify initial data, such as solving for $x \geq a$, given y, y' at $x = a$.

To solve the inhomogeneous equation, we want to use eigenfunction expansions such as Fourier series. In order to do this, we must first solve the related eigenvalue problem. In this case, that is

$$\alpha(x)y'' + \beta(x)y' + \gamma(x)y = -\lambda\rho(x)y$$

We must solve this equation with the same boundary conditions as the original problem. This form of equation often arises as a result of applying a separation of variables, particularly for PDEs in several dimensions.

2.2. Sturm–Liouville form

For two complex-valued functions f, g on $[a, b]$, we define the inner product as

$$\langle f, g\rangle = \int_a^b f^\star(x)g(x)\,\mathrm{d}x$$

The eigenvalue problem above greatly simplifies if $\mathcal{L}$ is self-adjoint, that is, if it can be expressed in Sturm–Liouville form:

$$\mathcal{L}y \equiv (-py')' + qy = \lambda wy$$

λ is an eigenvalue, and w is the *weight function*, which must be non-negative.

2.3. Converting to Sturm–Liouville form

Suppose we have the eigenvalue problem

$$\alpha(x)y'' + \beta(x)y' + \gamma(x)y = -\lambda\rho(x)y$$

Multiply this by an integrating factor F to give

$$F\alpha y'' + F\beta y' + F\gamma y = -\lambda F\rho y$$
$$\frac{\mathrm{d}}{\mathrm{d}x}(F\alpha y') - F'\alpha y' - F\alpha' y + F\beta y' + F\gamma y = -\lambda F\rho y$$

To eliminate the y' term, we require $F'\alpha = F(\beta - \alpha')$. Thus,

$$\frac{F'}{F} = \frac{\beta - \alpha'}{\alpha} \implies F = \exp\int^x \frac{\beta - \alpha'}{\alpha}\,\mathrm{d}x$$

and further,

$$(F\alpha y')' + F\gamma y = -\lambda F\rho y$$

hence

$$p = F\alpha$$
$$q = F\gamma$$
$$w = F\rho$$

and $F(x) > 0$ hence $w > 0$.

Example. Consider the Hermite equation,

$$y'' - 2xy' + 2ny = 0$$

In this case,

$$F = \exp\int^x \frac{-2x}{1}\,\mathrm{d}x = e^{-x^2}$$

Then the equation, in Sturm–Liouville form, is

$$\mathcal{L}y \equiv -\left(e^{-x^2}y'\right)' = 2ne^{-x^2}y$$

2.4. Self-adjoint operators

$\mathcal{L}$ is a self-adjoint operator on $[a, b]$ for all pairs of functions y_1, y_2 satisfying appropriate boundary conditions if

$$\langle y_1, \mathcal{L}y_2\rangle = \langle \mathcal{L}y_1, y_2\rangle$$

Written explicitly,

$$\int_a^b y_1^\star(x)\mathcal{L}y_2(x)\,\mathrm{d}x = \int_a^b (\mathcal{L}y_1(x))^\star y_2(x)\,\mathrm{d}x$$

Substituting Sturm–Liouville form into the above,

$$\begin{aligned}\langle y_1, \mathcal{L}y_2\rangle - \langle \mathcal{L}y_1, y_2\rangle &= \int_a^b [-y_1(py_2')' + y_1 q y_2 + y_2(py_1')' - y_2 q y_1]\,\mathrm{d}x \\ &= \int_a^b [-y_1(py_2')' + y_1 q y_2 + y_2(py_1')' - y_2 q y_1]\,\mathrm{d}x \\ &= \int_a^b [-y_1(py_2')' + y_2(py_1')']\,\mathrm{d}x\end{aligned}$$

Adding $-y_1' p y_2' + y_1' p y_2'$,

$$\begin{aligned}&= \int_a^b [-(py_1y_2')' + (py_1'y_2)']\,\mathrm{d}x \\ &= [-py_1y_2' + py_1'y_2]_a^b\end{aligned}$$

which must be zero for an equation in Sturm–Liouville form to be self-adjoint.

2.5. Self-adjoint compatible boundary conditions

- Suppose $y(a) = y(b) = 0$. Then certainly the Sturm–Liouville form of the differential equation is self-adjoint. We could also choose $y'(a) = y'(b) = 0$. Collectively, the act of using homogeneous boundary conditions is known as the *regular* Sturm–Liouville problem.
- Periodic boundary conditions could also be used, such as $y(a) = y(b)$.
- If a and b are singular points of the equation, i.e. $p(a) = p(b) = 0$, this is self-adjoint compatible.
- We could also have combinations of the above properties, one at a and one at b.

2.6. Properties of self-adjoint operators

The following properties hold for any self-adjoint differential operator $\mathcal{L}$.

(i) The eigenvalues λ_n are real.

(ii) The eigenfunctions y_n are orthogonal.

(iii) The y_n are a complete set; they span the space of all functions hence our general solution can be written in terms of these eigenfunctions.

Each property is proven in its own subsection.

2.7. Real eigenvalues

Proof. Suppose we have some eigenvalue λ_n, so $\mathcal{L}y_n = \lambda_n w y_n$. Taking the complex conjugate, $\mathcal{L}y_n^\star = \lambda_n^\star w y_n^\star$, since $\mathcal{L}, w$ are real. Now, consider

$$\int_a^b (y_n^\star \mathcal{L} y_n - y_n \mathcal{L} y_n^\star)\,\mathrm{d}x$$

which must be zero if $\mathcal{L}$ is self-adjoint. This can be written as

$$(\lambda_n - \lambda_n^\star)\int_a^b w y_n^\star y_n\,\mathrm{d}x$$

The integral is nonzero, hence $\lambda_n - \lambda_n^\star = 0$ which implies λ_n is real. Note, if the λ_n are non-degenerate (simple), i.e. with a unique eigenfunction y_n, then $y_n^\star = y_n$ hence they are real. We can in fact show that (for a second-order equation) it is always possible to take linear combinations of eigenfunctions such that the result is linear, for example in the exponential form of the Fourier series. Hence, we can assume that y_n is real. We can further prove that the regular Sturm–Liouville problem must have simple (non-degenerate) eigenvalues λ_n, by considering two possible eigenfunctions u, v for the same λ, and use the expression for self-adjointness. We find $u\mathcal{L}v - (\mathcal{L}u)v = [-p(uv' - u'v)]'$ which contains the Wrońskian. We can integrate and impose homogeneous boundary conditions to get the required result. □

2.8. Orthogonality of eigenfunctions

Suppose $\mathcal{L}y_n = \lambda_n w y_n$, and $\mathcal{L}y_m = \lambda_m w y_m$ where $\lambda_n \neq \lambda_m$. Then, we can integrate to find

$$\int_a^b (y_m \mathcal{L} y_n - y_n \mathcal{L} y_m)\,\mathrm{d}x = (\lambda_n - \lambda_m)\int_a^b w y_n y_m\,\mathrm{d}x = 0 \text{ by self-adjointness}$$

Since $\lambda_n \neq \lambda_m$, we have

$$\forall n \neq m, \int_a^b w y_n y_m\,\mathrm{d}x = 0$$

Hence, y_n and y_m are orthogonal *with respect to* the weight function w on $[a, b]$.

Definition. We define the inner product with respect to w to be

$$\langle f, g\rangle_w = \int_a^b w f^\star g\,\mathrm{d}x$$

Note,

$$\langle f, g\rangle_w = \langle wf, g\rangle = \langle f, wg\rangle$$

Hence, the orthogonality relation becomes

$$\forall n \neq m, \langle y_n, y_m\rangle_w = 0$$

2.9. Eigenfunction expansions

The completeness of the family of eigenfunctions (which is not proven here) implies that we can approximate any 'well-behaved' $f(x)$ on $[a, b]$ by the series

$$f(x) = \sum_{n=1}^{\infty} a_n y_n(x)$$

This is comparable to Fourier series. To find the coefficients a_n, we will take the inner product with an eigenfunction. By orthogonality,

$$\int_a^b w y_m f \,\mathrm{d}x = \sum_{n=1}^{\infty} a_n \int_a^b w y_n y_m \,\mathrm{d}x = a_m \int_a^b w y_m^2 \,\mathrm{d}x$$

Hence,

$$a_n = \frac{\int_a^b w y_n f \,\mathrm{d}x}{\int_a^b w y_n^2 \,\mathrm{d}x}$$

We can normalise eigenfunctions, for instance

$$Y_n(x) = \frac{y_n(x)}{\left(\int_a^b w y_n^2 \,\mathrm{d}x\right)^{\frac{1}{2}}}$$

hence

$$\langle Y_n, Y_m \rangle_w = \delta_{nm}$$

giving an orthonormal set of eigenfunctions. In this case,

$$f(x) = \sum_{n=1}^{\infty} A_n Y_n$$

where

$$A_n = \int_a^b w Y_n f \,\mathrm{d}x$$

Example. Recall Fourier series in Sturm–Liouville form:

$$\mathcal{L} y_n \equiv -\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = \lambda_n y_n$$

where in this case we have

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2$$

2.10. Completeness and Parseval's identity

Consider

$$\int_a^b \left[f(x) - \sum_{n=1}^{\infty} a_n y_n \right]^2 w \, \mathrm{d}x$$

By orthogonality, this is equivalently

$$\int_a^b \left[f^2 - 2f \sum_n a_n y_n + \sum_n a_n^2 y_n^2 \right] w \, \mathrm{d}x$$

Note that the second term can be extracted using the definition of a_n, giving

$$\int_a^b w f^2 \, \mathrm{d}x - \sum_{n=1}^{\infty} a_n^2 \int_a^b w y_n^2 \, \mathrm{d}x$$

If the eigenfunctions are complete, then the result will be zero, showing that the series expansion converges.

$$\int_a^b w f^2 \, \mathrm{d}x = \sum_{n=1}^{\infty} a_n^2 \int_a^b w y_n^2 \, \mathrm{d}x = \sum_{n=1}^{\infty} A_n^2$$

If some eigenfunctions are missing, this is Bessel's inequality:

$$\int_a^b w f^2 \, \mathrm{d}x \geq \sum_{n=1}^{\infty} A_n^2$$

We define the partial sum to be

$$S_N(x) = \sum_{n=1}^{N} a_n y_n$$

with $f(x) = \lim_{N\to\infty} S_N(x)$. Convergence is defined in terms of the mean-square error. In particular, if we have a complete set of eigenfunctions,

$$\varepsilon_N = \int_a^b w[f(x) - S_n(x)]^2 \, \mathrm{d}x \to 0$$

This 'global' definition of convergence is convergence in the mean, not pointwise convergence as in Fourier series. The error in partial sum S_N is minimised by a_n above for the $N = \infty$ expansion.

$$\frac{\partial \varepsilon_N}{\partial a_n} = -2 \int_a^b y_n w \left[f - \sum_{n=1}^{N} a_n y_n \right] \mathrm{d}x = -2 \int_a^b \left(w f y_n - a_n w y_n^2 \right) \mathrm{d}x = 0$$

It is minimal because we can show $\frac{\partial^2 \varepsilon}{\partial a_n^2} = 2 \int_a^b w y_n^2 \, \mathrm{d}x \geq 0$. Thus the a_n given above is the best possible choice for the coefficient at all N.

2.11. Legendre's equation

Legendre's equation is

$$(1-x^2)y'' - 2xy' + \lambda y = 0$$

on $[-1,1]$, with boundary conditions that y is finite at $x = \pm 1$, at the regular singular points of the ODE. This equation is already in Sturm–Liouville form with

$$p = 1-x^2, q = 0, w = 1$$

We seek a power series solution centred on $x = 0$:

$$y = \sum_n c_n x^n$$

Substituting into the differential equation,

$$(1-x^2)\sum_n n(n-1)x_n x^{n-2} - 2x\sum_n c_n c^{n-1} + \lambda \sum_n c_n x^n = 0$$

Equating powers,

$$(n+2)(n+1)c_{n+2} - n(n-1)c_n - 2nc_n + \lambda c_n = 0$$

which gives a recursion relation between c_{n+2} and c_n.

$$c_{n+2} = \frac{n(n+1)-\lambda}{(n+1)(n+2)}c_n$$

Hence, specifying c_0, c_1 gives two independent solutions. In particular,

$$y_{\text{even}} = c_0\left[1 + \frac{(-\lambda)}{2!}x^2 + \frac{(6-\lambda)(-\lambda)}{4!}x^4 + \dots\right]$$

$$y_{\text{odd}} = c_1\left[x + \frac{(2-\lambda)}{3!}x^3 + \dots\right]$$

As $n \to \infty$, $\frac{c_{n+2}}{c_n} \to 1$. So these are geometric series, with radius of convergence $|x| < 1$, hence there is divergence at $x = \pm 1$. So taking a power series does not give a useful solution.

Suppose we chose $\lambda = \ell(\ell+1)$. Then eventually we have n such that the numerator vanishes. In particular, by taking $\lambda = \ell(\ell+1)$, either the series for y_{even} or y_{odd} terminates. These functions are called the Legendre polynomials, denoted $P_\ell(x)$, with the normalisation convention $P_\ell(1) = 1$.

- $\ell = 0, \lambda = 0, P_0(x) = 1$
- $\ell = 1, \lambda = 2, P_1(x) = x$
- $\ell = 2, \lambda = 6, P_2(x) = \frac{3x^2-1}{2}$
- $\ell = 3, \lambda = 12, P_3(x) = \frac{5x^3-3x}{2}$

Note, $P_\ell(x)$ has ℓ zeroes. The polynomials oscillate in parity.

2.12. Properties of Legendre polynomials

Since Legendre polynomials come from a self-adjoint operator, they must have certain conditions, such as orthogonality. For $n \neq m$,

$$\int_{-1}^{1} P_n P_m \,\mathrm{d}x = 0$$

They are also normalisable,

$$\int_{-1}^{1} P_n^2 \,\mathrm{d}x = \frac{2}{2n+1}$$

We can prove this with Rodrigues' formula:

$$P_n(x) = \frac{1}{2^n n!}\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^n (x^2-1)^n$$

Alternatively we could use a generating function:

$$\begin{aligned}\sum_{n=0}^{\infty} P_n(x)t^n = \frac{1}{\sqrt{1-2xt+t^2}} &= 1 + \frac{1}{2}\left(2xt - t^2\right) + \frac{3}{8}\left(2xt-t^2\right)^2 + \dots \\ &= 1 + xt + \frac{1}{2}\left(3x^2-1\right)t^2 + \dots\end{aligned}$$

There are some useful recursion relations.

$$\ell(\ell+1)P_{\ell+1} = (2\ell+1)xP_\ell(x) - \ell P_{\ell-1}(x)$$

Also,

$$(2\ell+1)P_\ell(x) = \frac{\mathrm{d}}{\mathrm{d}x}[P_{\ell+1}(x) - P_{\ell-1}(x)]$$

2.13. Legendre polynomials as eigenfunctions

Any (well-behaved) function on $[-1,1]$ can be expressed as

$$f(x) = \sum_{\ell=0}^{\infty} a_\ell P_\ell(x)$$

where

$$a_\ell = \frac{2\ell+1}{2}\int_{-1}^{1} f(x)P_\ell(x)\,\mathrm{d}x$$

with no boundary conditions (e.g. periodicity conditions) on f.

2.14. Solving inhomogeneous differential equations

This can be thought of as the general case of Fourier series discussed previously.

Consider the problem

$$\mathcal{L}y = f(x) \equiv w(x)F(x)$$

on $x \in [a, b]$ assuming homogeneous boundary conditions. Given eigenfunctions $y_n(x)$ satisfying $\mathcal{L}y_n = \lambda_n w y_n$, we wish to expand this solution as

$$y(x) = \sum_n c_n y_n(x)$$

and

$$F(x) = \sum_n a_n y_n(x)$$

where a_n are known and c_n are unknown:

$$a_n = \frac{\int_a^b wFy_n \,\mathrm{d}x}{\int_a^b wy_n^2 \,\mathrm{d}x}$$

Substituting,

$$\mathcal{L}y = \mathcal{L}\sum_n c_n y_n = w\sum_n c_n \lambda_n y_n = w\sum_n a_n y_n$$

By orthogonality,

$$c_n\lambda_n = a_n \implies c_n = \frac{a_n}{\lambda_n}$$

In particular,

$$y(x) = \sum_{n=1}^{\infty} \frac{a_n}{\lambda_n} y_n(x)$$

We can further generalise; we can permit a driving force, which often induces a linear response term $\tilde{\lambda}wy$.

$$\mathcal{L}y - \tilde{\lambda}wy = f(x)$$

where $\tilde{\lambda}$ is fixed. The solution becomes

$$y(x) = \sum_{n=1}^{\infty} \frac{a_n}{\lambda_n - \tilde{\lambda}} y_n(x)$$

2.15. Integral solutions

Recall that

$$y(x) = \sum_{n=1}^{\infty} \frac{a_n}{\lambda_n} y_n(x) = \sum_n \frac{y_n(x)}{\lambda_n \lambda_n N_n} \int_a^b w(\xi)F(\xi)y_n(\xi)\,\mathrm{d}\xi$$

where

$$N_n = \int w y_n^2 \, \mathrm{d}x$$

This then gives

$$y(x) = \int_a^b \underbrace{\sum_{n=1}^{\infty} \frac{y_n(x) y_n(\xi)}{\lambda_n N_n}}_{G(x,\xi)} \underbrace{w(\xi) F(\xi)}_{f(\xi)} \mathrm{d}\xi = \int_a^b G(x;\xi) f(\xi) \, \mathrm{d}\xi$$

where

$$G(x,\xi) = \sum_{n=1}^{\infty} \frac{y_n(x) y_n(\xi)}{\lambda_n N_n}$$

is the eigenfunction expansion of the Green's function. Note that the Green's function does not depend on f, but only on $\mathcal{L}$ and the boundary conditions. In this sense, it acts like an inverse operator

$$\mathcal{L}^- \equiv \int \mathrm{d}\xi \, G(x,\xi)$$

analogously to how $Ax = b \implies x = A^{-1}b$ for matrix equations.

2.16. Waves on an elastic string

Consider a small displacement $y(x,t)$ on a stretched string with fixed ends at $x = 0$ and $x = L$, that is, with boundary conditions $y(0,t) = y(L,t) = 0$. We can determine the string's motion for specified initial conditions $y(x,0) = p(x)$ and $\frac{\partial y}{\partial t} = q(x)$. We derive the equation of motion governing the motion of the string by balancing forces on a string segment $(x, x + \delta x)$ and take the limit as $\delta x \to 0$. Let T_1 be the tension force acting to the left at angle θ_1 from the horizontal. Analogously, let T_2 be the rightwards tension force at angle θ_2. We assume at any point on the string that $\left|\frac{\partial y}{\partial x}\right| \ll 1$, so the angles of the forces are small. In the x dimension,

$$T_1 \cos\theta_1 = T_2 \cos\theta_2 \implies T_1 \approx T_2 = T$$

So the tension T is constant up to an error of order $O\left(\left|\frac{\partial y}{\partial x}\right|^2\right)$. In the y dimension, since θ are small,

$$F_T = T_2 \sin\theta_2 - T_1 \sin\theta_1 \approx T\left(\left.\frac{\partial y}{\partial x}\right|_{x+\delta x} - \left.\frac{\partial y}{\partial x}\right|_x\right) \approx T\frac{\partial^2 y}{\partial x^2}\delta x$$

By $F = ma$,

$$F_T + F_g = (\mu\delta x)\frac{\partial^2 y}{\partial t^2} = T\frac{\partial^2 y}{\partial x^2}\delta x - g\mu\delta x$$

where F_g is the gravitational force and μ is the linear mass density. We define the wave speed as

$$c = \sqrt{\frac{T}{\mu}}$$

and find

$$\frac{\partial^2 y}{\partial t^2} = \frac{T}{\mu}\frac{\partial^2 y}{\partial x^2} - g = c^2\frac{\partial^2 y}{\partial x^2}$$

We often assume gravity is negligible to produce the pure wave equation

$$\frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} = \frac{\partial^2 y}{\partial x^2}$$

3. Separation of variables

3.1. Separation of variables

We wish to solve the wave equation subject to certain boundary and initial conditions. Consider a possible solution of separable form:

$$y(x,t) = X(x)T(t)$$

Substituting into the wave equation,

$$\frac{1}{c^2}\ddot{y} = y'' \implies \frac{1}{c^2}X\ddot{T} = X''T$$

Then

$$\frac{1}{c^2}\frac{\ddot{T}}{T} = \frac{X''}{X}$$

However, $\frac{\ddot{T}}{T}$ depends only on t and $\frac{X''}{X}$ depends only on x. Thus, both sides must be equal to some *separation constant* $-\lambda$.

$$\frac{1}{c^2}\frac{\ddot{T}}{T} = \frac{X''}{X} = -\lambda$$

Hence,

$$X'' + \lambda X = 0; \quad \ddot{T} + \lambda c^2 T = 0$$

3.2. Boundary conditions and normal modes

We will begin by first solving the spatial part of the solution. One of $\lambda > 0, \lambda < 0, \lambda = 0$ must be true. The boundary conditions restrict the possible λ.

(i) First, suppose $\lambda < 0$. Take $\chi^2 = -\lambda$. Then,

$$X(x) = Ae^{\chi x} + Be^{-\chi x} = C\cosh(\chi x) + D\sinh(\chi x)$$

The boundary conditions are $x(0) = x(L) = 0$, so only the trivial solution is possible: $C = D = 0$.

(ii) Now, suppose $\lambda = 0$. Then

$$X(x) = Ax + B$$

Again, the boundary conditions impose $A = B = 0$ giving only the trivial solution.

(iii) Finally, the last possibility is $\lambda > 0$.

$$X(x) = A\cos\left(\sqrt{\lambda}x\right) + B\sin\left(\sqrt{\lambda}x\right)$$

The boundary conditions give

$$A = 0; \quad B\sin\left(\sqrt{\lambda}L\right) = 0 \implies \sqrt{\lambda}L = n\pi$$

The following are the eigenfunctions and eigenvalues.

$$X_n(x) = B_n \sin \frac{n\pi x}{L}; \quad \lambda_n = \left(\frac{n\pi}{L}\right)^2$$

These are also called the 'normal modes' of the system. The spatial shape in x does not change in time, but the amplitude may vary. The fundamental mode is the lowest frequency of vibration, given by

$$n = 1 \implies \lambda_1 = \frac{\pi^2}{L^2}$$

The second mode is the first overtone, and is given by

$$n = 2 \implies \lambda_2 = \frac{4\pi^2}{L^2}$$

3.3. Initial conditions and temporal solutions

Substituting λ_n into the time ODE,

$$\ddot{T} + \frac{n^2\pi^2c^2}{L^2} T = 0$$

Hence,

$$T_n(t) = C_n \cos \frac{n\pi ct}{L} + D_n \sin \frac{n\pi ct}{L}$$

Therefore, a specific solution of the wave equation satisfying the boundary conditions is (absorbing the B_n into the C_n, D_n):

$$y_n(x,t) = T_n(t)X_n(x) = \left(C_n \cos \frac{n\pi ct}{L} + D_n \sin \frac{n\pi ct}{L}\right) \sin \frac{n\pi x}{L}$$

To find a particular solution for a given set of initial conditions, we must consider a linear superposition of all possible y_n.

$$y(x,t) = \sum_{n=1}^{\infty} \left(C_n \cos \frac{n\pi ct}{L} + D_n \sin \frac{n\pi ct}{L}\right) \sin \frac{n\pi x}{L}$$

By construction, this $y(x,t)$ satisfies the boundary conditions, so now we can impose the initial conditions.

$$y(x,0) = p(x) = \sum_{n=1}^{\infty} C_n \sin \frac{n\pi x}{L}$$

We can find the C_n using standard Fourier series techniques, since this is exactly a half-range sine series. Further,

$$\frac{\partial y(x,0)}{\partial t} = q(x) = \sum_{n=1}^{\infty} \frac{n\pi c}{L} D_n \sin \frac{n\pi x}{L}$$

Again we can solve for the D_n in a similar way. In particular,

$$C_n = \frac{2}{L}\int_0^L p(x)\sin\frac{n\pi x}{L}\,\mathrm{d}x$$

$$D_n = \frac{2}{n\pi c}\int_0^L q(x)\sin\frac{n\pi x}{L}\,\mathrm{d}x$$

Example. Consider the initial condition of a see-saw wave parametrised by ξ, and let $L = 1$. This can be visualised as plucking the string at position ξ.

$$y(x,0) = p(x) = \begin{cases} x(1-\xi) & 0 \le x < \xi \\ \xi(1-x) & \xi \le x < 1 \end{cases}$$

We also define

$$\frac{\partial y(x,0)}{\partial t} = q(x) = 0$$

The Fourier series for p is given by

$$C_n = \frac{2\sin n\pi\xi}{(n\pi)^2}; \quad D_n = 0$$

Hence the solution to the wave equation is

$$y(x,t) = \sum_{n=1}^{\infty} \frac{2}{(n\pi)^2}\sin n\pi\xi \sin n\pi x \cos n\pi ct$$

3.4. Separation of variables methodology

A general strategy for solving higher-dimensional partial differential equations is as follows.

(i) Obtain a linear PDE system, using boundary and initial conditions.

(ii) Separate variables to yield decoupled ODEs.

(iii) Impose homogeneous boundary conditions to find eigenvalues and eigenfunctions.

(iv) Use these eigenvalues (constants of separation) to find the eigenfunctions in the other variables.

(v) Sum over the products of separable solutions to find the general series solution.

(vi) Determine coefficients for this series using the initial conditions.

Example. We will solve the wave equation instead in characteristic coordinates. Recall the sine and cosine summation identities:

$$\begin{aligned} y(x,t) &= \frac{1}{2}\sum_{n=1}^{\infty}\left[\left(C_n \sin\frac{n\pi}{L}(x-ct) + D_n \cos\frac{n\pi}{L}(x-ct)\right)\right. \\ &\quad \left. + \left(C_n \sin\frac{n\pi}{L}(x+ct) - D_n \cos\frac{n\pi}{L}(x+ct)\right)\right] \\ &= f(x-ct) + g(x+ct) \end{aligned}$$

The standing wave solution can be interpreted as a superposition of a right-moving wave and a left-moving wave. A special case is $q(x) = 0$, implying $f = g = \frac{1}{2}p$. Then,

$$y(x,t) = \frac{1}{2}[p(x-ct) + p(x+ct)]$$

3.5. Energy of oscillations

A vibrating string has kinetic energy due to its motion.

$$\text{Kinetic energy} = \frac{1}{2}\mu\int_0^L \left(\frac{\partial y}{\partial t}\right)^2 \mathrm{d}x$$

It has potential energy given by

$$\text{Potential energy} = T\Delta x = T\int_c^T \left(\sqrt{1+\left(\frac{\partial y}{\partial x}\right)^2} - 1\right)\mathrm{d}x \approx \frac{1}{2}T\int_0^L \left(\frac{\partial y}{\partial x}\right)^2 \mathrm{d}x$$

assuming that the disturbances on the string are small, that is, $\left|\frac{\partial y}{\partial x}\right| \ll 1$. The total energy on the string, given $c^2 = T/\mu$, is given by

$$E = \frac{1}{2}\mu\int_0^L \left[\left(\frac{\partial y}{\partial t}\right)^2 + c^2\left(\frac{\partial y}{\partial x}\right)^2\right]\mathrm{d}x$$

Substituting the solution, using the orthogonality conditions,

$$\begin{aligned} E &= \frac{1}{2}\mu\sum_{n=1}^{\infty}\int_0^L \left[-\left(\frac{n\pi c}{L}C_n \sin\frac{n\pi ct}{L} + \frac{n\pi c}{L}D_n \cos\frac{n\pi ct}{L}\right)^2 \sin^2\frac{n\pi x}{L}\right. \\ &\quad \left. + c^2\left(C_n \cos\frac{n\pi ct}{L} + D_n \sin\frac{n\pi ct}{L}\right)^2 \frac{n^2\pi^2}{L^2}\cos^2\frac{n\pi x}{L}\right]\mathrm{d}x \\ &= \frac{1}{4}\mu\sum_{n=1}^{\infty}\frac{n^2\pi^2c^2}{L}(C_n^2 + D_n^2) \end{aligned}$$

which is an analogous result to Parseval's theorem. This is true since

$$\int \cos^2 \frac{n\pi x}{L}\,\mathrm{d}x = \frac{1}{2}$$

and $\cos^2 + \sin^2 = 1$. We can think of this energy as the sum over all the normal modes of the energy in that specific mode. Note that this quantity is constant over time.

3.6. Wave reflection and transmission

The travelling wave has left-moving and right-moving modes. A *simple harmonic* travelling wave is

$$y = \operatorname{Re}\left[Ae^{i\omega(t-x/c)}\right] = A\cos\left[\omega(t-x/c)+\phi\right]$$

where the phase ϕ is equal to $\arg A$, and the wavelength λ is $2\pi c/\omega$. In further discussion, we assume only the real part is used. Consider a density discontinuity on the string at $x = 0$ with the following properties.

$$\mu = \begin{cases} \mu_- & \text{for } x < 0 \\ \mu_+ & \text{for } x > 0 \end{cases} \implies c = \begin{cases} c_- = \sqrt{\frac{T}{\mu_-}} & \text{for } x < 0 \\ c_+ = \sqrt{\frac{T}{\mu_+}} & \text{for } x > 0 \end{cases}$$

assuming a constant tension T. As a wave from the negative direction approaches the discontinuity, some of the wave will be reflected, given by $Be^{i\omega(t+x/c_-)}$, and some of the wave will be transmitted, given by $De^{i\omega(t-x/c_+)}$. The boundary conditions at $x = 0$ are

(i) y is continuous for all t (the string does not break), so

$$A + B = D \tag{$*$}$$

(ii) The forces balance, $T\left.\frac{\partial y}{\partial x}\right|_{x=0^-} = T\left.\frac{\partial y}{\partial x}\right|_{x=0^+}$ which means $\frac{\partial y}{\partial x}$ must be continuous for all t. This gives

$$\frac{-i\omega A}{c_-} + \frac{i\omega B}{c_-} = \frac{-i\omega D}{c_+} \tag{$\dagger$}$$

We can eliminate B from $(*)$ by subtracting $\frac{c_-}{i\omega}(\dagger)$.

$$2A = D + D\frac{c_-}{c_+} = \frac{D}{c_+}(c_+ + c_-)$$

Hence, given A, we have the solution for the transmitted amplitude and reflected amplitude to be

$$D = \frac{2c_+}{c_- + c_+}A; \quad B = \frac{c_+ - c_-}{c_- + c_+}$$

In general A, B, D are complex, hence different phase shifts are possible.

There are a number of limiting cases, for example

(i) If $c_- = c_+$ we have $D = A$ and $B = 0$ so we have full transmission and no reflection.

(ii) (Dirichlet boundary conditions) If $\frac{\mu_+}{\mu_-} \to \infty$, this models a fixed end at $x = 0$. We have $\frac{c_+}{c_-} \to 0$ giving $D = 0$ and $B = -A$. Notice that the reflection has occurred with opposite phase, $\phi = \pi$.

(iii) (Neumann boundary conditions) Consider $\frac{\mu_+}{\mu_-} \to 0$, this models a free end. Then $\frac{c_+}{c_-} \to \infty$ giving $D = 2A$, $B = A$. This gives total reflection but with the same phase.

3.7. Wave equation in plane polar coordinates

Consider the two-dimensional wave equation for $u(r, \theta, t)$ given by

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \nabla^2 u$$

with boundary conditions at $r = 1$ on a unit disc given by

$$u(1, \theta, t) = 0$$

and initial conditions for $t = 0$ given by

$$u(r, \theta, 0) = \phi(r, \theta); \quad \frac{\partial u}{\partial t} = \psi(r, \theta)$$

Suppose that this equation is separable. First, let us consider temporal separation. Suppose that

$$u(r, \theta, t) = T(t)V(r, \theta)$$

Then we have

$$\ddot{T} + \lambda c^2 T = 0; \quad \nabla^2 V + \lambda V = 0$$

In plane polar coordinates, we can write the spatial equation as

$$\frac{\partial^2 V}{\partial r^2} + \frac{1}{r}\frac{\partial V}{\partial r} + \frac{1}{r^2}\frac{\partial^2 V}{\partial \theta^2} + \lambda V = 0$$

We will perform another separation, supposing

$$V(r, \theta) = R(r)\Theta(\theta)$$

to give

$$\Theta'' + \mu\Theta = 0; \quad r^2 R'' + rR' + \big(\lambda r^2 - \mu\big)R = 0$$

where λ, μ are the separation constants. The polar solution is constrained by periodicity $\Theta(0) = \Theta(2\pi)$, since we are working on a disc. We also consider only $\mu > 0$. The eigenvalue is then given by $\mu = m^2$, where $m \in \mathbb{N}$.

$$\Theta_m(\theta) = A_m \cos m\theta + B_m \sin m\theta$$

Or, in complex exponential form,

$$\Theta_m(\theta) = C_m e^{im\theta}; \quad m \in \mathbb{Z}$$

3.8. Bessel's equation

We can solve the radial equation (in the previous subsection) by converting it first into Sturm–Liouville form, which can be accomplished by dividing by r.

$$\frac{\mathrm{d}}{\mathrm{d}r}(rR') - \frac{m^2}{r} = -\lambda rR$$

where $p(r) = r, q(r) = \frac{m^2}{r}, w(r) = r$, with self-adjoint boundary conditions with $R(1) = 0$. We will require R is bounded at $R(0)$, and since $p(0) = 0$ there is a regular singular point at $r = 0$. This particular equation for R is known as Bessel's equation. We will first substitute $z \equiv \sqrt{\lambda} r$, then we find the usual form of Bessel's equation,

$$z^2 \frac{\mathrm{d}^2 R}{\mathrm{d}z^2} + z \frac{\mathrm{d}R}{\mathrm{d}z} + (z^2 - m^2)R = 0$$

We can use the method of Frobenius by substituting the following power series:

$$R = z^p \sum_{n=0}^{\infty} a_n z^n$$

to find

$$\sum_{n=0}^{\infty} \left[a_n (n+p)(n+p-1) z^{n+p} + (n+p) z^{n+p} + z^{n+p+2} + m^2 z^{n+p} \right] = 0$$

Equating powers of z, we can find the indicial equation

$$p^2 - m^2 = 0 \implies p = m, -m$$

The regular solution, given by $p = m$, has recursion relation

$$(n+m)^2 a_n + a_{n-2} - m^2 a_n = 0$$

which gives

$$a_n = \frac{-1}{n(n+2m)} a_{n-2}$$

Hence, we can find

$$a_{2n} = a_0 \frac{(-1)^n}{2^{2n} n! (n+m)(n+m-1) \dots (m+1)}$$

If, by convention, we let

$$a_0 = \frac{1}{2^m m!}$$

we can then write the *Bessel function of the first kind* by

$$J_m(z) = \left(\frac{z}{2}\right)^m \sum_{n=0}^{\infty} \frac{(-1)^n}{n!(n+m)!} \left(\frac{z}{2}\right)^{2n}$$

3.9. Asymptotic behaviour of Bessel functions

If z is small, the leading-order behaviour of $J_m(z)$ is

$$J_0(z) \approx 1$$
$$J_m(z) \approx \frac{1}{m!}\left(\frac{z}{2}\right)^m$$

Now, let us consider large z. In this case, the function becomes oscillatory;

$$J_m(z) \approx \sqrt{\frac{2}{\pi z}} \cos\left(z - \frac{m\pi}{2} - \frac{\pi}{4}\right)$$

3.10. Zeroes of Bessel functions

We can see from the asymptotic behaviour that there are infinitely many zeroes of the Bessel functions of the first kind as $z \to \infty$. We define j_{mn} to be the nth zero of J_m, for $z > 0$. Approximately,

$$\cos\left(z - \frac{m\pi}{2} - \frac{\pi}{4}\right) = 0 \implies z - \frac{m\pi}{2} - \frac{\pi}{4} = n\pi - \frac{\pi}{2}$$

Hence

$$z \approx n\pi + \frac{m\pi}{2} - \frac{\pi}{4} \equiv \tilde{j}_{mn}$$

3.11. Solving the vibrating drum

Recall that the radial solutions become

$$R_m(z) = R_m(\sqrt{\lambda}x) = AJ_m(\sqrt{\lambda}x) + BY_m(\sqrt{\lambda}x)$$

Imposing the boundary condition of boundedness at $r = 0$, we must have $B = 0$. Further imposing $r = 1$ and $R = 0$ gives $J_m(\sqrt{\lambda}) = 0$. These zeroes occur at $j_{mn} \approx n\pi + \frac{m\pi}{2} - \frac{\pi}{4}$. Hence, the eigenvalues must be j_{mn}^2. Therefore, the spatial solution is

$$V_{mn}(r,\theta) = \Theta_m(\theta)R_{mn}(\sqrt{\lambda_{mn}}r) = (A_{mn}\cos m\theta + B_{mn}\sin m\theta)J_m(j_{mn}r)$$

The temporal solution is

$$\ddot{T} = -\lambda c T \implies T_{mn}(t) = \cos(j_{mn}ct), \sin(j_{mn}ct)$$

Combining everything together, the full solution is

$$\begin{aligned} u(r,\theta,t) = &\sum_{n=1}^{\infty} J_0(j_{0n}r)(A0n\cos j_{0n}ct + C_{0n}\sin j_{0n}ct) \\ &+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} J_m(j_{mn}r)(A_{mn}\cos m\theta + B_{mn}\sin m\theta)\cos j_{mn}ct \\ &+ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} J_m(j_{mn}r)(C_{mn}\cos m\theta + D_{mn}\sin m\theta)\sin j_{mn}ct \end{aligned}$$

Now, we impose the boundary conditions

$$u(r,\theta,0) = \phi(r,\theta) = \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} J_m(j_{mn}r)(A_{mn}\cos m\theta + B_{mn}\sin m\theta)$$

and

$$\frac{\partial u}{\partial t}(r,\theta,0) = \psi(r,\theta) = \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} j_{mn}cJ_m(j_{mn}r)(C_{mn}\cos m\theta + D_{mn}\sin m\theta)$$

We need to find the coefficients by multiplying by J_m, cos, sin and using the orthogonality relations, which are

$$\int_0^1 J_m(j_{mn}r)J_m(j_{mk}r)r\,\mathrm{d}r = \frac{1}{2}[J_m'(j_{mn})]^2\delta_{nk} = \frac{1}{2}[J_{m+1}(j_{mn})]^2\delta_{nk}$$

by using a recursion relation of the Bessel functions. We can then integrate to obtain the coefficients A_{mn}.

$$\int_0^{2\pi}\mathrm{d}\theta\cos p\theta\int_0^1 r\,\mathrm{d}r\,J_p(j_{pq}r)\phi(r,\theta) = \frac{\pi}{2}\big[J_{p+1}(j_{pq})\big]^2A_{pq}$$

where the $\frac{\pi}{2}$ coefficient is 2π for $p=0$. We can find analogous results for the B_{mn}, C_{mn}, D_{mn}.

Example. Consider an initial radial profile $u(r,\theta,0) = \phi(r) = 1-r^2$. Then, $m=0, B_{mn}=0$ for all m and $A_{mn}=0$ for all $m\neq 0$. Then

$$\frac{\partial u}{\partial t}(r,0,0) = 0$$

hence $C_{mn}, D_{mn} = 0$. We just now need to find

$$A_{0n} = \frac{2}{J_0(j_{0n})^2}\int_0^1 J_0(j_{0n}r)(1-r)^2r\,\mathrm{d}r = \frac{2}{J_0(j_{0n})^2}\frac{J_2(j_{0n})}{j_{0n}^2} \approx \frac{J_2(j_{0n})}{n} \text{ as } n\to\infty$$

Then the approximate solution is

$$u(r,\theta,t) = \sum_{n=1}^{\infty} A_{0n}J_0(j_{0n}r)\cos j_{0n}ct$$

The fundamental frequency is $\omega_d = j_{01}c\frac{2}{d} \approx 4.8\frac{c}{d}$ where d is the diameter of the drum. Comparing this to a string with length d, this has a fundamental frequency of $\omega_s = \frac{\pi c}{d} \approx 0.77\omega_d$.

3.12. Diffusion equation derivation with Fourier's law

In a volume V, the overall heat energy Q is given by

$$Q = \int_V c_V \rho \theta \, \mathrm{d}V$$

where c_V is the specific heat of the material, ρ is the mass density, and θ is the temperature. The rate of change due to heat flow is

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \int_V c_V \rho \frac{\partial \theta}{\partial t} \, \mathrm{d}V$$

Fourier's law for heat flow is

$$q = -k \nabla \theta$$

where q is the heat flux. We will integrate this over the surface $S = \partial V$, giving

$$-\frac{\mathrm{d}Q}{\mathrm{d}t} = \int_S q \cdot \hat{n} \, \mathrm{d}S$$

The negative sign is due to the normals facing outwards. This is exactly

$$-\frac{\mathrm{d}Q}{\mathrm{d}t} = \int_S (-k \nabla \theta) \cdot \hat{n} \, \mathrm{d}S = \int_V -k \nabla^2 \theta \, \mathrm{d}V$$

Equating these two forms for $\frac{\mathrm{d}Q}{\mathrm{d}t}$, we find

$$\int_V (c_V \rho \frac{\partial \theta}{\partial t} - k \nabla^2 \theta) \, \mathrm{d}V = 0$$

Since V was arbitrary, the integrand must be zero. So we have

$$\frac{\partial \theta}{\partial t} - \frac{k}{c_V \rho} \nabla^2 \theta = 0$$

Let $D = \frac{k}{c_V \rho}$ be the diffusion constant. Then we have the diffusion equation

$$\frac{\partial \theta}{\partial t} - D \nabla^2 \theta = 0$$

3.13. Diffusion equation derivation with statistical dynamics

We can derive this equation in another way, using statistical dynamics. Gas particles diffuse by scattering every fixed time step Δt with probability density function $p(\xi)$ of moving by a displacement ξ. On average, we have

$$\langle \xi \rangle = \int p(\xi) \xi \, \mathrm{d}\xi = 0$$

since there is no bias the direction in which any given particle is travelling. Suppose that the probability density function after $N\Delta t$ time is described by $P_{N\Delta t}(x)$. Then, for the next time step,

$$P_{(N+1)\Delta t}(x) = \int_{-\infty}^{\infty} p(\xi) P_{N\Delta t}(x-\xi)\,\mathrm{d}\xi$$

Using the Taylor expansion,

$$\begin{aligned} P_{(N+1)\Delta t}(x) &\approx \int_{-\infty}^{\infty} p(\xi)\left[P_{N\Delta t}(x) + P'_{N\Delta t}(x)(-\xi) + P''_{N\Delta t}(x)\frac{\xi^2}{2} + \cdots\right]\mathrm{d}\xi \\ &\approx P_{N\Delta t}(x) - P'_{N\Delta t}(x)\langle\xi\rangle + P''_{N\Delta t}(x)\frac{\langle\xi^2\rangle}{2} + \cdots \\ &\approx P_{N\Delta t}(x) + P''_{N\Delta t}(x)\frac{\langle\xi^2\rangle}{2} + \cdots \end{aligned}$$

since $\int p(\xi)\,\mathrm{d}\xi = 1$. Identifying $P_{N\Delta t}(x) = P(x, N\Delta t)$, we can write

$$P(x,(N+1)\Delta t) - P(x,N\Delta t) = \frac{\partial^2}{\partial x^2}P(x,N\Delta t)\frac{\langle\xi^2\rangle}{2}$$

Assuming that the variance $\frac{\langle\xi^2\rangle}{2}$ is proportional to $D\Delta t$, then for small Δt, we find

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2}$$

which is exactly the diffusion equation.

3.14. Similarity solutions

The characteristic relation between the variance and time suggests that we seek solutions with a dimensionless parameter. If we can a change of variables of the form $\theta(\eta) = \theta(x,t)$, then it will likely be easier to solve. Consider

$$\eta \equiv \frac{x}{2\sqrt{Dt}}$$

Then,

$$\frac{\partial\theta}{\partial t} = \frac{\partial\eta}{\partial t}\frac{\partial\theta}{\partial\eta} = \frac{-1}{2}\frac{x}{\sqrt{D}t^{3/2}}\theta' = \frac{-1}{2}\frac{\eta}{t}\theta'$$

and

$$D\frac{\partial^2\theta}{\partial x^2} = D\frac{\partial}{\partial x}\left(\frac{\partial\eta}{\partial x}\frac{\partial\theta}{\partial\eta}\right) = D\frac{\partial}{\partial x}\left(\frac{1}{2\sqrt{Dt}}\theta'\right) = \frac{D}{4Dt}\theta'' = \frac{1}{4t}\theta''$$

Substituting into the diffusion equation,

$$\theta'' = -2\eta\theta'$$

Let $\psi = \theta'$. Then

$$\frac{\psi'}{\psi} = -2\eta \implies \ln\psi = -\eta^2 + \text{constant}$$

Then, choosing a constant of $c\frac{2}{\sqrt{\pi}}$,

$$\psi = c\frac{2}{\sqrt{\pi}}e^{-\eta^2} \implies \theta(\eta) = c\frac{2}{\sqrt{\pi}}\int_0^\eta e^{-u^2}\,\mathrm{d}u = c\,\mathrm{erf}(\eta) = c\,\mathrm{erf}(\frac{x}{2\sqrt{Dt}})$$

where

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2}\,\mathrm{d}u$$

This describes discontinuous initial conditions that spread over time.

3.15. Heat conduction in a finite bar

Suppose we have a bar of length $2L$ with $-L \le x \le L$ and initial temperature

$$\theta(x,0) = H(x) = \begin{cases} 1 & \text{if } 0 \le x \le L \\ 0 & \text{if } -L \le x < 0 \end{cases}$$

with boundary conditions $\theta(L,t) = 1$, $\theta(-L,t) = 0$. Currently the boundary conditions are not homogeneous, so Sturm–Liouville theory cannot be used directly. If we can identify a steady-state solution (time-independent) that reflects the late-time behaviour, then we can turn it into a homogeneous set of boundary conditions. We will try a solution of the form

$$\theta_s(x) = Ax + B$$

since this certainly satisfies the diffusion equation. To satisfy the boundary conditions,

$$A = \frac{1}{2L}; \quad B = \frac{1}{2}$$

Hence we have a solution

$$\theta_s = \frac{x+L}{2L}$$

We will subtract this solution from our original equation for θ, giving

$$\hat{\theta}(x,t) = \theta(x,t) - \theta_s(x)$$

with homogeneous boundary conditions

$$\hat{\theta}(-L,t) = \hat{\theta}(L,t) = 0$$

and initial conditions

$$\theta(x,0) = H(x) - \frac{x+L}{2L}$$

We will now separate variables in the usual way. We will consider the ansatz

$$\hat{\theta}(x,t) = X(x)T(t) \implies X'' = -\lambda X; \dot{T} = -D\lambda T$$

The boundary conditions imply $\lambda > 0$ and give the Fourier modes $X(x) = A\cos\sqrt{\lambda}x + B\sin\sqrt{\lambda}x$. For $\cos\sqrt{\lambda}L = 0$, we require $\sqrt{\lambda_m} = \frac{m\pi}{2L}$ for m odd. Also, $\sin\sqrt{\lambda}L = 0$ gives $\sqrt{\lambda_n} = \frac{n\pi}{L}$ for n even. Since $\hat{\theta}$ is odd due to our initial conditions, we can take

$$X_n = B_n \sin\frac{n\pi x}{L}; \quad \lambda_n = \frac{n^2\pi^2}{L^2}$$

Substituting into $\dot{T} = -D\lambda T$, we have

$$T_n(t) = c_n \exp\left(-\frac{Dn^2\pi^2}{L^2}t\right)$$

In general, the solution is

$$\hat{\theta}(x,t) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}\exp\left(-\frac{Dn^2\pi^2}{L^2}t\right)$$

3.16. Particular solution to diffusion equation

Recall that

$$\hat{\theta}(x,t) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}\exp\left(-\frac{Dn^2\pi^2}{L^2}t\right)$$

At $t = 0$, we have a pure Fourier sine series. We can then impose the initial conditions, to give

$$b_n = \frac{1}{L}\int_{-L}^{L} \hat{\phi}(x,0)\sin\frac{n\pi x}{L}\,\mathrm{d}x$$

where

$$\hat{\phi}(x,0) = H(x) - \frac{x+L}{2L}$$

Hence, we can use the half-range sine series and find

$$b_n = \underbrace{\frac{2}{L}\int_0^L \left(H(x) = \frac{1}{2}\right)\sin\frac{n\pi x}{L}\,\mathrm{d}x}_{\text{square wave}/2} - \underbrace{\frac{2}{L}\frac{x}{2L}\sin\frac{n\pi x}{L}\,\mathrm{d}x}_{\text{sawtooth}/2L}$$

which gives

$$b_n = \frac{2}{(2m-1)\pi} - \frac{(-1)^{n+1}}{n\pi}$$

where $n = 2m - 1$, and the first term vanishes for n even. For n odd or even, we find the same result

$$b_n = \frac{1}{n\pi}$$

Hence

$$\hat{\theta}(x,t) = \sum_{n=1}^{\infty} \frac{1}{n\pi} \sin \frac{n\pi x}{L} e^{-D\frac{n^2\pi^2}{L^2}t}$$

For the inhomogeneous boundary conditions,

$$\theta(x,t) = \frac{x+L}{2L} + \sum_{n=1}^{\infty} \frac{1}{n\pi} \sin \frac{n\pi x}{L} e^{-D\frac{n^2\pi^2}{L^2}t}$$

The similarity solution $\frac{1}{2}\left(1 + \operatorname{erf}(\frac{x}{2\sqrt{Dt}})\right)$ is a good fit for early t, but it does not necessarily satisfy the boundary conditions, so for large t it is a bad approximation.

3.17. Laplace's equation

Laplace's equation is

$$\nabla^2 \phi = 0$$

This equation describes (among others) steady-state heat flow, potential theory $F = -\nabla\phi$, and incompressible fluid flow $v = \nabla\phi$. The equation is solved typically on a domain D, where boundary conditions are specified often on the boundary surface. The Dirichlet boundary conditions fix ϕ on the boundary surface ∂D. The Neumann boundary conditions fix $\hat{n} \cdot \nabla\phi$ on ∂D.

3.18. Laplace's equation in three-dimensional Cartesian coordinates

In $\mathbb{R}^3$ with Cartesian coordinates, Laplace's equation becomes

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = 0$$

We seek separable solutions in the usual way:

$$\phi(x,y,z) = X(x)Y(y)Z(z)$$

Substituting,

$$X''YZ + XY''Z + XYZ'' = 0$$

Dividing by XYZ as usual,

$$\begin{aligned}
\frac{X''}{X} &= \frac{-Y''}{Y} - \frac{Z''}{Z} = -\lambda_\ell \\
\frac{Y''}{Y} &= \frac{-Z''}{Z} - \frac{X''}{X} = -\lambda_m \\
\frac{Z''}{Z} &= \frac{-X''}{X} - \frac{Y''}{Y} = -\lambda_n = \lambda_\ell + \lambda_m
\end{aligned}$$

From the eigenmodes, our general solution will be of the form

$$\phi(x,y,z) = \sum_{\ell,m,n} a_{\ell mn} X_\ell(x) Y_m(y) Z_n(z)$$

Consider steady ($\frac{\partial \phi}{\partial t} = 0$) heat flow in a semi-infinite rectangular bar, with boundary conditions $\phi = 0$ at $x = 0$, $x = a$, $y = 0$ and $y = b$; and $\phi = 1$ at $z = 0$ and $\phi \to 0$ as $z \to \infty$. We will solve for each eigenmode successively. First, consider $X'' = -\lambda_\ell X$ with $X(0) = X(a) = 0$. This gives

$$\lambda_\ell = \frac{l^2\pi^2}{a^2}; \quad X_\ell = \sin\frac{\ell\pi x}{a}$$

where $\ell > 0, \ell \in \mathbb{N}$. By symmetry,

$$\lambda_m = \frac{m^2\pi^2}{b^2}; \quad Y_m = \sin\frac{m\pi y}{b}$$

For the z mode,

$$Z'' = -\lambda_n Z = (\lambda_\ell + \lambda_m) Z = \pi^2\left(\frac{\ell^2}{a^2} + \frac{m^2}{b^2}\right) Z$$

Since $\phi \to 0$ as $z \to \infty$, the growing exponentials must vanish. Therefore,

$$Z_{\ell m} = \exp\left[-\left(\frac{\ell^2}{a^2} + \frac{m^2}{b^2}\right)^{1/2} \pi z\right]$$

Thus the general solution is

$$\phi(x,y,z) = \sum_{\ell,m} a_{\ell m} \sin\frac{\ell\pi x}{a} \sin\frac{m\pi y}{b} \exp\left[-\left(\frac{\ell^2}{a^2} + \frac{m^2}{b^2}\right)^{1/2} \pi z\right]$$

Now, we will fix $a_{\ell m}$ using $\phi(x,y,0) = 1$ using the Fourier sine series.

$$a_{\ell m} = \frac{2}{b}\int_0^b \frac{2}{a}\int_0^a \underbrace{1 \sin\frac{\ell\pi x}{a}}_{\text{square wave}} \underbrace{\sin\frac{m\pi y}{b}}_{\text{square wave}} \mathrm{d}x\,\mathrm{d}y$$

So only the odd terms remain, giving

$$a_{\ell m} = \frac{4a}{a(2k-1)\pi} \cdot \frac{4b}{b(2p-1)\pi}$$

where $\ell = 2k - 1$ is odd and $m = 2p - 1$ is odd. Simplifying,

$$a_{\ell m} = \frac{16}{\pi^2 \ell m} \quad \text{for } \ell, m \text{ odd}$$

So the heat flow solution is

$$\phi(x,y,z) = \sum_{\ell,m \text{ odd}} \frac{16}{\pi^2\ell m} \sin\frac{\ell\pi x}{a} \sin\frac{\ell\pi y}{b} \exp\left[-\left(\frac{\ell^2}{a^2} + \frac{m^2}{b^2}\right)^{1/2} \pi z\right]$$

As z increases, every contribution but the lowest mode will be very small. So low ℓ, m dominate the solution.

3.19. Laplace's equation in plane polar coordinates

In plane polar coordinates, Laplace's equation becomes

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 \phi}{\partial \theta^2}=0$$

Consider a separable form of the answer, given by

$$\phi(r,\theta)=R(r)\Theta(\theta)$$

We then have

$$\Theta''+\mu\Theta=0;\quad r(rR')'-\mu R=0$$

The polar equation can be solved easily by considering periodic boundary conditions. This gives $\mu=m^2$ and the eigenmodes

$$\Theta_m(\theta)=\cos m\theta,\sin m\theta$$

The radial equation is *not* Bessel's equation, since there is no second separation constant. We simply have

$$r(rR')'-m^2R=0$$

We will try a power law solution, $r=\alpha r^\beta$. We find

$$\beta^2-m^2=0\implies \beta=\pm m$$

So the eigenfunctions are

$$R_m(r)=r^m,r^{-m}$$

which is one regular solution at the origin and one singular solution. In the case $m=0$, we have

$$(rR')=0\implies rR'=\text{constant}\implies R=\log r$$

So

$$R_0(r)=\text{constant},\log r$$

The general solution is therefore

$$\phi(r,\theta)=\frac{a_0}{2}+c_0\log r+\sum_{m=1}^{\infty}(a_m\cos m\theta+b_m\sin m\theta)r^m+\sum_{m=1}^{\infty}(c_m\cos m\theta+d_m\sin m\theta)r^{-m}$$

Example. Consider a soap film on a unit disc. We wish to solve Laplace's equation with a vertically distorted circular wire of radius $r=1$ with boundary conditions $\phi(1,\theta)=f(\theta)$. The z displacement of the wire produces the $f(\theta)$ term. We wish to find $\phi(r,\theta)$ for $r<1$, assuming regularity at $r=0$. Then, $c_m=d_m=0$ and the solution is of the form

$$\phi(r,\theta)=\frac{a_0}{2}+\sum_{m=1}^{\infty}(a_m\cos m\theta+b_m\sin m\theta)r^m$$

At $r = 1$,

$$\phi(1,\theta) = f(\theta) = \frac{a_0}{2} + \sum_{m=1}^{\infty} (a_m \cos m\theta + b_m \sin m\theta)$$

which is exactly the Fourier series. Thus,

$$a_m = \frac{1}{\pi}\int_0^{2\pi} f(\theta)\cos m\theta \,\mathrm{d}\theta\,; \quad b_m = \frac{1}{\pi}\int_0^{2\pi} f(\theta)\sin m\theta \,\mathrm{d}\theta$$

We can see from the equation that high harmonics are confined to have effects only near $r = 1$.

3.20. Laplace's equation in cylindrical polar coordinates

In cylindrical coordinates,

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right) + \frac{1}{4^2}\frac{\partial^2 \phi}{\partial \theta^2} + \frac{\partial^2 \phi}{\partial z^2} = 0$$

With $\phi = R(r)\Theta(\theta)Z(z)$, we find

$$\Theta'' = -\mu\Theta; \quad Z'' = \lambda Z; \quad r(rR')' + (\lambda r^2 - \mu)R = 0$$

The polar equation can be easily solved by

$$\mu_m = m^2; \quad \Theta_m(\theta) = \cos m\theta, \sin m\theta$$

The radial equation is Bessel's equation, giving solutions

$$R = J_m(kr), Y_m(kr)$$

Setting boundary conditions in the usual way, defining $R = 0$ at $r = a$ means that

$$J_m(ka) = 0 \implies k = \frac{j_{mn}}{a}$$

The radial solution is

$$R_{mn}(r) = J_m\left(\frac{j_{mn}}{a}r\right)$$

We have eliminated the Y_n term since we require $r = 0$ to give a finite ϕ. Finally, the z equation gives

$$Z'' = k^2 Z \implies Z = e^{-kz}, e^{kz}$$

We typically eliminate the e^{kz} mode due to boundary conditions, such as $Z \to 0$ as $z \to \infty$. The general solution is therefore

$$\phi(r,\theta,z) = \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} (a_{mn}\cos m\theta + b_{mn}\sin m\theta) J_m\left(\frac{j_{mn}}{a}r\right) e^{-frac j_{mn} r a}$$

3.21. Laplace's equation in spherical polar coordinates

In spherical polar coordinates,

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \Phi}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial\phi^2}=0$$

We will consider the *axisymmetric case*; supposing that there is no ϕ dependence. We seek a separable solution of the form

$$\Phi(r,\theta)=R(r)\Theta(\theta)$$

which gives

$$(\sin\theta\Theta')'+\lambda\sin\theta\Theta=0;\quad (r^2R')'-\lambda R=0$$

Consider the substitution $x=\cos\theta, \frac{\mathrm{d}x}{\mathrm{d}\theta}=-\sin\theta$ in the polar equation. This gives $\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}=-\sin\theta\frac{\mathrm{d}\Theta}{\mathrm{d}x}$ and hence

$$-\sin\theta\frac{\mathrm{d}}{\mathrm{d}x}\left[-\sin^2\theta\frac{\mathrm{d}\Theta}{\mathrm{d}x}\right]+\lambda\sin\theta\Theta=0\implies\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\frac{\mathrm{d}\Theta}{\mathrm{d}x}\right]+\lambda\Theta=0$$

This gives Legendre's equation, so it has solutions of eigenvalues $\lambda_\ell=\ell(\ell+1)$ and eigenfunctions

$$\Theta_\ell(\theta)=P_\ell(x)=P_\ell(\cos\theta)$$

The radial equation then gives

$$(r^2R')'-\ell(\ell+1)R=0$$

We will seek power law solutions: $R=\alpha r^\beta$. This gives

$$\beta(\beta+1)-\ell(\ell+1)=0\implies\beta=\ell,\beta=-\ell-1$$

Thus the radial eigenmodes are

$$R_\ell=r^\ell,r^{-\ell-1}$$

Therefore the general axisymmetric solution for spherical polar coordinates is

$$\Phi(r,\theta)=\sum_{\ell=0}^{\infty}(a_\ell r^\ell+b_\ell r^{-\ell-1})P_\ell(\cos\theta)$$

The a_ℓ, b_ℓ are determined by the boundary conditions. Orthogonality conditions for the P_ℓ can be used to determine coefficients. Consider a solution to Laplace's equation on the unit sphere with axisymmetric boundary conditions given by

$$\Phi(1,\theta)=f(\theta)$$

Given that we wish to find the interior solution, $b_n=0$ by regularity. Then,

$$f(\theta)=\sum_{\ell=0}^{\infty}a_\ell P_\ell(\cos\theta)$$

By defining $f(\theta) = F(\cos\theta)$,

$$F(x) = \sum_{\ell=0}^{\infty} a_\ell P_\ell(x)$$

We can then find the coefficients in the usual way, giving

$$a_\ell = \frac{2\ell+1}{2}\int_{-1}^{1} F(x)P_\ell(x)\,\mathrm{d}x$$

3.22. Generating function for Legendre polynomials

Consider a charge at $r_0 = (x, y, z) = (0, 0, 1)$. Then, the potential at a point P becomes

$$\begin{aligned}
\Phi(r) &= \frac{1}{|r - r_0|} = \frac{1}{(x^2 + y^2 + (x-1)^2)^{1/2}} \\
&= \frac{1}{(r^2(\sin^2\phi + \cos^2\phi)\sin^2\theta + r^2\cos^2\theta - 2r\cos\theta + 1)^{1/2}} \\
&= \frac{1}{(r^2\sin^2\theta + r^2\cos^2\theta - 2r\cos\theta + 1)^{1/2}} \\
&= \frac{1}{(r^2 - 2r\cos\theta + 1)^{1/2}} \\
&= \frac{1}{(r^2 - 2r\overline{x} + 1)^{1/2}}
\end{aligned}$$

where $\overline{x} \equiv \cos\theta$. This function Φ is a solution to Laplace's equation where $r \neq r_0$. Note that we can represent any axisymmetric solution as a sum of Legendre polynomials. Now,

$$\frac{1}{\sqrt{r^2 - 2rx + 1}} = \sum_{\ell=0}^{\infty} a_\ell P_\ell(x) r^\ell$$

With the normalisation condition for the Legendre polynomials $P_\ell(1) = 1$, we find

$$\frac{1}{1-r} = \sum_{\ell=0}^{\infty} a_\ell r^\ell$$

Using the geometric series expansion, we arrive at $a_\ell = 1$. This gives

$$\frac{1}{\sqrt{r^2 - 2rx + 1}} = \sum_{\ell=0}^{\infty} P_\ell(x) r^\ell$$

which is the generating function for the Legendre polynomials.

4. Green's functions

4.1. Dirac δ function

Definition. We define a generalised function $\delta(x-\xi)$ such that

(i) $\delta(x-\xi) = 0$ for all $x \neq \xi$;

(ii) $\int_{-\infty}^{\infty} \delta(x-\xi)\,\mathrm{d}x = 1$.

This acts as a linear operator $\int \mathrm{d}x\,\delta(x-\xi)$ on some function $f(x)$ to produce a number $f(\xi)$.

$$\int_{-\infty}^{\infty} \mathrm{d}x\,\delta(x-\xi)f(x) = f(\xi)$$

This relationship holds provided that $f(x)$ is sufficiently 'well-behaved' at $x = \xi$ and $x \to \pm\infty$.

Remark. Strictly, the δ 'function' is classified as a distribution, not as a function. For this reason, we will never use δ outside an integral, although such an integral may be implied. The δ function represents a unit point source or impulse.

We can approximate the δ function using a Gaussian approximation.

$$\delta_\varepsilon(x) = \frac{1}{\varepsilon\sqrt{\pi}} \exp\left[-\frac{x^2}{\varepsilon^2}\right]$$

Therefore,

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\delta(x)\,\mathrm{d}x &= \lim_{\varepsilon\to 0} \int_{-\infty}^{\infty} \frac{1}{\varepsilon\sqrt{\pi}} \exp\left[-\frac{x^2}{\varepsilon^2}\right] f(x)\,\mathrm{d}x \\
&= \lim_{\varepsilon\to 0} \int_{-\infty}^{\infty} \frac{1}{\varepsilon\sqrt{\pi}} \exp\left[-y^2\right] f(\varepsilon y)\,\mathrm{d}y \\
&= \lim_{\varepsilon\to 0} \int_{-\infty}^{\infty} \frac{1}{\varepsilon\sqrt{\pi}} \exp\left[-y^2\right] [f(0) + \varepsilon y f'(0) + \cdots]\,\mathrm{d}y \\
&= f(0)
\end{aligned}$$

for all well-behaved functions f at $0, \pm\infty$. We could alternatively use the Dirichlet kernel

$$\delta_n(x) = \frac{\sin nx}{\pi x} = \frac{1}{2\pi} \int_{-n}^{n} e^{ikx}\,\mathrm{d}k$$

or even

$$\delta_n(x) = \frac{n}{2} \operatorname{sech}^2 nx$$

4.2. Integral and derivative of δ function

We define the Heaviside step function by

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

For $x \neq 0$, we have

$$H(x) = \int_{-\infty}^{x} \delta(t)\,\mathrm{d}t$$

Thus,

$$\frac{\mathrm{d}}{\mathrm{d}x} H(x) = \delta(x)$$

where this identification takes place under an implied integral. We define $\delta'(x)$ using integration by parts.

$$\begin{aligned} \int_{-\infty}^{\infty} \delta'(x-\xi) f(x)\,\mathrm{d}x &= [\delta(x-\xi)f(x)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \delta(x-\xi) f'(x)\,\mathrm{d}x \\ &= -\int_{-\infty}^{\infty} \delta(x-\xi) f'(x)\,\mathrm{d}x \\ &= -f'(\xi) \end{aligned}$$

This is valid for all f that are smooth at $x = \xi$.

Example. Consider the Gaussian approximation:

$$\delta_\varepsilon(x) = \frac{1}{\varepsilon\sqrt{\pi}} \exp\left[-\frac{x^2}{\varepsilon^2}\right]$$

Then,

$$\delta_\varepsilon'(x) = \frac{-2x}{\varepsilon^3\sqrt{\pi}} \exp\left[-\frac{x^2}{\varepsilon^2}\right]$$

4.3. Properties of δ function

Note that

$$\int_a^b f(x)\delta(x-\xi)\,\mathrm{d}x = \begin{cases} f(\xi) & a < \xi < b \\ 0 & \text{otherwise} \end{cases}$$

So the δ function only 'samples' values within the integral range. This is known as the sampling property. Let $u = -(x-\xi)$, and consider

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\delta(-(x-\xi))\,\mathrm{d}x &= \int_{\infty}^{-\infty} f(\xi-u)\delta(u)(-\,\mathrm{d}u) \\ &= \int_{-\infty}^{\infty} f(\xi-u)\delta(u)\,\mathrm{d}u \\ &= f(\xi) \end{aligned}$$

Hence,

$$\int_{-\infty}^{\infty} f(x)\delta(-(x-\xi))\,\mathrm{d}x = \int_{-\infty}^{\infty} f(x)\delta(x-\xi)\,\mathrm{d}x$$

This is called the even property. Now, consider

$$\int_{-\infty}^{\infty} f(x)\delta(a(x-\xi))\,\mathrm{d}x = \frac{1}{|a|}f(\xi)$$

This is the scaling property. Let $g(x)$ be a function with n isolated roots at $x_1, \dots, x_n$. Then, assuming $g'(x)$ does not vanish at the x_i,

$$\delta(g(x)) = \sum_{i=1}^{n} \frac{\delta(x-x_i)}{|g'(x_i)|}$$

This is a generalisation of the above, known as the advanced scaling property. Now, if $g(x)$ is continuous at $x = 0$, then $g(x)\delta(x)$ equivalent to $g(0)\delta(x)$ inside an integral. This is known as the isolation property.

4.4. Fourier series expansion of δ function

Consider a complex Fourier series expansion,

$$\delta(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L}; \quad c_n = \frac{1}{2L}\int_{-L}^{L} \delta(x)e^{-in\pi x/L}\,\mathrm{d}x = \frac{1}{2L}$$

Hence,

$$\delta(x) = \frac{1}{2L}\sum_{n=-\infty}^{\infty} e^{in\pi x/L}$$

Let $f(x)$ be a function, so $f(x) = \sum_{n=-\infty}^{\infty} d_n e^{in\pi x/L}$. Then, their inner product is given by

$$\int_{-L}^{L} f^\star(x)\delta(x)\,\mathrm{d}x = \frac{1}{2L}\sum_{n=-\infty}^{\infty} d_n \int_{-L}^{L} e^{in\pi x/L}e^{in\pi x/L}\,\mathrm{d}x = \sum_{n=-\infty}^{\infty} d_n = f(0)$$

The Fourier expansion of the δ function can be extended periodically to the whole real line. This infinite set of δ functions is known as the Dirac comb, given by

$$\sum_{m=-\infty}^{\infty} \delta(x-2mL) = \sum_{n=-\infty}^{\infty} e^{in\pi x/L}$$

4.5. Arbitrary eigenfunction expansion of δ function

In general, suppose

$$\delta(x-\xi) = \sum_{n=1}^{\infty} a_n y_n(x)$$

with coefficients

$$a_n = \frac{\int_a^b w(x) y_n(x) \delta(x-\xi)\,\mathrm{d}x}{\int_a^b w(x) y_n(x)^2\,\mathrm{d}x} = \frac{w(\xi) y_n(\xi)}{\int_a^b w(x) y_n(x)^2\,\mathrm{d}x} = w_n(\xi) Y_n(\xi)$$

Then,

$$\delta(x-\xi) = w(\xi) \sum_{n=1}^{\infty} Y_n(\xi) Y_n(x) = w(x) \sum_{n=1}^{\infty} Y_n(\xi) Y_n(x)$$

since $\frac{w(x)}{w(\xi)}\delta(x-\xi) = \delta(x-\xi)$. Hence,

$$\delta(x-\xi) = w(x) \sum_{n=1}^{\infty} \frac{y_n(\xi) y_n(x)}{N_n}$$

where $N_n = \int_a^b w y_n^2\,\mathrm{d}x$ is a normalisation factor.

Example. Consider a Fourier series for $y(0) = y(1) = 0$, with $y_n(x) = \sin n\pi x$. From the sine series coefficient expression,

$$\delta(x-\xi) = 2 \sum_{n=1}^{\infty} \sin n\pi\xi \sin n\pi x$$

where $0 < \xi < 1$.

4.6. Motivation for Green's functions

Consider a massive static string with tension T and linear mass density μ, suspended between fixed ends $y(0) = y(1) = 0$. By resolving forces, we have the time independent form

$$T\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} - \mu g = 0$$

We will solve the inhomogeneous ODE $-\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = f(x)$ with $f(x) = -\frac{\mu g}{T}$. This has been placed in Sturm–Liouville form. We can integrate directly and find

$$-y = -\frac{\mu g}{2T}x^2 + k_1 x + k_2$$

Imposing boundary conditions,

$$y(x) = \left(-\frac{\mu g}{T}\right) \cdot \frac{1}{2} x(1-x)$$

Consider alternatively a solution obtained by solving the equation for a single point mass $\delta m = \mu \delta x$ suspended at $x = \xi$ on an very light string. We can then superimpose the solutions for each point mass to find the overall solution. For a single point mass, the solution is given by two straight lines from $(0,0)$ and $(1,0)$ to the point mass $(\xi_i, y_i(\xi_i))$. The angles of these straight lines from the horizontal are given by θ_1, θ_2. Resolving in the y direction,

$$\begin{aligned} 0 &= T(\sin\theta_1 + \sin\theta_2) - \delta m g \\ &= T\left(\frac{-y_i}{\xi_i} + \frac{-y_i}{1-\xi_i}\right) - \delta m g \\ \therefore -T(y_i(1-\xi_i) + y_i\xi_i) &= \delta m g \xi_i (1-\xi_i) \\ \therefore y_i(\xi_i) &= \frac{-\delta m g}{T}\xi_i(1-\xi_i) \end{aligned}$$

So the solution is

$$y_i(x) = \frac{-\delta m g}{T}\begin{cases} x(1-\xi_i) & x < \xi_i \\ \xi_i(1-x) & x > \xi_i \end{cases}$$

which is the generalised sawtooth. This can alternatively be written

$$f_i(\xi) G(x,\xi)$$

where f_i is a source term, and $G(x,\xi)$ is the Green's function, the solution for a unit point source. Since the differential equation is linear, we can sum the solutions, giving

$$y(x) = \sum_{i=1}^{N} f_i(\xi) G(x, \xi_i)$$

Taking a continuum limit,

$$f_i(\xi) = \frac{-\delta m g}{T} = \frac{-\mu \delta x g}{T} \equiv f(x)\,\mathrm{d}x \implies f(x) = \frac{-\mu g}{T}$$

which gives

$$y(x) = \int_0^1 f(\xi) G(x,\xi)\,\mathrm{d}\xi$$

Substituting the Green's function,

$$\begin{aligned} y(x) &= \left(\frac{-\mu g}{T}\right)\left[\int_0^x \xi(1-x)\,\mathrm{d}\xi + \int_x^1 x(1-\xi)\,\mathrm{d}\xi\right] \\ &= \left(\frac{-\mu g}{T}\right)\left\{\left[\frac{\xi^2}{2}(1-x)\right]_0^x + \left[x(\xi - \frac{\xi^2}{2})\right]_x^1\right\} \\ &= \left(\frac{-\mu g}{T}\right)\left(\frac{x^2}{2}(1-x) - 0 + \frac{x}{2} - x\left(x - \frac{x^2}{2}\right)\right) \\ &= \left(\frac{-\mu g}{T}\right)\cdot\frac{1}{2}x(1-x) \end{aligned}$$

So we have found the correct solution in two ways; once by direct integration, and once by superimposing point solutions. In general, direct integration is not trivial, and Green's functions are useful in this case.

4.7. Definition of Green's function

We wish to solve the inhomogeneous ODE

$$\mathcal{L}y \equiv \alpha(x)y'' + \beta(x)y' + \gamma(x)y = f(x)$$

on $a \leq x \leq b$, where $\alpha \neq 0$ and α, β, γ are continuous and bounded, taking homogeneous boundary conditions $y(a) = y(b) = 0$. The Green's function for $\mathcal{L}$ in this case is defined to be the solution for a unit point source at $x = \xi$. That is, $G(x, \xi)$ is the function that satisfies the boundary conditions and

$$\mathcal{L}G(x, \xi) = \delta(x - \xi)$$

so $G(a, \xi) = G(b, \xi) = 0$. Then, by linearity, the general solution is given by

$$y(x) = \int_a^b f(\xi)G(x, \xi)\,\mathrm{d}\xi$$

where $y(x)$ satisfies the homogeneous boundary conditions. We can verify this by checking

$$\mathcal{L}y = \int_a^b \mathcal{L}G(x, \xi)f(\xi)\,\mathrm{d}\xi = \int_a^b \delta(x - \xi)f(\xi)\,\mathrm{d}\xi = f(x)$$

So the solution is given by the inverse operator

$$y = \mathcal{L}^{-1}f; \quad \mathcal{L}^{-1} = \int_a^b \mathrm{d}\xi\, G(x, \xi)$$

The Green's function spits into two parts;

$$G(x, \xi) = \begin{cases} G_1(x, \xi) & a \leq x < \xi \\ G_2(x, \xi) & \xi < x < b \end{cases}$$

For all $x \neq \xi$, we have $\mathcal{L}G_1 = \mathcal{L}G_2 = 0$, so the parts are homogeneous solutions. G satisfies the homogeneous boundary conditions, so $G_1(a, \xi) = 0$ and $G_2(b, \xi) = 0$. G must be continuous at $x = \xi$, hence $G_1(\xi, \xi) = G_2(\xi, \xi)$. There is a jump condition; the derivative of G is discontinuous at $x = \xi$. This satisfies

$$[G']_{\xi_-}^{\xi_+} = \left.\frac{\mathrm{d}G_2}{\mathrm{d}x}\right|_{x=\xi_+} - \left.\frac{\mathrm{d}G_1}{\mathrm{d}x}\right|_{x=\xi_-} = \frac{1}{\alpha(\xi)}$$

4.8. Explicit form for Green's functions

We want to solve

$$\mathcal{L}G(x,\xi) = \delta(x-\xi)$$

on $a \le x \le b$, subject to homogeneous boundary conditions $G(a,\xi) = G(b,\xi) = 0$. The functions G_1, G_2 satisfy the homogeneous equation, so $\mathcal{L}G_i(x,\xi) = 0$. Suppose there exist two independent homogeneous solutions $y_1(x), y_2(x)$ to $\mathcal{L}y = 0$. Then, $G_1 = Ay_1 + By_2$, such that $Ay_1(a) + By_2(a) = 0$, which gives a constraint between A and B. This defines a complementary function $y_-(x)$ such that $y_-(a) = 0$. The general homogeneous solution with $G_1(a) = 0$ is

$$G_1 = Cy_-$$

C will be found later. Similarly we can define y_+ as a linear combination of y_1, y_2 such that $y_+(b) = 0$.

$$G_2 = Dy_+$$

We require $G_1(\xi,\xi) = G_2(\xi,\xi)$ for continuity, hence

$$Cy_-(\xi) = Dy_+(\xi)$$

Since $[G']_{\xi_-}^{\xi+} = \frac{1}{\alpha(\xi)}$, we have

$$Dy_+'(\xi) - CY_-'(\xi) = \frac{1}{\alpha(\xi)}$$

We can solve these equations for C, D simultaneously to find

$$C(\xi) = \frac{y_+(\xi)}{\alpha(\xi)W(\xi)}; \quad D(\xi) = \frac{y_-(\xi)}{\alpha(\xi)W(\xi)}$$

where $W(\xi)$ is the Wrońskian

$$W(\xi) = y_-(\xi)y_+'(\xi) - y_+(\xi)y_-'(\xi)$$

which is nonzero if y_-, y_+ are linearly independent. Hence,

$$G(x,\xi) = \begin{cases} \frac{y_-(x)y_+(\xi)}{\alpha(\xi)W(\xi)} & a \le x \le \xi \\ \frac{y_-(\xi)y_+(x)}{\alpha(\xi)W(\xi)} & \xi \le x \le b \end{cases}$$

4.9. Solving boundary value problems

We know that the solution of $\mathcal{L}y = f$ is

$$y(x) = \int_a^b G(x,\xi)f(\xi)\,\mathrm{d}\xi$$

We can split this into two intervals given that $G = G_1$ for $\xi > x$ and $G = G_2$ for $\xi < x$.

$$\begin{aligned} y(x) &= \int_a^x G_2(x,\xi) f(\xi)\,\mathrm{d}\xi + \int_x^b G_1(x,\xi) f(\xi)\,\mathrm{d}\xi \\ &= y_+(x) \int_a^x \frac{y_-(\xi) f(\xi)}{\alpha(\xi) W(\xi)}\,\mathrm{d}\xi + y_-(x) \int_a^x \frac{y_+(\xi) f(\xi)}{\alpha(\xi) W(\xi)}\,\mathrm{d}\xi \end{aligned}$$

Note that if $\mathcal{L}$ is in Sturm–Liouville form, so $\beta = \alpha'$, then the denominator $\alpha(\xi)W(\xi)$ is a constant. Further, G is symmetric; $G(x,\xi) = G(\xi,x)$. Often, by convention, we take $\alpha = 1$ (however Sturm–Liouville form typically takes $\alpha < 0$).

Example. Consider $y'' - y = f(x)$ with $y(0) = y(1) = 0$. Homogeneous solutions are $y_1 = e^x$, $y_2 = e^{-x}$. Imposing boundary conditions,

$$G = \begin{cases} C \sinh x & 0 \le x < \xi \\ D \sinh(1-x) & \xi < x \le b \end{cases}$$

Continuity at $x = \xi$ implies

$$C \sinh \xi = D \sinh(1-\xi) \implies C = D \frac{\sinh(1-\xi)}{\sinh \xi}$$

The jump condition is

$$-D \cosh(1-\xi) - C \cosh \xi = 1$$

Hence,

$$\begin{aligned} -D[\cosh(1-\xi)\sinh\xi + \sinh(1-\xi)\cosh\xi] &= \sinh\xi \\ -D[\sinh((1-\xi)+\xi)] &= \sinh\xi \\ -D\sinh 1 &= \sinh\xi \\ D &= \frac{\sinh\xi}{\sinh 1} \\ \therefore C &= \frac{-\sinh(1-\xi)}{\sinh 1} \end{aligned}$$

Therefore,

$$y(x) = \frac{-\sinh(1-x)}{\sinh 1} \int_0^x \sinh\xi f(\xi)\,\mathrm{d}\xi - \frac{\sinh x}{\sinh 1} \int_x^1 \sinh(1-\xi) f(\xi)\,\mathrm{d}\xi$$

Suppose we have inhomogeneous boundary conditions. In this case, we want to find a homogeneous solution y_p that solves the inhomogeneous boundary conditions. That is, $\mathcal{L} y_p = 0$ but $y_p(a), y_p(b)$ are as required for the boundary conditions. Then, by subtracting this solution from the original equation, we can solve using a homogeneous set of boundary conditions. For instance, in the above example, suppose $y(0) = 0, y(1) = 1$. We can find a solution $y_p = \frac{\sinh x}{\sinh 1}$ which has the inhomogeneous boundary conditions but solves the homogeneous problem.

4.10. Higher-order ODEs

Suppose $\mathcal{L}y = f(x)$ where $\mathcal{L}$ is an nth order linear differential operator, and $\alpha(x)$ is the coefficient for the highest degree derivative. Suppose that homogeneous boundary conditions are satisfied. Then we can define the Green's function in this case to be the function that solves

$$\mathcal{L}G(x,\xi) = \delta(x-\xi)$$

which has the properties:

(i) G_1, G_2 are homogeneous solutions satisfying the homogeneous boundary conditions;

(ii) $G_1^{(k)}(\xi) = G_2^{(k)}(\xi)$ for $k \in \{0, \dots, n-2\}$;

(iii) $G_2^{(n-1)}(\xi^+) - G_1^{(n-1)}(\xi^-) = \frac{1}{\alpha(\xi)}$.

4.11. Eigenfunction expansions of Green's functions

Suppose $\mathcal{L}$ is in Sturm–Liouville form with eigenfunctions $y_n(x)$ and eigenvalues λ_n. We seek $G(x,\xi) = \sum_{n=1}^{\infty} A_n y_n(x)$ satisfying $\mathcal{L}G = \delta(x-\xi)$.

$$\begin{aligned}\mathcal{L}G &= \sum_n A_n \mathcal{L}y_n \\ &= \sum_n A_n \lambda_n w(x) y_n(x)\end{aligned}$$

The δ function has expansion

$$\delta(x-\xi) = w(x)\sum_n \frac{y_n(\xi)y_n(x)}{N_n}; \quad N_n = \int w y_n^2 \,\mathrm{d}x$$

Hence,

$$A_n(\xi) = \frac{y_n(\xi)}{\lambda_n N_n}$$

Thus,

$$G(x,\xi) = \sum_{n=1}^{\infty} \frac{y_n(\xi)y_n(x)}{\lambda_n \int w y_n^2 \,\mathrm{d}x} = \sum_{n=1}^{\infty} \frac{Y_n(\xi)Y_N(x)}{\lambda_n}$$

which was already obtained earlier in the course when studying Sturm–Liouville theory.

4.12. Constructing Green's function for an initial value problem

Suppose we want to solve $\mathcal{L}y = f(t)$ for $t \geq a$ with $y(a) = y'(a) = 0$, using $G(t,\tau)$ satisfying $\mathcal{L}g = \delta(t-\tau)$. For $t < \tau$, we have

$$G_1 = Ay_1(t) + By_2(t); \quad Ay_1(a) + By_2(a) = 0; \quad Ay_1'(a) + By_2'(a) = 0$$

If $A \neq B \neq 0$, then we can solve this by dividing out A, B and find $y_1 y_2' - y_2 y_1' = 0$. Since the Wrońskian at a cannot be zero, $A = B = 0$. So $G_1(t, \tau) \equiv 0$ for $a \leq t < \tau$, so there is no change until the 'impulse' at $t = \tau$.

For $t > \tau$, by continuity we must have $G_2(\tau, \tau) = 0$. So we choose a complementary function $G_2 = Dy_+(t)$ with $y_+(t) = Ay_1(t) + By_2(t)$, and $y_+(\tau) = 0$. The discontinuity in the derivative implies that

$$G_2'(\tau, \tau) = Dy_+'(\tau) = \frac{1}{\alpha(\tau)}$$

Hence,

$$Ay_1'(\tau) + By_2'(\tau) = \frac{1}{\alpha(\tau)} \implies D(\tau) = \frac{1}{\alpha(\tau) y_+'(\tau)}$$

Hence we have a non-trivial solution

$$G(t, \tau) = \begin{cases} 0 & t < \tau \\ \frac{y_+(t)}{\alpha(\tau) y_+'(\tau)} & t > \tau \end{cases}$$

The initial value problem has solution

$$y(t) = \int_a^t G_2(t, \tau) f(\tau) \, \mathrm{d}\tau = \int_a^t \frac{y_+(t) f(\tau)}{y_+'(\tau)} \, \mathrm{d}\tau$$

Causality is 'built in' to this solution. Only forces which occur before t may have an impact on $y(t)$.

Example. Let us solve $y'' - y = f(t)$ with $y(0) = y'(0) = 0$. The homogeneous solution and initial conditions are

$$t < \tau \implies G_1 \equiv 0$$

and

$$t > \tau \implies G_2 = Ae^t + Be^{-t} = D \sinh(t - \tau)$$

Now,

$$[G']_{\tau_-}^{\tau_+} = \frac{1}{\alpha(\tau)} = 1 \implies G'(\tau, \tau) = D \cosh 0 = D = 1$$

Hence, the solution is

$$y(t) = \int_0^t f(\tau) \sinh(t - \tau) \, \mathrm{d}\tau$$

5. Fourier transforms

5.1. Definitions

Definition. The *Fourier transform* of a function $f(x)$ is

$$\widetilde{f}(k) = \mathcal{F}(f)(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} \, \mathrm{d}x$$

The *inverse Fourier transform* is

$$f(x) = \mathcal{F}^{-1}\big(\widetilde{f}\big)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widetilde{f}(k) e^{ikx} \, \mathrm{d}k$$

Different internally-consistent definitions exist, which distribute the multiplicative constants in different ways.

Theorem (Fourier inversion theorem)**.** For a function $f(x)$,

$$\mathcal{F}^{-1}(\mathcal{F}(f))(x) = f(x)$$

with a sufficient condition that f and $\widetilde{f}$ are absolutely integrable, so

$$\int_{-\infty}^{\infty} |f(x)| \, \mathrm{d}x = M < \infty$$

In particular, $f \to 0$ as $x \to \pm\infty$.

Example. Consider the Gaussian,

$$f(x) = \frac{1}{\sigma\sqrt{\pi}} \exp\left[-\frac{x^2}{\sigma^2}\right]$$

We wish to compute its Fourier transform. Since $i \sin kx$ is an odd function,

$$\widetilde{f}(k) = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{\sigma^2}\right] \exp[-ikx] \, \mathrm{d}x = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{\sigma^2}\right] \cos(kx) \, \mathrm{d}x$$

Consider, using Leibniz' rule,

$$\frac{\mathrm{d}\widetilde{f}}{\mathrm{d}k} = \frac{-1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} x \exp\left[\frac{-x^2}{\sigma^2}\right] \sin kx \, \mathrm{d}x$$

Integrating by parts,

$$\begin{aligned}
\frac{\mathrm{d}\widetilde{f}}{\mathrm{d}k} &= \frac{1}{\sigma\sqrt{\pi}} \left[\frac{\sigma^2}{2} \exp\left[\frac{-x^2}{\sigma^2}\right] \sin kx\right]_{-\infty}^{\infty} - \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{k\sigma^2}{2} \exp\left[\frac{-x^2}{\sigma^2}\right] \cos kx \, \mathrm{d}x \\
&= \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{k\sigma^2}{2} \exp\left[\frac{-x^2}{\sigma^2}\right] \cos kx \, \mathrm{d}x \\
&= -\frac{k\sigma^2}{2} \widetilde{f}(k)
\end{aligned}$$

This is a differential equation for $\widetilde{f}$, which gives

$$\widetilde{f}(k) = C \exp\left[-\frac{k^2\sigma^2}{4}\right]$$

Suppose $k = 0$. Then, in the original expression for the Fourier transform, we can directly find $\widetilde{f}(0) = 1$. Hence $C \exp\left[-\frac{0^2\sigma^2}{4}\right] = 1 \implies C = 1$. Hence,

$$\widetilde{f}(k) = \exp\left[-\frac{k^2\sigma^2}{4}\right]$$

which is another Gaussian with the width parameter inverted.

5.2. Converting Fourier series into Fourier transforms

Recall that the complex form of the Fourier series is

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{ik_n x}$$

where $k_n = \frac{n\pi}{L}$. We can write in particular $k_n = n\Delta k$ where $\Delta k = \frac{\pi}{L}$. Then,

$$c_n = \frac{1}{2L}\int_{-L}^{L} f(x)e^{-ik_n x}\,\mathrm{d}x = \frac{\Delta k}{2\pi}\int_{-L}^{L} f(x)e^{-ik_n x}\,\mathrm{d}x$$

Now, re-substituting into the Fourier series,

$$f(x) = \sum_{n=-\infty}^{\infty} \frac{\Delta k}{2\pi} e^{ik_n x}\int_{-L}^{L} f(x')e^{-ik_n x'}\,\mathrm{d}x'$$

Interpreting the sum multiplied by Δk as a Riemann integral,

$$f(x) \to \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{ik_n x}\int_{-L}^{L} f(x')e^{-ikx'}\,\mathrm{d}x'\,\mathrm{d}k$$

Taking the limit $L \to \infty$,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k\, e^{ikx}\int_{-\infty}^{\infty} \mathrm{d}x'\, f(x')e^{-ik_n x'}$$

which is the inverse Fourier transform of the Fourier transform of f, which gives the Fourier inversion theorem. Note that when $f(x)$ is discontinuous at x, the Fourier transform gives

$$\mathcal{F}^{-1}(\mathcal{F}(f))(x) = \frac{1}{2}(f(x_-) + f(x_+))$$

which is analogous to the result for Fourier series.

5.3. Properties of Fourier series

Recall the definition of the Fourier transform.

$$\widetilde{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-ikx}\,\mathrm{d}x$$

The (inverse) Fourier transform is linear.

$$h(x) = \lambda f(x) + \mu g(x) \iff \widetilde{h}(k) = \lambda\widetilde{f}(k) + \mu\tilde{g}(k)$$

Translated functions transform to multiplicative factors.

$$h(x) = f(x-\lambda) \iff \widetilde{h}(k) = e^{-i\lambda k}\widetilde{f}(k)$$

This is because

$$\widetilde{h}(k) = \int f(x-\lambda)e^{-ikx}\,\mathrm{d}x = \int f(y)e^{-ik(y+\lambda)}\,\mathrm{d}y = e^{-i\lambda k}\widetilde{f}(k)$$

Frequency shifts transform to translations in frequency space.

$$h(x) = e^{i\lambda x}f(x) \implies \widetilde{h}(k) = \widetilde{f}(k-\lambda)$$

A scalar multiple applied to the argument transforms into an inverse scalar multiple.

$$h(x) = f(\lambda x) \iff \widetilde{h}(k) = \frac{1}{|\lambda|}\widetilde{f}\left(\frac{k}{\lambda}\right)$$

Multiplication by x transforms into an imaginary derivative.

$$h(x) = xf(x) \iff \widetilde{h}(k) = i\widetilde{f}'(k)$$

This is because

$$\int_{-\infty}^{\infty} f(x)e^{-ikx}\,\mathrm{d}x = \frac{-1}{i}\frac{\mathrm{d}}{\mathrm{d}k}\int_{-\infty}^{\infty} f(x)e^{-ikx}\,\mathrm{d}x$$

Derivatives transform into a muliplication by ik.

$$h(x) = f'(x) \iff \widetilde{h}(k) = ik\widetilde{f}(k)$$

This is because we can integrate by parts and find

$$\widetilde{h}(k) = \int_{-\infty}^{\infty} f'(x)e^{-ikx}\,\mathrm{d}x = \underbrace{[f(x)e^{-ikx}]_{-\infty}^{\infty}}_{=0} + ik\int_{-\infty}^{\infty} f(x)e^{-ikx}\,\mathrm{d}x$$

The *general duality* property states that by mapping $x \mapsto -x$, we have

$$f(-x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(k)e^{-ikx}\,\mathrm{d}k$$

hence mapping $k \leftrightarrow x$, treating $\widetilde{f}$ now as a function in position space, we have

$$f(-k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(x)e^{-ikx}\,\mathrm{d}x$$

Thus

$$g(x) = \widetilde{f}(x) \iff \tilde{g}(k) = 2\pi f(-k)$$

We can then write the corollary that

$$f(-x) = \frac{1}{2\pi}\mathcal{F}(\mathcal{F}(f))(x)$$

Finally,

$$\mathcal{F}^4(f)(x) = 4\pi^2 f(x)$$

Example. Consider a function defined by

$$f(x) = \begin{cases} 1 & |x| \le a \\ 0 & \text{otherwise} \end{cases}$$

for some $a > 0$. By the definition of the Fourier transform,

$$\widetilde{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-ikx}\,\mathrm{d}x = \int_{-a}^{a} e^{-ikx}\,\mathrm{d}x = \int_{-a}^{a} \cos kx\,\mathrm{d}x = \frac{2}{k}\sin ka$$

By the Fourier inversion theorem,

$$\frac{1}{\pi}\int_{-\infty}^{\infty} e^{ikx}\frac{1}{k}\sin ka\,\mathrm{d}k = f(x)$$

for $x \neq a$. Now, in this expression, let $x = 0$ and let $k \mapsto x$. We arrive at the Dirichlet discontinuous formula.

$$\int_0^{\infty} \frac{\sin ax}{x}\,\mathrm{d}x = \frac{\pi}{2}\operatorname{sgn} a = \begin{cases} \frac{\pi}{2} & a > 0 \\ 0 & a = 0 \\ -\frac{\pi}{2} & a < 0 \end{cases}$$

5.4. Convolution theorem

We want to multiply Fourier transforms in the frequency domain (transformed space). This is useful for filtering or processing signals.

$$\widetilde{h}(k) = \widetilde{f}(k)\tilde{g}(k)$$

Consider the inverse.

$$\begin{aligned}
h(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(k)\tilde{g}(k)e^{ikx}\,\mathrm{d}k \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(y)e^{-iky}\,\mathrm{d}y\right)\tilde{g}(k)e^{ikx}\,\mathrm{d}k \\
&= \int_{-\infty}^{\infty} f(y)\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iky}\tilde{g}(k)e^{ikx}\,\mathrm{d}k\right)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} f(y)\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{g}(k)e^{ik(x-y)}\,\mathrm{d}k\right)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} f(y)g(x-y)\,\mathrm{d}y \\
&= (f * g)(x)
\end{aligned}$$

where $f * g$ is called the *convolution* of f and g. By duality, we also have

$$h(x) = f(x)g(x) \implies \widetilde{h}(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(p)\tilde{g}(k-p)\,\mathrm{d}p = \frac{1}{2\pi}\big(\widetilde{f} * \tilde{g}\big)(k)$$

5.5. Parseval's theorem

Consider $h(x) = g^\star(-x)$. Then, by letting $x = -y$,

$$\begin{aligned}
\widetilde{h}(k) &= \int_{-\infty}^{\infty} g^\star(-x)e^{-ikx}\,\mathrm{d}x \\
&= \left[\int_{-\infty}^{\infty} g(-x)e^{ikx}\,\mathrm{d}x\right]^\star \\
&= \left[\int_{-\infty}^{\infty} g(y)e^{-iky}\,\mathrm{d}y\right]^\star \\
&= \tilde{g}^\star(k)
\end{aligned}$$

Substituting this into the convolution theorem, with $g(x) \mapsto g^\star(-x)$, we have

$$\int_{-\infty}^{\infty} f(y)g^\star(y-x)\,\mathrm{d}y = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(k)\tilde{g}^\star(k)e^{ikx}\,\mathrm{d}x$$

Taking $x = 0$ in this expression and mapping $y \mapsto x$, we find

$$\int_{-\infty}^{\infty} f(x)g^\star(x)\,\mathrm{d}x = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(k)\tilde{g}^\star(k)\,\mathrm{d}x$$

Equivalently,

$$\langle g, f\rangle = \frac{1}{2\pi}\left\langle \tilde{g}, \widetilde{f}\right\rangle$$

So the inner product is conserved under the Fourier transform (up to a factor of 2π). Now, by setting $g^\star = f^\star$, we have

$$\int_{-\infty}^{\infty} |f(x)|^2 \,\mathrm{d}x = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left|\widetilde{f}(k)\right|^2 \mathrm{d}k$$

This is Parseval's theorem.

5.6. Fourier transforms of generalised functions

We can apply Fourier transforms to generalised functions by considering limiting distributions. Consider the inversion

$$\begin{aligned} f(x) &= \mathcal{F}^{-1}(\mathcal{F}(f))(x) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(u) e^{-iku} \,\mathrm{d}u \right] e^{ikx} \,\mathrm{d}k \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(u) \underbrace{\left[\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ik(x-u)} \,\mathrm{d}k \right]}_{\delta(x-u)} \mathrm{d}u \end{aligned}$$

In order to reconstruct $f(x)$ on the right hand side for any function f, we must have that the bracketed term is $\delta(x-u)$. So we identify

$$\delta(x-u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ik(x-u)} \,\mathrm{d}k$$

If $f(x) = \delta(x)$,

$$\widetilde{f}(k) = \int_{-\infty}^{\infty} \delta(x) e^{ikx} \,\mathrm{d}x = 1$$

This can be thought of as the Fourier transform of an infinitely thin Gaussian, which becomes an infinitely wide Gaussian (a constant). If $f(x) = 1$, then

$$\widetilde{f}(k) = \int_{-\infty}^{\infty} e^{-ikx} \,\mathrm{d}x = 2\pi\delta(k)$$

This can also be found by the duality formula. If $f(x) = \delta(x-a)$, we have

$$\widetilde{f}(k) = e^{-ika}$$

This is a translation of the original Fourier transform for the δ function above.

5.7. Trigonometric functions

Let $f(x) = \cos \omega x = \frac{1}{2}\left(e^{ix} + e^{-ix}\right)$. Then,

$$\widetilde{f}(k) = \pi(\delta(k+\omega) + \delta(k-\omega))$$

For $f(x) = \sin \omega x$, we have

$$\widetilde{f}(k) = i\pi(\delta(k+\omega) - \delta(k-\omega))$$

Using duality,

$$f(x) = \frac{1}{2}(\delta(x+a) + \delta(x-a)) \implies \widetilde{f}(k) = \cos ka$$
$$f(x) = \frac{1}{2i}(\delta(x+a) - \delta(x-a)) \implies \widetilde{f}(k) = \sin ka$$

5.8. Heaviside functions

Let $H(x)$ be the Heaviside function, such that $H(0) = \frac{1}{2}$. Then, $H(x) + H(-x) = 1$ for all x. We can take the Fourier transform of this and find

$$\widetilde{H}(k) + \widetilde{H}(-k) = 2\pi\delta(k)$$

Recall that $H'(x) = \delta(x)$. Thus,

$$ik\widetilde{H}(x) = \tilde{\delta}(k) = 1$$

Since $k\delta(k) = 0$, the two equations for $\widetilde{H}$ can be consistent if we take

$$\widetilde{H}(k) = \pi\delta(k) + \frac{1}{ik}$$

5.9. Dirichlet discontinuous formula

Recall the Dirichlet discontinuous formula:

$$\int_0^\infty \frac{\sin ax}{x}\,\mathrm{d}x = \frac{\pi}{2}\operatorname{sgn} a = \begin{cases} \frac{\pi}{2} & a > 0 \\ 0 & a = 0 \\ -\frac{\pi}{2} & a < 0 \end{cases}$$

We can rewrite this as

$$\frac{1}{2}\operatorname{sgn} x = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{ikx}}{ik}\,\mathrm{d}k$$

since the cosine term divided by ik is odd. Hence,

$$f(x) = \frac{1}{2}\operatorname{sgn} x \iff \widetilde{f}(k) = \frac{1}{ik}$$

This is the preferred form for a Heaviside-type function when used in Fourier transforms.

5.10. Solving ODEs for boundary value problems

Consider $y'' - y = f(x)$ with homogeneous boundary conditions $y \to 0$ as $x \to \pm\infty$. Taking the Fourier transform of this expression, we find

$$(-k^2 - 1)\tilde{y} = \widetilde{f}$$

Thus, the solution is

$$\tilde{y}(k) = \frac{-\widetilde{f}(k)}{1 + k^2} \equiv \widetilde{f}(k)\tilde{g}(k)$$

where $\tilde{g}(k) = \frac{-1}{1+k^2}$. Note that $\tilde{g}(k)$ is the Fourier transform of $g(x) = -\frac{1}{2}e^{-|x|}$. Applying the convolution theorem,

$$\begin{aligned} y(x) &= \int_{-\infty}^{\infty} f(u)g(x-u)\,\mathrm{d}u \\ &= -\frac{1}{2}\int_{-\infty}^{\infty} f(u)e^{-|x-u|}\,\mathrm{d}u \\ &= -\frac{1}{2}\left[\int_{-\infty}^{x} f(u)e^{u-x}\,\mathrm{d}u + \int_{x}^{\infty} f(u)e^{x-u}\,\mathrm{d}u\right] \end{aligned}$$

This is in the form of a boundary value problem Green's function. We can construct the same results by constructing the Green's function directly.

5.11. Signal processing

Suppose we have an input signal $\mathcal{I}(t)$, which is acted on by some linear operator $\mathcal{L}_{\text{in}}$ to yield an output $\mathcal{O}(t)$. The Fourier transform of the input $\tilde{\mathcal{I}}(\omega)$ is called the *resolution*.

$$\tilde{\mathcal{I}}(\omega) = \int_{-\infty}^{\infty} \mathcal{I}(t)e^{-i\omega t}\,\mathrm{d}t$$

In the frequency domain, the action of $\mathcal{L}_{\text{in}}$ on $\mathcal{I}(t)$ means that $\tilde{\mathcal{I}}(\omega)$ is multiplied by a transfer function $\widetilde{\mathcal{R}}(\omega)$. Thus,

$$\mathcal{O}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathcal{R}}(\omega)\tilde{\mathcal{I}}(\omega)e^{i\omega t}\,\mathrm{d}\omega$$

The inverse Fourier transform of the transfer function, $\mathcal{R}$, is called the *response function*, which is given by

$$\mathcal{R}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathcal{R}}(\omega)e^{i\omega t}\,\mathrm{d}\omega$$

By the convolution theorem,

$$\mathcal{O}(t) = \int_{-\infty}^{\infty} \mathcal{I}(u)\mathcal{R}(t-u)\,\mathrm{d}u$$

Suppose there is no input ($\mathcal{I}(t) = 0$) for $t < 0$. By causality, there should be zero output for the response function ($\mathcal{R}(t) = 0$) for $t < 0$. Therefore, we require $0 < u < t$ and hence

$$\mathcal{O}(t) = \int_0^t \mathcal{I}(u)\mathcal{R}(t-u)\,\mathrm{d}u$$

which resembles an initial value problem Green's function.

5.12. General transfer functions for ODEs

Suppose an input-output relationship is given by a linear ODE.

$$\mathcal{L}\mathcal{O}(t) \equiv \left(\sum_{i=0}^{n} a_i \frac{\mathrm{d}^i}{\mathrm{d}x^i}\right)\mathcal{O}(t) \equiv \mathcal{I}(t)$$

Here, $\mathcal{L}_{\text{in}} = 1$. We want to solve this ODE using a Fourier transform.

$$(a_0 + a_1 i\omega - a_2\omega^2 - a_3 i\omega^3 + \cdots + a_n(i\omega)^n)\widetilde{\mathcal{O}}(\omega) = \tilde{\mathcal{I}}(\omega)$$

We can solve this algebraically in Fourier transform space. The transfer function is

$$\widetilde{\mathcal{R}}(\omega) = \frac{1}{a_0 + \cdots + a_n(i\omega)^n}$$

We factorise the denominator to find partial fractions. Suppose there are J distinct roots $(i\omega - c_j)^{k_j}$, where k_j is the algebraic multiplicity of the jth root, so $\sum_{j=1}^{J} k_j = n$. So we can write

$$\widetilde{\mathcal{R}}(\omega) = \frac{1}{(i\omega - c_1)^{k_1} \ldots (i\omega - c_J)^{k_J}}$$

Expressing this as partial fractions,

$$\widetilde{\mathcal{R}}(\omega) = \sum_{j=1}^{J} \sum_{m=1}^{k_i} \frac{\Gamma_{jm}}{(i\omega - c_j)^m}$$

The Γ_{jm} terms are constant. To solve this, we must find the inverse Fourier transform of $(i\omega - a)^{-m}$. Recall that

$$\mathcal{F}^{-1}\left(\frac{1}{i\omega - a}\right) = \begin{cases} e^{at} & t > 0 \\ 0 & t < 0 \end{cases}$$

for $\operatorname{Re} a < 0$. So we will require $\operatorname{Re} c_j < 0$ for all j to eliminate exponentially growing solutions. Note that for $n = 2$,

$$i\frac{\mathrm{d}}{\mathrm{d}\omega}\left(\frac{1}{(i\omega - a)^2}\right)$$

and recall that

$$\mathcal{F}(tf(t)) = i\mathcal{F}'(\omega)$$

Hence,

$$\mathcal{F}^{-1}\left(\frac{1}{(i\omega-a)^2}\right)=\begin{cases}te^{at} & t>0\\ 0 & t<0\end{cases}$$

Inductively, we arrive at

$$\mathcal{F}^{-1}\left(\frac{1}{(i\omega-a)^m}\right)=\begin{cases}\frac{t^{m-1}}{(m-1)!}e^{at} & t>0\\ 0 & t<0\end{cases}$$

We can therefore invert any transfer function to obtain the response function. Thus the response function takes the form

$$\mathcal{R}(t)=\sum_{j=1}^{J}\sum_{m=1}^{k_i}\Gamma_{jm}\frac{t^{m-1}}{(m-1)!}e^{c_j t};\quad t>0$$

and zero for $t<0$. We can now solve such differential equations in Green's function form, or directly invert $\widetilde{\mathcal{R}}(\omega)\tilde{\mathcal{J}}(\omega)$ for a polynomial $\tilde{\mathcal{J}}(\omega)$.

5.13. Damped oscillator

We can use the Fourier transform method to solve the differential equation

$$\mathcal{L}y\equiv y''+2py'+(p^2+q^2)y=f(t)$$

where $p>0$. Consider homogeneous boundary conditions $y(0)=y'(0)=0$. The Fourier transform is

$$(i\omega)^2\tilde{y}+2ip\omega\tilde{y}+(p^2+q^2)\tilde{y}=\widetilde{f}$$

Hence,

$$\tilde{y}=\frac{\widetilde{f}}{-\omega^2+2ip\omega+p^2+q^2}\equiv\widetilde{R}\widetilde{f}$$

We can invert this using the convolution theorem by inverting $\widetilde{R}$.

$$y(t)=\int_0^t\mathcal{R}(t-\tau)f(\tau)\,\mathrm{d}\tau$$

where the response function is

$$\mathcal{R}(t-\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{i\omega(t-\tau)}}{p^2+q^2+2ip\omega-\omega^2}\,\mathrm{d}\omega$$

We can show that $\mathcal{L}\mathcal{R}(t-\tau)=\delta(t-\tau)$; in other words, $\mathcal{R}$ is the Green's function.

5.14. Discrete sampling and the Nyquist frequency

Suppose a signal $h(t)$ is sampled at equal times $t_n = n\Delta$ with a time step Δ and values $h_n = h(t_n) = h(n\Delta)$, for all $n \in \mathbb{Z}$. The sampling frequency is therefore Δ^{-1}, so the sampling angular velocity is $\omega_s = 2\pi f_s = \frac{2\pi}{\Delta}$. The Nyquist frequency is $f_c = \frac{1}{2\Delta}$, which is the highest frequency actually sampled at Δ. Suppose we have a signal g_f with a given frequency f. We will write

$$g_f(t) = A\cos(2\pi f t + \varphi) = \operatorname{Re}\left(Ae^{2\pi i f t+\varphi}\right) = \frac{1}{2}\left(Ae^{2\pi i f t+\varphi}\right) + \frac{1}{2}\left(Ae^{-2\pi i f t+\varphi}\right)$$

where $A \in \mathbb{R}$. Note that this signal has two 'frequencies'; a positive and a negative frequency. The combination of these frequencies gives the full wave. Suppose we sample $g_f(t)$ at the Nyquist frequency, so $f = f_c$. Then,

$$\begin{aligned} g_{f_c}(t_n) &= A\cos\left(2\pi\frac{1}{2\Delta}n\Delta + \varphi\right) \\ &= A\cos(\pi n + \varphi) \\ &= A\cos\pi n\cos\phi + A\sin\pi n\sin\phi \\ &= A'\cos(2\pi f_c f_n) \end{aligned}$$

where $A' = A\cos\phi$. This has removed half of the information about the wave; the amplitude and the phase have become degenerate. We can identify f_c with $-f_c$ when considering the remaining information; we say that the two frequencies are *aliased* together. Now, suppose we sample at greater than the Nyquist frequency, in particular $f = f_c + \delta f > f_c$, where for simplicity we let $\delta f < f_c$. We have

$$\begin{aligned} g_f(t_n) &= A\cos(2\pi(f_c + \delta f)t_n + \varphi) \\ &= A\cos(2\pi(f_c - \delta f)t_n - \varphi) \end{aligned}$$

So frequencies above the Nyquist frequency are reinterpreted after the sampling as a frequency lower than the Nyquist frequency. This aliases $f_c + \delta f$ with $f_c - \delta f$.

5.15. Nyquist–Shannon sampling theorem

Definition. A signal $g(t)$ is *bandwidth-limited* if it contains no frequencies above $\omega_{\max} = 2\pi f_{\max}$. In other words, $\tilde{g}(\omega) = 0$ for all $|\omega| > \omega_{\max}$. In this case,

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{g}(\omega)e^{i\omega t}\,\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\omega_{\max}}^{\omega_{\max}}\tilde{g}(\omega)e^{i\omega t}\,\mathrm{d}\omega$$

Suppose we set the sampling rate to the Nyquist frequency, so $\Delta = \frac{1}{2f_{\max}}$. Then,

$$g_n \equiv g(t_n) = \frac{1}{2\pi}\int_{-\omega_{\max}}^{\omega_{\max}}\tilde{g}(\omega)e^{i\pi n\omega/\omega_{\max}}\,\mathrm{d}\omega$$

This is a complex Fourier series coefficient c_n, multiplied by $\frac{\omega_{\max}}{\pi}$. The Fourier series is periodic in ω with period $2\omega_{\max}$, not in space or time.

$$\tilde{g}_{\text{per}}(\omega) = \frac{\pi}{\omega_{\max}} \sum_{n=-\infty}^{\infty} g_n e^{-i\pi n\omega/\omega_{\max}}$$

The actual Fourier transform $\tilde{g}$ is found by multiplying by a top hat window function

$$\widetilde{h}(\omega) = \begin{cases} 1 & |\omega| \leq \omega_{\max} \\ 0 & \text{otherwise} \end{cases}$$

Hence,

$$\tilde{g}(\omega) = \tilde{g}_{\text{per}}(\omega)\widetilde{h}(\omega)$$

Note that this relation is exact. Inverting this expression,

$$\begin{aligned} g(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{g}_{\text{per}}(\omega)\widetilde{h}(\omega) e^{i\omega t} \, \mathrm{d}\omega \\ &= \frac{1}{2\omega_{\max}} \sum_{n=-\infty}^{\infty} g_n \int_{-\omega_{\max}}^{\omega_{\max}} \exp\left(i\omega\left(t - \frac{n\pi}{\omega_{\max}}\right)\right) \mathrm{d}\omega \end{aligned}$$

Only the cosine term is even, hence

$$g(t) = \frac{1}{2\omega_{\max}} \sum_{n=-\infty}^{\infty} g_n \frac{\sin(\omega_{\max} t - \pi n)}{\omega_{\max} t - \pi n}$$

Hence, $g(t)$ can be written *exactly* as a combination of countably many discrete sample points.

5.16. Discrete Fourier transform

Suppose we have a finite number of samples $h_m = h(t_m)$ for $t_m = m\Delta$, where $m = 0, \ldots, N-1$. We will approximate the Fourier transform for N frequencies within the Nyquist frequency $f_c = \frac{1}{2\Delta}$, using equally-spaced frequencies, given by $\Delta_f = \frac{1}{N\Delta}$ in the range $-f_c \leq f \leq f_c$. We could take the convention $f_n = n\Delta_f = \frac{n}{N\Delta}$ for $n = -\frac{N}{2}, \ldots, -\frac{N}{2}$. However, this overcounts the Nyquist frequency (which is aliased), giving $N + 1$ frequencies instead of the desired N. Since frequencies above the Nyquist frequency are aliased to below it:

$$\left(\frac{N}{2} + m\right)\Delta f = f_c + \delta f \mapsto \left(\frac{N}{2} - m\right)\Delta f = -(f_c - \delta f)$$

we can instead use the convention $f_n = n\Delta_f = \frac{n}{N\Delta}$ for $n = 0, \ldots, N-1$. This counts the Nyquist frequency only once. The Fourier transform at a frequency f_n becomes

$$\begin{aligned}\widetilde{h}(f_n) &= \int_{-\infty}^{\infty} h(t) e^{-2\pi i f_n t}\,\mathrm{d}t \\ &\approx \Delta \sum_{m=0}^{N-1} h_m e^{-2\pi i f_n t_m} \\ &= \Delta \sum_{m=0}^{N-1} h_m e^{-2\pi i m n/N} \\ &= \Delta \widetilde{h}_d(f_n)\end{aligned}$$

where the function $\widetilde{h}_d(f_n)$ is the *discrete Fourier transform*. The matrix $[\mathrm{DFT}]_{mn} = e^{-2\pi i m n/N}$ defines the discrete Fourier transform for the vector $h = \{h_m\}$. The discrete Fourier transform is then

$$\widetilde{h}_d = [\mathrm{DFT}]h$$

By inverting the discrete Fourier transform matrix, we find

$$h = [\mathrm{DFT}]^{-1}\widetilde{h}_d = \frac{1}{N}[\mathrm{DFT}]^{\dagger}\widetilde{h}_d$$

since the inverse of the discrete Fourier transform matrix is its adjoint. The matrix is built from roots of unity $\omega = e^{-2\pi i/N}$. So, for instance, $n = 4$ gives $\omega = e^{-2\pi i/4} = -i$ giving

$$[\mathrm{DFT}] = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}$$

The inverse discrete Fourier transform is

$$\begin{aligned}h_m &= h(t_m) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{h}(\omega) e^{i\omega t_m}\,\mathrm{d}\omega \\ &= \int_{-\infty}^{\infty} \widetilde{h}(f) e^{2\pi i f t_m}\,\mathrm{d}f \\ &\approx \frac{1}{\Delta N}\sum_{n=0}^{N-1} \Delta\widetilde{h}_d(f_n) e^{2\pi i m n/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} \widetilde{h}_n e^{2\pi i m n/N}\end{aligned}$$

Hence, we can interpolate the initial function from its samples.

$$h(t) = \frac{1}{N}\sum_{n=0}^{N-1} \widetilde{h}_n e^{2\pi i n t/N}$$

Parseval's theorem becomes

$$\sum_{m=0}^{N-1} |h_m|^2 = \frac{1}{N} \sum_{n=0}^{N-1} \left|\widetilde{h}_n\right|^2$$

and the convolution theorem is

$$c_k = \sum_{m=0}^{N-1} g_m h_{k-m} \iff \tilde{c}_k = \tilde{g}_k \widetilde{h}_k$$

5.17. Fast Fourier transform (non-examinable)

While the discrete Fourier transform is an order $O(N^2)$ operation, we can reduce this into an order $O(n \log N)$ operation. Such a simplification is called the *fast Fourier transform*. We can split the discrete Fourier transform into even and odd parts, noting that $\omega_N = e^{-2\pi i/N}$ implies $\omega_N^2 = e^{-2\pi i/(N/2)} = \omega_{N/2}$

$$\begin{aligned}
\widetilde{h}_k &= \sum_{n=0}^{N-1} h_n \omega_N^{nk} \\
&= \sum_{m=0}^{N/2-1} h_{2m} \omega_N^{2mk} + \sum_{m=0}^{N/2-1} h_{2m+1} \omega_N^{(2m+1)k} \\
&= \sum_{m=0}^{N/2-1} h_{2m} (\omega_N^2)^{mk} + \omega_N^k \sum_{m=0}^{N/2-1} h_{2m+1} (\omega_N^2)^{mk} \\
&= \sum_{m=0}^{N/2-1} h_{2m} (\omega_{N/2})^{mk} + \omega_N^k \sum_{m=0}^{N/2-1} h_{2m+1} (\omega_{N/2})^{mk}
\end{aligned}$$

This algorithm iteratively reduces the Fourier transform's complexity by a factor of two, until the trivial case of finding the discrete Fourier transform of two data points.

6. Method of characteristics

6.1. Well-posed Cauchy problems

Solving partial differential equations depends on the nature of the equations in combination with the boundary or initial data. A *Cauchy problem* is the partial differential equation for some function ϕ together with the auxiliary data (in ϕ and its derivatives) specified on a surface (or a curve in two dimensions), which is called *Cauchy data*. For a Cauchy problem to be *well-posed*, we require that

(i) a solution exists (we do not have excessive auxiliary data);

(ii) the solution is unique (we do not have insufficient auxiliary data); and

(iii) the solution depends continuously on the auxiliary data.

6.2. Method of characteristics

Consider a parametrised curve C given by Cartesian coordinates $(x(s), y(s))$. The tangent vector is

$$v = \left(\frac{\mathrm{d}x(s)}{\mathrm{d}s}, \frac{\mathrm{d}y(s)}{\mathrm{d}s}\right)$$

We then define the directional derivative of a function $\phi(x, y)$ by

$$\left.\frac{\mathrm{d}\phi}{\mathrm{d}s}\right|_C = \frac{\mathrm{d}x(s)}{\mathrm{d}s}\frac{\partial\phi}{\partial x} + \frac{\mathrm{d}y(s)}{\mathrm{d}s}\frac{\partial\phi}{\partial y} = v \cdot \nabla\phi$$

Suppose $v \cdot \nabla\phi = 0$ then $\frac{\mathrm{d}\phi}{\mathrm{d}s} = 0$ and hence ϕ is constant along the curve. Suppose there exists a vector field

$$u = (\alpha(x, y), \beta(x, y))$$

with a family of non-intersecting integral curves C which fill the plane (or domain of the function more generally), such that at a point (x, y) the integral curve has tangent vector $u(x, y)$. Now, define a curve B by $(x(t), y(t))$ such that B is transverse to u; its tangent is nowhere parallel to u.

$$w = \left(\frac{\mathrm{d}x(t)}{\mathrm{d}t}, \frac{\mathrm{d}y(t)}{\mathrm{d}t}\right) \nparallel (\alpha(x, y), \beta(x, y)) = u$$

This can be used to parametrise the family of curves by labelling each curve C with the value of t at the intersection point between it and B. Along the curve, we use s such that $s = 0$ at the intersection. The integral curves $(x(s, t), y(s, t))$ satisfy

$$\frac{\mathrm{d}x}{\mathrm{d}s} = \alpha(x, y); \quad \frac{\mathrm{d}y}{\mathrm{d}s} = \beta(x, y)$$

We can solve these equations to find a family of characteristic curves, along which t remains constant. This yields a new coordinate system (s, t) associated with a differential equation we wish to solve.

6.3. Characteristics of a first order PDE

Consider

$$\alpha(x,y)\frac{\partial \phi}{\partial x} + \beta(x,y)\frac{\partial \phi}{\partial y} = 0$$

with Cauchy data on an initial curve B, defined by $(x(t), y(t))$:

$$\phi(x(t), y(t)) = f(t)$$

Note,

$$\alpha\phi_x + \beta\phi_y = u \cdot \nabla\phi = \left.\frac{\mathrm{d}\phi}{\mathrm{d}s}\right|_C$$

This is exactly the directional derivative along the integral curve C, defined by $u = (\alpha, \beta)$. Since $\frac{\mathrm{d}\phi}{\mathrm{d}s} = \alpha\phi_x + \beta\phi_y = 0$ from the original PDE, the function $\phi(x,y)$ is constant along this curve C. In other words, the Cauchy data $f(t)$ defined on B at $s = 0$ is propagated constantly along the integral curves. This gives the solution

$$\phi(s,t) = \phi(x(s,t), y(s,t)) = f(t)$$

To obtain ϕ in the original coordinates, we need to transform from s,t-space into x,y-space. Provided that the Jacobian $J = x_t y_s - x_s y_t$ is nonzero, we can invert the transformation and find s,t as functions of x,y. This gives

$$\phi(x,y) = f(t(x,y))$$

To solve such a PDE, we will typically use the following steps.

(i) Find the characteristic equations $\frac{\mathrm{d}x}{\mathrm{d}s} = \alpha$, $\frac{\mathrm{d}y}{\mathrm{d}s} = \beta$.

(ii) Parametrise the initial conditions on B by $(x(t), y(t))$.

(iii) Solve the characteristic equations to find $x = x(s,t)$ and $y = y(s,t)$ subject to the initial conditions at $s = 0$.

(iv) Solve the equation for ϕ given by $\frac{\mathrm{d}\phi}{\mathrm{d}s} = \alpha\phi_x + \beta\phi_y = 0$, so ϕ is constant along the integral curves, giving $\phi(s,t) = f(t)$.

(v) Invert the relations $s = s(x,y)$ and $t = t(x,y)$, then find ϕ in terms of x,y.

Example. Consider the equation

$$\frac{\mathrm{d}\phi(x,y)}{\mathrm{d}x} = 0$$

such that

$$\phi(0,y) = h(y)$$

The characteristic equations are given by

$$\frac{\mathrm{d}x}{\mathrm{d}s} = \alpha = 1; \quad \frac{\mathrm{d}y}{\mathrm{d}s} = \beta = 0$$

The initial curve B is given by

$$(x(t), y(t)) = (0, t)$$

Solving the characteristic equations,

$$x = s + c(t); \quad y = d(t)$$

At $x = 0$, we must have $s = 0$, so $c = 0$. Further, $y = t$ hence $d = t$. Thus,

$$x = s; \quad y = t$$

Thus,

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} = 0 \implies \phi(s, t) = h(t) \implies \phi(x, y) = h(y)$$

Example. Consider

$$e^x \phi_x + \phi_y = 0; \quad \phi(x, 0) = \cosh x$$

The characteristic equations are

$$\frac{\mathrm{d}x}{\mathrm{d}s} = e^x; \quad \frac{\mathrm{d}y}{\mathrm{d}s} = 1$$

The initial conditions are

$$x(t) = t; \quad y(t) = 0$$

We solve the characteristic equation subject to these initial conditions, giving

$$-e^{-x} = s + c(t); \quad y = s + d(t)$$

$s = 0$ implies $-e^{-t} = c(t)$ and $y = 0 = d(t)$. Hence

$$e^{-x} = e^{-t} - s; \quad y = s$$

Now,

$$\frac{\mathrm{d}\phi}{\mathrm{d}s} = 0 \implies \phi(s, t) = \cosh t$$

Since $s = y, e^{-t} = y + e^{-x}$, we have $t = -\log(y + e^{-x})$. Thus,

$$\phi(x, y) = \cosh\left[-\log(y + e^{-x})\right]$$

6.4. Inhomogeneous first order PDEs

Suppose we now wish to solve

$$\alpha(x, y)\phi_x + \beta(x, y)\phi_y = \gamma(x, y)$$

with Cauchy data $\phi(x(t), y(t)) = f(t)$ along a curve B. The characteristic curves are the same as the homogeneous case. However, the directional derivative no longer vanishes:

$$\left.\frac{\mathrm{d}\phi}{\mathrm{d}s}\right|_C = u \cdot \nabla\phi = \gamma(x, y)$$

where $\phi = f(t)$ at $s = 0$ on B. So $f(t)$ is no longer propagated constantly across characteristic polynomials, but is instead propagated according to the ODE in s above. We must therefore solve this ODE along C before reverting to x, y coordinates.

Example. Consider

$$\phi_x + 2\phi_y = ye^x; \quad \phi(x, x) = \sin x$$

The characteristic equation is given by

$$\frac{\mathrm{d}x}{\mathrm{d}s} = 1; \quad \frac{\mathrm{d}y}{\mathrm{d}s} = 2$$

The initial conditions are

$$x(t) = y(t) = t$$

From the characteristic equations,

$$x = s + c(t); \quad y = 2s + d(t)$$

Thus,

$$x = t = c(t); \quad y = t = d(t)$$

So the solutions to the characteristics are

$$x = s + t; \quad y = 2s + t$$

Now we solve

$$\frac{\mathrm{d}\phi}{\mathrm{d}s} = \gamma = ye^x = (2s + t)e^{s+t}$$

Note that $\frac{\mathrm{d}}{\mathrm{d}s}(2se^s) = 2e^s + 2se^s$, so the solution is

$$\phi(s, t) = (2s - 2 + t)e^{s+t} + c(s)$$

for some constant term $c(s)$. But $\phi(0, t) = \sin t$, hence

$$\sin t = (t - 2)e^t + c(s) \implies \phi(s, t) = (2s - 2 + t)e^{s+t} + \sin t - (2 - t)e^t$$

Inverting into x, y space,

$$\phi(x, y) = (y - 2)e^x + (y - 2x + 2)e^{2x-y} + \sin(2x - y)$$

6.5. Classification of second order PDEs

In two dimensions, the general second order PDE is

$$\begin{aligned}\mathcal{L}\phi \equiv a(x, y)\frac{\partial^2\phi}{\partial x^2} + 2b(x, y)\frac{\partial^2\phi}{\partial x\partial y} + c(x, y)\frac{\partial^2\phi}{\partial y^2} \\ + d(x, y)\frac{\partial\phi}{\partial x} + e(x, y)\frac{\partial\phi}{\partial y} + f(x, y)\phi(x, y)\end{aligned}$$

The *principal part* is given by

$$\sigma_P(x, y, k_x, k_y) \equiv k^\intercal A k = \begin{pmatrix} k_x & k_y \end{pmatrix} \begin{pmatrix} a(x,y) & b(x,y) \\ b(x,y) & c(x,y) \end{pmatrix} \begin{pmatrix} k_x \\ k_y \end{pmatrix}$$

The PDE is classified by the properties of the eigenvalues of A.

(i) If $b^2 - ac < 0$, the equation is *elliptic*. The eigenvalues have the same sign. An example is the Laplace equation.

(ii) If $b^2 - ac > 0$, the equation is *hyperbolic*. The eigenvalues have opposite signs. An example is the wave equation.

(iii) If $b^2 - ac = 0$, the equation is *parabolic*, where at least one eigenvalue is zero. An example is the heat equation.

Note that a differential equation may have different classifications at different points (x, y) in space.

6.6. Characteristic curves of second order PDEs

A curve defined by $f(x, y)$ constant is a characteristic if

$$\begin{pmatrix} f_x & f_y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} f_x \\ f_y \end{pmatrix} = 0$$

This is a generalisation of the first order case $u \cdot \nabla f = 0$ where $u = (\alpha, \beta)$. The curve can be written as $y = y(x)$ by the chain rule.

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = 0 \implies \frac{f_x}{f_y} = -\frac{\mathrm{d}y}{\mathrm{d}x}$$

Substituting into the quadratic form,

$$a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - 2b\frac{\mathrm{d}y}{\mathrm{d}x} + c = 0$$

for which we have a quadratic solution given by

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b \pm \sqrt{b^2 - ac}}{a}$$

(i) Hyperbolic equations have two such solutions, since $b^2 - ac > 0$.

(ii) Parabolic equations have one solution.

(iii) Elliptic equations have no real characteristics.

6.7. Characteristic coordinates

Transforming to characteristic coordinates u, v will set $a = 0$ and $c = 0$. Hence, the PDE will take the canonical form

$$\frac{\partial^2 \phi}{\partial u \partial v} + \cdots + = 0$$

where the omitted terms are lower order.

Example. Consider

$$-y\phi_{xx} + \phi_{yy} = 0$$

Here, $a = -y, b = 0, c = 1$ hence $b^2 - ac = y$. For $y > 0$, the equation is hyperbolic, for $y < 0$ it is elliptic, and for $y = 0$ it is parabolic. Consider the characteristics for $y > 0$.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b \pm \sqrt{b^2 - ac}}{a} = \pm \frac{1}{\sqrt{y}}$$

Hence,

$$\int \sqrt{y}\, \mathrm{d}y = \pm \int \mathrm{d}x \implies \frac{2}{3} y^{\frac{3}{2}} \pm x = C_{\pm}$$

Therefore, the characteristic curves are

$$u = \frac{2}{3} y^{\frac{3}{2}} + x; \quad v = \frac{2}{3} y^{\frac{3}{2}} - x$$

Taking derivatives,

$$u_x = 1; \quad u_y = \sqrt{y}; \quad v_x = -1; \quad v_y = \sqrt{y}$$

Hence,

$$\begin{aligned}
\phi_x &= \phi_u u_x + \phi_v v_x = \phi_u - \phi_v \\
\phi_y &= \sqrt{y}(\phi_u + \phi_v) \\
\phi_{xx} &= \phi_{uu} - 2\phi_{uv} + \phi_{vv} \\
\phi_{yy} &= y(\phi_{uu} + 2\phi_{uv} + \phi_{vv}) + \frac{1}{2\sqrt{y}}(\phi_u + \phi_v)
\end{aligned}$$

Substituting into the original PDE,

$$-y\phi_{xx} + \phi_{yy} = y\left(4\phi_{uv} + \frac{1}{2y^{\frac{3}{2}}}(\phi_u + \phi_v)\right)$$

Note, $u + v = \frac{4}{3} y^{\frac{3}{2}}$, hence we have the canonical form

$$4\phi_{uv} + \frac{1}{6(u+v)}(\phi_u + \phi_v) = 0$$

6.8. General solution to wave equation

The wave equation is

$$\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} = 0$$

We wish to solve this with initial conditions $\phi(x,0) = f(x)$, and $\phi_t(x,0) = g(x)$. Here, $a = \frac{1}{c^2}, b = 0, c = -1$ hence $b^2 - ac > 0$. The characteristic equation is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{0 \pm \sqrt{0 + \frac{1}{c^2}}}{\frac{1}{c^2}} = \pm c$$

Hence the characteristic coordinates are

$$u = x - ct; \quad v = x + ct$$

This yields the canonical form

$$\frac{\partial^2\phi}{\partial u \partial v} = 0$$

This may be integrated directly to find

$$\frac{\partial\phi}{\partial v} = F(v) \implies \phi = G(u) + \int^v F(y)\,\mathrm{d}y = G(u) + H(v)$$

Imposing the initial conditions at $t = 0$, we find

$$G(x) + H(x) = f(x); \quad -cG'(x) + cH'(x) = g(x)$$

Differentiating the first equation, we find

$$G'(x) + H'(x) = f'(x)$$

We can combine this with the second equation to give

$$H'(x) = \frac{1}{2}\left(f'(x) + \frac{1}{c}g(x)\right) \implies H(x) = \frac{1}{2}(f(x) - f(0)) + \frac{1}{2c}\int_0^x g(y)\,\mathrm{d}y$$

Similarly,

$$G'(x) = \frac{1}{2}\left(f'(x) - \frac{1}{c}g(x)\right) \implies G(x) = \frac{1}{2}(f(x) - f(0)) - \frac{1}{2c}\int_0^x g(y)\,\mathrm{d}y$$

The final solution is therefore

$$\phi(x,t) = G(x-ct) + H(x+ct) = \frac{1}{2}(f(x-ct) + f(x+ct)) + \frac{1}{2c}\int_{x-ct}^{x+ct} g(y)\,\mathrm{d}y$$

Waves propagate at a velocity c, hence $\phi(x,t)$ is fully determined by values of f, g in the interval $[x-ct, x+ct]$.

7. Solving partial differential equations with Green's functions

7.1. Diffusion equation and Fourier transform

Recall the heat equation for a conducting wire given by

$$\frac{\partial \Theta}{\partial t}(x,t) - D\frac{\partial^2 \Theta}{\partial x^2}(x,t) = 0$$

with initial conditions $\Theta(x,0) = h(x)$ and boundary conditions $\Theta \to 0$ as $x \to \pm\infty$. Taking the Fourier transform with respect to x,

$$\frac{\partial}{\partial t}\widetilde{\Theta}(k,t) = -Dk^2\widetilde{\Theta}(k,t)$$

Integrating, we find

$$\widetilde{\Theta}(k,t) = Ce^{-Dk^2t}$$

The initial conditions give $\widetilde{\Theta}(k,0) = \widetilde{h}(k)$ and therefore

$$\widetilde{\Theta}(k,t) = \widetilde{h}(k)e^{-Dk^2t}$$

We take the inverse Fourier transform to find

$$\Theta(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{h}(k)\underbrace{e^{-Dk^2t}}_{\text{FT of Gaussian}}e^{ikx}\,\mathrm{d}k$$

Hence, by the convolution theorem,

$$\begin{aligned}\Theta(x,t) &= \frac{1}{\sqrt{4\pi Dt}}\int_{-\infty}^{\infty} h(u)\exp\left(-\frac{(x-u)^2}{4Dt}\right)\mathrm{d}u \\ &\equiv \int_{-\infty}^{\infty} h(u)S_d(x-u,t)\,\mathrm{d}u\end{aligned}$$

where the *fundamental solution* is

$$S_d(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp\left(-\frac{x^2}{4Dt}\right)$$

which is the Fourier transform of $\exp\left(-Dk^2t\right)$. Note, with localised initial conditions $\Theta(x,0) = \Theta_0\delta(x)$, the solution is exactly the fundamental solution:

$$\Theta(x,t) = \Theta_0 S_d(x,t) = \frac{\Theta_0}{\sqrt{4\pi Dt}}\exp\left(-\eta^2\right); \quad \eta = \frac{x}{2\sqrt{Dt}}$$

where η is the similarity parameter.

7.2. Gaussian pulse for heat equation

Suppose that the initial conditions for the head equation are given by

$$f(x) = \sqrt{\frac{a}{\pi}}\Theta_0 e^{-ax^2}$$

Then, our previous solution gives

$$\begin{aligned}\Theta(x,t) &= \frac{\Theta_0\sqrt{a}}{\sqrt{4\pi^2 Dt}}\int_{-\infty}^{\infty}\exp\left[-au^2 - \frac{(x-u)^2}{4Dt}\right]\mathrm{d}u \\ &= \frac{\Theta_0\sqrt{a}}{\sqrt{4\pi^2 Dt}}\int_{-\infty}^{\infty}\exp\left[-\frac{(1+4aDt)u^2 - 2xu + x^2}{4Dt}\right]\mathrm{d}u \\ &= \frac{\Theta_0\sqrt{a}}{\sqrt{4\pi^2 Dt}}\int_{-\infty}^{\infty}\exp\left[-\frac{1+4aDt}{4Dt}\left(u - \frac{x}{1+4aDt}\right)\right]\exp\left[\frac{-ax^2}{1+4aDt}\right]\mathrm{d}u\end{aligned}$$

Recall that

$$\int_{-\infty}^{\infty}\exp\left[\frac{-(u-\mu)^2}{\sigma^2}\right]\mathrm{d}u = \sigma\sqrt{\pi}$$

The integral above is a Gaussian, so its solution can be read off directly as

$$\Theta(x,t) = \frac{\Theta_0\sqrt{a}}{\sqrt{\pi(1+4\pi^2 Dt)}}\exp\left[\frac{-ax^2}{1+4aDt}\right]$$

So the width of the Gaussian pulse will get wider over time, according to $\sigma^2 \sim t$, as it evolves according to the heat equation. The area is constant, so heat energy is conserved in the system.

7.3. Forced diffusion equation

Consider the equation

$$\frac{\partial}{\partial t}\Theta(x,t) - D\frac{\partial^2\Theta}{\partial x^2} = f(x,t)$$

subject to homogeneous initial conditions $\Theta(x,0) = 0$. We construct a two-dimensional Green's function $G(x,t;\xi,\tau)$ such that

$$\frac{\partial}{\partial t}G(x,t) - D\frac{\partial^2 G}{\partial x^2} = \delta(x-\xi)\delta(t-\tau)$$

subject to the same homogeneous boundary conditions $G(x,0;\xi,\tau) = 0$. Consider the Fourier transform with respect to x.

$$\frac{\partial\widetilde{G}}{\partial t} + Dk^2\widetilde{G} = e^{-ik\xi}\delta(t-\tau)$$

We can solve this using an integrating factor e^{Dk^2t} and integrating with respect to time. Since $G = 0$ at $t = 0$,

$$\frac{\partial}{\partial t}\left[e^{Dk^2t}\widetilde{G}\right] = e^{-ik\xi+Dk^2t}\delta(t-\tau)$$
$$\int_0^t \frac{\partial}{\partial t'}\left[e^{Dk^2t'}\widetilde{G}\right]\mathrm{d}t' = \int_0^t e^{-ik\xi+Dk^2t'}\delta(t'-\tau)\,\mathrm{d}t'$$
$$e^{Dk^2t}\widetilde{G} = e^{-ik\xi}\int_0^t e^{Dk^2t'}\delta(t'-\tau)\,\mathrm{d}t'$$
$$e^{Dk^2t}\widetilde{G} = e^{-ik\xi}e^{Dk^2\tau}H(t-\tau)$$

where H is the Heaviside step function. Thus,

$$\widetilde{G}(k,t;\xi,\tau) = e^{-ik\xi}e^{-Dk^2(t-\tau)}H(t-\tau)$$

The inverse Fourier transform gives the Green's function.

$$G(x,t;\xi,\tau) = \frac{H(t-\tau)}{2\pi}\int_{-\infty}^{\infty} e^{-ik\xi}e^{-Dk^2(t-\tau)}e^{ikx}\,\mathrm{d}k$$

This is a Gaussian; by changing variables into $x' = x - \xi$ and $t' = t - \tau$ we find

$$G(x,t;\xi,\tau) = \frac{H(t')}{2\pi}\int_{-\infty}^{\infty} e^{ikx'}e^{-Dk^2t'}\,\mathrm{d}k = \frac{H(t')}{\sqrt{4\pi Dt'}}\exp\left[-\frac{(x')^2}{4Dt'}\right]$$

Converting back,

$$G(x,t;\xi,\tau) = \frac{H(t-\tau)}{\sqrt{4\pi D(t-\tau)}}\exp\left[-\frac{(x-\xi)^2}{4D(t-\tau)}\right] = H(t-\tau)S_d(x-\xi,t-\tau)$$

where S_d is the fundamental solution as above. Thus, the general solution is

$$\Theta(x,t) = \int_0^\infty \mathrm{d}\tau\int_{-\infty}^{\infty}\mathrm{d}\xi\, G(x,t;\xi,\tau)f(\xi,\tau)$$

Let $\xi = u$, then

$$\Theta(x,t) = \int_0^t \mathrm{d}\tau\int_{-\infty}^{\infty}\mathrm{d}u\, f(u,\tau)S_d(x-u,t-\tau)$$

7.4. Duhamel's principle

In the above equation, omitting the integral over time, this is exactly the solution as found earlier with initial conditions at $t = \tau$, which was

$$\Theta(x,t) = \int_{-\infty}^{\infty}\mathrm{d}u\, f(u)S_d(x-u,t-\tau)$$

The forced PDE with homogeneous boundary conditions can be related to solutions of the homogeneous PDE with inhomogeneous boundary conditions. The forcing term $f(x,t)$ at $t=\tau$ acts as an initial condition for subsequent evolution. Thus, the solution is a superposition of the effects of the initial conditions integrated over $0<\tau<t$. This relation between the homogeneous and inhomogeneous problems is known as *Duhamel's principle*.

7.5. Forced wave equation

Consider the forced wave equation, given by

$$\frac{\partial^2\phi}{\partial t^2}-c^2\frac{\partial^2\phi}{\partial x^2}=f(x,t)$$

with $\phi(x,0)=\phi_t(x,0)=0$. We construct the Green's function using

$$\frac{\partial^2 G}{\partial t^2}-c^2\frac{\partial^2 G}{\partial x^2}=\delta(x-\xi)\delta(t-\tau)$$

with $G(x,0)=\phi_t(x,0)=0$. We take the Fourier transform with respect to x, and find

$$\frac{\partial^2\widetilde{G}}{\partial t^2}+c^2k^2\widetilde{G}=e^{-ik\xi}\delta(t-\tau)$$

We can solve this by inspection by comparing with the corresponding initial value problem Green's function, and find

$$\widetilde{G}=\begin{cases}0 & t<\tau\\ e^{-ik\xi}\frac{\sin kc(t-\tau)}{kc} & t>\tau\end{cases}$$

Using the Heaviside function.

$$\widetilde{G}=e^{-ik\xi}\frac{\sin kc(t-\tau)}{kc}H(t-\tau)$$

We invert the Fourier transform.

$$G(x,t;\xi,\tau)=\frac{H(t-\tau)}{2\pi c}\int_{-\infty}^{\infty}e^{ik(x-\xi)}\frac{\sin kc(t-\tau)}{k}\,\mathrm{d}k$$

Let $A=x-\xi$, and $B=ct-\tau$. By oddness of sine, only the cosine term of the complex exponential remains. Noting the similarity to the Dirichlet discontinuous function,

$$\begin{aligned}G(x,t;\xi,\tau)&=\frac{H(t-\tau)}{\pi c}\int_0^{\infty}\frac{\cos(kA)\sin(kB)}{k}\,\mathrm{d}k\\&=\frac{H(t-\tau)}{2\pi c}\int_0^{\infty}\frac{\sin k(A+B)-\sin k(A-B)}{k}\,\mathrm{d}k\\&=\frac{H(t-\tau)}{4c}[\operatorname{sgn}(A+B)-\operatorname{sgn}(A-B)]\end{aligned}$$

Since the $H(t-\tau)$ term is nonzero only for $t > \tau$, we must have $B = c(t-\tau) > 0$. The only way that the bracketed term can be nonzero is when $|A| < B$; so $|x-\xi| < c(t-\tau)$. This is the domain of dependence as found before, demonstrating the causality of the relation. Hence,

$$G(x,t;\xi,\tau) = \frac{1}{2c} H(c(t-\tau) - |x-\xi|)$$

Thus, the solution is

$$\begin{aligned}\phi(x,t) &= \int_0^\infty \mathrm{d}\tau \int_{-\infty}^\infty \mathrm{d}\xi\, f(\xi,t) G(x,t;\xi,\tau) \\ &= \frac{1}{2c} \int_0^t \mathrm{d}\tau \int_{x-c(t-\tau)}^{x+c(t-\tau)} \mathrm{d}\xi\, f(\xi,\tau)\end{aligned}$$

7.6. Poisson's equation

Consider

$$\nabla^2 \phi = -\rho(r)$$

defined on a three-dimensional domain D, with Dirichlet boundary conditions $\phi = 0$ on a boundary ∂D. The Dirac δ function, when defined in $\mathbb{R}^3$, has the following properties.

(i) $\delta(r - r') = 0$ for all $r \neq r'$;

(ii) $\int_D \delta(r-r')\,\mathrm{d}^3 r = 1$ if $r' \in D$, and zero otherwise;

(iii) $\int_D f(r)\delta(r-r')\,\mathrm{d}^3 r = f(r')$.

First, we consider $D = \mathbb{R}^3$ with the homogeneous boundary conditions that $G \to 0$ as $\|r\| \to \infty$. This is known as the *free-space* Green's function, denoted G_{FS}. The potential here is spherically symmetric, so the Green's function is a function only of the distance between the point and the source. WIthout loss of generality, let $r' = 0$, so G is a function only of the radius, now denoted r. Integrating the left hand side of Poisson's equation over a ball B with radius r around zero, we find

$$\int_B \nabla^2 G_{\mathrm{FS}}\,\mathrm{d}^3 r = \int_{\partial B} \nabla G_{\mathrm{FS}} \cdot \hat{n}\,\mathrm{d}S = \int_{\partial B} \frac{\partial G}{\partial r} r^2\,\mathrm{d}\Omega$$

where $\mathrm{d}\Omega$ is the angle element. This gives

$$\int_B \nabla^2 G_{\mathrm{FS}}\,\mathrm{d}^3 r = 4\pi r^2 \frac{\partial G_{\mathrm{FS}}}{\partial r}$$

The right hand side of Poisson's equation gives unity, since zero is contained in the ball. Therefore,

$$\frac{\partial G_{\mathrm{FS}}}{\partial r} = \frac{1}{4\pi r^2} \implies G_{\mathrm{FS}} = \frac{-1}{4\pi r} + c$$

Since $G \to 0$ as $r \to \infty$, we must have $c = 0$. The fundamental solution is therefore the free-space Green's function given by

$$G(r; r') = \frac{-1}{4\pi \|r - r'\|}$$

Thus, Poisson's equation is solved by

$$\Phi(r) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{\rho(r')}{\|r - r'\|} \, \mathrm{d}^3 r'$$

7.7. Green's identities

Consider scalar functions ϕ, ψ which are twice differentiable on a domain D. By the divergence theorem, *Green's first identity* is

$$\int_D \nabla \cdot (\phi \nabla \psi) \, \mathrm{d}^3 r = \int_D \left(\phi \nabla^2 \psi + \nabla \phi \cdot \nabla \psi \right) \mathrm{d}^3 r = \int_{\partial D} \phi \nabla \psi \cdot \hat{n} \, \mathrm{d}S$$

Switching ψ and ϕ and subtracting from the above, we arrive at *Green's second identity*:

$$\int_{\partial D} \left(\phi \frac{\partial \psi}{\partial \hat{n}} - \psi \frac{\partial \phi}{\partial \hat{n}} \right) \mathrm{d}S = \int_D \left(\phi \nabla^2 \psi - \psi \nabla^2 \phi \right) \mathrm{d}^3 r$$

Suppose we remove a ball $\mathcal{B}_\varepsilon(r')$ from the domain. Without loss of generality let $r' = 0$. Let ϕ be a solution to Poisson's equation, so $\nabla^2 \phi = -\rho$ and let ψ be the free-space Green's function. Thus, the right hand side of the second identity becomes

$$\int_{D \setminus \mathcal{B}_\varepsilon} \left(\phi \nabla^2 G_{\mathrm{FS}} - G_{\mathrm{FS}} \nabla^2 \phi \right) \mathrm{d}^3 r = \int_{D \setminus \mathcal{B}_\varepsilon} G_{\mathrm{FS}} \rho \, \mathrm{d}^3 r$$

The left hand side is

$$\int_{\partial D} \left(\phi \frac{\partial G_{\mathrm{FS}}}{\partial \hat{n}} - G_{\mathrm{FS}} \frac{\partial \phi}{\partial \hat{n}} \right) \mathrm{d}S + \int_{\partial \mathcal{B}_\varepsilon} \left(\phi \frac{\partial G_{\mathrm{FS}}}{\partial \hat{n}} - G_{\mathrm{FS}} \frac{\partial \phi}{\partial \hat{n}} \right) \mathrm{d}S$$

For the second integral, we take the limit as $\varepsilon \to 0$. Let ϕ be regular, and let $\overline{\phi}$ be the average value and $\overline{\frac{\partial \phi}{\partial \hat{n}}}$ be the average derivative. This integral then becomes

$$\left(\overline{\phi} \frac{-1}{4\pi\varepsilon^2} - \frac{1}{4\pi\varepsilon} \overline{\frac{\partial \phi}{\partial \hat{n}}} \right) 4\pi\varepsilon^2 \to -\phi(0)$$

Combining the above, we find *Green's third identity*, which is

$$\phi(r') = \int_D G_{\mathrm{FS}}(r; r')(-\rho(r)) \, \mathrm{d}^3 r + \int_{\partial D} \left(\phi(r) \frac{\partial G_{\mathrm{FS}}}{\partial \hat{n}}(r; r') - G_{\mathrm{FS}}(r; r') \frac{\partial \phi}{\partial \hat{n}}(r) \right) \mathrm{d}S$$

The second integral provides the ability to use inhomogeneous boundary conditions

7.8. Dirichlet Green's function

We will solve Poisson's equation $\nabla^2\phi = -\rho$ on D with inhomogeneous boundary conditions $\phi(r) = h(r)$ on ∂D. The Dirichlet Green's function satisfies

(i) $\nabla^2 G(r;r') = 0$ for all $r \neq r'$;

(ii) $G(r;r') = 0$ on ∂D;

(iii) $G(r;r') = G_{\text{FS}}(r;r') + H(r;r')$ where H satisfies Laplace's equation, the homogeneous version of Poisson's equation, for all $r \in D$.

Green's second identity with $\nabla^2\phi = -\rho$, $\nabla^2 H = 0$ gives

$$\int_{\partial D}\left(\phi\frac{\partial H}{\partial \hat{n}} - H\frac{\partial \phi}{\partial \hat{n}}\right)\mathrm{d}S = \int_D H\rho\,\mathrm{d}^3r$$

Now, we set $G_{\text{FS}} = G - H$ into Green's third identity to find

$$\phi(r') = \int_D (G-H)(-\rho)\,\mathrm{d}^3r + \int_{\partial D}\left(\phi\frac{\partial(G-H)}{\partial \hat{n}} - (G-H)\frac{\partial\phi}{\partial n}\right)\mathrm{d}S$$

All of the H terms can be cancelled by substituting the form of the second identity the derived above. Now, given $G = 0, \phi = h$ on ∂D, we have

$$\phi(r') = \int_D G(r;r')(-\rho(r))\,\mathrm{d}^3r + \int_{\partial D} h(r)\frac{\partial G(r;r')}{\partial \hat{n}}\,\mathrm{d}S$$

This is the general solution. The first integral is the Green's function solution, and the second integral yields the inhomogeneous boundary conditions.

7.9. Method of images for Laplace's equation

For symmetric domains D, we can construct Green's functions with $G = 0$ on ∂D by cancelling the boundary potential out by using an opposite 'mirror image' Green's function placed outside the domain. Consider Laplace's equation $\nabla^2\phi = 0$ on half of $\mathbb{R}^3$, in particular, the subset of $\mathbb{R}^3$ such that $z > 0$. Let $\phi(x,y,0) = h(x,y)$ and $\phi \to 0$ as $r \to \infty$. The free space Green's function satisfies $G_{\text{FS}} \to 0$ as $r \to \infty$, but does not satisfy the boundary condition that $G_{\text{FS}} = 0$ at $z = 0$. For G_{FS} at $r' = (x', y', z')$, we will subtract a copy of G_{FS} located at $r'' = (x', y', -z')$. This gives

$$\begin{aligned} G(r,r') &= \frac{-1}{4\pi|r-r'|} - \frac{-1}{4\pi|r-r''|} \\ &= \frac{-1}{4\pi\sqrt{(x-x')^2+(y-y')^2+(z-z')^2}} + \frac{1}{4\pi\sqrt{(x-x')^2+(y-y')^2+(z+z')^2}} \end{aligned}$$

Hence $G((x,y,0),r') = 0$, so this function satisfies the Dirichlet boundary conditions on all of the boundary ∂D. We have

$$\left.\frac{\partial G}{\partial \hat{n}}\right|_{z=0} = \left.\frac{\partial G}{\partial z}\right|_{z=0} = \frac{-1}{4\pi}\left(\frac{z-z'}{|r-r'|^3} - \frac{z+z'}{|r-r'|^3}\right) = \frac{z'}{2\pi}\left((x-x')^2+(y-y')^2+(z')^2\right)^{-3/2}$$

The solution is then given by

$$\Phi(x', y', z') = \frac{z'}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[(x - x')^2 + (y - y')^2 + (z')^2\right]^{-3/2} h(x, y)\, \mathrm{d}x\, \mathrm{d}y$$

7.10. Method of images for wave equation

Consider the one-dimensional wave equation

$$\ddot{\phi} - c^2 \phi'' = f(x, t)$$

with Dirichlet boundary conditions $\phi(0, t) = 0$. We create matching Green's functions with an opposite sign centred at $-\xi$.

$$G(x, t; \xi, \tau) = \frac{1}{2c} H(c(t - \tau) - |x - \xi|) - \frac{1}{2c} H(c, (t - \tau) - |x + \xi|)$$

We can replace the addition of the two terms with a subtraction to instead use Neumann boundary conditions. Suppose we wish to solve the homogeneous problem with $f = 0$ for initial conditions of a Gaussian pulse. Here, for $x > 0$ we have

$$\phi(x, t) = \exp\left[-(x - \xi + ct)^2\right] - \exp\left[-(-x - \xi + ct)^2\right]$$

The solution travels to the left, cancelling with the image at $t = \frac{\xi}{c}$, which emerges and travels right as the reflected wave.

VI. Quantum Mechanics

Lectured in Michaelmas 2021 by Dr. M. Ubiali

In this course, we explore the basics of quantum mechanics using the Schrödinger equation. This equation explains how a quantum wavefunction changes over time. By solving the Schrödinger equation with different inputs and boundary conditions, we can understand some of the ways in which quantum mechanics differs from classical physics, explaining some of the scientific discoveries of the past century. We prove some theoretical facts about quantum operators and observables, such as the uncertainty theorem, which roughly states that it is impossible to know both the position and momentum of a particle.

Contents

1. Historical introduction 300
1.1. Timeline 300
1.2. Particles and waves in classical mechanics 300
1.3. Black-body radiation 301
1.4. Planck's constant 301
1.5. Photoelectric effect 302
1.6. Compton scattering 302
1.7. Atomic spectra 303
2. Wavefunctions 305
2.1. Wave-like behaviour of particles 305
2.2. Probabilistic interpretation of wavefunctions 305
2.3. Bases and equivalence classes 306
2.4. Hilbert spaces 306
2.5. Inner product 307
2.6. Normalisation 307
2.7. Time-dependent Schrödinger equation 308
2.8. Normalisation and time evolution 309
2.9. Conserved probability current 309
3. Observables and operators 310
3.1. Expectation and operators 310
3.2. Dynamical observables 310
3.3. Hamiltonian operator 311
3.4. Time-independent Schrödinger equation 312
4. One-dimensional solutions to the Schrödinger equation 313
4.1. Stationary states 313
4.2. Infinite potential well 313
4.3. Finite potential well 314
4.4. Free particles 316
4.5. Gaussian wavepacket 317
4.6. Beam interpretation 318
4.7. Scattering states 319
4.8. Scattering off potential step 319
4.9. Scattering off a potential barrier 321
4.10. Harmonic oscillator 322
5. Operators and measurements 325
5.1. Hermitian operators 325
5.2. Postulates of quantum mechanics 326

5.3. Expectation of operators . . . 327
5.4. Commutators . . . 328
5.5. Simultaneously diagonalisable operators . . . 328
5.6. Uncertainty . . . 329
5.7. Schwarz inequality . . . 330
5.8. Generalised uncertainty theorem . . . 330
5.9. Consequences of uncertainty relation . . . 332
5.10. States of minimal uncertainty . . . 332
5.11. Ehrenfest theorem . . . 332
6. Three-dimensional solutions to the Schrödinger equation . . . 335
6.1. Time-independent Schrödinger equation in spherical polar coordinates . . . 335
6.2. Spherically symmetric potential well . . . 336
7. Solution to hydrogen atom . . . 337
7.1. Radial wavefunction of hydrogen atom . . . 337
7.2. Angular momentum . . . 339
7.3. Commutativity of angular momentum operators . . . 340
7.4. Joint eigenfunctions of angular momentum . . . 341
7.5. Full solution to hydrogen atom . . . 342
7.6. Comparison to Bohr model . . . 344
7.7. Other elements of the periodic table . . . 345

1. Historical introduction

1.1. Timeline

- (1801–3) Particles were shown to have wave-like properties using Young's double slit experiment.
- (1862–4) Electromagnetism was conceived by Maxwell. Light was discovered to be an electromagnetic wave.
- (1897) Discovery of the electron by Thomson.
- (1900) The Planck law was discovered, which explains black-body radiation.
- (1905) The photoelectric effect was discovered by Einstein.
- (1909) Wave-light interference patterns were shown to exist with only one photon recorded at a time.
- (1911) Rutherford created his atomic model.
- (1913) Bohr created his atomic model.
- (1923) The Compton experiment showed x-ray scattering off electrons.
- (1923–4) De Broglie discovered the concept of wave-particle duality.
- (1925–30) The theory of quantum mechanics emerged at this time.
- (1927–8) The diffraction experiment was carried out with electrons.

1.2. Particles and waves in classical mechanics

In classical mechanics, a point-particle is an object with energy and momentum in an infinitesimally small point of space. Therefore, a particle is determined by the three-dimensional vectors $\mathbf{x}, \mathbf{v} = \dot{\mathbf{x}}$. The motion of a particle is governed by Newton's second law,

$$m\ddot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, \dot{\mathbf{x}})$$

Solving this equation involves determination of $\mathbf{x}, \dot{\mathbf{x}}$ for all $t > t_0$, once initial conditions $\mathbf{x}(t_0), \dot{\mathbf{x}}(t_0)$ are known.

Waves are classically defined as any real- or complex-valued function with periodicity in time and/or space. For instance, consider a function f such that $f(t+T) = f(t)$, which is a wave with period T. The frequency ν is defined to be $\frac{1}{T}$, and the angular frequency ω is defined as $2\pi\nu = \frac{2\pi}{T}$. Suppose we have a function in one dimension obeying $f(x+\lambda) = f(x)$. This has wavelength λ and wave number $k = \frac{2\pi}{\lambda}$.

Consider $f(x) = \exp(\pm ikx)$. In three dimensions, this becomes $f(x) = \exp(\pm i\mathbf{k} \cdot \mathbf{x})$. This is called a 'plane wave'; the one-dimensional wave number k has been transformed into a three-dimensional wave vector $\mathbf{k}$. λ is now defined as $\frac{2\pi}{|k|}$.

The wave equation in one dimension is

$$\frac{\partial^2 f(x,t)}{\partial t^2} - c^2 \frac{\partial^2 f(x,t)}{\partial x^2} = 0; \quad c \in \mathbb{R}$$

The solutions to this equation are

$$f_\pm(x,t) = A_\pm \exp(\pm ikx - i\omega t)$$

where $\omega = ck; \lambda = \frac{c}{\nu}$. The two conditions are known as the dispersion relations. $A_\pm$ is the amplitude of the waves.

In three dimensions,

$$\frac{\partial^2 f(\mathbf{x},t)}{\partial t^2} - c^2 \nabla^2 f(\mathbf{x},t) = 0; \quad c \in \mathbb{R}$$

The solution is

$$f(\mathbf{x},t) = A \exp(\pm i\mathbf{k} \cdot \mathbf{x} - i\omega t)$$

where $\omega = c|\mathbf{k}|; \lambda = \frac{c}{\nu}$.

Note. Other kinds of waves are solutions to other governing equations, provided that another dispersion relation $\omega(\mathbf{k})$ is given. Also, for any governing equation linear in f, the superposition principle holds: if f_1, f_2 are solutions then so is $f_1 + f_2$.

1.3. Black-body radiation

Several experiments have shown that light behaves with some particle-like characteristics. For example, consider a body heated at some temperature T. Any such body will emit radiation. The simplest body to study is called a 'black-body', which is a totally absorbing surface. The intensity of light emitted by a black body was modelled as a function of the frequency. The classical prediction for the spectrum of emitted radiation was that as the frequency increased, the intensity would also increase. A curve with a clear maximum point was observed. Planck's law was found to be the equation of this curve, which can be derived from the equation $E = \hbar\omega$ involving the Planck constant, instead of the classical energy equation $E = k_B T$ involving the Boltzmann constant. This then implies that light was 'quantised' into particles.

1.4. Planck's constant

The Planck constant is $h \approx 6.61 \times 10^{-34}\,\mathrm{J\,s}$. The reduced Planck constant is $\hbar = \frac{h}{2\pi}$. Quantum mechanics typically uses the reduced Planck constant over the normal Planck constant. The dimensionality of h is energy multiplied by time, or position multiplied by momentum.

1.5. Photoelectric effect

Consider a metal surface in a vacuum, which is hit by light with angular frequency ω. When the radiation hits the surface of the metal, electrons were emitted. Classically, we would expect that:

(i) Since the incident light carries energy proportional to its intensity, increasing the intensity we should have sufficient energy to break the bonds of the electrons with the atoms of the metal.

(ii) Since the intensity and frequency are independent, light of any ω would eventually cause electrons to be emitted, given a high enough intensity.

(iii) The emission rate should be constant.

In fact, the experiment showed that

(i) The maximum energy $E_{\max}$ of emitted electrons depended on ω, and not on the intensity.

(ii) Below a given threshold $\omega_{\min}$, there was no electron emission.

(iii) The emission rate increased with the intensity.

Einstein's explanation for this phenomenon was that the light was quantised into small quanta, called photons. Photons each carry an energy $E = \hbar\omega$. Each photon could liberate only one electron. Thus,

$$E_{\max} = \hbar\omega - \phi$$

where ϕ is the binding energy of the electron with the metal. The higher the intensity, the more photons hit the metal. This implies that more electrons will be scattered.

1.6. Compton scattering

X-rays were emitted towards a crystal, scattering free electrons. The X-ray should then be deflected by some angle θ. Classically, for a given θ we would expect that the intensity as a function of ω would have a maximum at ω_0, the frequency of the incoming X-rays. This is because we would not expect ω to change much after scattering an electron. However, there was another peak at ω', which was dependent on the angle θ. In fact, considering the photon and electron as a relativistic system of particles, we can derive (from IA Dynamics and Relativity),

$$2\sin^2\frac{\theta}{2} = \frac{mc}{|\mathbf{q}|} - \frac{mc}{|\mathbf{p}|}$$

where $\mathbf{p}$ is the initial momentum and $\mathbf{q}$ is the final momentum. Assuming $E = \hbar\omega$ and $\mathbf{p} = \hbar\mathbf{k}$,

$$|\mathbf{p}| = \hbar|\mathbf{k}| = \hbar\frac{\omega}{c}; \quad |\mathbf{q}| = \hbar|\mathbf{k}'| = \hbar\frac{\omega'}{c}$$

Hence,

$$\frac{1}{\omega} = \frac{1}{\omega'} + \frac{\hbar}{mc}(1 - \cos\theta)$$

So the frequency of the outgoing X-ray should have an angular frequency which is shifted from the original. The expected peak was actually caused by X-rays simply not interacting with the electrons.

1.7. Atomic spectra

The Rutherford scattering experiment involved shooting α particles at some thin gold foil. Most particles travelled through the foil, some were slightly deflected, and some were deflected completely back. This indicated that the gold foil was mostly comprised of vacuum and there was a high density of positive charge within the atom. Electrons would orbit around the nucleus. However, there were problems with this model:

(i) If the electrons in orbits moved, they would radiate and lose energy. However if the electrons were static, they would simply collapse and fall into the nucleus.

(ii) This model did not explain the atomic spectra, the observed frequencies of light absorbed or emitted by an atom when electrons change energy levels.

The spectra had frequency

$$\omega_{mn} = 2\pi c R_0\left(\frac{1}{n^2} - \frac{1}{mc}\right); \quad m, n \in \mathbb{N}, m > n$$

where R_0 is the Rydberg constant, approximately $1 \times 10^7\,\mathrm{m}^{-1}$. Bohr theorised that the electron orbits themselves are quantised, so L (the orbital angular momentum) is an integer multiple of $\hbar$; $L_n = n\hbar$. First, the quantisation of L implies the quantisation of v and r. Indeed, given that $L \equiv m_e v r$, we have that v is quantised: $v_n = \frac{n\hbar}{m_e r}$. Further, by the Coulomb force, $F = \frac{e^2}{4\pi\varepsilon^2}\frac{1}{r^2}\mathbf{e}_r = m_e a_r \mathbf{e}_r$ where a_r is the radial acceleration. Then $\frac{e^2}{4\pi\varepsilon^2}\frac{1}{r^2} = m_e\frac{v^2}{r} \implies r = r_n = \frac{4\pi\varepsilon_0\hbar^2}{m_e e^2}n^2$. The coefficient on n^2 is known as the Bohr radius. Immediately then the energy levels E of the atom can be shown to be quantised, since

$$E = \frac{1}{2}m_e v^2 - \frac{e^2}{4\pi\varepsilon_0}\frac{1}{r}$$

giving

$$E_n = -\frac{e^2}{8\pi\varepsilon_0 a_0}\frac{1}{n^2} = \frac{-e^4 m_e}{32\pi^2\varepsilon_0^2\hbar^2}\frac{1}{n^2}$$

The ground energy level is at $n = 1$, giving

$$E_1 = -13.6\,\mathrm{eV}$$

The excited states are E_n for $n > 1$. The energy emitted when descending from E_n to E_1 are the spectral lines:

$$\Delta E = \hbar\omega$$

The Bohr model gives

$$\omega_{mn} = \frac{\Delta E_{mn}}{\hbar} = 2\pi c\left(\frac{e^2}{4\pi\varepsilon_0\hbar c}\right)^2\left(\frac{1}{n^2} - \frac{1}{m^2}\right)$$

which agrees with the Rydberg constant R_0 defined earlier.

2. Wavefunctions

2.1. Wave-like behaviour of particles

De Broglie hypothesised that any particle of any mass is associated with a wave with

$$\omega = \frac{E}{\hbar}; \quad \mathbf{k} = \frac{\mathbf{p}}{\hbar}$$

This hypothesis made sense of the quantisation of electron angular momentum; if the electron lies on a circular orbit then $2\pi r = n\lambda$ where λ is the wavelength of the electron. However,

$$p = \hbar k = \hbar\frac{2\pi}{\lambda} \implies L = m_e v r = pr = \hbar\frac{2\pi}{\lambda}\frac{n\lambda}{2\pi} = n\hbar$$

Hence the angular momentum must be quantised. The electron diffraction experiment showed that this hypothesis was true, by showing that electrons behaved sufficiently like waves. Interference patterns were observed with $\lambda = \frac{2\pi}{|\mathbf{k}|} = \frac{2\pi k}{|\mathbf{p}|}$ compatible with the De Broglie hypothesis.

2.2. Probabilistic interpretation of wavefunctions

In classical mechanics, we can describe a particle with $\mathbf{x}, \dot{\mathbf{x}}$ or $\mathbf{p} = m\dot{\mathbf{x}}$. In quantum mechanics, we need the state ψ described by $\psi(\mathbf{x}, t)$ called the wavefunction.

Remark. Note that the state is an abstract entity, while $\psi(\mathbf{x}, t)$ is the representation of ψ in the space of $\mathbf{x}$. In some sense, $\psi(\mathbf{x}, t)$ is the complex coefficient of ψ in the continuous basis of $\mathbf{x}$. In other words, $\psi(\mathbf{x}, t)$ is ψ in the $\mathbf{x}$ representation. In this course, we always work in the $\mathbf{x}$ representation.

Definition. A wavefunction is a function $\psi(\mathbf{x}, t) : \mathbb{R}^3 \to \mathbb{C}$ that satisfies certain mathematical properties (defined later) dictated by its physical interpretation. t is considered a fixed external parameter, so it is not included in the function's type.

The physical interpretation of a wavefunction is called Born's rule. The probability density for a particle to be at some point $\mathbf{x}$ at t is given by $|\psi(\mathbf{x}, t)|^2$. We write the probability density as ρ, hence $\rho(\mathbf{x}, t)\,\mathrm{d}V$ is the probability that the particle lies in some small volume V centred at $\mathbf{x}$. Now, since the particle must be somewhere, the wave function must be *normalisable*, or *square-integrable* in $\mathbb{R}^3$:

$$\int_{\mathbb{R}^3} \psi^\star(\mathbf{x}, t)\psi(\mathbf{x}, t)\,\mathrm{d}V = \int_{\mathbb{R}^3} |\psi(\mathbf{x}, t)|^2\,\mathrm{d}V = N \in (0, \infty)$$

Since we want the total probability to be 1, we must normalise the wavefunction by defining

$$\overline{\psi}(\mathbf{x}, t) = \frac{1}{\sqrt{N}}\psi(\mathbf{x}, t) \iff \int_{\mathbb{R}^3} \left|\overline{\psi}(\mathbf{x}, t)\right|^2 \mathrm{d}V = 1$$

Hence, $\rho(\mathbf{x}, t) = \left|\overline{\psi}(\mathbf{x}, t)\right|^2$ really is a probability density. From now, we will not use the bar for denoting normalisation, since normalisation is evident from context.

2.3. Bases and equivalence classes

In linear algebra, we consider vectors in some vector space such as $\mathbb{R}^n$. In quantum mechanics, we instead consider states in a space of wave functions. The analogous concept to vector components is to represent a state ψ in an infinite-dimensional x axis basis $\psi(x, t)$. Note that if two wavefunctions differ by a constant phase, that is, $\exists \alpha \in \mathbb{R}$ such that

$$\widetilde{\psi}(x, t) = e^{i\alpha}\psi(x, t)$$

then the states are equivalent in terms of probability, since the probability density is given by the norm of ψ, not its angle. We can think of states as arrays in the vector space of wavefunctions. We can then describe the equivalence class $[\psi]$ as the set of all functions ϕ such that $\phi = \lambda\psi$, for some $\lambda \in \mathbb{C}\backslash\{0\}$, since we must retain the condition that ϕ is normalisable.

2.4. Hilbert spaces

In quantum mechanics, we are interested in the functional space of square-integrable functions on $\mathbb{R}^3$, which is a type of *Hilbert space* and denoted $\mathcal{H}$.

Remark. Since the set of wavefunctions form a vector space, $\psi_1, \psi_2 \in \mathcal{H}$ implies that $\psi = \lambda_1\psi_1 + \lambda_2\psi_2 \in \mathcal{H}$ for constants $\lambda_1, \lambda_2 \in \mathbb{C}$ provided this ψ is nonzero. For waves, this is the well-known superposition principle. Note that this exact formulation of linearity is unique to quantum mechanics; for example, in classical mechanics, two solutions to Newton's equations may not be combined into a new solution by taking their sum.

Proposition. If $\psi_1(x, t), \psi_2(x, t)$ are normalisable, then $\psi = \lambda_1\psi_i(x, t) + \lambda_2\psi_2(x, t)$ is also normalisable.

Proof. Recall the inequality

$$2|z_1||z_2| \leq |z_1|^2 + |z_2|^2$$

Then we can show

$$\begin{aligned}
\int_{\mathbb{R}^3} |\lambda_1\psi_1 + \lambda_2\psi_2|^2 \,\mathrm{d}V &= \int_{\mathbb{R}^3} \left(|\lambda_1\psi_1| + |\lambda_2\psi_2|\right)^2 \mathrm{d}V \\
&= \int_{\mathbb{R}^3} \left(|\lambda_1\psi_1|^2 + 2|\lambda_1\psi_1||\lambda_2\psi_2| + |\lambda_2\psi_2|^2\right) \mathrm{d}V \\
&= \int_{\mathbb{R}^3} \left(2|\lambda_1\psi_1|^2 + 2|\lambda_2\psi_2|^2\right) \mathrm{d}V < \infty
\end{aligned}$$

so the norm is non-infinite. □

2.5. Inner product

We define the inner product between two wavefunctions to be

$$\langle \psi, \phi \rangle = \int_{\mathbb{R}^3} \psi^\star \phi \, \mathrm{d}V$$

The following statements hold.

(i) $\langle \psi, \phi \rangle$ exists for all wave functions $\psi, \phi \in \mathcal{H}$;

(ii) $\langle \psi, \phi \rangle^\star = \langle \phi, \psi \rangle$;

(iii) the inner product is antilinear in the first entry, and linear in the second entry; and

(iv) for continuous ψ, $\langle \psi, \psi \rangle = 0$ is true if and only if ψ is identically zero.

We prove the first statement, since the others are obvious from the definition. By the Cauchy–Schwarz inequality,

$$\begin{aligned} \int_{\mathbb{R}^3} |\psi|^2 \, \mathrm{d}V &\leq N_1; \\ \int_{\mathbb{R}^3} |\phi|^2 \, \mathrm{d}V &\leq N_2; \\ \therefore \int_{\mathbb{R}^3} |\psi\phi| \, \mathrm{d}V &\leq \sqrt{\int_{\mathbb{R}^3} |\psi|^2 \, \mathrm{d}V \cdot \int_{\mathbb{R}^3} |\phi|^2 \, \mathrm{d}V} < \infty \end{aligned}$$

2.6. Normalisation

Definition. We define the norm of a wavefunction to be $\|\psi\| \equiv \langle \psi, \psi \rangle$. A wavefunction ψ is *normalised* if $\|\psi\| = 1$.

Definition. A set of wavefunctions $\{\psi_n\}$ is *orthonormal* if $\langle \psi_m, \psi_n \rangle = \delta_{mn}$. A set of wavefunctions $\{\psi_n\}$ is *complete* if for any $\psi \in \mathcal{H}$, we can write

$$\psi = \sum_n \lambda_n \psi_n$$

for $\lambda_n \in \mathbb{C}$.

Proposition. If $\{\psi_n\}$ is a complete and orthonormal basis of $\mathcal{H}$, then

$$\phi = \sum_{k=0}^{n} c_k \psi_k$$

where

$$c_k = \langle \psi_k, \phi \rangle$$

Proof. Suppose we can write ϕ in this form. Then,

$$\begin{aligned}\langle \psi_n, \phi \rangle &= \left\langle \psi_n, \sum_m c_m \psi_m \right\rangle \\ &= \sum_m c_m \langle \psi_n, \psi_m \rangle \\ &= \sum_m c_m \delta_{mn} \\ &= c_n\end{aligned}$$

□

Remark. If ϕ is the desired outcome of a measurement for a particle described by ψ, then the probability of observing ϕ given ψ at some time t is

$$|\langle \psi, \phi \rangle|^2 = \left| \int_{\mathbb{R}^3} \psi^* \phi \, \mathrm{d}V \right|^2$$

2.7. Time-dependent Schrödinger equation

Definition. The evolution of the wavefunction over time is given by the *time-dependent Schrödinger equation (TDSE)*,

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \nabla^2 \psi + U\psi$$

where $U = U(x)$ is a real potential energy term.

Remark. This equation is a first-order differential equation in t. Contrast this to Newton's second law, which is a second-order differential equation in t. This implies that we only need a single initial condition $\psi(x, t_0)$ to determine all future behaviour.

Remark. Note the asymmetry between the spatial and temporal components: there is only a first derivative in time but a second derivative in space. This implies that this equation is incompatible with relativity, where time and space must be treated equitably.

One way to conceptualise the TDSE is by letting ψ be some wave defined by

$$\psi(x, t) = \exp[i(k \cdot x - \omega t)]$$

Then, the De Broglie hypothesis ($k = p/\hbar, \omega = E/m$) implies that

$$\psi(x, t) = \exp\left[\frac{i}{\hbar}\left(p \cdot x - \frac{p^2}{2m} t\right)\right]$$

which is a solution to the TDSE.

2.8. Normalisation and time evolution

Because of the TDSE, we can show that the norm N of a wavefunction ψ is independent of t.

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \int_{\mathbb{R}^3} \frac{\partial}{\partial t} |\psi(x,t)|^2 \,\mathrm{d}V$$

Now, note that

$$\frac{\partial}{\partial t}|\psi|^2 = \frac{\partial}{\partial t}\langle \psi^\star, \psi\rangle = \psi^\star \frac{\partial \psi}{\partial t} + \psi \frac{\partial \psi^\star}{\partial t}$$

The TDSE then gives

$$\begin{aligned}
\frac{\partial \psi}{\partial t} &= \frac{i\hbar}{2m}\nabla^2\psi^2 + \frac{i}{k}U\psi; \\
\frac{\partial \psi^\star}{\partial t} &= -\frac{i\hbar}{2m}\nabla^2\psi^2 - \frac{i}{k}U\psi^\star \\
\therefore \frac{\partial |\psi|^2}{\partial t} &= \nabla \cdot \left[\frac{i\hbar}{2m}(\psi^\star \nabla\psi - \psi\nabla\psi^\star)\right]
\end{aligned}$$

Finally,

$$\int_{\mathbb{R}^3} \frac{\partial |\psi|^2}{\partial t}\,\mathrm{d}V = \int_{\mathbb{R}^3} \nabla \cdot \left[\frac{i\hbar}{2m}(\psi^\star \nabla\psi - \psi\nabla\psi^\star)\right] = 0$$

since $\psi, \psi^\star$ are such that $|\psi| \to 0$ as $|x| \to \infty$.

2.9. Conserved probability current

We have proven that the normalisation of wavefunctions are constant in time. Hence, we can derive the probability conservation law:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot J = 0; \quad J(x,t) = \frac{-i\hbar}{2m}(\psi^\star \nabla\psi - \psi\nabla\psi^\star)$$

This is the conserved probability current.

3. Observables and operators

3.1. Expectation and operators

Given the wavefunction, we would like to extract some information about the particle it represents.

Definition. An *observable* is a property of the particle that can be measured.

Definition. An *operator* is any linear map $\mathcal{H} \to \mathcal{H}$ such that

$$\hat{O}(a_1\psi_1 + a_2\psi_2) = a_1\hat{O}(\psi_1) + a_2\hat{O}(\psi_2)$$

where $a_1, a_2 \in \mathbb{C}, \psi_1, \psi_2 \in \mathcal{H}$.

In quantum mechanics, each observable is represented by an operator acting on the state ψ. Each measurement is represented by an expectation value of the operator. In comparison, in linear algebra we would often use a linear transformation for a similar purpose. Once we have a basis for a linear transformation, we have a matrix. In quantum mechanics, we use the x basis, so we can write

$$\widetilde{\psi} = (\hat{O})(x, t)$$

Example. Consider the class of finite differential operators

$$\sum_{n=0}^{N} p_n(x)\frac{\partial^n}{\partial x^n}$$

This includes, for example, position, momentum, and energy.

Example. A translation is an operator:

$$s_a : \psi(x) \mapsto \psi(x - a)$$

Example. The parity operator is

$$P : \psi(x) \mapsto \psi(-x)$$

3.2. Dynamical observables

In general, to calculate the expectation value of an observable, we place the operator between $\psi^\star$ and ψ and integrate over the whole space. From the probabilistic interpretation of the Born rule, the position of the particle can be interpreted as

$$\langle x \rangle = \int_{-\infty}^{+\infty} x|\psi(x,t)|^2 \,\mathrm{d}x = \int_{-\infty}^{+\infty} \psi^\star x \psi \,\mathrm{d}x$$

Hence, we can write the coefficient x as the operator $\hat{x}$. Now, consider the momentum. By considering the time-dependent Schrödinger equation with $U = 0$, and then integrating by parts,

$$\begin{aligned}
\langle p \rangle &= m \frac{\mathrm{d}}{\mathrm{d}t} \langle x \rangle \\
&= m \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{+\infty} x \psi^\star \psi \, \mathrm{d}x \\
&= m \int_{-\infty}^{+\infty} x \frac{\partial}{\partial t} (\psi^\star \psi) \, \mathrm{d}x \\
&= m \cdot \frac{i\hbar}{2m} \int_{-\infty}^{+\infty} x \frac{\partial}{\partial x} \left(\psi^\star \frac{\partial \psi}{\partial x} - \psi \frac{\partial \psi^\star}{\partial x} \right) \mathrm{d}x \\
&= \frac{-i\hbar}{2} \int_{-\infty}^{+\infty} x \frac{\partial}{\partial x} \left(\psi^\star \frac{\partial \psi}{\partial x} - \psi \frac{\partial \psi^\star}{\partial x} \right) \mathrm{d}x \\
&= \frac{-i\hbar}{2} \int_{-\infty}^{+\infty} \left(\psi^\star \frac{\partial \psi}{\partial x} - \psi \frac{\partial \psi^\star}{\partial x} \right) \mathrm{d}x \\
&= -i\hbar \int_{-\infty}^{+\infty} \psi^\star \frac{\partial \psi}{\partial x} \, \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \psi^\star \left(-i\hbar \frac{\partial}{\partial x} \right) \psi \, \mathrm{d}x
\end{aligned}$$

So the operator $\hat{p}$ is $-i\hbar \frac{\partial}{\partial x}$. Given x and p, we can write many classical dynamical observables. The classical notion is written in parentheses. The symbol $\mapsto$ is used instead of equality since we are representing the observable in the x basis.

$$\begin{aligned}
\hat{x} &\mapsto x \\
\hat{p} &\mapsto -i\hbar \frac{\partial}{\partial x} \\
\left(T = \frac{p^2}{2m} \right) \quad \hat{T} &\mapsto \frac{\hat{p}^2}{2m} = \frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \\
\hat{U} &\mapsto U(\hat{x}) = U(x)
\end{aligned}$$

3.3. Hamiltonian operator

The total energy is

$$E = T + U$$

given by the Hamiltonian operator

$$\hat{H} = \hat{T} + \hat{U}$$

In one dimension,

$$\hat{H} \mapsto \frac{-\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + U(x)\psi$$

In three dimensions,

$$\hat{H} \mapsto \frac{-\hbar^2}{2m}\nabla^2\psi + U(x)\psi$$

We can now represent the time-dependent Schrödinger equation in a more compact form:

$$i\hbar\frac{\partial\psi}{\partial t} = \hat{H}\psi$$

We can now prove that for a particle in a potential $U(x) \neq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle p\rangle = -\left\langle\frac{\partial U}{\partial x}\right\rangle$$

3.4. Time-independent Schrödinger equation

From the time-dependent version of the equation,

$$i\hbar\frac{\partial\psi}{\partial t} = \hat{H}\psi$$

we can try a solution of the form

$$\psi(x,t) = T(t)\chi(x)$$

Then, we can find

$$i\hbar\frac{\partial T(t)}{\partial t}\chi(x) = T(t)\hat{H}\chi(x)$$

Then, dividing by $T\chi$,

$$\frac{1}{T(t)}\left(i\hbar\frac{\partial T}{\partial t}\right) = \frac{\hat{H}\chi(x)}{\chi}$$

Since the left and right hand sides depend only on x and t respectively but are equal, they must be equal to a separation constant $E \in \mathbb{R}$. Solving for time,

$$\frac{1}{T}i\hbar\frac{\partial T}{\partial t} = E \implies T(t) = e^{\frac{-iEt}{\hbar}}$$

If E were complex, T would diverge. Solving for space, we have the time-independent Schrödinger equation as follows.

$$\hat{H}\chi(x) = E\chi(x)$$

Explicitly,

$$-\frac{\hbar^2}{2m}\nabla^2\chi(x) + U(x)\chi(x) = E\chi(x)$$

This is an eigenvalue equation for $\hat{H}$; we wish to find the eigenvalues for $\hat{H}$ in the x basis. Note that the factorised solution $\psi = T\chi$ is just a particular class of solutions for the time-dependent Schrödinger equation. However, it can be shown that any solution to the time-dependent equation can be written as a linear combination of the time-independent equation solutions.

4. One-dimensional solutions to the Schrödinger equation

4.1. Stationary states

Definition. With the ansatz $\psi(x,t) = \chi(x)T(t)$, we have found a particular class of solutions of the time-independent Schrödinger equation:

$$\psi(x,t) = \chi(x)e^{-\frac{iEt}{\hbar}}$$

where $\chi(x)$ are the eigenfunctions of $\hat{H}$ with eigenvalue E. Such solutions are called stationary states.

Note,

$$\rho(x,t) = |\psi(x,t)|^2 = |\chi(x)|^2$$

This explains the naming of the states as 'stationary', as their probability density is independent of time. Now, suppose E is quantised. Then, the general solution to the system is

$$\psi(x,t) = \sum_{n=1}^{N} a_n \chi_n(x) e^{-\frac{iE_n t}{\hbar}}$$

where N can be finite or infinite. In principle, we can also have a continuous energy state $E_\alpha, \alpha \in \mathbb{R}$. We can still use the same idea:

$$\psi(x,t) = \int_{\Delta\alpha} A(\alpha)\chi_\alpha(x)e^{-\frac{iE_\alpha t}{\hbar}}\,\mathrm{d}\alpha$$

Note that $|a_n|^2$ and $A(\alpha)\,\mathrm{d}\alpha$ give the probability of measuring the particle energy to be E_n or E_α.

4.2. Infinite potential well

We define

$$U(x) = \begin{cases} 0 & \text{for } |x| \le a \\ \infty & \text{for } |x| > a \end{cases}$$

For $|x| > a$, we must have $\chi(a) = 0$. Otherwise, $\chi \cdot U = \infty$. This gives us a boundary condition, $\chi(\pm a) = 0$. For $|x| \le a$, we seek solutions of the form

$$-\frac{\hbar^2}{2m}\chi''(x) = E\chi(x); \quad \chi(\pm a) = 0$$

Equivalently,

$$\chi''(x) + k^2\chi(x) = 0; \quad k = \sqrt{\frac{2mE}{\hbar^2}}$$

Since $E > 0$,

$$\chi(x) = A\sin kx + B\cos kx$$

Imposing boundary conditions,

$$A \sin ka + B \cos ka = 0; \quad A \sin ka - B \cos ka = 0$$

Suppose $A = 0$, giving $\chi(x) = B \cos kx$. Then, imposing boundary conditions, $\chi_n(x) = B \cos k_n x$ where $k_n = \frac{n\pi}{2a}$, and n are odd positive integers. These are even solutions.

Alternatively, suppose $B = 0$. In this case, $\chi(x) = A \sin kx$. Thus, $\chi_n(x) = A \sin k_n x$ where $k_n = \frac{n\pi}{2a}$, and n are even nonzero positive integers. These provide odd solutions.

We can also determine the normalisation constants by defining that the eigenfunctions of the Hamiltonian are normalised to unity. Thus,

$$\int_{-a}^{a} |\chi_n(x)|^2 = 1 \implies A = B = \sqrt{\frac{1}{a}}$$

Hence, the general solution is given by the eigenvalues

$$E_n = \frac{\hbar^2}{2n} k_n^2 = \frac{\hbar^2 \pi^2 n^2}{2ma^2}$$

and eigenfunctions

$$\chi_n(x) = \sqrt{\frac{1}{a}} \begin{cases} \cos\left(\frac{n\pi x}{2a}\right) & \text{if } n \text{ odd} \\ \sin\left(\frac{n\pi x}{2a}\right) & \text{if } n \text{ even} \end{cases}$$

Remark. Note that unlike classical mechanics, the ground state energy is not zero. Note also that χ_n have $(n+1)$ nodes in which $\rho(x) = 0$. When $n \to \infty$, $\rho_n(x)$ tends to a constant, which is like in classical mechanics. Eigenfunctions of the Hamiltonian in this case were either odd or even; we can in fact prove that this is the case in general.

Proposition. If we have a system of non-degenerate eigenstates ($E_i \neq E_j$), then if $U(x) = U(-x)$ the eigenfunctions of $\hat{H}$ must be either odd or even.

Proof. The time-independent Schrödinger equation is invariant under $x \mapsto -x$ if U is even. Hence, if $\chi(x)$ is a solution with eigenvalue E, then $\chi(-x)$ is also a solution. Since we have a non-degenerate solution, $\chi(-x) = \chi(x)$ hence the solutions must be the same up to a normalisation factor. For consistency, $\chi(x) = \chi(-(-x)) = \alpha\chi(-x) = \alpha^2\chi(x)$. Hence $\alpha = \pm 1$, so χ is either odd or even. □

4.3. Finite potential well

We define

$$U(x) = \begin{cases} 0 & \text{for } |x| \leq a \\ U_0 & \text{for } |x| > a \end{cases}$$

Classically, if $E < U_0$, the particle has insufficient energy to escape the well. We will only consider eigenstates with $E < U_0$ here, but we will find that it is possible in quantum mechanics to escape the well with positive probability. We will search for even functions only, odd functions can be solved independently. If $|x| \leq a$,

$$-\frac{\hbar^2}{2m}\chi''(x) = E\chi(x)$$

Equivalently,

$$\chi''(x) + k^2\chi(x) = 0; \quad k = \sqrt{\frac{2mE}{\hbar^2}}$$

The solution becomes

$$\chi(x) = A\sin kx + B\cos kx \implies \chi(x) = B\cos kx$$

since we are only looking for even solutions. In the region $|x| > a$,

$$-\frac{\hbar^2}{2m}\chi''(x) + U_0\chi(x) = E\chi(x)$$

giving

$$\chi''(x) - \overline{k}^2\chi(x) = 0; \quad \overline{k} = \sqrt{\frac{2m(U_0 - E)}{\hbar^2}}$$

This yields exponential solutions:

$$\chi(x) = Ce^{\overline{k}x} + De^{-\overline{k}x}$$

Imposing the normalisability constraints, for $x > a$ we have $C = 0$, and for $x < -a$ we have $D = 0$. Imposing even parity, $C = D$ when nonzero. Thus,

$$\chi(x) = \begin{cases} Ce^{\overline{k}x} & x < -a \\ B\cos(kx) & |x| \leq a \\ Ce^{-\overline{k}x} & x > a \end{cases}$$

Now we must impose continuity of $\chi(x)$ and its derivative at $x = \pm a$. First,

$$Ce^{-\overline{k}a} = B\cos(ka)$$

The other gives

$$-\overline{k}Ce^{-\overline{k}a} = -kB\sin(ka)$$

From the ratio of both constraints,

$$k\tan(ka) = \overline{k}$$

From the definition of $k, \overline{k}$,

$$k^2 + \overline{k}^2 = \frac{2mU_0}{\hbar^2}$$

We will define some rescaled variables for convenience: $\xi = ka$, $\eta = \bar{k}a$. Rewriting,

$$\xi \tan \xi = \eta; \quad \xi^2 + \eta^2 = r_0^2; \quad r_0 = \frac{2mU}{\hbar}$$

This may be solved graphically. The eigenvalues of the system correspond to the points of intersection between the two equations. There are always a finite number of possible intersections, regardless of the value of r_0. The eigenvalues are

$$E_n = \frac{\hbar^2}{2na^2}\xi_n^2; \quad \xi \in \{\xi_1, \dots, \xi_n\}; \quad n = 1, \dots, p$$

When $U_0 \to \infty$, $r_0 \to \infty$. At this point, there are an infinite amount of intersections, so the eigenvalues of the Hamiltonian become that of the infinite well. Further $\chi(x)$ tends to the eigenfunctions of the infinite well. Note that the $\chi_n(x)$ have some positive region outside the well. We can use the unused condition above to write C in terms of B, and then we can use the normalisation condition to find B.

4.4. Free particles

A free particle is under no potential. The time-independent Schrödinger equation is

$$-\frac{\hbar}{2m}\chi''(x) = E\chi(x)$$

This has solutions

$$\chi_k(x) = Ae^{ikx}; \quad k = \sqrt{\frac{2mE}{\hbar^2}}$$

The complete solution, adding $T(t)$, is thus

$$\psi_k(x,t) = \chi_k(x)e^{-iE_k t/\hbar} = Ae^{i\left(kx - \frac{\hbar k^2}{2m}t\right)}$$

which are called De Broglie plane waves. This is not a solution since

$$\int_{-\infty}^{\infty} |\phi_k(x,t)|\,\mathrm{d}x = |A|^2 \int_{-\infty}^{\infty} 1\,\mathrm{d}x$$

which diverges. In general, any non-bound solution is non-normalisable. This is true since $\int_{-\infty}^{\infty} |\chi(x)|^2\,\mathrm{d}x < \infty$ requires $\lim_{R\to\infty} \int_{|x|>R} |\chi(x)|\,\mathrm{d}x = 0$. So, to solve the free particle system, we will build a linear combination of plane waves χ to yield a normalisable solution. This is called the Gaussian wavepacket. Alternatively, we can simply ignore the problem of normalisability, and change the interpretation of $\chi_n(x)$.

4.5. Gaussian wavepacket

Due to the superposition principle, we can take a continuous linear combination of the ψ_k functions.

$$\psi(x,t) = \int_0^\infty A(k)\psi_k(x,t)\,\mathrm{d}k$$

We can construct a suitable $A(k)$ such that ψ is normalisable. Choosing

$$A(k) = A_{\mathrm{GP}}(k) = \exp\Big[-\frac{\sigma}{2}(k-k_0)^2\Big]; \quad k_0 \in \mathbb{R}, \sigma \in \mathbb{R}^+$$

produces a solution called the Gaussian wavepacket. Substituting into the above,

$$\psi_{\mathrm{GP}}(x,t) = \int_0^\infty \exp\Big[-\frac{\sigma}{2}(k-k_0)^2\Big]\psi_k(x,t)\,\mathrm{d}k = \int_0^\infty \exp[F(k)]\,\mathrm{d}k$$
$$F(k) = -\frac{\sigma}{2}(k-k_0)^2 + ikx - i\frac{\hbar k^2}{2m}t$$

We can rewrite this as

$$F(k) = -\frac{1}{2}\Big(\sigma + \frac{i\hbar t}{m}\Big)k^2 + (k_0\sigma + ix)k - \frac{\sigma}{2}k_0^2$$

We define further

$$\alpha \equiv \sigma + \frac{i\hbar t}{m}; \quad \beta = k_0\sigma + ix; \quad \delta = -\frac{\sigma}{2}k_0^2$$

Completing the square,

$$F(k) = -\frac{\alpha}{2}\Big(k - \frac{\beta}{\alpha}\Big)^2 + \frac{\beta^2}{2\alpha} + \delta$$

We arrive at the solution

$$\psi_{\mathrm{GP}}(x,t) = \exp\Big[\frac{\beta^2}{2\alpha} + \delta\Big]\int_{-\infty}^\infty \exp\Big[-\frac{\alpha}{2}\Big(k - \frac{\beta}{\alpha}\Big)^2\Big]\,\mathrm{d}k$$

Under a change of variables $\tilde{k} = k - \frac{\beta}{\alpha}, u = \mathrm{Im}(\frac{\beta}{\alpha})$,

$$\psi_{\mathrm{GP}}(x,t) = \exp\Big[\frac{\beta^2}{2\alpha} + \delta\Big]\int_{\infty - iu}^{\infty - iu} \exp\Big[-\frac{\alpha}{2}\tilde{k}\Big]\,\mathrm{d}\tilde{k}$$

We arrive at the usual Gaussian integral:

$$I(a) = \int_{-\infty}^\infty \exp\big[-ax^2\big]\,\mathrm{d}x = \sqrt{\frac{\pi}{2}}$$

giving

$$\psi_{\mathrm{GP}}(x,t) = \sqrt{\frac{2\pi}{\alpha}}\exp\Big[\frac{\beta^2}{2\alpha} + \delta\Big] = \sqrt{\frac{2\pi}{\alpha}}\exp\left[-\frac{\sigma}{2}\frac{\Big(x - \frac{\hbar k_0}{m}t\Big)^2}{\Big(\sigma^2 + \frac{\hbar^2 t^2}{m^2}\Big)}\right]$$

We define $\overline{\psi}_{\text{GP}}$ to be the normalised Gaussian wavefunction, so $\overline{\psi}_{\text{GP}} = C\psi_{\text{GP}}$. We can find that

$$\rho_{\text{GP}}(x,t) = \left|\overline{\psi}_{\text{GP}}(x,t)\right|^2 = \sqrt{\frac{\sigma}{\pi\left(\sigma^2 + \frac{\hbar^2 t^2}{m^2}\right)}} \exp\left[-\frac{\sigma\left(x - \frac{\hbar k}{m}t\right)^2}{\sigma^2 + \frac{\hbar^2 t^2}{m^2}}\right]$$

This is a wavefunction whose probability density distribution resembles a Gaussian e^{-x^2} term, with a maximum point at

$$\langle x\rangle = \int_{-\infty}^{\infty} \psi_{\text{GP}}^{\star} x \psi_{\text{GP}}\,\mathrm{d}x = \int_{-\infty}^{\infty} x\rho_{\text{GP}}\,\mathrm{d}x = \frac{\hbar k_0}{m}t$$

and a width of

$$\Delta x = \sqrt{\langle x^2\rangle - \langle x\rangle^2} = \sqrt{\frac{1}{2}\left(\sigma + \frac{\hbar^2 t^2}{m^2\sigma}\right)}$$

The physical interpretation is that the uncertainty of the particle's position grows with time. In this case, we can find

$$\langle p\rangle = \int_{-\infty}^{\infty} \psi_{\text{GP}}^{\star} i\hbar \frac{\partial}{\partial x}\psi_{\text{GP}}\,\mathrm{d}x = \hbar k_0$$

which is constant. The uncertainty in the momentum can be found to be

$$\Delta p = \sqrt{\langle p^2\rangle - \langle p\rangle^2} = \frac{\hbar}{\sqrt{\frac{1}{2}\left(\sigma + \frac{\hbar^2 t^2}{m\sigma}\right)}}$$

Thus,

$$\Delta x \Delta p = \frac{\hbar}{2}$$

We can find for a single plane wave that

$$\Delta x = \infty; \quad \Delta p = 0$$

4.6. Beam interpretation

We can choose to ignore the normalisation problem and take the plane waves as the eigenfunctions of the Hamiltonian:

$$\chi_k(x) = Ae^{ikx}; \quad \psi_k(x,t) = Ae^{ikx}e^{-\frac{i\hbar^2 k^2}{2m}t}$$

Instead of $\chi_k(x)$ describing a single particle, we can interpret it as a beam of particles with momentum $p = \hbar k$ and $E = \frac{\hbar^2 k^2}{2m}$ with probability density

$$\rho_k(x) = \left|\chi_k(x)e^{-\frac{i\hbar^2 k^2}{2m}t}\right|^2 = |A|^2$$

which here is interpreted as a constant average density of particles. The probability current is given by

$$J_k(x,t) = -\frac{i\hbar}{2m}\left(\psi_k^\star \frac{\partial \psi_k}{\partial x} - \psi_k \frac{\partial \psi_k^\star}{\partial x}\right) = -\frac{i\hbar}{2m}|A|^2 2ik = |A|^2 \frac{\hbar k}{m} = |A|^2 \underbrace{\frac{p}{m}}_{\text{velocity}}$$

This is interpreted as the average flux of particles.

4.7. Scattering states

We wish to investigate what happens when a particle, or beam of particles, is thrown onto a potential $U(x)$. In this case, suppose we have a step function

$$U(x) = \begin{cases} U_0 & \text{if } 0 \leq x < a \\ 0 & \text{otherwise} \end{cases}$$

and a Gaussian wavepacket which is centred at $x_0 \ll 0$ moving in the $+x$ direction, towards the spike in potential. As $t \gg 0$, we end up with a probability density given by two wavepackets; one will be moving left from the spike and one will have cleared the spike and continues moving to the right.

Definition. The reflection coefficient R is

$$R = \lim_{t\to\infty} \int_{-\infty}^{0} |\psi_{\text{GP}}(x,t)|^2 \, \mathrm{d}x$$

which is the probability for the particle to be reflected. The transmission coefficient is

$$T = \lim_{t\to\infty} \int_{0}^{\infty} |\psi_{\text{GP}}(x,t)|^2 \, \mathrm{d}x$$

By definition, $R + T = 1$.

In practice, working with Gaussian packets is mathematically challenging, although not impossible. The beam interpretation, by allowing us to use non-normalisable stationary state wavefunctions, greatly simplifies the computation.

4.8. Scattering off potential step

Consider a potential

$$U(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ U_0 & \text{if } x > 0 \end{cases}$$

We want to solve

$$-\frac{\hbar^2}{2m}\chi_k''(x) + U(x)\chi_k(x) = E\chi_k(x)$$

We split the problem into two regions: $x \leq 0, x > 0$. For $x \leq 0$, the TISE becomes

$$\chi_k''(x) + k^2\chi_k(x) = 0; \quad k = \sqrt{\frac{2mE}{\hbar^2}}$$

The solution is

$$\chi(x) = Ae^{ikx} + Be^{-ikx}$$

This is a superposition of two beams; the beam of incident particles Ae^{ikx} and the beam of reflected particles Be^{-ikx} which are travelling in the opposite direction. In the region $x > 0$, we have

$$\chi_{\bar{k}}''(x) + \bar{k}^2\chi_{\bar{k}}(x) = 0; \quad \bar{k} = \sqrt{\frac{2m(E-U_0)}{\hbar^2}}$$

where $\bar{k}$ is real if $E > U_0$, and $\bar{k}$ is pure-imaginary if $E < U_0$. Therefore, for $E > U_0$ we have

$$\chi_{\bar{k}}(x) = Ce^{i\bar{k}x} + De^{-i\bar{k}x}$$

which is a beam of particles moving towards the right and an incident beam of particles from the right moving towards the left. Since no such incident beam exists, we can set $D = 0$. If $E < U_0$, the solution is

$$\bar{k} \equiv i\eta \implies \chi_{\bar{k}}(x) = Ce^{-\eta x} + De^{\eta x}$$

$D \neq 0$ would give infinite values of $\chi_{\bar{k}}(x)$ as $x \to \infty$. In either case, the eigenfunctions are

$$\chi_{k,\bar{k}}(x) = \begin{cases} Ae^{ikx} + Be^{-ikx} & x \leq 0 \\ Ce^{i\bar{k}x} & x > 0 \end{cases}$$

By imposing the boundary conditions, specifically the continuity of χ, we can determine the constants.

$$A + B = C; \quad ikA - ikB = i\bar{k}C$$

which gives

$$B = \frac{k-\bar{k}}{k+\bar{k}}A; \quad C = \frac{2k}{k+\bar{k}}A$$

We can view these solutions in terms of particle flux.

$$J_k(x,t) = -\frac{i\hbar}{2m}\left(\psi_k^\star\frac{\partial\psi_k}{\partial x} - \psi_k\frac{\partial\psi_k^\star}{\partial x}\right)$$

If $E > U_0$, we find

$$J(x,t) = \begin{cases} \frac{\hbar k}{m}(|A|^2 - |B|^2) & x < 0 \\ \frac{\hbar\bar{k}}{m}|C|^2 & x \geq 0 \end{cases}$$

The incident flux is $\frac{\hbar k}{m}|A|^2$, the reflected flux is $\frac{\hbar k}{m}|B|^2$, and the transmitted flux is $\frac{\hbar k}{m}|C|^2$. We can define

$$R = \frac{J_{\text{ref}}}{J_{\text{inc}}} = \frac{|B|^2}{|A|^2} = \left(\frac{k-\bar{k}}{k+\bar{k}}\right)^2$$

We can also define

$$T = \frac{J_{\text{trans}}}{J_{\text{inc}}} = \frac{k|C|^2}{\overline{k}|A|^2} = \frac{4k\overline{k}}{(k+\overline{k})^2}$$

We can check that our original interpretation makes sense; for example, $R + T = 1$, and $E \to U_0, \overline{k} \to 0$ implies $T \to 0, R \to 1$. If $E \to \infty$, $T \to 1$ and $R \to 0$. If $E < U_0$,

$$J(x,t) = \begin{cases} \frac{\hbar k}{m}(|A|^2 + |B|^2) & x < 0 \\ 0 & x \geq 0 \end{cases}$$

since $\chi_{\overline{k}} = \chi^\star_{\overline{k}}$. Here, $T = 0$ but $\chi_{\overline{k}}(x) \neq 0$.

4.9. Scattering off a potential barrier

Consider the potential

$$U(x) = \begin{cases} 0 & x \leq 0, x \geq a \\ U_0 & 0 < x < a \end{cases}$$

When $E < U_0$, we define

$$k = \sqrt{\frac{2mE}{\hbar^2}} > 0; \quad \eta = \sqrt{\frac{2m(U_0 - E)}{\hbar^2}} > 0$$

The solution is then

$$\chi(x) = \begin{cases} e^{ikx} + Ae^{-ikx} & x \leq 0 \\ Be^{-\eta x} + Ce^{\eta x} & 0 < x < a \\ De^{ikx} & x \geq a \end{cases}$$

since we can normalise the incoming flux to one. The boundary conditions are that $\chi(x) = \chi'(x)$ are both continuous at $x = 0, x = a$. This gives four conditions, which are enough to solve the problem. $\chi(x)$ and its derivative at zero give

$$1 + A = B + C; \quad ik - ikA = -\eta B + \eta C$$

and the continuity at a gives

$$Be^{-\eta a} + Ce^{\eta a} = De^{ika}; \quad -\eta Be^{-\eta a} + \eta Ce^{\eta a} = ikDe^{ika}$$

Solving the system gives

$$D = \frac{-4i\eta k}{(\eta - ik)^2 \exp[(\eta + ik)a] - (\eta + ik)^2 \exp[-(\eta - ik)a]}$$

The transmitted flux is $j_{\text{tr}} = \frac{\hbar k}{m}|D|^2$ and the incident flux is $j_{inc} = \frac{\hbar k}{m}$. Hence, the transmission coefficient is $T = |D|^2$. This is

$$T = \frac{4k^2\eta^2}{(k^2 + \eta^2)^2 \sinh^2(\eta a) + 4k^2\eta^2}$$

If we take the limit as $U_0 \gg E$, we have $\eta a \gg 1$. Then

$$T \to \frac{16k^2\eta^2}{(\eta^2+k^2)^2}\exp[-2\eta a] \propto \exp\left[-\frac{2a}{k}\sqrt{2m(U_0-E)}\right]$$

So the probability decreases exponentially with the width of the barrier.

4.10. Harmonic oscillator

Consider a parabolic potential

$$U(x) = \frac{1}{2}kx^2 = \frac{1}{2}m\omega^2x^2$$

where k is an elastic constant and $\omega = \sqrt{\frac{k}{m}}$ is the angular frequency of the harmonic oscillator. Classically, we find the solution $x = A\cos\omega t + B\sin\omega t$. This gives a continuous energy spectrum. The TDSE gives

$$-\frac{\hbar^2}{2m}\chi''(x) + \frac{1}{2}m\omega^2x^2\chi(x) = E\chi(X)$$

Since this is a bound system, we will have a discrete set of eigenvalues. The potential is symmetric so the eigenfunctions are odd or even. We will make the change of variables

$$\xi^2 = \frac{m\omega}{\hbar}x^2; \quad \varepsilon = \frac{2E}{\hbar\omega}$$

which reformulates the TDSE as

$$-\frac{\mathrm{d}^2\chi}{\mathrm{d}\xi^2} + \xi^2\chi = \varepsilon\chi$$

We will start by considering the solution for $\varepsilon = 1$. In this case, $E = \frac{\hbar\omega}{2}$. The solution in this case is

$$\chi_0(\xi) = \exp\left[-\frac{\xi^2}{2}\right]$$

So the first eigenfunction, χ_0, is known in terms of x, given by

$$\chi_0(x) = A\exp\left[-\frac{m\omega}{2\hbar}x^2\right]; \quad E_0 = \frac{\hbar\omega}{2}$$

To find the other eigenfunctions, we will take the general form

$$\chi(\xi) = f(\xi)\exp\left[-\frac{\xi^2}{2}\right]$$

This works because we know we have a bound solution and χ must tend to zero quickly as ξ tends to infinity, due to the differential equation in terms of ξ, ε. Using the above ansatz for χ in the Schrödinger equation,

$$-\frac{\mathrm{d}^2f}{\mathrm{d}\xi^2} + 2\xi\frac{\mathrm{d}f}{\mathrm{d}\xi} + (1-\varepsilon)f = 0$$

Note that if $\varepsilon = 1$, a solution is $f = 1$. We can find a power series solution to this differential equation, with $\xi = 0$ as a regular point.

$$f(\xi) = \sum_{n=0}^{\infty} a_n \xi^n$$

We find

$$\xi \frac{\mathrm{d}f}{\mathrm{d}\xi} = \sum_{n=0}^{\infty} n a_n \xi^n; \quad \frac{\mathrm{d}^2 f}{\mathrm{d}\xi^2} = \sum_{n=0}^{\infty} n(n-1) a_n \xi^{n-2} = \sum_{n=0}^{\infty} (n+1)(n+2) a_{n+2} \xi^n$$

Comparing coefficients of ξ^n,

$$(n+1)(n+2)a_{n+2} - 2n a_n + (\varepsilon - 1)a_n = 0$$

Hence,

$$a_{n+2} = \frac{2n - \varepsilon + 1}{(n+1)(n+2)} a_n$$

Since the function must be either even or odd, exactly one of a_0 and a_1 must be zero.

Proposition. If the series for f does not terminate, χ is not normalisable.

Proof. Suppose the series does not terminate. We will consider the asymptotic behaviour as $n \to \infty$.

$$\frac{a_{n+2}}{a_n} \to \frac{2}{n}$$

But this is the same asymptotic behaviour as the function $g(\xi)$ given by

$$g(\xi) = \exp\left[\xi^2\right] = \sum_{m=0}^{\infty} \frac{\xi^{2m}}{m!} = \sum_{n=0}^{\infty} b_n \xi^n$$

with

$$b_n = \begin{cases} \frac{1}{m!} & n = 2m \\ 0 & n = 2m+1 \end{cases}$$

So asymptotically,

$$\frac{b_{n+2}}{b_n} = \frac{\left(\frac{n}{2}\right)!}{\left(\frac{n}{2}+1\right)!} = \frac{2}{n+2} \to \frac{2}{n}$$

Hence χ would have a form asymptotically equal to

$$\chi(\xi) \sim \exp\left[\frac{\xi^2}{2}\right]$$

Hence $\chi(\xi)$ would be not normalisable. □

Hence f must be a polynomial. So there exists N such that $a_{N+2} = 0$ and $a_N \neq 0$. So for this value,

$$2N - \varepsilon + 1 = 0 \implies \varepsilon = 2N + 1$$

By the definition of ε,

$$E_N = \left(N + \frac{1}{2}\right)\hbar\omega$$

In particular, $E_{N+1} - E_N = \hbar\omega$. The eigenfunctions are

$$\chi_N(\xi) = f_N(\xi) \exp\left[-\frac{\xi^2}{2}\right]$$

with the property that

$$\chi_N(-\xi) = (-1)^N \chi_N(\xi)$$

$$\begin{aligned} f_0(\xi) &= 1 \\ f_1(\xi) &= \xi \\ f_2(\xi) &= 1 - 2\xi^2 \\ f_3(\xi) &= \xi - \frac{2}{3}\xi^3 \\ &\vdots \end{aligned}$$

5. Operators and measurements

5.1. Hermitian operators

Definition. The *Hermitian conjugate* of an operator $\hat{A}$ is written $\hat{A}^\dagger$, and is defined such that

$$\langle \hat{A}^\dagger \psi_1, \psi_2 \rangle = \langle \psi_1, \hat{A}\psi_2 \rangle$$

where $\psi_1, \psi_2 \in \mathcal{H}$.

We can verify that for $a_1, a_2 \in \mathbb{C}$,

(i) $(a_1\hat{A}_1 + a_2\hat{A}_2)^\dagger = a_1^\star \hat{A}_1^\dagger + a_2^\star \hat{A}_2^\dagger$;

(ii) $(\hat{A}\hat{B})^\dagger = \hat{B}^\dagger \hat{A}^\dagger$

Definition. A *Hermitian operator* is a linear operator $\hat{O} : \mathcal{H} \to \mathcal{H}$ such that

$$\hat{A}^\dagger = \hat{A}$$

Equivalently,

$$\langle \hat{A}\psi_1, \psi_2 \rangle = \langle \psi_1, \hat{A}\psi_2 \rangle$$

Example. The familiar operators $\hat{x}$, $\hat{p}$ are Hermitian.

$$\begin{aligned}
\langle \hat{x}\psi_1, \psi_2 \rangle &= \int_{\mathbb{R}^3} (x\psi_1)^\star \psi_2 \, \mathrm{d}V \\
&= \int_{\mathbb{R}^3} \psi_1^\star x \psi_2 \, \mathrm{d}V \\
&= \langle \psi_1, \hat{x}\psi_2 \rangle
\end{aligned}$$

For $\hat{p}$, integrating by parts, we have

$$\begin{aligned}
\langle \hat{p}\psi_1, \psi_2 \rangle &= \int_{-\infty}^{\infty} \left(-i\hbar \frac{\partial}{\partial x} \psi_1 \right)^\star \psi_2 \, \mathrm{d}x \\
&= i\hbar \int_{-\infty}^{\infty} \frac{\partial \psi_1^\star}{\partial x} \psi_2 \, \mathrm{d}x \\
&= -i\hbar \int_{-\infty}^{\infty} \psi_1^\star \frac{\partial \psi_2}{\partial x} \, \mathrm{d}x \\
&= \langle \psi_1, \hat{p}\psi_2 \rangle
\end{aligned}$$

Theorem. The eigenvalues of a Hermitian operator are real.

Proof. Let $\hat{A}$ be a Hermitian operator, and ψ a normalised eigenfunction with eigenvalue a.

$$\langle \psi, \hat{A}\psi \rangle = \langle \psi, a\psi \rangle = a \langle \psi, \psi \rangle = a$$

Since $\hat{A}$ is Hermitian,

$$\langle \psi, \hat{A}\psi \rangle = \langle \hat{A}\psi, \psi \rangle = \langle a\psi, \psi \rangle = a^\star \langle \psi, \psi \rangle = a^\star$$

Hence $a = a^\star$ so $a \in \mathbb{R}$. □

Theorem. Let $\hat{A}$ be a Hermitian operator, and ψ_1, ψ_2 normalised eigenfunctions with distinct eigenvalues a_1, a_2. Then ψ_1, ψ_2 are orthogonal.

Proof. We have $\hat{A}\psi_1 = a_1\psi_1$ and $\hat{A}\psi_2 = a_2\psi_2$. Then,

$$\langle \hat{A}\psi_1, \psi_2 \rangle = a_1 \langle \psi_1, \psi_2 \rangle$$

But also,

$$\langle \psi_1, \hat{A}\psi_2 \rangle = a_2 \langle \psi_1, \psi_2 \rangle$$

These two values must be the same, so $\langle \psi_1, \psi_2 \rangle = 0$. □

Theorem. The discrete and continuous set of eigenfunctions of any Hermitian operator form a complete orthogonal basis for the Hilbert space. This theorem is stated without proof.

Corollary. Every solution of the time-dependent Schrödinger can be written as a superposition of stationary states.

$$\psi(x,t) = \sum_{n=1}^{\infty} a_n \chi_n(x) e^{-iE_n t/\hbar}; \quad a_n = \langle \chi_n, \psi \rangle$$

In the continuous case,

$$\psi(x,t) = \int_{\Delta_\alpha} A(\alpha)\chi_\alpha(x) e^{-iE_n t/\hbar} \, \mathrm{d}\alpha; \quad A(\alpha) = \langle \chi_\alpha, \psi \rangle$$

5.2. Postulates of quantum mechanics

The following postulates are used to interpret measurements in quantum systems.

(i) Any observable O is represented by a Hermitian operator $\hat{O}$.

(ii) The possible outcomes of O are the eigenvalues of $\hat{O}$. Since $\hat{O}$ is Hermitian, we can only ever observe real values.

(iii) Let $\hat{O}$ have a discrete set of normalised eigenfunctions $\{\psi_i\}$ with distinct eigenvalues $\{\lambda_i\}$. Let ψ be a state, written in terms of the eigenfunctions of $\hat{O}$.

$$\psi = \sum a_i \psi_i$$

Suppose we measure O on a particle in the state ψ. Then, the probability that O takes value λ_i is

$$\mathbb{P}(O = \lambda_i) = |a_i|^2 = a_i^\star a_i$$

(iv) The above postulate can be generalised to the case where $\hat{O}$ has degenerate eigenvalues. Let $\{\psi_i\}$ be a discrete set of normalised eigenfunctions with not necessarily distinct eigenvalues $\{\lambda_i\}$. If $\{\psi_i\}_{i\in I}$ is a complete set of orthonormal eigenfunctions with the same eigenvalue λ, then

$$\mathbb{P}(O=\lambda)=\sum_{i\in I}|a_i|^2=\sum_{i\in I}a^\star a$$

(v) We can verify from the postulates above that the sum of all probabilities is unity.

$$\sum_i|a_i|^2=\sum_i\langle a_i\psi_i,a_i\psi_i\rangle=\sum_i\sum_j\langle a_i\psi_i,a_j\psi_j\rangle=\langle\psi,\psi\rangle=1$$

(vi) If O is measured on a state ψ at time t, and the outcome is λ_i, then the wavefunction instantaneously 'collapses' into the measured state after the measurement.

$$\psi\mapsto\psi_i$$

This is called the *projection postulate*.

(vii) If $\hat{O}$ has degenerate eigenfunctions all with eigenvalue λ, then instead we find

$$\psi\mapsto\sum_{i\in I}a_i\psi_i$$

So in this case, the wavefunction collapses to a linear combination of the eigenfunctions that give this eigenvalue.

5.3. Expectation of operators

Definition.

$$\psi=\sum_i a_i\psi_i=\sum_i\langle\psi_i,\psi\rangle\psi_i$$

The *projector* operator projects ψ onto a specific eigenfunction.

$$\hat{P}:\ \psi\mapsto\langle\psi_i,\psi\rangle\psi_i$$

Definition. The expectation value of an observable $\hat{O}$ on a state ψ is

$$\begin{aligned}\langle O\rangle_\psi&=\sum_i\lambda_i\mathbb{P}(O=\lambda_i)\\&=\sum_i\lambda_i|\langle\psi_i,\psi\rangle|^2\\&=\left\langle\sum_i\langle\psi_i,\psi\rangle\psi_i,\sum_j\lambda_j\langle\psi_j,\psi\rangle\psi_j\right\rangle\\&=\langle\psi,\hat{O}\psi\rangle\end{aligned}$$

5.4. Commutators

Definition. The *commutator* of two operators $\hat{A}$ and $\hat{B}$ is the operator given by

$$\left[\hat{A}, \hat{B}\right] = \hat{A}\hat{B} - \hat{B}\hat{A}$$

We observe the following properties of the commutator.

(i) $\left[\hat{A}, \hat{B}\right] = -\left[\hat{B}, \hat{A}\right]$;

(ii) $\left[\hat{A}, \hat{A}\right] = 0$;

(iii) $\left[\hat{A}, \hat{B}\hat{C}\right] = \left[\hat{A}, \hat{B}\right]\hat{C} + \hat{B}\left[\hat{A}, \hat{C}\right]$;

(iv) $\left[\hat{A}\hat{B}, \hat{C}\right] = \hat{A}\left[\hat{B}, \hat{C}\right] + \left[\hat{A}, \hat{C}\right]\hat{B}$;

Example. The commutator $[\hat{x}, \hat{p}]$ in one dimension is given by, for every $\psi \in \mathcal{H}$,

$$\begin{aligned} \hat{x}\hat{p}\psi &= x\left(-i\hbar\frac{\partial}{\partial x}\right)\psi(x) = -i\hbar x\frac{\partial \psi}{\partial x} \\ \hat{p}\hat{x}\psi &= \left(-i\hbar\frac{\partial}{\partial x}\right)x\psi(x) = -i\hbar\psi - i\hbar x\frac{\partial \psi}{\partial x} \\ \therefore [\hat{x}, \hat{p}]\psi &= i\hbar\psi \end{aligned}$$

Hence,

$$[\hat{x}, \hat{p}] = i\hbar\hat{I}$$

where $\hat{I}$ is the identity operator. This specific commutator is known as the canonical commutator relation.

5.5. Simultaneously diagonalisable operators

Definition. Hermitian operators $\hat{A}$ and $\hat{B}$ are said to be *simultaneously diagonalisable* if there exists a complete basis of joint eigenfunctions $\{\psi_i\}$ such that $\hat{A}\psi_i = \lambda_i\psi_i$ and $\hat{B}\psi_i = \mu_i\psi_i$ for $\lambda_i, \mu_i \in \mathbb{R}$.

Theorem. Hermitian operators $\hat{A}$ and $\hat{B}$ are simultaneously diagonalisable if and only if $\left[\hat{A}, \hat{B}\right] = 0$.

Proof. Suppose $\hat{A}$ and $\hat{B}$ are simultaneously diagonalisable. Then, by definition, there exists a complete basis $\{\psi_i\}$ with eigenvalues λ_i, μ_i for $\hat{A}, \hat{B}$. Now, for any element ψ_i of this basis, the commutator is

$$\left[\hat{A}, \hat{B}\right]\psi_i = \hat{A}\hat{B}\psi_i - \hat{B}\hat{A}\psi_i = \hat{A}\mu_i\psi_i - \hat{B}\lambda_i\psi_i = \mu_i\hat{A}\psi_i - \lambda_i\hat{B}\psi_i = \lambda_i\mu_i\psi_i - \mu_i\lambda_i\psi_i = 0$$

Let ψ be an arbitrary function in the Hilbert space $\mathcal{H}$. Then by linearity,

$$\left[\hat{A}, \hat{B}\right]\psi = \sum_i c_i\left[\hat{A}, \hat{B}\right]\psi_i = 0$$

Conversely, suppose that the commutator is zero. Let ψ_i be an eigenfunction of $\hat{A}$ with eigenvalue λ_i. Then, since the commutator is zero, we have

$$0 = [\hat{A}, \hat{B}]\psi_i = \hat{A}\hat{B}\psi_i - \hat{B}\hat{A}\psi_i \implies \hat{A}(\hat{B}\psi_i) = \lambda_i(\hat{B}\psi_i)$$

Hence, $\hat{B}$ maps the eigenspace E_i of $\hat{A}$ with eigenvalue λ_i into itself. So $\hat{B}|_{E_i}$ is a Hermitian operator on E_i. Since this holds for any eigenfunction and eigenvalue, we can find a complete basis of simultaneous eigenfunctions of $\hat{A}$ and $\hat{B}$. □

5.6. Uncertainty

Definition. The *uncertainty* in a measurement of an observable A on a state ψ is defined as

$$\Delta_\psi A = \sqrt{(\Delta_\psi A)^2}$$

where

$$(\Delta_\psi A)^2 = \left\langle \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)^2 \right\rangle_\psi = \langle \hat{A}^2 \rangle_\psi - \left(\langle \hat{A} \rangle_\psi\right)^2$$

The two definitions are equivalent:

$$\begin{aligned}
\left\langle \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)^2 \right\rangle_\psi &= \int_{\mathbb{R}^3} \psi^\star \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)^2 \psi \, \mathrm{d}V \\
&= \int_{\mathbb{R}^3} \psi^\star \hat{A}^2 \psi \, \mathrm{d}V + \left(\langle \hat{A} \rangle_\psi\right)^2 \int_{\mathbb{R}^3} \psi^\star \psi \, \mathrm{d}V - 2 \langle \hat{A} \rangle_\psi \int_{\mathbb{R}^3} \psi^\star A \psi \, \mathrm{d}V \\
&= \langle \hat{A}^2 \rangle_\psi + \left(\langle \hat{A} \rangle_\psi\right)^2 - 2\left(\langle \hat{A} \rangle_\psi\right)^2 \\
&= \langle \hat{A}^2 \rangle_\psi - \left(\langle \hat{A} \rangle_\psi\right)^2
\end{aligned}$$

Lemma. $(\Delta_\psi A)^2 \geq 0$, and $\Delta_\psi A = 0$ if and only if ψ is an eigenfunction of $\hat{A}$.

Proof. Since $\hat{A}$ is Hermitian,

$$\begin{aligned}
(\Delta_\psi A)^2 &= \left\langle \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)^2 \right\rangle_\psi \\
&= \left\langle \psi, \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)^2 \psi \right\rangle \\
&= \left\langle \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)\psi, \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)\psi \right\rangle \\
&= \left\| \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)\psi \right\|
\end{aligned}$$

Let $\phi = \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)\psi$. The norm of any function is non-negative, so the square uncertainty is non-negative. Now, suppose this norm $\|\phi\|$ is zero. Then, $\phi = 0$. Hence,

$$\hat{A}\psi = \langle \hat{A} \rangle_\psi \psi$$

so it is an eigenfunction of $\hat{A}$. If ψ is conversely an eigenfunction of $\hat{A}$ with eigenvalue a, then

$$\langle \hat{A} \rangle_\psi = \langle \psi, \hat{A}\psi \rangle = a\|\psi\| = a$$

Further,

$$\langle \hat{A}^2 \rangle_\psi = \langle \psi, \hat{A}^2\psi \rangle = a^2$$

Hence,

$$\left(\Delta_\psi A\right)^2 = a^2 - a^2 = 0$$

□

5.7. Schwarz inequality

Theorem. Let $\psi, \phi \in \mathcal{H}$. Then,

$$|\langle \psi, \phi \rangle|^2 \leq \langle \phi, \phi \rangle \langle \psi, \psi \rangle$$

and

$$|\langle \psi, \phi \rangle|^2 = \langle \phi, \phi \rangle \langle \psi, \psi \rangle \iff \exists a \in \mathbb{C}, \phi = a\psi$$

Proof. For all $a \in \mathbb{C}$, we have

$$0 \leq \langle \phi - a\psi, \phi - a\psi \rangle$$

In particular, let

$$a = \frac{\langle \psi, \phi \rangle}{\langle \psi, \psi \rangle}$$

Then,

$$0 \leq \langle \phi, \phi \rangle - \frac{2|\langle \psi, \phi \rangle|^2}{\langle \psi, \psi \rangle} + \frac{|\langle \psi, \phi \rangle|^2}{\langle \psi, \psi \rangle} = \langle \phi, \phi \rangle - \frac{|\langle \psi, \phi \rangle|^2}{\langle \psi, \psi \rangle}$$

Hence,

$$|\langle \psi, \phi \rangle|^2 \leq \langle \psi, \psi \rangle \langle \phi, \phi \rangle$$

Equality holds if and only if $\phi - a\psi = 0$. □

5.8. Generalised uncertainty theorem

Theorem. Let A and B be observables, and $\psi \in \mathcal{H}$. Then

$$\left(\Delta_\psi A\right)\left(\Delta_\psi B\right) \geq \frac{1}{2}|\langle \psi, [\hat{A}, \hat{B}]\psi \rangle|$$

Proof.

$$\left(\Delta_\psi A\right)^2 = \left\langle \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)\psi, \left(\hat{A} - \langle \hat{A} \rangle_\psi \hat{I}\right)\psi \right\rangle$$

Defining $\hat{A}' = \hat{A} - \langle \hat{A} \rangle_\psi \hat{I}$ and $\hat{B}' = \hat{B} - \langle \hat{B} \rangle_\psi \hat{I}$,

$$\left(\Delta_\psi \hat{A}'\right)^2 = \langle \hat{A}'\psi, \hat{A}'\psi \rangle$$

and analogously for $\hat{B}'$. Now,

$$\left(\Delta_\psi \hat{A}'\right)^2 \left(\Delta_\psi \hat{B}'\right)^2 = \langle \hat{A}'\psi, \hat{A}'\psi \rangle \langle \hat{B}'\psi, \hat{B}'\psi \rangle \geq \left| \langle \hat{A}'\psi, \hat{B}'\psi \rangle \right|^2$$

Since $\hat{A}'$ is Hermitian,

$$(\Delta_\psi \hat{A}')(\Delta_\psi \hat{B}') \geq \left| \langle \psi, \hat{A}'\hat{B}'\psi \rangle \right|$$

By definition, $\left[\hat{A}, \hat{B}\right] = \hat{A}\hat{B} - \hat{B}\hat{A}$ and let the anticommutator be $\left\{\hat{A}, \hat{B}\right\} = \hat{A}\hat{B} + \hat{B}\hat{A}$. If $\hat{A}'$ and $\hat{B}'$ are Hermitian,

$$\left[\hat{A}', \hat{B}'\right]^\dagger = -\left[\hat{A}', \hat{B}'\right]$$

and

$$\left\{\hat{A}', \hat{B}'\right\}^\dagger = \left\{\hat{A}', \hat{B}'\right\}$$

So the anticommutator is Hermitian. Now, we can write

$$\hat{A}'\hat{B}' = \frac{1}{2}\left[\hat{A}', \hat{B}'\right] + \frac{1}{2}\left\{\hat{A}', \hat{B}'\right\}$$

Hence,

$$\begin{aligned}
(\Delta_\psi \hat{A}')(\Delta_\psi \hat{B}') &\geq \left| \left\langle \psi, \left(\frac{1}{2}\left[\hat{A}', \hat{B}'\right] + \frac{1}{2}\left\{\hat{A}', \hat{B}'\right\} \right) \psi \right\rangle \right| \\
&= \left| \left\langle \psi, \frac{1}{2}\left[\hat{A}', \hat{B}'\right]\psi \right\rangle + \left\langle \psi, \frac{1}{2}\left\{\hat{A}', \hat{B}'\right\}\psi \right\rangle \right|
\end{aligned}$$

We can prove that $\left\langle \psi, \left\{\hat{A}', \hat{B}'\right\}\psi \right\rangle \in \mathbb{R}$. Since the anticommutator is Hermitian,

$$\left\langle \psi, \left\{\hat{A}', \hat{B}'\right\}\psi \right\rangle = \left\langle \left\{\hat{A}', \hat{B}'\right\}\psi, \psi \right\rangle = \left\langle \psi, \left\{\hat{A}', \hat{B}'\right\}\psi \right\rangle^\star$$

Analogously we can prove that $\left\langle \psi, \left[\hat{A}', \hat{B}'\right]\psi \right\rangle \in i\mathbb{R}$.

$$\left\langle \psi, \left[\hat{A}', \hat{B}'\right]\psi \right\rangle = \left\langle \left[\hat{A}', \hat{B}'\right]^\star \psi, \psi \right\rangle = -\left\langle \psi, \left[\hat{A}', \hat{B}'\right]\psi \right\rangle^\star$$

Hence,

$$\begin{aligned}
\left(\Delta_\psi \hat{A}'\right)^2 \left(\Delta_\psi \hat{B}'\right)^2 &\geq \left| \left\langle \psi, \frac{1}{2}\left[\hat{A}', \hat{B}'\right]\psi \right\rangle + \left\langle \psi, \frac{1}{2}\left\{\hat{A}', \hat{B}'\right\}\psi \right\rangle \right|^2 \\
&= \frac{1}{4}\left| \langle \psi, \left[\hat{A}', \hat{B}'\right]\psi \rangle \right|^2 + \frac{1}{4}\left| \langle \psi, \left\{\hat{A}', \hat{B}'\right\}\psi \rangle \right|^2 \\
&\geq \frac{1}{4}\left| \langle \psi, \left\{\hat{A}', \hat{B}'\right\}\psi \rangle \right|^2 \\
\therefore \left(\Delta_\psi \hat{A}'\right)^2 \left(\Delta_\psi \hat{B}'\right)^2 &\geq \frac{1}{4}\left| \langle \psi, \{\hat{A}, \hat{B}\}\psi \rangle \right|^2
\end{aligned}$$

□

5.9. Consequences of uncertainty relation

(i) $\left[\hat{A}, \hat{B}\right] = 0$ implies that there exists a joint set of eigenfunctions which is a complete basis of $\mathcal{H}$. In particular, $\hat{A}$ and $\hat{B}$ can be measured simulaneously with arbitrary precision. For instance, we can measure E, $\left|\overline{L}\right|$ and L_z simultaneously for an electron on a hydrogen atom.

(ii) We cannot simultaneously measure position and momentum of a particle with arbitrary precision. In particular,

$$\Delta_\psi x \Delta_\psi p \geq \frac{\hbar}{2}$$

This is Heisenberg's uncertainty principle.

5.10. States of minimal uncertainty

The Gaussian wavepacket was a state of minimal uncertainty:

$$\Delta_\psi x \Delta_\psi p = \frac{\hbar}{2}$$

We would like to analyse the conditions for a state ψ to have minimal uncertainty.

Lemma. ψ is a state of minimal uncertainty if and only if

$$\hat{x}\psi = ia\hat{p}\psi$$

for some $a \in \mathbb{R}$. A non-example is the De Broglie plane waves.

Lemma. The condition for the above lemma to hold is that

$$\psi(x) = ce^{-bx^2}; \quad b, c \in \mathbb{R}, b > 0, c \neq 0$$

The Gaussian wavepacket is an example of this form.

5.11. Ehrenfest theorem

Theorem. The time evolution of a Hermitian operator $\hat{A}$ is governed by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{A} \right\rangle_\psi = \frac{i}{\hbar}\left\langle \left[\hat{H}, \hat{A}\right] \right\rangle_\psi + \left\langle \frac{\partial \hat{A}}{\partial t} \right\rangle_\psi$$

In this course, we will not see any operators with time dependence, so the last term will not be needed.

Proof.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{A}\right\rangle_{\psi} &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty}\psi^{\star}\hat{A}\psi\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\frac{\partial}{\partial t}\left(\psi^{\star}\hat{A}\psi\right)\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\left[\frac{\partial\psi^{\star}}{\partial t}\hat{A}\psi + \psi^{\star}\frac{\partial\hat{A}}{\partial t}\psi + \psi^{\star}\hat{A}\frac{\partial\psi}{\partial t}\right]\mathrm{d}x\end{aligned}$$

The time-dependent Schrödinger equation gives

$$\left(i\hbar\frac{\partial\psi}{\partial t}\right)^{\star} = \left(\hat{H}\psi\right)^{\star} \implies -i\hbar\frac{\partial\psi^{\star}}{\partial t} = \psi^{\star}\hat{H}^{\star} = \psi^{\star}\hat{H}$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{A}\right\rangle_{\psi} &= \frac{i}{\hbar}\int_{-\infty}^{\infty}\left[\psi^{\star}\hat{H}\hat{A}\psi - \psi^{\star}\hat{A}\hat{H}\psi\right]\mathrm{d}x + \int_{-\infty}^{\infty}\psi^{\star}\frac{\partial\hat{A}}{\partial t}\psi\,\mathrm{d}x \\ &= \frac{i}{\hbar}\left\langle\left[\hat{H},\hat{A}\right]\right\rangle_{\psi} + \left\langle\frac{\partial\hat{A}}{\partial t}\right\rangle_{\psi}\end{aligned}$$

□

Example. Let $\hat{A} = \hat{H}$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{H}\right\rangle_{\psi} = 0$$

This corresponds to the classical notion of conservation of energy.

Example. Let $\hat{A} = \hat{p}$. First, note

$$\begin{aligned}\left[\hat{H},\hat{p}\right]\psi &= \left[\frac{\hat{p}^2}{2m} + U(\hat{x}),\hat{p}\right]\psi \\ &= \left[U(\hat{x}),\hat{p}\right]\psi \\ &= U(x)\left(-i\hbar\frac{\partial}{\partial x}\right)\psi - \left(-i\hbar\frac{\partial}{\partial x}\right)U(x)\psi \\ &= i\hbar\frac{\partial U(x)}{\partial x}\psi\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\hat{p}\right\rangle_{\psi} = \frac{i}{\hbar}\left\langle\left[\hat{H},\hat{p}\right]\right\rangle_{\psi} = -\left\langle\frac{\partial U}{\partial x}\right\rangle_{\psi}$$

This corresponds exactly to Newton's second law,

$$\dot{p} = -\frac{\mathrm{d}U}{\mathrm{d}x}$$

Example. Let $\hat{A} = \hat{x}$. We have

$$\begin{aligned}\left[\hat{H}, \hat{x}\right]\psi &= \left[\frac{\hat{p}^2}{2m} + U(\hat{x}), \hat{x}\right]\psi \\ &= \frac{1}{2m}\left[\hat{p}^2, \hat{x}\right]\psi \\ &= \frac{1}{2m}(\hat{p}[\hat{p}, \hat{x}] + [\hat{p}, \hat{x}]\hat{p})\psi \\ &= \frac{-i\hbar}{m}\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{x}\right\rangle_\psi = \frac{i}{\hbar}\left\langle \left[\hat{H}, \hat{x}\right]\right\rangle_\psi = \frac{\left\langle \hat{p}\right\rangle_\psi}{m}$$

which aligns with the classical equation

$$\dot{x} = \frac{p}{m}$$

6. Three-dimensional solutions to the Schrödinger equation

6.1. Time-independent Schrödinger equation in spherical polar coordinates

For a spherically symmetric potential in $\mathbb{R}^3$, the time-independent Schrödinger equation is

$$-\frac{\hbar^2}{2m}\nabla^2\chi(x) + U(x)\chi(x) = E\chi(x)$$

Recall that the Laplacian operator can be expanded in spherical polar coordinates as

$$-\frac{\hbar^2}{2m}\left(\frac{1}{r}\frac{\partial^2}{\partial r^2}r + \frac{1}{r^2\sin^2\theta}\left[\sin\theta\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{\partial^2}{\partial\phi^2}\right]\right)\chi(x) + U(x)\chi(x) = E\chi(x)$$

where

$$x = r\cos\phi\sin\theta; \quad y = r\sin\phi\sin\theta; \quad z = r\cos\theta$$

Definition. A *spherically symmetric potential* is a potential U which depends only on r.

We search for the particular solutions of the time-dependent Schrödinger equation with spherically symmetric potential that are radial eigenfunctions. If $\chi(r)$ is a function of r alone,

$$\nabla^2\chi(r) = \frac{1}{r}\frac{\partial^2}{\partial r^2}(r\chi(r))$$

Hence,

$$-\frac{\hbar^2}{2mr}\frac{\partial^2}{\partial r^2}(r\chi(r)) + U(r)\chi(r) = E\chi(r)$$

This is equivalent to

$$-\frac{\hbar^2}{2m}\left(\chi''(r) + \frac{2}{r}\chi'(r)\right) + U(r)\chi(r) = E\chi(r)$$

The normalisation condition is

$$\int_0^\infty |\chi(r)|^2 r^2\,\mathrm{d}r < N$$

The eigenfunctions $\chi(r)$ must converge to zero sufficiently fast as $r \to \infty$ in order to be normalisable. To solve the time-independent Schrödinger equation, we will define

$$\sigma(r) = r\chi(r)$$

Then,

$$-\frac{\hbar^2}{2m}\sigma''(r) + U(r)\sigma(r) = E\sigma(r)$$

This is defined for $r \geq 0$. The normalisation condition here is

$$\int_0^\infty |\sigma(r)|^2\,\mathrm{d}r < N; \quad \sigma(0) = 0; \quad \sigma'(0) < \infty$$

The conditions at zero force χ to be defined and have finite derivative at zero. To solve the equation for σ, we solve on $\mathbb{R}$ and search for odd solutions $\sigma^{(-)}$, so

$$\sigma^{(-)}(-r) = -\sigma^{(-)}(r)$$

6.2. Spherically symmetric potential well

Consider the potential well given by

$$U(r) = \begin{cases} 0 & r \le a \\ U_0 & r > a \end{cases}$$

where $a, U_0 > 0$. The time-independent Schrödinger equation is

$$-\frac{\hbar^2}{2m}\sigma''(r) + U(r)\sigma(r) = E\sigma(r)$$

We search for odd-parity bound states, so $0 < E < U_0$. Let

$$k = \sqrt{\frac{2mE}{\hbar^2}}; \quad \bar{k} = \sqrt{\frac{2m(U_0 - E)}{\hbar^2}}$$

The solution for σ is

$$\sigma(r) = \begin{cases} A\sin(kr) & r \le a \\ Be^{-\bar{k}r} & r > a \end{cases}$$

The continuity condition at $r = a$ can be imposed to find $A\sin ka = Be^{-\bar{k}a}$. The continuity of the derivative gives $kA\cos ka = -\bar{k}Be^{-\bar{k}a}$. Therefore,

$$-k\cot(ka) = \bar{k}; \quad k^2 + \bar{k}^2 = \frac{2mU_0}{\hbar^2}$$

Hence,

$$-\xi\cot\xi = \eta; \quad \xi^2 + \eta^2 = r_0^2$$

where $\xi = ka$ and $\eta = \bar{k}a$, and $r_0 = a\sqrt{2mU_0/\hbar}$. If $r_0 < \frac{\pi}{2}$, we have no solutions because $\xi \ge 0$. Equivalently, there are no solutions if

$$U_0 < \frac{\pi^2\hbar^2}{8ma^2}$$

7. Solution to hydrogen atom

7.1. Radial wavefunction of hydrogen atom

The hydrogen atom is comprised of a nucleus and a single electron. The nucleus has a positive charge and the electron has a negative charge. We will model the proton to be stationary at the origin. The Coulomb force experienced by the electron is given by

$$F = -\frac{e^2}{4\pi\varepsilon_0}\frac{1}{r^2} = -\frac{\partial U}{\partial r} \implies U = -\frac{e^2}{4\pi\varepsilon_0}\frac{1}{r}$$

Since zero potential is achieved only at infinity, we search for bound states with $E < 0$. We will search for the radial symmetric eigenfunctions. We have

$$-\frac{\hbar^2}{2m_e}\left(\chi''(r) + \frac{2}{r}\chi'(r)\right) - \frac{e^2}{4\pi\varepsilon_0}\frac{1}{r}\chi(r) = E\chi(r)$$

We define

$$\nu^2 = -\frac{2mE}{\hbar^2} > 0; \quad \beta = \frac{e^2 m_e}{2\pi\varepsilon_0\hbar^2} > 0$$

The Schrödinger equation becomes

$$\chi''(r) + \frac{2}{r}\chi'(r) + \left(\frac{\beta}{r} - \nu^2\right)\chi(r) = 0$$

Asymptotically as $r \to \infty$, we can see that $\chi'' \sim \nu^2\chi$. Since $\nu^2 > 0$, this yields solutions that asymptotically behave similarly to $e^{-r\nu}$, where the positive exponential solution is not applicable due to the normalisation condition. For $r = 0$, the eigenfunction should be finite. We will consider an ansatz educated by the asymptotical behaviour. Suppose

$$\chi(r) = f(r)e^{-\nu r}$$

and we solve for $f(r)$. The Schrödinger equation is

$$f''(r) + \frac{2}{r}(1 - \nu r)f'(r) + \frac{1}{r}(\beta - 2\nu)f(r) = 0$$

This is a homogeneous linear ordinary differential equation with a regular point at $r = 0$. Suppose there exist series solutions.

$$f(r) = r^c \sum_{n=0}^{\infty} a_n r^n$$

We can differentiate and find

$$f'(r) = \sum_{n=0}^{\infty} a_n(c+n)r^{c+n-1}; \quad f''(r) = \sum_{n=0}^{\infty} a_n(c+n)(c+n-1)r^{c+n-2}$$

Hence,

$$\sum_{n=0}^{\infty}\left[a_n(c+n)(c+n-1)r^{c+n-2}+\frac{2}{r}(1-\nu r)a_n(c+n)r^{c+n-1}+(\beta-2\nu)r^{c+n-1}\right]=0$$

By comparing coefficients of the lowest power of r,

$$a_0c(c-1)+2a_0c=0 \implies a_0c(c+1)=0 \implies c=-1,0$$

The solution $c=-1$ implies $\chi(r)\sim\frac{1}{r}$ which is invalid at $r=0$. So we require $c=0$. Then the power series becomes

$$\sum_{n=0}^{\infty}a_n[n(n-1)+2n]r^{n-2}+\sum_{n=0}^{\infty}a_n(-2\nu n+\beta-2\nu)r^{n-1}=0$$

Comparing coefficients of equal powers of r,

$$a_nn(n+1)+a_{n-1}(-2\nu n+2\nu+\beta-2\nu)=0$$

Hence, we arrive at the recurrence relation

$$a_n=\frac{2\nu n-\beta}{n(n+1)}a_{n-1}$$

Suppose this series were infinite. Asymptotically, the behaviour of $f(r)$ is determined by $\frac{a_n}{a_{n-1}}\sim\frac{2\nu}{n}$. We can compare this behaviour to the asymptotic behaviour of $g(r)=e^{2\nu r}$. In this case, the series expansion with coefficients b_n satisfies

$$b_n=\frac{(2\nu)^n}{n!} \implies \frac{b_n}{b_{n-1}}=\frac{2\nu}{n}$$

Hence, asymptotically $f(r)\sim e^{2\nu r}$ if the series does not terminate. Since $\chi(r)=f(r)e^{-\nu r}$, we have $\chi(r)\sim e^{\nu r}$ which is not normalisable. Hence the series is finite. So there exists an integer $N>0$ such that $a_N=0$ and $a_{N-1}\neq 0$. This implies $2\nu N-\beta=0$ hence $\nu=\frac{\beta}{2N}$. Substituting ν^2 and β, we find

$$E=E_N=-\frac{e^4m_e}{32\pi^2\varepsilon_0^2\hbar^2N^2}$$

So the eigenvalues are equivalent to those found in the Bohr model. We now wish to find the radial eigenfunctions. Note, $\frac{\beta}{2\nu}=N$ hence we can substitute and find

$$\frac{a_n}{a_{n-1}}=-2\nu\frac{N-n}{n(n+1)}$$

This recursion can be used to find the coefficients of the polynomial $f_N(r)$.

$$\begin{aligned} f_1(r) &= 1 \\ f_2(r) &= 1 - \nu r \\ f_3(r) &= 1 - 2\nu r + \frac{2}{3}\nu^2 r^2 \end{aligned}$$

These are called the Laguerre polynomials of order $N-1$ (for example, the first order Laguerre polynomial is f_2). We can then multiply the Laguerre polynomials by $e^{-\nu r}$ and normalise over $\mathbb{R}^3$ to find the normalised eigenfunctions $\chi_N(r)$. For example,

$$\chi_1(r) = \frac{\nu^{3/2}}{\sqrt{\pi}} = \frac{1}{\sqrt{\pi}}\left(\frac{e^2 m_e}{4\pi\varepsilon_0\hbar^2}\right)^{3/2} e^{-\nu r}$$

Recall that the Bohr model implied that the ground state has radius a_0, known as the Bohr radius, given in terms of ν by $a_0 = \frac{1}{\nu}$. Using quantum mechanics, we instead find

$$\begin{aligned} \langle r \rangle_{\chi_1} &= \int_{\mathbb{R}^3} \chi_1^*(r) r \chi_1(r)\,\mathrm{d}V \\ &= \int_0^{2\pi} \mathrm{d}\phi \int_{-1}^{1} \mathrm{d}\cos\theta \int_0^{\infty} \frac{\nu^3}{\pi} r^3 e^{-2\nu r}\,\mathrm{d}r \\ &= 4\pi\frac{\nu^3}{\pi}\int_0^{\infty} r^3 e^{-2\nu r}\,\mathrm{d}r \\ &= \frac{3}{2}a_0 \end{aligned}$$

We have verified with physical experiments that this larger expected radius is physically accurate.

7.2. Angular momentum

Recall that classically the angular momentum L is given by

$$L = x \times p$$

Spherically symmetric potentials conserve classical angular momentum:

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \dot{x} \times p + x \times \dot{p} = 0$$

Solving classical problems in this way allows us to reduce a three-dimensional problem into a two-dimensional problem, by considering motion on the plane $L \cdot x = 0$. Then we reduce to one dimension by considering $\hat{e}_r$. In quantum mechanics, we can do an analogous simplification.

Definition. In quantum mechanics, the angular momentum is given by

$$\hat{L} = \hat{x} \times \hat{p} = i\hbar x \times \nabla$$

In Cartesian coordinates, this reduces to

$$\hat{L}_i = -i\hbar\varepsilon_{ijk} x_j \frac{\partial}{\partial x_k}$$

Each component $\hat{L}_i$ is a Hermitian operator. Note,

$$\begin{aligned}
[\hat{L}_1, \hat{L}_2]\psi(x_1, x_2, x_3) &= -\hbar^2 \Bigg[\left(x_2 \frac{\partial}{\partial x_3} - x_3 \frac{\partial}{\partial x_2} \right) \left(x_3 \frac{\partial}{\partial x_1} - x_1 \frac{\partial}{\partial x_3} \right) \\
&\quad - \left(x_3 \frac{\partial}{\partial x_1} - x_1 \frac{\partial}{\partial x_3} \right) \left(x_2 \frac{\partial}{\partial x_3} - x_3 \frac{\partial}{\partial x_2} \right) \Bigg] \psi \\
&= -\hbar^2 \left[x_2 \frac{\partial}{\partial x_1} - x_1 \frac{\partial}{\partial x_2} \right] \psi \\
&= -i\hbar \hat{L}_3 \psi
\end{aligned}$$

Hence the commutator $[\hat{L}_i, \hat{L}_j] = i\hbar\varepsilon_{ijk}\hat{L}_k$ is nonzero for $i \neq j$. In particular, we cannot measure each component of the angular momentum simultaneously.

Definition. The *total angular momentum* is

$$\hat{L}^2 = \hat{L}_1^2 + \hat{L}_2^2 + \hat{L}_3^2$$

We can find that $[\hat{L}^2, \hat{L}_i] = 0$, so we can measure both the total angular momentum and a specific component of angular momentum simultaneously. For a spherically symmetric potential, given by $\hat{H} = \frac{\hat{p}^2}{2m} + U(\hat{r})$, we can find

$$[\hat{H}, \hat{L}^2] = [\hat{H}, \hat{L}_i] = 0$$

7.3. Commutativity of angular momentum operators

The set $\{\hat{H}, \hat{L}^2, \hat{L}_i\}$ commutes pairwise. By convention, we choose $i = 3$ to extract the z component of the angular momentum. Hence,

(i) We can find joint eigenstates of the three operators, and such eigenstates can be chosen to form a basis for the Hilbert space $\mathcal{H}$.

(ii) The corresponding eigenvalues $|L|, L_z, E$ can be measured simultaneously to an arbitrary precision.

(iii) The set of operators is maximal; there exists no operator (other than a linear combination of the above) that commutes with all three.

7.4. Joint eigenfunctions of angular momentum

We search for joint eigenfunctions of $\hat{L}_z$ and $\hat{L}^2$. We will write $\hat{L}$ in spherical coordinates. In Cartesian coordinates,

$$\hat{L} = -i\hbar x \cdot \nabla$$

Hence,

$$\hat{L}_3 = -i\hbar \frac{\partial}{\partial \phi}; \quad \hat{L}^2 = -\frac{\hbar^2}{\sin^2\theta}\left[\sin\theta \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial}{\partial\theta}\right) + \frac{\partial^2}{\partial\phi^2}\right]$$

Now we search for eigenfunctions of these operators.

$$\hat{L}^2 Y(\theta,\phi) = \lambda Y(\theta,\phi); \quad \hat{L}_3 Y(\theta,\phi) = \hbar m Y(\theta,\phi)$$

Solving the equation in $\hat{L}_3$,

$$-i\hbar \frac{\partial}{\partial\phi} Y(\theta,\phi) = \hbar m Y(\theta,\phi)$$

We can find solutions of the form $Y(\theta,\phi) = y(\theta)x(\phi)$. We find

$$-i\hbar y(\theta)x'(\phi) = \hbar m y(\theta)x(\phi)$$

Hence $y(\theta)$ is arbitrary, and further

$$-i\hbar x'(\phi) = \hbar m x(\phi) \implies x(\phi) = e^{im\phi}$$

Given that the wavefunctions must be single-valued on $\mathbb{R}^3$, we must have $x(\phi)$ invariant under the choice of $\phi = \phi + 2\pi k$. Hence m must be an integer. Since this must also be an eigenfunction of $\hat{L}^2$, we have further

$$-\frac{\hbar^2}{\sin^2\theta}\left[\sin\theta \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial}{\partial\theta}\right) + \frac{\partial^2}{\partial\phi^2}\right][y(\theta)x(\phi)] = \lambda y(\theta)x(\phi)$$

Hence, substituting $x(\phi) = e^{im\phi}$, we find

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta y'(\theta)) - \frac{m^2}{\sin^2\theta} y(\theta) = -\frac{\lambda}{\hbar^2} y(\theta)$$

This is the associate Legendre equation. The solutions of $y(\theta)$ are the associate Legendre functions.

$$y(\theta) = P_{\ell,m}(\cos\theta) = (\sin\theta)^{|m|} \frac{\mathrm{d}^{|m|}}{\mathrm{d}(\cos\theta)^{|m|}} P_\ell(\cos\theta)$$

where the P_ℓ are the Legendre polynomials. Since the ordinary Legendre polynomials are of degree ℓ, we must have $|m| \leq \ell$ to obtain a nonzero solution. This corresponds to the classical notion that $|L_z| \leq |L|$ for a physical solution. The eigenvalues of $\hat{L}^2$ are

$$\lambda = \ell(\ell+1)\hbar^2$$

with $\ell \in \{0, 1, 2, \dots\}$. Thus,

$$Y_{\ell,m}(\theta,\phi) = P_{\ell,m}(\cos\theta)e^{im\phi}$$

The Y functions are called the spherical harmonics. The parameters ℓ, m are known as the quantum numbers of the eigenfunction; ℓ is the total angular momentum quantum number and m is the azimuthal quantum number. Examples of normalised eigenfunctions are

$$Y_{0,0} = \frac{1}{\sqrt{4\pi}}$$
$$Y_{1,0} = \sqrt{\frac{3}{4\pi}}\cos\theta$$
$$Y_{1,\pm1} = \mp\sqrt{\frac{3}{8\pi}}\sin\theta e^{-i\phi}$$

All spherical harmonics can be shown to be orthogonal.

7.5. Full solution to hydrogen atom

The time-independent Schrödinger equation for the hydrogen atom is

$$-\frac{\hbar^2}{2m_e}\nabla^2\chi(r,\theta,\phi) - \frac{e^2}{4\pi\varepsilon_0}\frac{1}{r}\chi(r,\theta,\phi) = E\chi(r,\theta,\phi)$$

Writing the Laplacian in spherical polar coordinates,

$$\nabla^2 = \frac{1}{r}\frac{\partial^2}{\partial r^2} + \frac{1}{r^2\sin^2\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + \frac{\partial^2}{\partial\phi^2}\right)$$

Hence,

$$\hat{L}^2 = \frac{\hbar^2}{\sin^2\theta}\left[\sin\theta\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + \frac{\partial^2}{\partial\phi^2}\right] \implies -\hbar^2\nabla^2 = -\frac{\hbar^2}{r}\frac{\partial^2}{\partial r^2}r + \frac{\hat{L}^2}{r^2}$$

Thus we can rewrite the TISE as

$$-\frac{\hbar^2}{2m_e}\frac{1}{r}\left(\frac{\partial^2}{\partial r^2}(r\chi)\right) + \frac{\hat{L}^2}{2m_e r^2}\chi - \frac{e^2}{4\pi\varepsilon_0 r}\chi = E\chi$$

Since $\hat{L}^2, \hat{L}_3, \hat{H}$ are a maximal set of pairwise commuting operators, we know that the eigenfunctions of the Hamiltonian χ must also be eigenfunctions of $\hat{L}^2, \hat{L}_3$. Hence,

$$\chi(r,\theta,\phi) = R(r)Y_{\ell,m}(\theta,\phi)$$

Since χ is an eigenfunction of $\hat{L}^2$,

$$\hat{L}^2(R(r)Y_{\ell,m}(\theta,\phi)) = R(r)\hbar^2\ell(\ell+1)Y_{\ell,m}(\theta,\phi)$$

Substituting into the TISE, we find

$$-\frac{\hbar^2}{2m_e}\left(\frac{\partial^2 R}{\partial r^2} + \frac{2}{r}\frac{\partial R}{\partial r}\right)Y_{\ell,m}(\theta,\phi) + \frac{\hbar^2}{2m_e r^2}\ell(\ell+1)R(r)Y_{\ell,m}(\theta,\phi) - \frac{e^2}{4\pi\varepsilon_0 r}R(r)Y_{\ell,m}(\theta,\phi)$$
$$= ER(r)Y_{\ell,m}(\theta,\phi)$$

Cancelling the spherical harmonic,

$$-\frac{\hbar^2}{2m_e}\left(\frac{\partial^2 R}{\partial r^2}+\frac{2}{r}\frac{\partial R}{\partial r}\right)+\underbrace{\left(\frac{\hbar^2}{2m_e r^2}\ell(\ell+1)-\frac{e^2}{4\pi\varepsilon_0 r}\right)}_{U_{\text{eff}}=\text{ effective potential}}R(r)=ER(r)$$

This is an equation for the radial part of the solution. We have already solved this equation for $\ell = 0$ to find $\chi(r)$, the radial wavefunction. Note that the azimuthal quantum number does not appear in the effective potential, giving a degeneracy of order at least $2\ell + 1$. We define

$$\nu^2 = -\frac{2m_e E}{\hbar^2} > 0; \quad \beta = \frac{e^2 m_e}{2\pi\varepsilon_0\hbar^2}$$

Hence,

$$R'' + \frac{2}{r}R' + \left(\frac{\beta}{r} - \nu^2 - \frac{\ell(\ell+1)}{r^2}\right)R = 0$$

The asymptotic limit is as before in the radial case, since the angular velocity dependence is suppressed by $\frac{1}{r^2}$. We have $R'' - \nu^2 R \to 0$ hence $R \propto e^{-\nu r}$ in the limit. We let $R(r) = g(r)e^{-\nu r}$. Then,

$$g'' + \frac{2}{r}(1-\nu r)g' + \left(\frac{\beta}{r} - 2\nu - \frac{\ell(\ell+1)}{r^2}\right)g = 0$$

Expanding in power series,

$$g(r) = r^\sigma \sum_{n=0}^{\infty} a_n r^n$$

Substituting and comparing the lowest power of r,

$$a_0[\sigma(\sigma-1) + 2\sigma - \ell(\ell+1)] = 0 \implies \sigma(\sigma+1) = \ell(\ell+1)$$

Hence, $\sigma = \ell$ or $\sigma = -\ell - 1$. If $\sigma = -\ell - 1$, we have $R(r) \sim \frac{1}{r^{\ell+1}}$ which cannot be the solution, so $\sigma = \ell$. Thus,

$$g(r) = r^\ell \sum_{n=0}^{\infty} a_n r^n$$

We can evaluate the recurrence relation between the coefficients as before to find

$$\sum_{n=0}^{\infty}[(n+\ell)(n+\ell-1)a_n + 2(n+1)a_n - \ell(\ell+1)a_n \\ - 2\nu(n+\ell-1)a_{n-1} + (\beta - 2\nu)a_{n-1}]r^{\ell+n-2} = 0$$

which gives

$$a_n = \frac{2\nu(n+\ell) - \beta}{n(n+2\ell-1)}$$

If $\ell = 0$ this yields the result for the radial solution. Unless the series terminates, it is possible to show that R diverges. Hence g must be a polynomial with first zero coefficient $a_{n_{\max}}$. Here,

$$2\nu(n_{\max} + \ell) - \beta = 0$$

We define $N = n_{\max} + \ell$, so $2\nu N - \beta = 0$ giving $\nu = \frac{\beta}{2N}$. Note that $N > \ell$ since $n_{\max} > 0$. We can then find the energy level to be

$$E_N = -\frac{e^4 m_e}{32\pi^2 \varepsilon_0^2 \hbar^2} \frac{1}{n^2}$$

which is an identical energy spectrum as we found before when not considering angular momentum (using the Bohr model). For each E_N, we have $N = n_{\max} + \ell$ so there can be $\ell = 0, \dots, N-1$ and $m = -\ell, \dots, \ell$. Hence, the degeneracy of the solution for each N is

$$D(N) = \sum_{\ell=0}^{N-1} \sum_{m=-\ell}^{\ell} 1 = N^2$$

So the degeneracy increases quadratically with the energy level. For example, for $N = 2$ there are four possible eigenfunctions with the same energy. The eigenfunctions are now dictated by three quantum numbers.

$$\chi_{N,\ell,m}(r,\theta,\phi) = R_{N,\ell}(r) Y_{\ell,m}(\theta,\phi) = r^\ell g_{N,\ell}(r) e^{-\frac{\beta r}{2N}} Y_{\ell,m}(\theta,\phi)$$

where $g_{N,\ell}$ is a polynomial of degree $N - \ell - 1$ defined by the recurrence relation

$$a_k = \frac{2\nu}{k} \frac{k + \ell - N}{k + 2\ell + 1} a_{n-1}$$

These are the generalised Laguerre polynomials, often written

$$g_{N,\ell}(r) = L_{N-\ell-1}^{2\ell+1}(2r)$$

The quantum number $N \in \{0, 1, \dots\}$ is known as the *principal* quantum number.

7.6. Comparison to Bohr model

In the Bohr model, the energy levels were predicted accurately. Further, the maximum of the radial probability corresponds to the orbits found in the Bohr model:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\left|\chi_{N,0,0}(r)\right|^2 r^2\right) = 0$$

The classical trajectory, and the assumption about the angular momentum $L^2 = N^2\hbar^2$, were incorrect. The angular momentum found in quantum mechanics is $L^2 = \ell(\ell+1)\hbar^2$, which corresponds closely with the Bohr model for large ℓ.

7.7. Other elements of the periodic table

The above solution does not hold for other elements of the periodic table. Generalising to a nucleus with charge $+ze$ and z orbiting electrons, we could model this as

$$\chi(x_1, \dots, x_z) = \chi(x_1) \dots \chi(x_N); \quad E = \sum_{j=1}^{N} e_j$$

This approximation can be acceptable for small z, but diverges very quickly from the true solution as z increases, due to the electron-electron interactions and the Pauli exclusion principle.

VII. Linear Algebra

Lectured in Michaelmas 2021 by PROF. P. RAPHAEL

Linear algebra is the field of study that deals with vector spaces and linear maps. A vector space can be thought of as a generalisation of $\mathbb{R}^n$ or $\mathbb{C}^n$, although they can be based off any field (not just $\mathbb{R}$ or $\mathbb{C}$), and may have infinitely many dimensions. In this course, we mainly study finite-dimensional vector spaces and the linear functions between them. Any linear map between finite-dimensional vector spaces can be encoded as a matrix. Such maps have properties such as their trace and determinant, which can be easily obtained from a matrix representing them. As was shown for real matrices in Vectors and Matrices, if the determinant of a matrix is nonzero it can be inverted.

Contents

1. Vector spaces and linear dependence 351
1.1. Vector spaces 351
1.2. Subspaces 351
1.3. Sum of subspaces 352
1.4. Quotients 352
1.5. Span 353
1.6. Dimensionality 353
1.7. Linear independence 354
1.8. Bases 354
1.9. Steinitz exchange lemma 355
1.10. Consequences of Steinitz exchange lemma 356
1.11. Dimensionality of sums 356
1.12. Direct sums 357
2. Linear maps 360
2.1. Linear maps 360
2.2. Isomorphism 361
2.3. Kernel and image 362
2.4. Rank and nullity 363
2.5. Space of linear maps 363
2.6. Matrices 364
2.7. Linear maps as matrices 364
2.8. Change of basis 366
2.9. Equivalent matrices 367
2.10. Column rank and row rank 369
2.11. Conjugation and similarity 370
2.12. Elementary operations 370
2.13. Gauss' pivot algorithm 371
2.14. Representation of square invertible matrices 371
3. Dual spaces 373
3.1. Dual spaces 373
3.2. Annihilators 375
3.3. Dual maps 375
3.4. Properties of dual map 377
3.5. Double duals 378
4. Bilinear forms 380
4.1. Introduction 380
4.2. Change of basis for bilinear forms 382

5. Trace and determinant . . . 383
5.1. Trace . . . 383
5.2. Permutations and transpositions . . . 383
5.3. Determinant . . . 384
5.4. Volume forms . . . 385
5.5. Multiplicative property of determinant . . . 386
5.6. Singular and non-singular matrices . . . 387
5.7. Determinants of linear maps . . . 388
5.8. Determinant of block-triangular matrices . . . 388
6. Adjugate matrices . . . 390
6.1. Column and row expansions . . . 390
6.2. Adjugates . . . 391
6.3. Cramer's rule . . . 392
7. Eigenvectors and eigenvalues . . . 393
7.1. Eigenvalues . . . 393
7.2. Polynomials . . . 393
7.3. Characteristic polynomials . . . 394
7.4. Polynomials for matrices and endomorphisms . . . 395
7.5. Sharp criterion of diagonalisability . . . 396
7.6. Simultaneous diagonalisation . . . 398
7.7. Minimal polynomials . . . 399
7.8. Cayley–Hamilton theorem . . . 400
7.9. Algebraic and geometric multiplicity . . . 401
7.10. Characterisation of diagonalisable complex endomorphisms . . . 402
8. Jordan normal form . . . 404
8.1. Definition . . . 404
8.2. Similarity to Jordan normal form . . . 404
8.3. Direct sum of eigenspaces . . . 404
9. Properties of bilinear forms . . . 408
9.1. Changing basis . . . 408
9.2. Quadratic forms . . . 408
9.3. Diagonalisation of symmetric bilinear forms . . . 409
9.4. Sylvester's law . . . 410
9.5. Kernels of bilinear forms . . . 412
9.6. Sesquilinear forms . . . 413
9.7. Hermitian forms . . . 413
9.8. Polarisation identity . . . 414
9.9. Hermitian formulation of Sylvester's law . . . 414
9.10. Skew-symmetric forms . . . 415
9.11. Skew-symmetric formulation of Sylvester's law . . . 415

10. Inner product spaces 416
10.1. Definition 416
10.2. Cauchy–Schwarz inequality 416
10.3. Orthogonal and orthonormal sets 417
10.4. Parseval's identity 418
10.5. Gram–Schmidt orthogonalisation process 418
10.6. Orthogonality of matrices 419
10.7. Orthogonal complement and projection 419
10.8. Projection maps 420
10.9. Adjoint maps 421
10.10. Self-adjoint and isometric maps 421
10.11. Spectral theory for self-adjoint maps 422
10.12. Spectral theory for unitary maps 423
10.13. Application to bilinear forms 424
10.14. Simultaneous diagonalisation 425

1. Vector spaces and linear dependence

1.1. Vector spaces

Definition. Let F be an arbitrary field. An *F-vector space* is an abelian group $(V, +)$ equipped with a function

$$F \times V \to V; \quad (\lambda, v) \mapsto \lambda v$$

such that

(i) $\lambda(v_1 + v_2) = \lambda v_1 + \lambda v_2$

(ii) $(\lambda_1 + \lambda_2)v = \lambda_1 v + \lambda_2 v$

(iii) $\lambda(\mu v) = (\lambda\mu)v$

(iv) $1v = v$

Such a vector space may also be called a *vector space over F*.

Example. Let X be a set, and define $\mathbb{R}^X = \{f : X \to \mathbb{R}\}$. Then $\mathbb{R}^X$ is an $\mathbb{R}$-vector space, where $(f_1 + f_2)(x) = f_1(x) + f_2(x)$.

Example. Define $M_{n,m}(F)$ to be the set of $n \times m$ F-valued matrices. This is an F-vector space, where the sum of matrices is computed elementwise.

Remark. The axioms of scalar multiplication imply that $\forall v \in V, 0_F v = 0_V$.

1.2. Subspaces

Definition. Let V be an F-vector space. The subset $U \subseteq V$ is a vector subspace of V, denoted $U \leq V$, if

(i) $0_V \in U$

(ii) $u_1, u_2 \in U \implies u_1 + u_2 \in U$

(iii) $(\lambda, u) \in F \times U \implies \lambda u \in U$

Conditions (ii) and (iii) are equivalent to

$$\forall \lambda_1, \lambda_2 \in F, \forall u_1, u_2 \in U, \lambda_1 u_1 + \lambda_2 u_2 \in U$$

This means that U is *stable* by addition and scalar multiplication.

Proposition. If V is an F-vector space, and $U \leq V$, then U is an F-vector space.

Example. Let $V = \mathbb{R}^{\mathbb{R}}$ be the space of functions $\mathbb{R} \to \mathbb{R}$. The set $C(\mathbb{R})$ of continuous real functions is a subspace of V. The set $\mathbb{P}$ of polynomials is a subspace of $C(\mathbb{R})$.

Example. Consider the subset of $\mathbb{R}^3$ such that $x_1 + x_2 + x_3 = t$ for some real t. This is a subspace for $t = 0$ only, since no other t values yield the origin as a member of the subset.

Proposition. Let V be an F-vector space. Let $U, W \leq V$. Then $U \cap W$ is a subspace of V.

Proof. First, note $0_V \in U, 0_V \in W \implies 0_V \in U \cap W$. Now, consider stability:

$$\lambda_1, \lambda_2 \in F, v_1, v_2 \in U \cap W \implies \lambda_1 v_1 + \lambda_2 v_2 \in U, \lambda_1 v_1 \lambda_2 v_2 \in W$$

Hence stability holds. □

1.3. Sum of subspaces

Remark. The union of two subspaces is not, in general, a subspace. For instance, consider $\mathbb{R}, i\mathbb{R} \subset \mathbb{C}$. Their union does not span the space; for example, $1 + i \notin \mathbb{R} \cup i\mathbb{R}$.

Definition. Let V be an F-vector space. Let $U, W \leq V$. The sum $U + W$ is defined to be the set

$$U + W = \{u + w : u \in U, w \in W\}$$

Proposition. $U + W$ is a subspace of V.

Proof. First, note $0_{U+W} = 0_U + 0_W = 0_V$. Then, for $\lambda_1, \lambda_2 \in F$, and $u \in U, w \in W$,

$$\lambda_1 u + \lambda_2 w = u' + w' \in U + W$$

since $u' \in U, w' \in W$. We can decompose a vector from $U+W$ into its U and W components. Adding these components independently (noting that V is abelian) yields the requirements of a subspace. □

Proposition. The sum $U + W$ is the smallest subspace of V that contains both U and W.

1.4. Quotients

Definition. Let V be an F-vector space. Let $U \leq V$. The quotient space V/U is the abelian group V/U equipped with the scalar multiplication function

$$F \times V/U \to V/U; \quad (\lambda, v + U) \mapsto \lambda v + U$$

Proposition. V/U is an F-vector space.

Proof. We must check that the multiplication operation is well-defined. Indeed, suppose $v_1 + U = v_2 + U$. Then,

$$v_1 - v_2 \in U \implies \lambda(v_1 - v_2) \in U \implies \lambda v_1 + U = \lambda v_2 + U \in V/U$$

□

1.5. Span

Definition. Let V be an F-vector space. Let $S \subset V$. We define the span of S, written $\langle S \rangle$, as the set of finite linear combinations of elements of S. In particular,

$$\langle S \rangle = \left\{ \sum_{s \in S} \lambda_s v_s : \lambda_s \in F, v_s \in S, \text{only finitely many nonzero } \lambda_s \right\}$$

By convention, we specify

$$\langle \varnothing \rangle = \{0\}$$

so that all spans are subspaces.

Remark. $\langle S \rangle$ is the smallest vector subspace of V containing S.

Example. Let $V = \mathbb{R}^3$, and

$$S = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \right\}, \begin{pmatrix} 3 \\ -2 \\ -4 \end{pmatrix}$$

Then we can check that

$$\langle S \rangle = \left\{ \begin{pmatrix} a \\ b \\ 2b \end{pmatrix} : (a, b) \in \mathbb{R} \right\}$$

Example. Let $V = \mathbb{R}^n$. We define

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where the 1 is in the ith position. Then $V = \left\langle (e_i)_{1 \le i \le n} \right\rangle$.

Example. Let X be a set, and $\mathbb{R}^X = \{f : X \to \mathbb{R}\}$. Then let $S_x : X \to \mathbb{R}$ be defined by

$$S_x(y) = \begin{cases} 1 & y = x \\ 0 & \text{otherwise} \end{cases}$$

Then, $\langle (S_x)_{x \in X} \rangle = \left\{ f \in \mathbb{R}^X : f \text{ has finite support} \right\}$, where the support of f is defined to be $\{x : f(x) \neq 0\}$.

1.6. Dimensionality

Definition. Let V be an F-vector space. Let $S \subset V$. We say that S spans V if $\langle S \rangle = V$. If S spans V, we say that S is a generating family of V.

Definition. Let V be an F-vector space. V is finite-dimensional if it is spanned by a finite set.

Example. Consider the set $V = \mathbb{P}[x]$ which is the set of polynomials on $\mathbb{R}$. Further, consider $V_n = \mathbb{P}_n[x]$ which is the subspace with degree less than or equal to n. Then V_n is spanned by $\{1, x, x^2, \dots, x^n\}$, so V_n is finite-dimensional. Conversely, V is infinite-dimensional; there is no finite set S such that $\langle S \rangle = V$.

1.7. Linear independence

Definition. We say that $v_1, \dots, v_n \in V$ are linearly independent if, for $\lambda_i \in F$,

$$\sum_{i=1}^{n} \lambda_i v_i = 0 \implies \forall i, \lambda_i = 0$$

Definition. Similarly, $v_1, \dots, v_n \in V$ are linearly dependent if

$$\exists \boldsymbol{\lambda} \in F^n, \sum_{i=1}^{n} \lambda_i v_i = 0, \exists i, \lambda_i \neq 0$$

Equivalently, one of the vectors can be written as a linear combination of the remaining ones.

Remark. If $(v_i)_{1 \leq i \leq n}$ are linearly independent, then

$$\forall i \in \{1, \dots, n\}, v_i \neq 0$$

1.8. Bases

Definition. $S \subset V$ is a basis of V if

(i) $\langle S \rangle = V$

(ii) S is a linearly independent set

So, a basis is a linearly independent (also known as *free*) generating family.

Example. Let $V = \mathbb{R}^n$. The *canonical basis* (e_i) is a basis since we can show that they are free and span V.

Example. Let $V = \mathbb{C}$, considered as a $\mathbb{C}$-vector space. Then $\{1\}$ is a basis. If V is a $\mathbb{R}$-vector space, $\{1, i\}$ is a basis.

Example. Consider again $\mathbb{P}[x]$. Then $S = \{x^n : n \in \mathbb{N}\}$ is a basis of $\mathbb{P}$.

Lemma. Let V be an F-vector space. Then, $(v_1, \dots, v_n)$ is a basis of V if and only if any vector $v \in V$ has a unique decomposition

$$v = \sum_{i=1}^{n} \lambda_i v_i, \forall i, \lambda_i \in F$$

In the above definition, we call $(\lambda_1, \dots, \lambda_n)$ the *coordinates* of v in the basis $(v_1, \dots, v_n)$.

Proof. Suppose $(v_1, \dots, v_n)$ is a basis of V. Then $\forall v \in V$ there exists $\lambda_1, \dots, \lambda_n \in F$ such that

$$v = \sum_{i=1}^{n} \lambda_i v_i$$

So there exists a tuple of λ values. Suppose two such λ tuples exist. Then

$$v = \sum_{i=1}^{n} \lambda_i v_i = \sum_{i=1}^{n} \lambda_i' v_i \implies \sum_{i=1}^{n} (\lambda_i - \lambda_i') v_i = 0 \implies \lambda_i = \lambda_i'$$

The converse is left as an exercise. □

Lemma. If $\langle \{v_1, \dots, v_n\} \rangle = V$, then some subset of this set is a basis of V.

Proof. If $(v_1, \dots, v_n)$ are linearly independent, this is a basis. Otherwise, one of the vectors can be written as a linear combination of the others. So, up to reordering,

$$v_n \in \langle \{v_1, \dots, v_{n-1}\} \rangle = V$$

So we have removed a vector from this set and preserved the span. By induction, we will eventually reach a basis. □

1.9. Steinitz exchange lemma

Theorem. Let V be a finite dimensional F-vector space. Let $(v_1, \dots, v_m)$ be linearly independent, and $(w_1, \dots, w_n)$ which spans V. Then,

(i) $m \leq n$; and

(ii) up to reordering, $(v_1, \dots, v_m, w_{m+1}, \dots w_n)$ spans V.

Proof. Suppose that we have replaced $\ell \geq 0$ of the w_i.

$$\langle v_1, \dots, v_\ell, w_{\ell+1}, \dots w_n \rangle = V$$

If $m = \ell$, we are done. Otherwise, $\ell < m$. Then, $v_{\ell+1} \in V = \langle v_1, \dots, v_\ell, w_{\ell+1}, \dots w_n \rangle$ Hence $v_{\ell+1}$ can be expressed as a linear combination of the generating set. Since the $(v_i)_{1 \leq i \leq m}$ are linearly independent (free), one of the coefficients on the w_i are nonzero. In particular, up to reordering we can express $w_{\ell+1}$ as a linear combination of $v_1, \dots, v_{\ell+1}, w_{\ell+2}, \dots, w_n$. Inductively, we may replace m of the w terms with v terms. Since we have replaced m vectors, necessarily $m \leq n$. □

1.10. Consequences of Steinitz exchange lemma

Corollary. Let V be a finite-dimensional F-vector space. Then, any two bases of V have the same number of vectors. This number is called the dimension of V, $\dim_F V$.

Proof. Suppose the two bases are $(v_1, \dots, v_n)$ and $(w_1, \dots, w_m)$. Then, $(v_1, \dots, v_n)$ is free and $(w_1, \dots, w_m)$ is generating, so the Steinitz exchange lemma shows that $n \leq m$. Vice versa, $m \leq n$. Hence $m = n$. □

Corollary. Let V be an F-vector space with finite dimension n. Then,

(i) Any independent set of vectors has at most n elements, with equality if and only if it is a basis.

(ii) Any spanning set of vectors has at least n elements, with equality if and only if it is a basis.

Proof. Exercise. □

1.11. Dimensionality of sums

Proposition. Let V be an F-vector space. Let U, W be subspaces of V. If U, W are finite-dimensional, then so is $U + W$, with

$$\dim_F(U + W) = \dim_F U + \dim_F W - \dim_F(U \cap W)$$

Proof. Consider a basis $(v_1, \dots, v_n)$ of the intersection. Extend this basis to a basis

$$(v_1, \dots, v_n, u_1, \dots, u_m) \text{ of } U; \quad (v_1, \dots, v_n, w_1, \dots, w_k) \text{ of } W$$

Then, we will show that $(v_1, \dots, v_n, u_1, \dots, u_m, w_1, \dots, w_k)$ is a basis of $\dim_F(U + W)$, which will conclude the proof. Indeed, since any component of $U + W$ can be decomposed as a sum of some element of U and some element of W, we can add their decompositions together.

Now we must show that this new basis is free.

$$\sum_{i=1}^{n} \alpha_i v_i + \sum_{i=1}^{m} \beta_i u_i + \sum_{i=1}^{k} \gamma_i w_i = 0$$

$$\underbrace{\sum_{i=1}^{n} \alpha_i v_i + \sum_{i=1}^{m} \beta_i u_i}_{\in U} = \underbrace{\sum_{i=1}^{k} \gamma_i w_i}_{\in W}$$

$$\sum_{i=1}^{k} \gamma_i w_i \in U \cap W$$

$$\sum_{i=1}^{k} \gamma_i w_i = \sum_{i=1}^{n} \delta_i v_i$$

$$\sum_{i=1}^{n} (\alpha_i + \delta_i) v_i + \sum_{i=1}^{m} \beta_i u_i = 0$$

$$\beta_i = 0, \alpha_i = -\delta_i$$

$$\sum_{i=1}^{n} \alpha_i v_i + \sum_{i=1}^{k} \gamma_i w_i = 0$$

$$\alpha_i = 0, \gamma_i = 0$$

□

Proposition. If V is a finite-dimensional F-vector space, and $U \leq V$, then U and V/U are also finite-dimensional. In particular, $\dim_F V = \dim_F U + \dim_F(V/U)$.

Proof. Let $(u_1, \dots, u_\ell)$ be a basis of U. We extend this basis to a basis of V, giving

$$(u_1, \dots, u_\ell, w_{\ell+1}, \dots, w_n)$$

We claim that $(w_{\ell+1} + U, \dots, w_n + U)$ is a basis of the vector space V/U. □

Remark. If V is an F-vector space, and $U \leq V$, then we say U is a proper subspace if $U \neq V$. Then if U is proper, then $\dim_F U < \dim_F V$ and $\dim_F(V/U) > 0$ because $(V/U) \neq \varnothing$.

1.12. Direct sums

Definition. Let V be an F-vector space and U, W be subspaces of V. We say that $V = U \oplus W$, read as the direct sum of U and W, if $\forall v \in V, \exists! u \in U, \exists! w \in W, u + w = v$. We say that W is *a* direct complement of U in V; there is no uniqueness of such a complement.

Lemma. Let V be an F-vector space, and $U, W \leq V$. Then the following statements are equivalent.

(i) $V = U \oplus W$

(ii) $V = U + W$ and $U \cap W = \{0\}$

(iii) For any basis B_1 of U and B_2 of W, $B_1 \cup B_2$ is a basis of V

Proof. First, we show that (ii) implies (i). If $V = U + W$, then certainly $\forall v \in V, \exists u \in U, \exists w \in W, v = u + w$, so it suffices to show uniqueness. Note, $u_1 + w_1 = u_2 + w_2 \implies u_1 - u_2 = w_2 - w_1$. The left hand side is an element of U and the right hand side is an element of W, so they must be the zero vector; $u_1 = u_2, w_1 = w_2$.

Now, we show (i) implies (iii). Suppose B_1 is a basis of U and B_2 is a basis of W. Let $B = B_1 \cup B_2$. First, note that B is a generating family of $U + W$. Now we must show that B is free.

$$\underbrace{\sum_{u \in B_1} \lambda_u u}_{\in U} + \underbrace{\sum_{w \in B_2} \lambda_w w}_{\in W} = 0$$

Hence both sums must be zero. Since B_1, B_2 are bases, all λ are zero, so B is free and hence a basis.

Now it remains to show that (iii) implies (ii). We must show that $V = U+W$ and $U \cap W = \{0\}$. Now, suppose $v \in V$. Then, $v = \sum_{u \in B_1} \lambda_u u + \sum w \in B_2 \lambda_w w$. In particular, $V = U + W$, since the λ_u, λ_w are arbitrary. Now, let $v \in U \cap W$. Then

$$v = \sum_{u \in B_1} \lambda_u u = \sum_{w \in B_2} \lambda_w w \implies \lambda_u = \lambda_w = 0$$

□

Definition. Let V be an F-vector space, with subspaces $V_1, \dots, V_p \leq V$. Then

$$\sum_{i=1}^{p} V_i = \{v_1, \dots, v_\ell, v_i \in V_i, 1 \leq i \leq \ell\}$$

We say the sum is direct, written

$$\bigoplus_{i=1}^{p} V_i$$

if the decomposition is unique. Equivalently,

$$V = \bigoplus_{i=1}^{p} V_i \iff \exists! v_1 \in V_1, \dots, v_n \in V_n, v = \sum_{i=1}^{n} v_i$$

Lemma. The following are equivalent:

(i) $\sum_{i=1}^{p} V_i = \bigoplus_{i=1}^{p} V_i$

(ii) $\forall 1 \leq i \leq l, V_i \cap \left(\sum_{j \neq i} V_j\right) = \{0\}$

(iii) For any basis B_i of V_i, $B = \bigcup_{i=1}^{n} B_i$ is a basis of $\sum_{i=1}^{n} V_i$.

Proof. Exercise. □

2. Linear maps

2.1. Linear maps

Definition. If V, W are F-vector spaces, a map $\alpha : V \to W$ is *linear* if

$$\forall \lambda_1, \lambda_2 \in F, \forall v_1, v_2 \in V, \alpha(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 \alpha(v_1) + \lambda_2 \alpha(v_2)$$

Example. Let M be a matrix with n rows and m columns. Then the map $\alpha : \mathbb{R}^m \to \mathbb{R}^n$ defined by $x \mapsto Mx$ is a linear map.

Example. Let $\alpha : \mathcal{C}([0,1], \mathbb{R}) \to \mathcal{C}([0,1], \mathbb{R})$ defined by $f \mapsto a(f)(x) = \int_0^x f(t)\,\mathrm{d}t$. This is linear.

Example. Let $x \in [a, b]$. Then $\alpha : \mathcal{C}([a,b], \mathbb{R}) \to \mathbb{R}$ defined by $f \mapsto f(x)$ is a linear map.

Remark. Let U, V, W be F-vector spaces. Then,

(i) The identity function $i_V : V \to V$ defined by $x \mapsto x$ is linear.

(ii) If $\alpha : U \to V$ and $\beta : V \to W$ are linear, then $\beta \circ \alpha$ is linear.

Lemma. Let V, W be F-vector spaces. Let B be a basis for V. If $\alpha_0 : B \to V$ is *any* map (not necessarily linear), then there exists a unique linear map $\alpha : V \to W$ extending α_0: $\forall v \in B, \alpha_0(v) = \alpha(v)$.

Proof. Let $v \in V$. Then, given $B = (v_1, \dots, v_n)$.

$$v = \sum_{i=1}^{n} \lambda_i v_i$$

By linearity,

$$\alpha(v) = \alpha\left(\sum_{i=1}^{n} \lambda_i v_i\right) = \sum_{i=1}^{n} \alpha(\lambda_i v_i) = \sum_{i=1}^{n} \alpha_0(\lambda_i v_i)$$

□

Remark. This lemma is also true in infinite-dimensional vector spaces. Often, to define a linear map, we instead define its action on the basis vectors, and then we 'extend by linearity' to construct the entire map.

Remark. If $\alpha_1, \alpha_2 : V \to W$ are linear maps, then if they agree on any basis of V then they are equal.

2.2. Isomorphism

Definition. Let V, W be F-vector spaces. A map $\alpha : V \to W$ is an *isomorphism* if and only if

(i) α is linear

(ii) α is bijective

If such an α exists, we say that V and W are isomorphic, written $V \simeq W$.

Remark. If α in the above definition is an isomorphism, then $\alpha^{-1} : W \to V$ is linear. Indeed, if $w_1, w_2 \in W$ with $w_1 = \alpha(v_1)$ and $w_2 = \alpha(v_2)$,

$$\alpha^{-1}(w_1 + w_2) = \alpha^{-1}(\alpha(v_1) + \alpha(v_2)) = \alpha^{-1}\alpha(v_1 + v_2) = v_1 + v_2 = \alpha^{-1}(w_1) + \alpha^{-1}(w_2)$$

Similarly, for $\lambda \in F, w \in W$,

$$\alpha^{-1}(\lambda w) = \lambda \alpha^{-1}(w)$$

Lemma. Isomorphism is an equivalence relation on the class of all vector spaces over F.

Proof. (i) $i_V : V \to V$ is an isomorphism

(ii) If $\alpha : V \to W$ is an isomorphism, $\alpha^{-1} : W \to V$ is an isomorphism.

(iii) If $\beta : U \to V, \alpha : V \to W$ are isomorphisms, then $\alpha \circ \beta : U \to W$ is an isomorphism.

The proofs of each part are left as an exercise. □

Theorem. If V is an F-vector space of dimension n, then $V \simeq F^n$.

Proof. Let $B = (v_1, \dots, v_n)$ be a basis for V. Then, consider $\alpha : V \to F^n$ defined by

$$v = \sum_{i=1}^{n} \lambda_i v_i \mapsto \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}$$

We claim that this is an isomorphism. This is left as an exercise. □

Remark. Choosing a basis for V is analogous to choosing an isomorphism from V to F^n.

Theorem. Let V, W be F-vector spaces with finite dimensions n, m. Then,

$$V \simeq W \iff n = m$$

Proof. If $\dim V = \dim W = n$, then there exist isomorphisms from both V and W to F^n. By transitivity, therefore, there exists an isomorphism between V and W.

Conversely, if $V \simeq W$ then let $\alpha : V \to W$ be an isomorphism. Let B be a basis of V, then we claim that $\alpha(B)$ is a basis of W. Indeed, $\alpha(B)$ spans W from the surjectivity of α, and $\alpha(B)$ is free due to injectivity. □

2.3. Kernel and image

Definition. Let V, W be F-vector spaces. Let $\alpha : V \to W$ be a linear map. We define the kernel and image as follows.

$$N(\alpha) = \ker \alpha = \{v \in V : \alpha(v) = 0\}$$

$$\mathrm{Im}(\alpha) = \{w \in W : \exists v \in V, w = \alpha(v)\}$$

Lemma. $\ker \alpha$ is a subspace of V, and $\mathrm{Im}\, \alpha$ is a subspace of W.

Proof. Let $\lambda_1, \lambda_2 \in F$ and $v_1, v_2 \in \ker \alpha$. Then

$$\alpha(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 \alpha(v_1) + \lambda_2 \alpha(v_2) = 0$$

Hence $\lambda_1 v_1 + \lambda_2 v_2 \in \ker \alpha$.

Now, let $\lambda_1, \lambda_2 \in F$, $v_1, v_2 \in V$, and $w_1 = \alpha(v_1), w_2 = \alpha(v_2)$. Then

$$\lambda_1 w_1 + \lambda_2 w_2 = \lambda_1 \alpha(v_1) + \lambda_2 \alpha(v_2) = \alpha(\lambda_1 v_1 + \lambda_2 v_2) \in \mathrm{Im}\, \alpha$$

□

Remark. $\alpha : V \to W$ is injective if and only if $\ker \alpha = \{0\}$. Further, $\alpha : V \to W$ is surjective if and only if $\mathrm{Im}\, \alpha = W$.

Theorem. Let V, W be F-vector spaces. Let $\alpha : V \to W$ be a linear map. Then $\overline{\alpha} : V/\ker \alpha \to \mathrm{Im}\, \alpha$ defined by

$$\overline{\alpha}(v + \ker \alpha) = \alpha(v)$$

is an isomorphism. *This is the isomorphism theorem from IA Groups.*

Proof. First, note that $\overline{\alpha}$ is well defined. Suppose $v + \ker \alpha = v' + \ker \alpha$. Then $v - v' \in \ker \alpha$, hence

$$\alpha(v - v') = 0 \implies \alpha(v) - \alpha(v') = 0$$

so $\overline{\alpha}$ is indeed well defined.

Now, we show $\overline{\alpha}$ is injective.

$$\overline{\alpha}(v + \ker \alpha) = 0 \implies \alpha(v) = 0 \implies v \in \ker \alpha$$

Hence, $v + \ker \alpha = 0 + \ker \alpha$.

Further, $\overline{\alpha}$ is surjective. This follows from the definition the image. □

2.4. Rank and nullity

Definition. The *rank* of α is

$$r(\alpha) = \dim \operatorname{Im} \alpha$$

The *nullity* of α is

$$n(\alpha) = \dim \ker \alpha$$

Theorem (Rank-nullity theorem)**.** Let U, V be F-vector spaces such that the dimension of U is finite. Let $\alpha : U \to V$ be a linear map. Then,

$$\dim U = r(\alpha) + n(\alpha)$$

Proof. We have proven that $U/\ker \alpha \simeq \operatorname{Im} \alpha$. Hence, the dimensions on the left and right match: $\dim(U/\ker \alpha) = \dim \operatorname{Im} \alpha$.

$$\dim U - \dim \ker \alpha = \dim \operatorname{Im} \alpha$$

and the result follows. □

Lemma (Characterisation of isomorphisms)**.** Let V, W be F-vector spaces with equal, finite dimension. Let $\alpha : V \to W$ be a linear map. Then, the following are equivalent.

(i) α is injective.

(ii) α is surjective.

(iii) α is an isomorphism.

Proof. Clearly, (iii) follows from (i) and (ii) and vice versa. The rest of the proof is left as an exercise, which follows from the rank-nullity theorem. □

2.5. Space of linear maps

Let V and W be F-vector spaces. Consider the space of linear maps from V to W. Then $L(V, W) = \{\alpha : V \to W \text{ linear}\}$.

Proposition. $L(V, W)$ is an F-vector space under the operation

$$(\alpha_1 + \alpha_2)(v) = \alpha_1(v) + \alpha_2(v);$$

$$(\lambda\alpha)(v) = \lambda(\alpha(v))$$

Further, if V and W are finite-dimensional, then so is $L(V, W)$ with

$$\dim_F L(V, W) = \dim_F V \dim_F W$$

Proof. Proving that $L(V, W)$ is a vector space is left as an exercise. The dimensionality part is proven later. □

2.6. Matrices

Definition. An $m \times n$ matrix over F is an array of m rows and n columns, with entries in F.

We write $M_{m\times n}(F)$ for the set of $m \times n$ matrices over F.

Proposition. $M_{m\times n}(F)$ is an F-vector space under

$$((a_{ij}) + (b_{ij})) = (a_{ij} + b_{ij});$$

$$\lambda(a_{ij}) = (\lambda a_{ij})$$

Proposition. $\dim_F M_{m,n}(F) = mn$.

Proof. Consider the basis defined by, the 'elementary matrix' for all i, j:

$$e_{pq} = \delta_{ip}\delta_{jq}$$

Then (e_{ij}) is a basis of $M_{m\times n}(F)$, since it spans $M_{m\times n}(F)$ and we can show that it is free. □

2.7. Linear maps as matrices

Consider bases B of V and C of W:

$$B = (v_1, \dots, v_n); C = (w_1, \dots, w_n)$$

Then let $v \in V$. We have

$$v = \sum_{j=1}^{n} \lambda_j v_j \equiv [v]_B = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \in F^n$$

where the vector given is the coordinates in basis B. We can equivalently find $[w]_C$, the coordinates of w in basis C. We can now define a matrix of some linear map α in the B, C basis.

Definition.

$$[\alpha]_{B,C} = \left([\alpha(v_1)]_C, \dots, [\alpha(v_n)]_C\right) \in M_{m\times n}(F)$$

Note that if $[\alpha]_{BC} = (a_{ij})$, then by definition

$$\alpha(v_j) = \sum_{i=1}^{n} a_{ij} w_i$$

Lemma. For all $v \in V$,

$$[\alpha(v)]_C = [\alpha]_{BC} \cdot [v]_B$$

Proof. We have

$$v = \sum_{i=1}^{n} \lambda_j v_j$$

Hence

$$\alpha\left(\sum_{i=1}^{n} \lambda_j v_j\right) = \sum_{j=1}^{n} \lambda_j \alpha(v_j) = \sum_{j=1}^{n} \lambda_i \sum_{i=1}^{m} a_{ij} w_i = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} \lambda_j\right) w_i$$

□

Lemma. Let $\beta : U \to V$ and $\alpha : V \to W$ be linear maps. Then, if A, B, C are bases of U, V, W respectively, then

$$[\alpha \circ \beta]_{A,C} = [\alpha]_{B,C} \cdot [\beta]_{A,B}$$

Proof. Consider $u \in A$. Then

$$(\alpha \circ \beta)(u) = \alpha(\beta(u))$$

giving

$$\alpha\left(\sum_{j} b_{jp} v_i\right) = \sum_{j} b_{jp} \alpha(v_j) = \sum_{j} b_{jp} \sum_{i} a_{ij} w_i = \sum_{i} (\sum_{j} a_{ij} b_{jp}) w_i$$

where $a_{ij} p_{jp}$ is the (i, j) element of AB by the definition of the product of matrices. □

Proposition. If V, W are F-vector spaces, and $\dim V = n$, $\dim W = m$, then

$$L(V, W) \simeq M_{m \times n}(F)$$

which implies the dimensionality of $L(V, W)$ in F is $m \times n$.

Proof. Consider two bases B, C of V, W. We claim that

$$\theta : L(V, W) \to M_{m \times n}(F)$$

defined by $\theta(\alpha) = [\alpha]_{B,C}$. is an isomorphism. First, note that θ is linear. Then, θ is surjective; consider any matrix $A = (a_{ij})$ and consider $\alpha : v_j \mapsto \sum_{i=1}^{m} a_{ij} w_i$. Then this is certainly a linear map which extends uniquely by linearity to A, giving $[\alpha]_{B,C} = (a_{ij}) = A$. Now, θ is injective since $[\alpha]_{B,C} = 0 \implies \alpha = 0$. □

Remark. If B, C are bases of V, W respectively, and $\varepsilon_B : V \to F^n$ is defined by $v \mapsto [v]_B$, and analogously for ε_C, then

$$[\alpha]_{B,C} \circ \varepsilon_B = \varepsilon_C \circ \alpha$$

so the operations commute.

Example. Let $\alpha : V \to W$ be a linear map and $Y \leq V$, where V, W are finite-dimensional. Then let $\alpha(Y) = Z \leq W$. Consider a basis B of V, such that $B' = (v_1, \dots, v_k)$ is a basis of Y completed by $B'' = (v_{k+1}, \dots, v_n)$ into $B = B' \cup B''$. Then let C be a basis of W, such that $C' = (w_1, \dots, w_\ell)$ is a basis of Z completed by $C'' = (w_{\ell+1}, \dots, w_m)$ into $C = C' \cup C''$. Then

$$[\alpha]_{B,C} = \begin{pmatrix} \alpha(v_1) & \dots & \alpha(v_k) & \alpha(v_{k+1}) & \dots & \alpha(v_n) \end{pmatrix}$$

For $1 \leq i \leq k$, $\alpha(v_i) \in Z$ since $v_i \in Y, \alpha(Y) = Z$. So the matrix has an upper-left $\ell \times k$ block A which is $\alpha : Y \to Z$ on the basis B', C'. We can show further that α induces a map $\overline{\alpha} : V/Y \to W/Z$ by $v + Y \mapsto \alpha(v) + Z$. This is well-defined; $v_1 + Y = v_2 + Y$ implies $v_1 - v_2 \in Y$ hence $\alpha(v_1 - v_2) \in Z$ as required. The bottom-right block is $[\overline{\alpha}]_{B'',C''}$.

2.8. Change of basis

Suppose we have two bases $B = \{v_1, \dots, v_n\}, B' = \{v'_1, \dots, v'_n\}$ of V and corresponding C, C' for W. If we have a linear map $[\alpha]_{B,C}$, we are interested in finding the components of this linear map in another basis, that is,

$$[\alpha]_{B,C} \mapsto [\alpha]_{B',C'}$$

Definition. The *change of basis* matrix P from B' to B is

$$P = \begin{pmatrix} [v'_1]_B & \cdots & [v'_n]_B \end{pmatrix}$$

which is the identity map in B', written

$$P = [I]_{B',B}$$

Lemma. For a vector v,

$$[v]_B = P[v]_{B'}$$

Proof. We have

$$[\alpha(v)]_C = [\alpha]_{B,C} \cdot [v]_C$$

Since $P = [I]_{B',B}$,

$$[I(v)]_B = [I]_{B',B} \cdot [v]_{B'} \implies [v]_B = P[v]_{B'}$$

as required. □

Remark. P is an invertible $n \times n$ square matrix. In particular,

$$P^{-1} = [I]_{B,B'}$$

Indeed,

$$I_n = [I \cdot I]_{B,B} = [I]_{B',B} \cdot [I]_{B',B}$$

where I_n is the $n \times n$ identity matrix.

Proposition. If α is a linear map from V to W, and $P = [I]_{B',B}, Q = [I]_{C',C}$, we have

$$A' = [\alpha]_{B',C'} = [I]_{C,C'}[\alpha]_{B,C}[I]_{B,'B} = Q^{-1}AP$$

where $A = [\alpha]_{B,C}, A' = [\alpha]_{B',C'}$.

Proof.

$$\begin{aligned} [\alpha(v)]_C &= Q[\alpha(v)]_{C'} \\ &= Q[\alpha]_{B',C'}[v]_{B'} \\ [\alpha(v)]_C &= [\alpha]_{B,C}[v]_B \\ &= AP[v]_{B'} \\ \therefore \forall v,\ QA[v]_{B'} &= AP[v]_{B'} \\ \therefore QA &= AP \end{aligned}$$

as required. □

2.9. Equivalent matrices

Definition. Matrices A, A' are called *equivalent* if

$$A' = Q^{-1}AP$$

for some invertible $m \times m, n \times n$ matrices Q, P.

Remark. This defines an equivalence relation on $M_{m,n}(F)$.

- $A = I_m^{-1}AI_n$;
- $A' = Q^{-1}AP \implies A = QA'P^{-1}$;
- $A' = Q^{-1}AP, A'' = (Q')^{-1}A'P' \implies A'' = (QQ')^{-1}A(PP')$.

Proposition. Let $\alpha : V \to W$ be a linear map. Then there exists a basis B of V and a basis C of W such that

$$[\alpha]_{B,C} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

so the components of the matrix are exactly the identity matrix of size r in the top-left corner, and zeroes everywhere else.

Proof. We first fix $r \in \mathbb{N}$ such that $\dim \ker \alpha = n - r$. Then we will construct a basis $\{v_{r+1}, \dots, v_n\}$ of the kernel. We extend this to a basis of the entirety of V, that is, $\{v_1, \dots, v_n\}$. Then, we want to show that

$$\{\alpha(v_1), \dots, \alpha(v_r)\}$$

is a basis of $\operatorname{Im}\alpha$. Indeed, it is a generating family:

$$\begin{aligned} v &= \sum_{i=1}^{n} \lambda_i v_i \\ \alpha(v) &= \sum_{i=1}^{n} \lambda_i \alpha(v_i) \\ &= \sum_{i=1}^{r} \lambda_i \alpha(v_i) \end{aligned}$$

Then if $y \in \operatorname{Im}\alpha$, there exists v such that $\alpha(v) = y$. Further, it is a free family:

$$\begin{gathered} \sum_{i=1}^{r} \lambda_i \alpha(v_i) = 0 \\ \alpha\left(\sum_{i=1}^{r} \lambda_i v_i\right) = 0 \\ \sum_{i=1}^{r} \lambda_i v_i \in \ker\alpha \\ \sum_{i=1}^{r} \lambda_i v_i = \sum_{i=r+1}^{n} \lambda_i v_i \\ \sum_{i=1}^{r} \lambda_i v_i - \sum_{i=r+1}^{n} \lambda_i v_i = 0 \end{gathered}$$

But since $\{v_1, \dots, v_n\}$ is a basis, $\lambda_i = 0$ for all i. Hence $\{\alpha(v_i)\}$ is a basis of $\operatorname{Im}\alpha$. Now, we wish to extend this basis to the whole of W to form

$$\{\alpha(v_1), \dots, \alpha(v_r), w_{r+1}, \dots, w_n\}$$

Now,

$$\begin{aligned} [\alpha]_{BC} &= \begin{pmatrix} \alpha(v_1) & \cdots & \alpha(v_r) & \alpha(v_{r+1}) & \cdots & \alpha(v_n) \end{pmatrix} \\ &= \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

□

Remark. This also proves the rank-nullity theorem:

$$\operatorname{rank}\alpha + \operatorname{null}\alpha = n$$

Corollary. Any $m \times n$ matrix A is equivalent to a matrix of the form

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

where $r = \operatorname{rank} A$.

2.10. Column rank and row rank

Definition. Let $A \in M_{m,n}(F)$. Then, the *column rank* of A, here denoted $r_c(A)$, is the dimension of the subspace of F^n spanned by the column vectors.

$$r_c(A) = \dim \operatorname{span}\{c_1, \dots, c_n\}$$

Remark. If α is a linear map, represented in bases B, C by the matrix A, then

$$\operatorname{rank} \alpha = r_c(A)$$

Proposition. Two matrices are equivalent if they have the same column rank:

$$r_c(A) = r_c(A')$$

Proof. If the matrices are equivalent, then $A = [\alpha]_{BC}, A' = [\alpha]_{B',C'}$. Then

$$r_c(A) = r_c(\alpha) = r_c(A')$$

Conversely, if $r_c(A) = r_c(A') = r$, then A, A' are equivalent to

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

By transitivity, A, A' are equivalent. □

Theorem. Column rank $r_c(A)$ and row rank $r_c(A^\intercal)$ are equivalent.

Proof. Let $r = r_C(A)$. Then,

$$Q^{-1}AP = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}_{m \times n}$$

Then, consider

$$P^\intercal A^\intercal \left(Q^{-1}\right)^\intercal = (Q^{-1}AP)^\intercal = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}^\intercal_{m \times n} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}_{n \times m}$$

Note that we can swap the transpose and inverse on Q because

$$\begin{aligned}(AB)^\intercal &= B^\intercal A^\intercal \\ \left(QQ^{-1}\right)^\intercal &= Q^\intercal\left(Q^{-1}\right)^\intercal \\ I &= Q^\intercal\left(Q^{-1}\right)^\intercal \\ \left(Q^\intercal\right)^{-1} &= \left(Q^{-1}\right)^\intercal\end{aligned}$$

Then $r_c(A) = \operatorname{rank}(A) = \operatorname{rank}(A^\intercal) = r_c(A^\intercal)$. □

So we can drop the concepts of column and row rank, and just talk about rank as a whole.

2.11. Conjugation and similarity

Consider the following special case of changing basis. If $\alpha : V \to V$ is linear, α is called an *endomorphism*. If $B = C, B' = C'$ then the special case of the change of basis formula is

$$[\alpha]_{B',B'} = P^{-1}[\alpha]_{B,B}P$$

Then, we say square matrices A, A' are *similar* or *conjugate* if there exists P such that $A' = P^{-1}AP$.

2.12. Elementary operations

Definition. An *elementary column operation* is

(i) swap columns i, j

(ii) replace column i by λ multiplied by the column

(iii) add λ multiplied by column i to column j

We define analogously the elementary row operations. Note that these elementary operations are invertible (for $\lambda \neq 0$). These operations can be realised through the action of elementary matrices. For instance, the column swap operation can be realised using

$$T_{ij} = \begin{pmatrix} I_n & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & I_m \end{pmatrix}; \quad A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & I_k & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

To multiply a column by λ,

$$n_{i,\lambda} = \begin{pmatrix} I_n & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & I_m \end{pmatrix}$$

To add a multiple of a column,

$$c_{ij,\lambda} = I + \lambda E_{ij}$$

where E_{ij} is the matrix defined by elements $(e_{ij})_{pq} = \delta_{ip}\delta_{jq}$. An elementary column (or row) operation can be performed by multiplying A by the corresponding elementary matrix from the right (on the left for row operations). This will essentially provide a constructive proof that any $n \times n$ matrix is equivalent to

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

We will start with a matrix A. If all entries are zero, we are done. So we will pick $a_{ij} = \lambda \neq 0$, and swap rows $i, 1$ and columns $j, 0$. This ensures that $a_{11} = \lambda \neq 0$. Now we multiply column 1 by $\frac{1}{\lambda}$. Finally, we can clear out row 1 and column 1 by subtracting multiples of the first row or column. Then we can perform similar operations on the $(n-1) \times (n-1)$ matrix in the bottom right block and inductively finish this process.

2.13. Gauss' pivot algorithm

If only row operations are used, we can reach the 'row echelon' form of the matrix, a specific case of an upper triangular matrix. On each row, there are a number of zeroes until there is a one, called the pivot. First, we assume that $a_{ij} \neq 0$. We swap rows $i, 1$. Then divide the first row by $\lambda = a_{i1}$ to get a one in the top left. We can use this one to clear the rest of the first column. Then, we can repeat on the next column, and iterate. This is a technique for solving a linear system of equations.

2.14. Representation of square invertible matrices

Lemma. If A is an $n \times n$ square invertible matrix, then we can obtain I_n using only row elementary operations, or only column elementary operations.

Proof. We show an algorithm that constructs this I_n. This is exactly going to invert the matrix, since the resultant operations can be combined to get the inverse matrix. We will show here the proof for column operations. We argue by induction on the number of rows. Suppose we can make the form

$$\begin{pmatrix} I_k & 0 \\ A & B \end{pmatrix}$$

We want to obtain the same structure with $k+1$ rows. We claim that there exists $j > k$ such that $a_{k+1,j} \neq 0$. Indeed, otherwise we can show that the vector

$$\begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \delta_{k+1,i}$$

is not in the span of the column vectors of A. This contradicts the invertibility of the matrix. Now, we will swap columns $k+1$, j and divide this column by λ. We can now use this 1 to clear the rest of the $k+1$ row.

Inductively, we have found $AE_1 \dots E_n = I_n$ where E_n are elementary. Thus, we can find A^{-1}. □

Proposition. Any invertible square matrix is a product of elementary matrices.

The proof is exactly the proof of the lemma above.

3. Dual spaces

3.1. Dual spaces

Definition. Let V be an F-vector space. Then $V^\star$ is the *dual* of V, defined by

$$V^\star = L(V, F) = \{\alpha : V \to F\}$$

where the α are linear. If $\alpha : V \to F$ is linear, then we say α is a linear form. So the dual of V is the set of linear forms on V.

Example. For instance, the trace $\operatorname{tr} : M_{n,n}(F) \to F$ is a linear form on $M_{n,n}(F)$.

Example. Consider functions $[0,1] \to \mathbb{R}$. We can define $T_f : \mathcal{C}^\infty([0,1], \mathbb{R}) \to \mathbb{R}$ such that $\phi \mapsto \int_0^1 f(x)\phi(x)\,\mathrm{d}x$. Then T_f is a linear form on $\mathcal{C}^\infty([0,1], \mathbb{R})$. We can then reconstruct f given T_f. This mathematical formulation is called distribution.

Lemma. Let V be an F-vector space with a finite basis $B = \{e_1, \dots, e_n\}$. Then there exists a basis $B^\star$ for $V^\star$ given by

$$B^\star = \{\varepsilon_1, \dots, \varepsilon_n\}; \quad \varepsilon_j\left(\sum_{i=1}^n a_i e_i\right) = a_j$$

We call $B^\star$ the *dual basis* for B.

Proof. We know

$$\varepsilon_j\left(\sum_{i=1}^n a_i e_i\right) = a_j$$

Equivalently,

$$\varepsilon_j(e_i) = \delta_{ij}$$

First, we will show that the set of linear forms as defined is free. For all i,

$$\begin{aligned} \sum_{j=1}^n \lambda_j \varepsilon_j &= 0 \\ \therefore \left(\sum_{j=1}^n \lambda_j \varepsilon_j\right) e_i &= 0 \\ \sum_{j=1}^n \lambda_j \varepsilon_j(e_i) &= 0 \\ \lambda_i &= 0 \end{aligned}$$

Now we show that the set spans $V^\star$. Suppose $\alpha \in V^\star$, $x \in V$.

$$\begin{aligned}\alpha(x) &= \alpha\left(\sum_{j=1}^{n} \lambda_j e_j\right) \\ &= \sum_{i=1}^{n} \lambda_j \alpha(e_j)\end{aligned}$$

Conversely, we can write

$$\sum_{i=1}^{n} \alpha(e_j)\varepsilon(j) \in V^\star$$

Thus,

$$\begin{aligned}\left(\sum_{i=1}^{n} \alpha(e_j)\varepsilon_j\right)(x) &= \sum_{j=1}^{n} \alpha(e_j)\varepsilon_j\left(\sum_{k=1}^{n} \lambda_k e_k\right) \\ &= \sum_{j=1}^{n} \alpha(e_j) \sum_{k=1}^{n} \lambda_k \varepsilon_j(e_k) \\ &= \sum_{j=1}^{n} \alpha(e_j) \sum_{k=1}^{n} \lambda_k \delta_{jk} \\ &= \sum_{j=1}^{n} \alpha(e_j)\lambda_j \\ &= \alpha(x)\end{aligned}$$

We have then shown that

$$\alpha = \sum_{j=1}^{n} \alpha(e_j)\varepsilon_j$$

as required. □

Corollary. If V is finite-dimensional, $V^\star$ has the same dimension.

Remark. It is sometimes convenient to think of $V^\star$ as the spaces of row vectors of length $\dim V$ over F. For instance, consider the basis $B = (e_1, \dots, e_n)$, so $x = \sum_{i=1}^{n} x_i e_i$. Then we can pick $(\varepsilon_1, \dots, \varepsilon_n)$ a basis of $V^\star$, so $\alpha = \sum_{i=1}^{n} \alpha_i \varepsilon_i$. Then

$$\alpha(x) = \sum_{i=1}^{n} \alpha_i \varepsilon_i(x) = \sum_{i=1}^{n} \alpha_i \varepsilon\left(\sum_{j=1}^{n} x_j e_j\right) = \sum_{i=1}^{n} \alpha_i x_i$$

This is exactly

$$\begin{pmatrix}\alpha_1 & \cdots & \alpha_n\end{pmatrix}\begin{pmatrix}x_1 \\ \vdots \\ x_n\end{pmatrix}$$

which essentially defines a scalar product between the two spaces.

3.2. Annihilators

Definition. Let $U \subseteq V$. Then the annihilator of U is

$$U^0 = \{\alpha \in V^\star : \forall u \in U, \alpha(u) = 0\}$$

Lemma. (i) $U^0 \leq V^\star$;

(ii) If $U \leq V$ and $\dim V < \infty$, then $\dim V = \dim U + \dim U^0$.

Proof. (i) First, note that $0 \in U^0$ since $\alpha(0) = 0$ by linearity. If $\alpha, \alpha' \in U^0$, then for all $u \in U$,

$$(\alpha + \alpha')(u) = \alpha(u) + \alpha'(u) = 0$$

Further, for all $\lambda \in F$,

$$(\lambda\alpha)(u) = \lambda\alpha(u) = 0$$

Hence $U^0 \leq V^\star$.

(ii) Let $(e_1, \dots, e_k)$ be a basis of U, completed into a basis $B = (e_1, \dots, e_k, e_{k+1}, \dots, e_n)$ of V. Let $(\varepsilon_1, \dots, \varepsilon_n)$ be the dual basis $B^\star$. We then will prove that

$$U^0 = \langle \varepsilon_{k+1}, \dots, \varepsilon_n \rangle$$

If $i > k$, then $\varepsilon_i(e_k) = \delta_{ik} = 0$. Hence $\varepsilon_i \in U^0$. Thus $\langle \varepsilon_{k+1}, \dots, \varepsilon_n \rangle \subset U^0$. Conversely, let $\alpha \in U^0$. Then $\alpha = \sum_{i=1}^{n} \alpha_i \varepsilon_i$. For $i \leq k$, $\alpha \in U^0$ hence $\alpha(e_i) = 0$. Hence,

$$\alpha = \sum_{i=k+1}^{n} \alpha_i \varepsilon_i$$

Thus

$$\alpha \in \langle \varepsilon_{k+1}, \dots, \varepsilon_n \rangle$$

as required.

□

3.3. Dual maps

Lemma. Let V, W be F-vector spaces. Let $\alpha \in L(V, W)$. Then there exists a unique $\alpha^\star \in L(W^\star, V^\star)$ such that

$$\varepsilon \mapsto \varepsilon \circ \alpha$$

called the dual map.

Proof. First, note $\varepsilon(\alpha) : V \to F$ is a linear map. Hence, $\varepsilon \circ \alpha \in V^\star$. Now we must show $\alpha^\star$ is linear.

$$\alpha^\star(\theta_1 + \theta_2) = (\theta_1 + \theta_2)(\alpha) = \theta_1 \circ \alpha + \theta_2 \circ \alpha = \alpha^\star(\theta_1) + \alpha^\star(\theta_2)$$

Similarly, we can show

$$\alpha^\star(\lambda\theta) = \lambda\alpha^\star(\theta)$$

as required. Hence $\alpha^\star \in L(W^\star, V^\star)$. □

Proposition. Let V, W be finite-dimensional F-vector spaces with bases B, C respectively. Then

$$[\alpha^\star]_{C^\star,B^\star} = [\alpha]_{B,C}^{\mathrm{T}}$$

Thus, we can think of the dual map as the *adjoint* of α.

Proof. This follows from the definition of the dual map. Let $B = (b_1, \dots, b_n)$, $C = (c_1, \dots, c_m)$, $B^\star = (\beta_1, \dots, \beta_n)$, $C^\star = (\gamma_1, \dots, \gamma_m)$. Let $[\alpha]_{B,C} = (a_{ij})$. Then, we compute

$$\begin{aligned}
\alpha^\star(\gamma_r)(b_s) &= \gamma_r \circ \alpha(b_s) \\
&= \gamma_r\left(\sum_t a_{ts} c_t\right) \\
&= \sum_t a_{ts}\gamma_r(c_t) \\
&= \sum_t a_{ts}\delta_{tr} \\
&= a_{rs}
\end{aligned}$$

We can conversely write $[\alpha^\star]_{C^\star,B^\star} = (m_{ij})$ and

$$\begin{aligned}
\alpha^\star(\gamma_r) &= \sum_{i=1}^{n} m_{ir}\beta_i \\
\alpha^\star(\gamma_r)(b_s) &= \sum_{i=1}^{n} m_{ir}\beta_i(b_s) \\
&= \sum_{i=1}^{n} m_{ir}\delta_{is} \\
&= m_{sr}
\end{aligned}$$

Thus,

$$a_{rs} = m_{sr}$$

as required. □

3.4. Properties of dual map

Let $\alpha \in L(V, W)$, and $\alpha^\star \in L(W^\star, V^\star)$. Let B and C be bases of V, W respectively, and $B^\star, C^\star$ be their duals. We have proven that

$$[\alpha]_{B,C} = [\alpha^\star]_{B,C}^\intercal$$

Lemma. Suppose that $E = (e_1, \dots, e_n)$ and $F = (f_1, \dots, f_n)$ are bases of V. Let $P = [I]_{F,E}$ be a change of basis matrix from F to E. The bases $E^\star = (\varepsilon_1, \dots, \varepsilon_n)$, $F^\star = (\eta_1, \dots, \eta_n)$ are the corresponding dual bases. Then, the change of basis matrix from $F^\star$ to $E^\star$ is

$$\left(P^{-1}\right)^\intercal$$

Proof. Consider

$$[I]_{F^\star,E^\star} = [I]_{E,F}^\intercal = \left([I]_{F,E}^{-1}\right)^\intercal = \left(P^{-1}\right)^\intercal$$

□

Lemma. Let V, W be F-vector spaces. Let $\alpha \in L(V, W)$. Let $\alpha^\star$ be the corresponding dual map. Then, denoting $N(\alpha)$ for the kernel of α,

(i) $N(\alpha^\star) = (\operatorname{Im}\alpha)^0$, so $\alpha^\star$ is injective if and only if α is surjective.

(ii) $\operatorname{Im}\alpha^\star \leq (N(\alpha))^0$, with equality if V, W are finite-dimensional. In this finite-dimensional case, $\alpha^\star$ is surjective if and only if α is injective.

Remark. In many applications, it is often simpler to understand the dual map $\alpha^\star$ than it is to understand α.

Proof. First, we prove (i). Let $\varepsilon \in W^\star$. Then, $\varepsilon \in N(\alpha^\star)$ means $\alpha^\star(\varepsilon) = 0$. Hence, $\alpha^\star(\varepsilon) = \varepsilon \circ \alpha = 0$ So for any $v \in V$, $\varepsilon(\alpha(v)) = 0$. Equivalently, ε is an element of the annihilator of $\operatorname{Im}\alpha$.

Now, we will show (ii). Let $\varepsilon \in \operatorname{Im}\alpha^\star$. Then $\alpha^\star(\phi) = \varepsilon$ for some $\phi \in W^\star$. Then, for all $u \in N(\alpha)$, $\varepsilon(u) = (\alpha^\star(\phi))(u) = \phi \circ \alpha(u) = \phi(\alpha(u)) = 0$. Certainly then $\varepsilon \in (N(\alpha))^0$. Then, $\operatorname{Im}\alpha^\star \leq (N(\alpha))^0$.

In the finite-dimensional case, we can compare the dimension of these two spaces.

$$\dim \operatorname{Im}\alpha^\star = r(\alpha^\star) = r\left([\alpha^\star]_{C^\star,B^\star}\right) = r\left([\alpha]_{B,C}^\intercal\right) = r\left([\alpha]_{B,C}\right) = r(\alpha) = \dim \operatorname{Im}\alpha$$

Due to the rank-nullity theorem, $\dim \operatorname{Im}\alpha^\star = \dim V - \dim N(\alpha) = \dim\left[(N(\alpha))^0\right]$. Hence,

$$\operatorname{Im}\alpha^\star \leq (N(\alpha))^0; \quad \dim \operatorname{Im}\alpha^\star = \dim(N(\alpha))^0$$

The dimensions are equal, and one is a subspace of the other, hence the spaces are equal. □

3.5. Double duals

Definition. Let V be an F-vector space. Let $V^\star$ be the dual of V. The *double dual* or bidual of V is

$$V^{\star\star} = L(V^\star, F) = (V^\star)^\star$$

Remark. In general, there is no obvious relation between V and $V^\star$. However, the following useful facts hold about V and $V^{\star\star}$.

(i) There is a *canonical embedding* from V to $V^{\star\star}$. In particular, there exists i in $L(V, V^{\star\star})$ which is injective.

(ii) There are examples of infinite-dimensional spaces where $V \simeq V^{\star\star}$. These are called reflexive spaces. Such spaces are investigated in the study of Banach spaces.

Theorem. V embeds into $V^{\star\star}$.

Proof. Choose a vector $v \in V$ and define the linear form $\hat{v} \in L(V^\star, F)$ such that

$$\hat{v}(\varepsilon) = \varepsilon(v)$$

So clearly $\hat{v}$ is linear. We want to show $\hat{v} \in V^{\star\star}$. If $\varepsilon \in V^\star, \varepsilon(v) \in F$. Further, $\lambda_1, \lambda_2 \in F$ and $\varepsilon_1, \varepsilon_2 \in V^\star$ give

$$\hat{v}(\lambda_1\varepsilon_1 + \lambda_2\varepsilon_2) = (\lambda_1\varepsilon_1 + \lambda_2\varepsilon_2)(v) = \lambda_1\varepsilon_1(v) + \lambda_2\varepsilon_2(v) = \lambda_1\hat{v}(\varepsilon_1) + \lambda_2\hat{v}(\varepsilon_2)$$

□

Theorem. If V is finite-dimensional, then $i : V \to V^{\star\star}$ given by $i(v) = \hat{v}$ is an isomorphism.

Proof. We will show i is linear. If $v_1, v_2 \in V, \lambda_1, \lambda_2 \in F$, then

$$i(\lambda_1 v_1 + \lambda_2 v_2)(\varepsilon) = \varepsilon(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1\varepsilon(v_1) + \lambda_2\varepsilon(v_2) = \lambda_1\hat{v}_1(\varepsilon) + \lambda_2\hat{v}_2(\varepsilon)$$

Now, we will show that i is injective for finite-dimensional V. Let $e \in V \setminus \{0\}$. We will show that $e \notin \ker i$. We extend e into a basis $(e, e_2, \dots, e_n)$ of V. Now, let $(\varepsilon, \varepsilon_2, \dots, \varepsilon_n)$ be the dual basis. Then $\hat{e}(\varepsilon) = \varepsilon(e) = 1$. In particular, $\hat{e} \neq 0$. Hence $\ker i = \{0\}$, so it is injective.

We now show that i is an isomorphism. We need to simply compute the dimension of the image under i. Certainly, $\dim V = \dim V^\star = \dim(V^\star)^\star = \dim V^{\star\star}$. Since i is injective, $\dim V = \dim V^{\star\star}$. So i is surjective as required. □

Lemma. Let V be a finite-dimensional F-vector space. Let $U \leq V$. Then,

$$\hat{U} = U^{00}$$

After identifying V and $V^{\star\star}$, we typically say

$$U = U^{00}$$

although this is is incorrect notation and not an equality.

Proof. We will show that $\hat{U} \leq U^{00}$. Indeed, let $u \in U$, then by definition

$$\forall \varepsilon \in U^0, \varepsilon(u) = 0 \implies \hat{u}(\varepsilon) = 0$$

Hence $\hat{u} \in U^{00}$ and so $\hat{U} \leq U^{00}$.

Now, we will compute dimension: $\dim U^{00} = \dim V - \dim U^0 = \dim U$. Since $\hat{U} \simeq U$, their dimensions are the same, so $U^{00} = \hat{U}$. □

Remark. Due to this identification of $V^{\star\star}$ and V, we can define

$$T \leq V^\star, T^0 = \{v \in V : \forall \theta \in T, \theta(v) = 0\}$$

Lemma. Let V be a finite-dimensional F-vector space. Let U_1, U_2 be subspaces of V. Then

(i) $(U_1 + U_2)^0 = U_1^0 \cap U_2^0$;

(ii) $(U_1 \cap U_2)^0 = U_1^0 + U_2^0$

Proof. Let $\theta \in V^\star$. Then $\theta \in (U_1 + U_2)^0 \iff \forall u_1 \in U_1, u_2 \in U_2, \theta(u_1 + u_2) = 0$. Hence $\theta(u) = 0$ for all $u \in U_1 \cup U_2$ by linearity. Hence $\theta \in U_1^0 \cap U_2^0$. Now, take the annihilator of (i) and $U^{00} = U$ to complete part (ii). □

4. Bilinear forms

4.1. Introduction

Definition. Let U, V be F-vector spaces. Then $\phi\colon U \times V \to F$ is a *bilinear form* if it is linear in both components. For example, ϕ at a fixed $u \in U$ is a linear form $V \to F$ and an element of $V^\star$.

Example. Consider the map $V \times V^\star \to F$ given by

$$(v, \theta) \mapsto \theta(v)$$

Example. The scalar product on $U = V = \mathbb{R}^n$ is given by

$$\psi(x, y) = \sum_{i=1}^{n} x_i y_i$$

Example. Let $U = V = C([0,1], \mathbb{R})$ and consider

$$\phi(f, g) = \int_0^1 f(t) g(t)\,\mathrm{d}t$$

Definition. If $B = (e_1, \dots, e_m)$ is a basis of U and $C = (f_1, \dots, f_n)$ is a basis of V, and $\phi\colon U \times V \to F$ is a bilinear form, then the matrix of the bilinear form in this basis is

$$[\phi]_{B,C} = \big(\phi(e_i, f_j)\big)_{1 \le i \le m, 1 \le j \le n}$$

Lemma. We can link ϕ with its matrix in a given basis as follows.

$$\phi(u, v) = [u]_B^{\mathrm{T}} [\phi]_{B,C} [v]_C$$

Proof. Let $u = \sum_{i=1}^m \lambda_i u_i$ and $v = \sum_{j=1}^n \mu_j v_j$. Then

$$\phi(u, v) = \phi\left(\sum_{i=1}^{m} \lambda_i u_i, \sum_{j=1}^{n} \mu_j v_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_i \mu_j \phi(u_i, v_j) = [u]_B^{\mathrm{T}} [\phi]_{B,C} [v]_C$$

□

Remark. Note that $[\phi]_{B,C}$ is the only matrix such that $\phi(u, v) = [u]_B^{\mathrm{T}} [\phi]_{B,C} [v]_C$.

Definition. Let $\phi\colon U \times V \to F$ be a bilinear form. Then ϕ induces two linear maps given by the partial application of a single parameter to the function.

$$\phi_L\colon U \to V^\star; \quad \phi_L(u)\colon V \to F; \quad v \mapsto \phi(u, v)$$

$$\phi_R\colon V \to U^\star; \quad \phi_R(v)\colon U \to F; \quad u \mapsto \phi(u, v)$$

In particular,

$$\phi_L(u)(v) = \phi(u, v) = \phi_R(v)(u)$$

Lemma. Let $B = (e_1, \dots, e_m)$ be a basis of U, and let $B^\star = (\varepsilon_1, \dots, \varepsilon_m)$ be its dual; and let $C = (f_1, \dots, f_n)$ be a basis of V, and let $C^\star = (\eta_1, \dots, \eta_n)$ be its dual. Let $A = [\phi]_{B,C}$. Then

$$[\phi_R]_{C,B^\star} = A; \quad [\phi_L]_{B,C^\star} = A^\intercal$$

Proof.

$$\phi_L(e_i)(f_j) = \phi(e_i, f_j) = A_{ij}$$

Since η_j is the dual of f_j,

$$\phi_L(e_i) = \sum_i A_{ij}\eta_j$$

Further,

$$\phi_R(f_j)(e_i) = \phi(e_i, f_j) = A_{ij}$$

and then similarly

$$\phi_R(f_j) = \sum_i A_{ij}\varepsilon_i$$

□

Definition. $\ker \phi_L$ is called the *left kernel* of ϕ. $\ker \phi_R$ is the *right kernel* of ϕ.

Definition. We say that ϕ is *non-degenerate* if $\ker \phi_L = \ker \phi_R = \{0\}$. Otherwise, ϕ is *degenerate*.

Theorem. Let B be a basis of U, and let C be a basis of V, where U, V are finite-dimensional. Let $\phi : U \times V \to F$ be a bilinear form. Let $A = [\phi]_{B,C}$. Then, ϕ is non-degenerate if and only if A is invertible.

Corollary. If ϕ is non-degenerate, then $\dim U = \dim V$.

Proof. Suppose ϕ is non-degenerate. Then $\ker \phi_L = \ker \phi_R = \{0\}$. This is equivalent to saying that $n(\phi_L) = n(\phi_R) = 0$. We can use the rank-nullity theorem to state that $r(A^\intercal) = \dim V$ and $r(A) = \dim V$. This is equivalent to saying that A is invertible. Note that this forces $\dim U = \dim V$. □

Remark. The canonical example of a non-degenerate bilinear form is the scalar product $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ represented by the identity matrix in the standard basis.

Corollary. If U and V are finite-dimensional with $\dim U = \dim V$, then choosing a non-degenerate bilinear form $\phi : U \times V \to F$ is equivalent to choosing an isomorphism $\phi_L : U \simeq V^\star$.

Definition. If $T \subset U$, then we define

$$T^\perp = \{v \in V : \forall t \in T, \phi(t, v) = 0\}$$

Further, if $S \subset V$, we define

$${}^\perp S = \{u \in U : \forall s \in S, \phi(u, s) = 0\}$$

These are called the *orthogonals* of T and S.

4.2. Change of basis for bilinear forms

Proposition. Let B, B' be bases of U and $P = [I]_{B',B}$, let C, C' be bases of V and $Q = [I]_{C',C}$, and finally let $\phi : U \times V \to F$ be a bilinear form. Then

$$[\phi]_{B',C'} = P^\intercal [\phi]_{B,C} Q$$

Proof. We have $\phi(u, v) = [u]_B^\intercal [\phi]_{B,C} [v]_C$. Changing coordinates, we have

$$\phi(u, v) = (P[u]_{B'})^\intercal [\phi]_{B,C} (Q[v]_{C'}) = [u]_{B'}^\intercal (P^\intercal [\phi]_{B,C} Q)[v]_{C'}$$

□

Lemma. The *rank* of a bilinear form ϕ, denoted $r(\phi)$ is the rank of any matrix representing ϕ. This quantity is well-defined.

Remark. $r(\phi) = r(\phi_R) = r(\phi_L)$, since $r(A) = r(A^\intercal)$.

Proof. For any invertible matrices P, Q, $r(P^\intercal A Q) = r(A)$. □

5. Trace and determinant

5.1. Trace

Definition. The *trace* of a square matrix $A \in M_{n,n}(F) \equiv M_n(F)$ is defined by

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}$$

The trace is a linear form.

Lemma. $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for any matrices $A, B \in M_n(F)$.

Proof. We have

$$\operatorname{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} = \operatorname{tr}(BA)$$

□

Corollary. Similar matrices have the same trace.

Proof.

$$\operatorname{tr}\left(P^{-1}AP\right) = \operatorname{tr}\left(AP^{-1}P\right) = \operatorname{tr} A$$

□

Definition. If $\alpha : V \to V$ is linear, we can define the trace of α as

$$\operatorname{tr}\alpha = \operatorname{tr}[\alpha]_B$$

for any basis B. This is well-defined by the corollary above.

Lemma. If $\alpha : V \to V$ is linear, $\alpha^\star : V^\star \to V^\star$ satisfies

$$\operatorname{tr}\alpha = \operatorname{tr}\alpha^\star$$

Proof.

$$\operatorname{tr}\alpha = \operatorname{tr}[\alpha]_B = \operatorname{tr}[\alpha]_B^{\mathrm{T}} = \operatorname{tr}[\alpha^\star]_{B^\star} = \operatorname{tr}\alpha^\star$$

□

5.2. Permutations and transpositions

Recall the following facts about permutations and transpositions. S_n is the group of permutations of the set $\{1, \dots, n\}$; the group of bijections $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$. A transposition $\tau_{k\ell} = (k, \ell)$ is defined by $k \mapsto \ell, \ell \mapsto k, x \mapsto x$ for $x \neq k, \ell$. Any permutation σ can be decomposed as a product of transpositions. This decomposition is not necessarily unique, but the parity of the number of transpositions is well-defined. We say that the signature of a permutation, denoted $\varepsilon : S_n \to \{-1, 1\}$, is 1 if the decomposition has even parity and -1 if it has odd parity. We can then show that ε is a homomorphism.

5.3. Determinant

Definition. Let $A \in M_n(F)$. We define

$$\det A = \sum_{\sigma \in S_n} \varepsilon(\sigma) A_{\sigma(1)1} \dots A_{\sigma(n)n}$$

Example. Let $n = 2$. Then,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \implies \det A = a_{11}a_{22} - a_{12}a_{21}$$

Lemma. If $A = (a_{ij})$ is an upper (or lower) triangular matrix (with zeroes on the diagonal), then $\det A = 0$.

Proof. Let $(a_{ij}) = 0$ for $i > j$. Then

$$\det A = \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{\sigma(1)1} \dots a_{\sigma(n)n}$$

For the summand to be nonzero, $\sigma(j) \leq j$ for all j. Thus,

$$\det A = a_{11} \dots a_{nn} = 0$$

□

Lemma. Let $A \in M_n(F)$. Then, $\det A = \det A^\intercal$.

Proof.

$$\begin{aligned} \det A &= \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{\sigma(1)1} \dots a_{\sigma(n)n} \\ &= \sum_{\sigma^{-1} \in S_n} \varepsilon(\sigma) a_{\sigma(1)1} \dots a_{\sigma(n)n} \\ &= \sum_{\sigma \in S_n} \varepsilon(\sigma^{-1}) a_{1\sigma(1)} \dots a_{n\sigma(n)} \\ &= \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{1\sigma(1)} \dots a_{n\sigma(n)} \\ &= \det A^\intercal \end{aligned}$$

□

5.4. Volume forms

Definition. A volume form d on F^n is a function $d : \underbrace{F^n \times \cdots \times F^n}_{n \text{ times}} \to F$ satisfying

(i) d is multilinear: for all $i \in \{1, \dots, n\}$ and for all $v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n \in F^n$, the map from F^n to F defined by

$$v \mapsto (v_1, \dots, v_{i-1}, v, v_{i+1}, \dots, v_n)$$

is linear. In other words, this map is an element of $(F^n)^\star$.

(ii) d is alternating: for $v_i = v_j$ for some $i \neq j$, $d = 0$.

So an alternating multilinear form is a volume form. We want to show that, up to multiplication by a scalar, the determinant is the only volume form.

Lemma. The map $(F^n)^n \to F$ defined by $(A^{(1)}, \dots, A^{(n)}) \mapsto \det A$ is a volume form. This map is the determinant of A, but thought of as acting on the column vectors of A.

Proof. We first show that this map is multilinear. Fix $\sigma \in S_n$, and consider $\prod_{i=1}^n a_{\sigma(i)i}$. This product contains exactly one term in each column of A. Thus, the map $(A^{(1)}, \dots, A^{(n)}) \mapsto \prod_{i=1}^n a_{\sigma(i)i}$ is multilinear. This then clearly implies that the determinant, a sum of such multilinear maps, is itself multilinear.

Now, we show that the determinant is alternating. Let $k \neq \ell$, and $A^{(k)} = A^{(\ell)}$. Let $\tau = (k\ell)$ be the transposition exchanging k and ℓ. Then, for all $i, j \in \{1, \dots, n\}$, $a_{ij} = a_{i\tau(j)}$. We can decompose permutations into two disjoint sets: $S_n = A_n \cup \tau A_n$, where A_n is the alternating group of order n. Now, note that $\prod_{i=1}^n a_{\sigma(i)i} + \prod_{i=1}^n a_{(\tau\circ\sigma)(i)i} = 0$. So the sum over all $\sigma \in A_n$ gives zero. So the determinant is alternating, and hence a volume form. □

Lemma. Let d be a volume form. Then, swapping two entries changes the sign.

Proof. Take the sum of these two results:

$$\begin{aligned} d(v_1, \dots, v_i, \dots, v_j, \dots, v_n) &+ d(v_1, \dots, v_j, \dots, v_i, \dots, v_n) \\ &= d(v_1, \dots, v_i, \dots, v_j, \dots, v_n) \\ &+ d(v_1, \dots, v_j, \dots, v_i, \dots, v_n) \\ &+ d(v_1, \dots, v_i, \dots, v_i, \dots, v_n) \\ &+ d(v_1, \dots, v_j, \dots, v_j, v_n) \\ &= 2d(v_1, \dots, v_i + v_j, \dots, v_i + v_j, \dots, v_n) \\ &= 0 \end{aligned}$$

as required. □

Corollary. If $\sigma \in S_n$ and d is a volume form, $d(v_{\sigma(1)}, \dots, v_{\sigma(n)}) = \varepsilon(\sigma) d(v_1, \dots, v_n)$.

Proof. We can decompose σ as a product of transpositions $\prod_{i=1}^{n_\sigma} e_i$. □

Theorem. Let d be a volume form on F^n. Let A be a matrix whose columns are $A^{(i)}$. Then

$$d(A^{(1)}, \dots, A^{(n)}) = \det A \cdot d(e_1, \dots, e_n)$$

So there is a unique volume form up to a constant multiple. We can then see that $\det A$ is the only volume form such that $d(e_1, \dots, e_n) = 1$.

Proof.

$$d(A^{(1)}, \dots, A^{(n)}) = d\left(\sum_{i=1}^{n} a_{i1} e_i, A^{(2)}, \dots, A^{(n)}\right)$$

Since d is multilinear,

$$d(A^{(1)}, \dots, A^{(n)}) = \sum_{i=1}^{n} a_{i1} d\big(e_i, A^{(2)}, \dots, A^{(n)}\big)$$

Inductively on all columns,

$$d(A^{(1)}, \dots, A^{(n)}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i1}a_{j2} d\big(e_i, e_j, A^{(3)}, \dots, A^{(n)}\big) = \cdots = \sum_{1 \le i_1, \le \cdots \le n} \prod_{k=1}^{n} a_{i_\ell k} d(e_{i_1}, \dots e_{i_n})$$

Since d is alternating, we know that for $d(e_{i_1}, \dots, e_{i_n})$ to be nonzero, the i_k must be different, so this corresponds to a permutation $\sigma \in S_n$.

$$d(A^{(1)}, \dots, A^{(n)}) = \sum_{\sigma \in S_n} \prod_{k=1}^{n} a_{\sigma(k)k} \varepsilon(\sigma) d(e_1, \dots, e_n)$$

which is exactly the determinant up to a constant multiple. □

5.5. Multiplicative property of determinant

Lemma. Let $A, B \in M_n(F)$. Then $\det(AB) = \det(A)\det(B)$.

Proof. Given A, we define the volume form $d_A : (F^n)^n \to F$ by

$$d_A(v_1, \dots, v_n) \mapsto \det(Av_1, \dots, Av_n)$$

$v_i \mapsto Av_i$ is linear, and the determinant is multilinear, so d_A is multilinear. If $i \neq j$ and $v_i = v_j$, then $\det\big(\dots, Av_i, \dots, Av_j, \dots\big) = 0$ so d_A is alternating. Hence d_A is a volume form. Hence there exists a constant C_A such that $d_A(v_1, \dots, v_n) = C_A \det(v_1, \dots, v_n)$. We can compute C_A by considering the basis vectors; $Ae_i = A_i$ where A_i is the ith column vector of A. Then,

$$C_A = d_A(e_1, \dots, e_n) = \det(Ae_1, \dots, Ae_n) = \det A$$

Hence,

$$\det(AB) = d_A(B) = \det A \det B$$

□

5.6. Singular and non-singular matrices

Definition. Let $A \in M_n(F)$. We say that

(i) A is *singular* if $\det A = 0$;

(ii) A is *non-singular* if $\det A \neq 0$.

Lemma. If A is invertible, it is non-singular.

Proof. If A is invertible, there exists A^{-1}. Then, since the determinant is a homomorphism,

$$\det\left(AA^{-1}\right) = \det I = 1$$

Thus $\det A \det A^{-1} = 1$ and hence neither of these determinants can be zero. □

Theorem. Let $A \in M_n(F)$. The following are equivalent.

(i) A is invertible;

(ii) A is non-singular;

(iii) $r(A) = n$.

Proof. We have already shown that (i) implies (ii). We have also shown that (i) and (iii) are equivalent by the rank-nullity theorem. So it suffices to show that (ii) implies (iii).

Suppose $r(A) < n$. Then we will show A is singular. We have $\dim \operatorname{span}(A_1, \dots, A_n) < n$. Therefore, since there are n vectors, $(A_1, \dots, A_n)$ is not free. So there exist scalars λ_i not all zero such that $\sum_i \lambda_i A_i = 0$. Choose j such that $\lambda_j \neq 0$. Then,

$$A_j = -\frac{1}{\lambda_j} \sum_{i \neq j} \lambda_i A_i$$

So we can compute the determinant of A by

$$\det A = \det\left(A_1, \dots, -\frac{1}{\lambda_j} \sum_{i \neq j} \lambda_i A_i, \dots, A_n\right)$$

Since the determinant is alternating and linear in the jth entry, its value is zero. So A is singular as required. □

Remark. The above theorem gives necessary and sufficient conditions for invertibility of a set of n linear equations with n unknowns.

5.7. Determinants of linear maps

Lemma. Similar matrices have the same determinant.

Proof.

$$\det\left(P^{-1}AP\right) = \det\left(P^{-1}\right)\det A \det P = \det A \det\left(P^{-1}P\right) = \det A$$

□

Definition. If α is an endomorphism, then we define

$$\det \alpha = \det[\alpha]_{B,B}$$

where B is any basis of the vector space. This is well-defined, since this value does not depend on the choice of basis.

Theorem. $\det : L(V,V) \to F$ satisfies the following properties.

(i) $\det I = 1$;

(ii) $\det(\alpha\beta) = \det\alpha \det\beta$;

(iii) $\det\alpha \neq 0$ if and only if α is invertible, and in this case, $\det\left(\alpha^{-1}\right)\det\alpha = 1$.

This is simply a reformulation of the previous theorem for matrices. The proof is simple, and relies on the invariance of the determinant under a change of basis.

5.8. Determinant of block-triangular matrices

Lemma. Let $A \in M_k(F)$, $B \in M_\ell(F)$, $C \in M_{k,\ell}(F)$. Consider the matrix

$$M = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$$

Then $\det M = \det A \det B$.

Proof. Let $n = k + \ell$, so $M \in M_n(F)$. Let $M = (m_{ij})$. We must compute

$$\det M = \sum_{\sigma \in S_n} \varepsilon(\sigma) \prod_{i=1}^{n} m_{\sigma(i)i}$$

Observe that $m_{\sigma(i)i} = 0$ if $i \leq k$ and $\sigma(i) > k$. Then, we need only sum over $\sigma \in S_n$ such that for all $j \leq k$, we have $\sigma(j) \leq k$. Thus, for all $j \in \{k+1, \dots, n\}$, we have $\sigma(j) \in \{k+1, \dots, n\}$.

We can then uniquely decompose σ into two permutations $\sigma = \sigma_1\sigma_2$, where σ_1 is restricted to $\{1, \dots, k\}$ and σ_2 is restricted to $\{k+1, \dots, n\}$. Hence,

$$\begin{aligned}
\det M &= \sum_{\sigma_1 \in S_k} \sum_{\sigma_2 \in S_{n-k}} \varepsilon(\sigma) \prod_{i=1}^{n} m_{\sigma(i)i} \\
&= \sum_{\sigma_1 \in S_k} \sum_{\sigma_2 \in S_{n-k}} \varepsilon(\sigma_1)\varepsilon(\sigma_2) \prod_{i=1}^{k} m_{\sigma(i)i} \prod_{i=k+1}^{n} m_{\sigma(i)i} \\
&= \sum_{\sigma_1 \in S_k} \varepsilon(\sigma_1) \prod_{i=1}^{k} m_{\sigma(i)i} \sum_{\sigma_2 \in S_{n-k}} \varepsilon(\sigma_2) \prod_{i=k+1}^{n} m_{\sigma(i)i} \\
&= \det A \det B
\end{aligned}$$

□

Corollary. We need not restrict ourselves to just two blocks, since we can apply the above lemma inductively. In particular, this implies that an upper-triangular matrix with diagonal elements λ_i has determinant $\prod_i \lambda_i$.

6. Adjugate matrices

6.1. Column and row expansions

Let $A \in M_n(F)$ with column vectors $A^{(i)}$. We know that

$$\det\big(A^{(1)}, \dots, A^{(j)}, \dots, A^{(k)}, \dots, A^{(n)}\big) = -\det\big(A^{(1)}, \dots, A^{(k)}, \dots, A^{(j)}, \dots, A^{(n)}\big)$$

Using the fact that $\det A = \det A^\intercal$ we can similarly see that swapping two rows will invert the sign of the determinant.

Remark. We could have proven all of the properties of the determinant above by using the decomposition of A into elementary matrices.

Definition. Let $A \in M_n(F)$. Let $i, j \in \{1, \dots, n\}$. We define the *minor* $A_{\widehat{ij}} \in M_{n-1}(F)$ to be the matrix obtained by removing the ith row and the jth column.

Lemma. Let $A \in M_n(F)$.

(i) Let $j \in \{1, \dots, n\}$. The determinant of A is given by the *column expansion with respect to the jth column*:

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{\widehat{ij}}$$

(ii) Let $i \in \{1, \dots, n\}$. The same determinant is also given by the *row expansion with respect to the ith row*:

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{\widehat{ij}}$$

This is a process of reducing the computation of $n \times n$ determinants to $(n-1) \times (n-1)$ determinants.

Proof. We will prove case (i), the column expansion with respect to the jth column. Then (ii) will follow from the transpose of the matrix. Let $j \in \{1, \dots, n\}$. We can write $A^{(j)} = \sum_{i=1}^{n} a_{ij} e_i$ where the e_i are the canonical basis. Then, by swapping rows and columns,

$$\begin{aligned}
\det A &= \det\left(A^{(1)}, \dots, \sum_{i=1}^{n} a_{ij} e_i, \dots, A^{(n)}\right) \\
&= \sum_{i=1}^{n} a_{ij} \det\left(A^{(1)}, \dots, e_i, \dots, A^{(n)}\right) \\
&= \sum_{i=1}^{n} a_{ij} (-1)^{j-1} \det\left(e_i, A^{(1)}, \dots, A^{(n)}\right) \\
&= \sum_{i=1}^{n} a_{ij} (-1)^{j-1} (-1)^{i-1} \det\left(e_1, \overline{A}^{(1)}, \dots, \overline{A}^{(n)}\right)
\end{aligned}$$

This has brought the matrix into block form, where there is an element of value 1 in the top left, and the matrix $A_{\widehat{ij}}$ in the bottom right. The bottom left block is entirely zeroes. Hence,

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{\widehat{ij}}$$

as required. □

Remark. We have proven that

$$\det\left(A^{(1)}, \dots, e_i, \dots, A^{(n)}\right) = (-1)^{i+j} \det A_{\widehat{ij}}$$

6.2. Adjugates

Definition. Let $A \in M_n(F)$. The *adjugate matrix* of A, denoted $\operatorname{adj} A$, is the $n \times n$ matrix given by

$$(\operatorname{adj} A)_{ij} = (-1)^{i+j} \det A_{\widehat{ji}}$$

Hence,

$$\det\left(A^{(1)}, \dots, e_i, \dots, A^{(n)}\right) = (\operatorname{adj} A)_{ji}$$

Theorem. Let $A \in M_n(F)$. Then

$$(\operatorname{adj} A)A = (\det A)I$$

In particular, when A is invertible,

$$A^{-1} = \frac{\operatorname{adj} A}{\det A}$$

Proof. We have

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{\widehat{ij}}$$

Hence,

$$\det A = \sum_{i=1}^{n} (\operatorname{adj} A)_{ji} a_{ij} = ((\operatorname{adj} A)A)_{jj}$$

So the diagonal terms match. Off the diagonal,

$$0 = \det\left(A^{(1)}, \dots, \underbrace{A^{(k)}}_{j\text{th position}}, \dots, A^{(k)}, \dots, A^{(n)}\right)$$

By linearity,

$$\begin{aligned}
0 &= \det\left(A^{(1)}, \dots, \underbrace{\sum_{i=1}^{n} a_{ik} e_i}_{j\text{th position}}, \dots, A^{(k)}, \dots, A^{(n)}\right) \\
&= \sum_{i=1}^{n} a_{ik} \det\left(A^{(1)}, \dots, \underbrace{e_i}_{j\text{th position}}, \dots, A^{(k)}, \dots, A^{(n)}\right) \\
&= \sum_{i=1}^{n} a_{ik} (\operatorname{adj} A)_{ji} \\
&= ((\operatorname{adj} A)A)_{jk}
\end{aligned}$$

□

6.3. Cramer's rule

Proposition. Let A be an invertible square matrix of dimension n. Let $b \in F^n$. Then the unique solution to $Ax = b$ is given by

$$x_i = \frac{1}{\det A} \det\left(A_{\widehat{ib}}\right)$$

where $A_{\widehat{ib}}$ is obtained by replacing the ith column of A by b. This is an algorithm to compute x, avoiding the computation of A^{-1}.

Proof. Let A be invertible. Then there exists a unique $x \in F^n$ such that $Ax = b$. Then, since the determinant is alternating,

$$\begin{aligned}
\det\left(A_{\widehat{ib}}\right) &= \det\left(A^{(1)}, \dots, A^{(i-1)}, b, A^{(i+1)}, \dots, A^{(n)}\right) \\
&= \det\left(A^{(1)}, \dots, A^{(i-1)}, \sum_{j=1}^{n} x_j A^{(j)}, A^{(i+1)}, \dots, A^{(n)}\right) \\
&= \det\left(A^{(1)}, \dots, A^{(i-1)}, x_i A^{(i)}, A^{(i+1)}, \dots, A^{(n)}\right) \\
&= x_i \det A
\end{aligned}$$

So the formula works. □

7. Eigenvectors and eigenvalues

7.1. Eigenvalues

Let V be an F-vector space. Let $\dim V = n < \infty$, and let α be an endomorphism of V. We wish to find a basis B of V such that, in this basis, $[\alpha]_B \equiv [\alpha]_{B,B}$ has a simple (e.g. diagonal, triangular) form. Recall that if B' is another basis and P is the change of basis matrix, $[\alpha]_{B'} = P^{-1}[\alpha]_B P$. Equivalently, given a square matrix $A \in M_n(F)$ we want to conjugate it by a matrix P such that the result is 'simpler'.

Definition. Let $\alpha \in L(V)$ be an endomorphism. We say that α is *diagonalisable* if there exists a basis B of V such that the matrix $[\alpha]_B$ is diagonal. We say that α is *triangulable* if there exists a basis B of V such that $[\alpha]_B$ is triangular.

Remark. We can express this equivalently in terms of conjugation of matrices.

Definition. A scalar $\lambda \in F$ is an *eigenvalue* of an endomorphism α if and only if there exists a vector $v \in V \setminus \{0\}$ such that $\alpha(v) = \lambda v$. Such a vector is an *eigenvector* with eigenvalue λ. $V_\lambda = \{v \in V : \alpha(v) = \lambda v\} \leq V$ is the *eigenspace* associated to λ.

Lemma. λ is an eigenvalue if and only if $\det(\alpha - \lambda I) = 0$.

Proof. If λ is an eigenvalue, there exists a nonzero vector v such that $\alpha(v) = \lambda v$, so $(\alpha - \lambda)(v) = 0$. So the kernel is non-trivial. So $\alpha - \lambda I$ is not injective, so it is not surjective by the rank-nullity theorem. Hence this matrix is not invertible, so it has zero determinant. □

Remark. If $\alpha(v_j) = \lambda v_j$ for $j \in \{1, \dots, m\}$, we can complete the family v_j into a basis $(v_1, \dots, v_n)$ of V. Then in this basis, the first m columns of the matrix α has diagonal entries λ_j.

7.2. Polynomials

Recall the following facts about polynomials on a field, for instance

$$f(t) = a_n t^n + \cdots + a_1 t + a_0$$

We say that the degree of f, written $\deg f$ is n. The degree of $f + g$ is at most the maximum degree of f and g. $\deg(fg) = \deg f + \deg g$. Let $F[t]$ be the vector space of polynomials with coefficients in F. If λ is a root of f, then $(t - \lambda)$ divides F.

Proof.

$$f(t) = a_n t^n + \cdots + a_1 t + a_0$$

Hence,

$$f(\lambda) = a_n \lambda^n + \cdots + a_1 \lambda + a_0 = 0$$

which implies that

$$f(t) = f(t) - f(\lambda) = a_n(t^n - \lambda^n) + \cdots + a_1(t - \lambda)$$

But note that, for all n,

$$t^n - \lambda^n = (1 - \lambda)(t^{n-1} + \lambda t^{n-2} + \cdots + \lambda^{n-2}t + \lambda^{n-1})$$

□

Remark. We say that λ is a root of *multiplicity* k if $(t - \lambda)^k$ divides f but $(t - \lambda)^{k+1}$ does not.

Corollary. A nonzero polynomial of degree n has at most n roots, counted with multiplicity.

Corollary. If f_1, f_2 are two polynomials of degree less than n such that $f_1(t_i) = f_2(t_i)$ for $i \in \{1, \dots, n\}$ and t_i distinct, then $f_1 \equiv f_2$.

Proof. $f_1 - f_2$ has degree less than n, but has n roots. Hence it is zero. □

Theorem. Any polynomial $f \in \mathbb{C}[t]$ of positive degree has a complex root. When counted with multiplicity, f has a number of roots equal to its degree.

Corollary. Any polynomial $f \in \mathbb{C}[t]$ can be factorised into an amount of linear factors equal to its degree.

7.3. Characteristic polynomials

Definition. Let α be an endomorphism. The *characteristic polynomial* of α is

$$\chi_\alpha(\lambda) = \det(\alpha - \lambda I)$$

Remark. χ_α is a polynomial because the determinant is defined as a polynomial in the terms of the matrix. Note further that conjugate matrices have the same characteristic polynomial, so the above definition is well defined in any basis. Indeed, $\det\big(P^{-1}\alpha P - \lambda I\big) = \det\big(P^{-1}(\alpha - \lambda I)P\big) = \det(\alpha - \lambda I)$.

Theorem. Let $\alpha \in L(V)$. α is triangulable if and only if χ_α can be written as a product of linear factors over F. In particular, all complex matrices are triangulable.

Proof. Suppose α is triangulable. Then for a basis B, $[\alpha]_B$ is triangulable with diagonal entries a_i. Then

$$\chi_\alpha(t) = (a_1 - t)(a_2 - t) \cdots (a_n - t)$$

Conversely, let $\chi_\alpha(t)$ be the characteristic polynomial of α with a root λ. Then, $\chi_\alpha(\lambda) = 0$ implies λ is an eigenvalue. Let V_λ be the corresponding eigenspace. Let $(v_1, \dots, v_k)$ be the

basis of this eigenspace, completed to a basis $(v_1, \dots, v_n)$ of V. Let $W = \operatorname{span}\{v_{k+1}, \dots, v_n\}$, and then $V = V_\lambda \oplus W$. Then

$$[\alpha]_B = \begin{pmatrix} \lambda I & \star \\ 0 & C \end{pmatrix}$$

where $\star$ is arbitrary, and C is a block of size $(n-k)\times(n-k)$. Then α induces an endomorphism $\overline{\alpha}\colon V/U \to V/U$ with respect to the basis $(v_{k+1}, \dots, v_n)$, where $U = V_\lambda$. By induction on the dimension, we can find a basis $(w_{k+1}, \dots, w_n)$ for which C has a triangular form. Then the basis $(v_1, \dots, v_k, w_{k+1}, \dots, w_n)$ is a basis for which α is triangular. □

Lemma. Let $n = \dim V$, and V be a vector space over $\mathbb{R}$ or $\mathbb{C}$. Let α be an endomorphism on V. Then

$$\chi_\alpha(t) = (-1)^n t^n + c_{n-1}t^{n-1} + \cdots + c_0$$

with

$$c_0 = \det A; \quad c_{n-1} = (-1)^{n-1} \operatorname{tr} A$$

Proof.

$$\chi_\alpha(t) = \det(\alpha - tI) \implies \chi_\alpha(0) = \det(\alpha)$$

Further, for $\mathbb{R}, \mathbb{C}$ we know that α is triangulable over $\mathbb{C}$. Hence $\chi_\alpha(t)$ is the determinant of a triangular matrix;

$$\chi_\alpha(t) = \prod_{i=1}^{n} (a_i - t)$$

Hence

$$c_{n-1} = (-1)^{n-1} a_i$$

Since the trace is invariant under a change of basis, this is exactly the trace as required. □

7.4. Polynomials for matrices and endomorphisms

Let $p(t)$ be a polynomial over F. We will write

$$p(t) = a_n t^n + \cdots + a_0$$

For a matrix $A \in M_n(F)$, we write

$$p(A) = a_n A^n + \cdots + a_0 \in M_n(F)$$

For an endomorphism $\alpha \in L(V)$,

$$p(\alpha) = a_n \alpha^n + \cdots + a_0 I \in L(V); \quad \alpha^k \equiv \underbrace{\alpha \circ \cdots \circ \alpha}_{k \text{ times}}$$

7.5. Sharp criterion of diagonalisability

Theorem. Let V be a vector space over F of finite dimension n. Let α be an endomorphism of V. Then α is diagonalisable if and only if there exists a polynomial p which is a product of *distinct* linear factors, such that $p(\alpha) = 0$. In other words, there exist distinct $\lambda_1, \dots, \lambda_k$ such that

$$p(t) = \prod_{i=1}^{n}(t - \lambda_i) \implies p(\alpha) = 0$$

Proof. Suppose α is diagonalisable in a basis B. Let $\lambda_1, \dots, \lambda_k$ be the $k \leq n$ *distinct* eigenvalues. Let

$$p(t) = \prod_{i=1}^{k}(t - \lambda_i)$$

Let $v \in B$. Then $\alpha(v) = \lambda_i v$ for some i. Then, since the terms in the following product commute,

$$(\alpha - \lambda_i I)(v) = 0 \implies p(\alpha)(v) = \left[\prod_{i=1}^{k}(\alpha - \lambda_i I)\right](v) = 0$$

So for all basis vectors, $p(\alpha)(v)$. By linearity, $p(\alpha) = 0$.

Conversely, suppose that $p(\alpha) = 0$ for some polynomial $p(t) = \prod_{i=1}^{k}(t - \lambda_i)$ with distinct λ_i. Let $V_{\lambda_i} = \ker(\alpha - \lambda_i I)$. We claim that

$$V = \bigoplus_{i=1}^{k} V_{\lambda_i}$$

Consider the polynomials

$$q_j(t) = \prod_{i=1, i\neq j}^{k} \frac{t - \lambda_i}{\lambda_j - \lambda_i}$$

These polynomials evaluate to one at λ_j and zero at λ_i for $i \neq j$. Hence $q_j(\lambda_i) = \delta_{ij}$. We now define the polynomial

$$q = q_1 + \cdots + q_k$$

The degree of q is at most $(k - 1)$. Note, $q(\lambda_i) = 1$ for all $i \in \{1, \dots, k\}$. The only polynomial that evaluates to one at k points with degree at most $(k - 1)$ is exactly given by $q(t) = 1$. Consider the endomorphism

$$\pi_j = q_j(\alpha) \in L(V)$$

These are called the 'projection operators'. By construction,

$$\sum_{j=1}^{k} \pi_j = \sum_{j=1}^{k} q_j(\alpha) = I$$

So the sum of the π_j is the identity. Hence, for all $v \in V$,

$$I(v) = v = \sum_{j=1}^{k} \pi_j(v) = \sum_{j=1}^{k} q_j(\alpha)(v)$$

So we can decompose any vector as a sum of its projections $\pi_j(v)$. Now, by definition of q_j and p,

$$\begin{aligned}(\alpha - \lambda_j I)q_j(\alpha)(v) &= \frac{1}{\prod_{i\neq j}(\lambda_j - \lambda_i)}(\alpha - \lambda_j I)\left[\prod_{i\neq j}(t - \lambda_i)\right](\alpha) \\ &= \frac{1}{\prod_{i\neq j}(\lambda_j - \lambda_i)} \prod_{i=1}^{k}(\alpha - \lambda_i I)(v) \\ &= \frac{1}{\prod_{i\neq j}(\lambda_j - \lambda_i)} p(\alpha)(v)\end{aligned}$$

By assumption, this is zero. For all v, we have $(\alpha - \lambda_j I)q_j(\alpha)(v)$. Hence,

$$(\alpha - \lambda_j I)\pi_j(v) = 0 \implies \pi_j(v) \in \ker(\alpha - \lambda_j I) = v_j$$

We have then proven that, for all $v \in V$,

$$v = \sum_{j=1}^{k} \underbrace{\pi_j(v)}_{\in V_j}$$

Hence,

$$V = \sum_{j=1}^{k} V_j$$

It remains to show that the sum is direct. Indeed, let

$$v \in V_{\lambda_j} \cap \left(\sum_{i\neq j} V_{\lambda_i}\right)$$

We must show $v = 0$. Applying π_j,

$$\pi_j(v) = q_j(\alpha)(v) = \prod_{i\neq j} \frac{(\alpha - \lambda_i I)(v)}{\lambda_j - \lambda_i}$$

Since $\alpha(v) = \lambda_j v$,

$$\pi_j(v) = \prod_{i\neq j} \frac{(\lambda_j - \lambda_i)v}{\lambda_j - \lambda_i} = v$$

Hence π_j really projects onto V_{λ_j}. However, we also know $v \in \sum_{i\neq j} V_{\lambda_i}$. So we can write $v = \sum_{i\neq j} w_i$ for $w \in V_{\lambda_i}$. Thus,

$$\pi_j(w_i) = \prod_{m\neq j} \frac{(\alpha - \lambda_m I)(v)}{\lambda_m - \lambda_j}$$

Since $\alpha(w_i) = \lambda_i w_i$, one of the factors will vanish, hence

$$\pi_j(w_i) = 0$$

So

$$v = \sum_{i\neq j} w_i \implies \pi_j(v) = \sum_{i\neq j} \pi_j(w_i) = 0$$

But $v = \pi_j(v)$ hence $v = 0$. So the sum is direct. Hence, $B = (B_1, \dots, B_k)$ is a basis of V, where the B_i are bases of V_{λ_i}. Then $[\alpha]_B$ is diagonal. □

Remark. We have shown further that if $\lambda_1, \dots, \lambda_k$ are distinct eigenvalues of α, then

$$\sum_{i=1}^{k} V_{\lambda_i} = \bigoplus_{i=1}^{k} V_{\lambda_i}$$

Therefore, the only way that diagonalisation fails is when this sum is not direct, so

$$\sum_{i=1}^{k} V_{\lambda_i} < V$$

Example. Let $F = \mathbb{C}$. Let $A \in M_n(F)$ such that A has finite order; there exists $m \in \mathbb{N}$ such that $A^m = I$. Then A is diagonalisable. This is because

$$t^m - 1 = p(t) = \prod_{j=1}^{m} (t - \xi_m^j); \quad \xi_m = e^{2\pi i/m}$$

and $p(A) = 0$.

7.6. Simultaneous diagonalisation

Theorem. Let α, β be endomorphisms of V which are diagonalisable. Then α, β are *simultaneously diagonalisable* (there exists a basis B of V such that $[\alpha]_B, [\beta]_B$ are diagonal) if and only if α and β commute.

Proof. Two diagonal matrices commute. If such a basis exists, $\alpha\beta = \beta\alpha$ in this basis. So this holds in any basis. Conversely, suppose $\alpha\beta = \beta\alpha$. We have

$$V = \bigoplus_{i=1}^{k} V_{\lambda_i}$$

where $\lambda_i, \dots, \lambda_k$ are the k distinct eigenvalues of α. We claim that $\beta\left(V_{\lambda_j}\right) \leq V_{\lambda_j}$. Indeed, for $v \in V_{\lambda_j}$,

$$\alpha\beta(v) = \beta\alpha(v) = \beta(\lambda_j v) = \lambda_j \beta(v) \implies \alpha(\beta(v)) = \lambda_j \beta(v)$$

Hence, $\beta(v) \in V_{\lambda_j}$. By assumption, β is diagonalisable. Hence, there exists a polynomial p with distinct linear factors such that $p(\beta) = 0$. Now, $\beta\left(V_{\lambda_j}\right) \leq V_{\lambda_j}$ so we can consider $\beta|_{V_{\lambda_j}}$. This is an endomorphism of V_{λ_j}. We can compute

$$p\left(\beta\Big|_{V_{\lambda_j}}\right) = 0$$

Hence, $\beta|_{V_{\lambda_j}}$ is diagonalisable. Let B_i be the basis of V_{λ_i} in which $\beta|_{V_{\lambda_j}}$ is diagonal. Since $V = \bigoplus V_{\lambda_i}, B = (B_1, \dots, B_k)$ is a basis of V. Then the matrices of α and β in V are diagonal. □

7.7. Minimal polynomials

Recall from IB Groups, Rings and Modules the Euclidean algorithm for dividing polynomials. Given a, b polynomials over F with b nonzero, there exist polynomials q, r over F with $\deg r < \deg b$ and $a = qb + r$.

Definition. Let V be a finite dimensional F-vector space. Let α be an endomorphism on V. The *minimal polynomial* m_α of α is the nonzero polynomial with smallest degree such that $m_\alpha(\alpha) = 0$.

Remark. If $\dim V = n < \infty$, then $\dim L(V) = n^2$. In particular, the family $\left\{I, \alpha, \dots, \alpha^{n^2}\right\}$ cannot be free since it has $n^2 + 1$ entries. This generates a polynomial in α which evaluates to zero. Hence, a minimal polynomial always exists.

Lemma. Let $\alpha \in L(V)$ and $p \in F[t]$ be a polynomial. Then $p(\alpha) = 0$ if and only if m_α is a factor of p. In particular, m_α is well-defined and unique up to a constant multiple.

Proof. Let $p \in F[t]$ such that $p(\alpha) = 0$. If $m_\alpha(\alpha) = 0$ and $\deg m_\alpha < \deg p$, we can perform the division $p = m_\alpha q + r$ for $\deg r < \deg m_\alpha$. Then $p(\alpha) = m_\alpha(\alpha)q(\alpha) + r(\alpha)$. But $m_\alpha(\alpha) = 0$. But $\deg r < \deg m_\alpha$ and m_α is the smallest degree polynomial which evaluates to zero for α, so $r \equiv 0$ so $p = m_\alpha q$. In particular, if m_1, m_2 are both minimal polynomials that evaluate to zero for α, we have m_1 divides m_2 and m_2 divides m_1. Hence they are equivalent up to a constant. □

Example. Let $V = F^2$ and

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

We can check $p(t) = (t-1)^2$ gives $p(A) = p(B) = 0$. So the minimal polynomial of A or B must be either $(t-1)$ or $(t-1)^2$. For A, we can find the minimal polynomial is $(t-1)$, and

for B we require $(t-1)^2$. So B is not diagonalisable, since its minimal polynomial is not a product of distinct linear factors.

7.8. Cayley–Hamilton theorem

Theorem. Let V be a finite dimensional F-vector space. Let $\alpha \in L(V)$ with characteristic polynomial $\chi_\alpha(t) = \det(\alpha - tI)$. Then $\chi_\alpha(\alpha) = 0$.

Two proofs will provided; one more physical and based on $F = \mathbb{C}$ and one more algebraic.

Proof. Let $B = \{v_1, \dots, v_n\}$ be a basis of V such that $[\alpha]_B$ is triangular. This can be done when $F = \mathbb{C}$. Note, if the diagonal entries in this basis are a_i,

$$\chi_\alpha(t) = \prod_{i=1}^{n}(a_i - t) \implies \chi_\alpha(\alpha) = (\alpha - a_1 I)\dots(\alpha - a_n I)$$

We want to show that this expansion evaluates to zero. Let $U_j = \operatorname{span}\{v_1, \dots, v_j\}$. Let $v \in V = U_n$. We want to compute $\chi_\alpha(\alpha)(v)$. Note, by construction of the triangular matrix.

$$\begin{aligned}
\chi_\alpha(\alpha)(v) &= (\alpha - a_1 I)\dots\underbrace{(\alpha - a_n I)(v)}_{\in U_{n-1}} \\
&= (\alpha - a_1 I)\dots\underbrace{(\alpha - a_{n-1} I)(\alpha - a_n I)(v)}_{\in U_{n-2}} \\
&= \dots \\
&\in U_0
\end{aligned}$$

Hence this evaluates to zero. □

The following proof works for any field where we can equate coefficients, but is much less intuitive.

Proof. We will write

$$\det(tI - \alpha) = (-1)^n \chi_\alpha(t) = t^n + a_{n-1}t^{n-1} + \dots + a_0$$

For any matrix B, we have proven $B \operatorname{adj} B = (\det B)I$. We apply this relation to the matrix $B = tI - A$. We can check that

$$\operatorname{adj} B = \operatorname{adj}(tI - A) = B_{n-1}t^{n-1} + \dots + B_1 t + B_0$$

since adjugate matrices are degree $(n-1)$ polynomials for each element. Then, by applying $B \operatorname{adj} B = (\det B)I$,

$$(tI - A)[B_{n-1}t^{n-1} + \dots + B_1 t + B_0] = (\det B)I = (t^n + \dots + a_0)I$$

Since this is true for all t, we can equate coefficients. This gives

$$\begin{aligned} t^n &: & I &= B_{n-1} \\ t^{n-1} &: & a_{n-1}I &= B_{n-2} - AB_{n-1} \\ &\vdots & &\vdots \\ t^0 &: & a_0 I &= -AB_1 \end{aligned}$$

Then, substituting A for t in each relation will give, for example, $A^n I = A^n B_{n-1}$. Computing the sum of all of these identities, we recover the original polynomial in terms of A instead of in terms of t. Many terms will cancel since the sum telescopes, yielding

$$A^n + a_{n-1}A^{n-1} + \cdots + a_0 I = 0$$

□

7.9. Algebraic and geometric multiplicity

Definition. Let V be a finite dimensional F-vector space. Let $\alpha \in L(V)$ and let λ be an eigenvalue of α. Then

$$\chi_\alpha(t) = (t-\lambda)^{a_\lambda} q(t)$$

where $q(t)$ is a polynomial over F such that $(t-\lambda)$ does not divide q. a_λ is known as the *algebraic multiplicity* of the eigenvalue λ. We define the *geometric multiplicity* g_λ of λ to be the dimension of the eigenspace associated with λ, so $g_\lambda = \dim\ker(\alpha - \lambda I)$.

Lemma. If λ is an eigenvalue of $\alpha \in L(V)$, then $1 \leq g_\lambda \leq a_\lambda$.

Proof. We have $g_\lambda = \dim\ker(\alpha - \lambda I)$. There exists a nontrivial vector $v \in V$ such that $v \in \ker(\alpha - \lambda I)$ since λ is an eigenvalue. Hence $g_\lambda \geq 1$. We will show that $g_\lambda \leq a_\lambda$. Indeed, let $v_1, \ldots, v_{g_\lambda}$ be a basis of $V_\lambda \equiv \ker(\alpha - \lambda I)$. We complete this into a basis $B \equiv (v_1, \ldots, v_{g_\lambda}, v_{g_\lambda+1}, \ldots, v_n)$ of V. Then note that

$$[\alpha]_B = \begin{pmatrix} \lambda I_{g_\lambda} & \star \\ 0 & A_1 \end{pmatrix}$$

for some matrix A_1. Now,

$$\det(\alpha - tI) = \det\begin{pmatrix} (\lambda - t)I_{g_\lambda} & \star \\ 0 & A_1 - tI \end{pmatrix}$$

By the formula for determinants of block matrices with a zero block on the off diagonal,

$$\det(\alpha - tI) = (\lambda - t)^{g_\lambda} \det(A_1 - tI)$$

Hence $g_\lambda \leq a_\lambda$ since the determinant is a polynomial that could have more factors of the same form. □

Lemma. Let V be a finite dimensional F-vector space. Let $\alpha \in L(V)$ and let λ be an eigenvalue of α. Let c_λ be the multiplicity of λ as a root of the minimal polynomial of α. Then $1 \leq c_\lambda \leq a_\lambda$.

Proof. By the Cayley–Hamilton theorem, $\chi_\alpha(\alpha) = 0$. Since m_α is linear, m_α divides χ_α. Hence $c_\lambda \leq a_\lambda$. Now we show $c_\lambda \geq 1$. Indeed, λ is an eigenvalue hence there exists a nonzero $v \in V$ such that $\alpha(v) = \lambda v$. For such an eigenvector, $\alpha^P(v) = \lambda^P v$ for $P \in \mathbb{N}$. Hence for $p \in F[t]$, $p(\alpha)(v) = [p(\lambda)](v)$. Hence $m_\alpha(\alpha)(v) = [m_\alpha(\lambda)](v)$. Since the left hand side is zero, $m_\alpha(\lambda) = 0$. So $c_\lambda \geq 1$. □

Example. Let

$$A = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

The minimal polynomial can be computed by considering the characteristic polynomial

$$\chi_A(t) = (t-1)^2(t-2)$$

So the minimal polynomial is either $(t-1)^2(t-2)$ or $(t-1)(t-2)$ We check $(t-1)(t-2)$. $(A-I)(A-2I)$ can be found to be zero. So $m_A(t) = (t-1)(t-2)$. Since this is a product of distinct linear factors, A is diagonalisable.

Example. Let A be a Jordan block of size $n \geq 2$. Then $g_\lambda = 1$, $a_\lambda = n$, and $c_\lambda = n$.

7.10. Characterisation of diagonalisable complex endomorphisms

Lemma. Let $F = \mathbb{C}$. Let V be a finite-dimensional $\mathbb{C}$-vector space. Let α be an endomorphism of V. Then the following are equivalent.

(i) α is diagonalisable;

(ii) for all λ eigenvalues of α, we have $a_\lambda = g_\lambda$;

(iii) for all λ eigenvalues of α, $c_\lambda = 1$.

Proof. First, the fact that (i) is true if and only if (iii) is true has already been proven. Now let us show that (i) is equivalent to (ii). Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of α. We have already found that α is diagonalisable if and only if $V = \bigoplus V_{\lambda_i}$. The sum was found to be always direct, regardless of diagonalisability. We will compute the dimension of V in two ways;

$$n = \dim V = \deg \chi_\alpha; \quad n = \dim V = \sum_{i=1}^{k} a_{\lambda_i}$$

since χ_α is a product of $(t - \lambda_i)$ factors as $F = \mathbb{C}$. Since the sum is direct,

$$\dim\left(\bigoplus_{i=1}^{k} V_{\lambda_i}\right) = \sum_{i=1}^{k} g_{\lambda_i}$$

α is diagonalisable if and only if the dimensions are equal, so

$$\sum_{i=1}^{k} g_{\lambda_i} = \sum_{i=1}^{k} a_{\lambda_i}$$

Conversely, we have proven that for all eigenvalues λ_i, we have $g_{\lambda_i} \leq a_{\lambda_i}$. Hence, $\sum_{i=1}^{k} g_{\lambda_i} = \sum_{i=1}^{k} a_{\lambda_i}$ holds if and only if $g_{\lambda_i} = a_{\lambda_i}$ for all i. □

8. Jordan normal form

For this section, let $F = \mathbb{C}$.

8.1. Definition

Definition. Let $A \in M_n(\mathbb{C})$. We say that A is in *Jordan normal form* if it is a block diagonal matrix, where each block is of the form

$$J_{n_i}(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}$$

We say that $J_{n_i}(\lambda) \in M_{n_i}(\mathbb{C})$ are *Jordan blocks*. The $\lambda_i \in \mathbb{C}$ need not be distinct.

Remark. In three dimensions,

$$A = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

is in Jordan normal form, with three one-dimensional Jordan blocks with the same λ value.

8.2. Similarity to Jordan normal form

Theorem. Any complex matrix $A \in M_n(\mathbb{C})$ is similar to a matrix in Jordan normal form, which is unique up to reordering the Jordan blocks.

The proof is non-examinable. This follows from IB Groups, Rings and Modules.

Example. Let $\dim V = 2$. Then any matrix is similar to one of

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}; \quad \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}; \quad \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

The minimal polynomials are

$$(t - \lambda_1)(t - \lambda_2); \quad (t - \lambda); \quad (t - \lambda)^2$$

8.3. Direct sum of eigenspaces

Theorem. Let V be a $\mathbb{C}$-vector space. Let $\dim V = n < \infty$. Then, the minimal polynomial $m_\alpha(t)$ of an endomorphism $\alpha \in L(V)$ satisfies

$$V = \bigoplus_{j=1}^{k} V_j$$

where $V_j = \ker[(\alpha - \lambda_j I)^{c_j}]$, and where

$$m_\alpha(t) = \prod_{i=1}^{k} (t - \lambda_i)^{c_i}$$

V_j is called a *generalised eigenspace* associated with λ_j.

Remark. Note that V_j is stable by α, that is, $\alpha(V_j) = V_j$. Note further that $(\alpha - \lambda_j I)\big|_{V_j} = \mu_j$ gives that μ_j is a nilpotent endomorphism; $\mu_j^{c_j} = 0$. So the Jordan normal form theorem is a statement about nilpotent matrices.

Note, when α is diagonalisable, $c_j = 1$ and hence we recover $V_j = \ker(\alpha - \lambda_j I)$ and $V = \bigoplus V_j$.

Proof. The key to this proof is that the projectors onto V_j are 'explicit'. First, recall

$$m_\alpha(t) = \prod_{j=1}^{k} (t - \lambda_j)^{c_j}$$

Then, let

$$p_j(t) = \prod_{i \neq j} (t - \lambda_i)^{c_i}$$

Then p_j have by definition no common factor. So by Euclid's algorithm, we can find polynomials q_i such that

$$\sum_{i=1}^{k} q_i p_i = 1$$

We define the projector $\pi_j = q_j p_j(\alpha)$, which is an endomorphism. By construction, for all $v \in V$, we have

$$\sum_{j=1}^{k} \pi_j(v) = \sum_{j=1}^{k} a_j p_j(\alpha(v)) = I(v) = v$$

Hence,

$$v = \sum_{i=1}^{k} \pi_i(v)$$

Observe further that $\pi_j(v) \in V_j$. Indeed,

$$(\alpha - \lambda_j I)^{c_j} \pi_j(v) = (\alpha - \lambda_j I)^{c_j} q_j p_j(\alpha(v)) = q_j m_\alpha(\alpha(v)) = 0$$

Hence $\pi_j(v) \in V_j$. In particular, $V = \sum_{j=1}^{k} V_j$. We need to show that this sum is direct. Note, for $i \neq j$, $\pi_i \pi_j = 0$ from the definition of π. Hence, observe that

$$\pi_i = \pi_i \left(\sum_{j=1}^{k} \pi_j \right) \implies \pi_i = \pi_i \pi_i$$

Thus, π is a projector. In particular, this implies that $\pi_i|_{V_j}$ is the identity if $i = j$ and zero if $i \neq j$. This immediately implies that th sum is direct;

$$V = \bigoplus_{j=1}^{k} V_j$$

Indeed, suppose

$$\sum_{j=1}^{k} \alpha_j v_j = 0; \quad v_j \in V_j; \quad \alpha_1 = 0$$

Then

$$v_1 = -\frac{1}{\alpha_1} \sum_{j=2}^{k} \alpha_j v_j$$

Applying π_1,

$$v_1 = -\frac{1}{\alpha_1} \sum_{j=2}^{k} \alpha_j \pi_1(v_j) = 0$$

Iterating, we find $v = 0$. □

Remark. We can compute the quantities $a_\lambda, g_\lambda, c_\lambda$ on the Jordan normal form of a matrix. Indeed, let $m \geq 2$ and consider a Jordan block $J_m(\lambda)$. Then $J_m(\lambda) - \lambda I$ is the zero matrix with ones on the off-diagonal. $(J_m(\lambda) - \lambda I)^k$ pushes the ones onto the next line iteratively, so

$$(J_m(\lambda) - \lambda I)^k = \begin{pmatrix} 0 & I_{m-k} \\ 0 & 0 \end{pmatrix}$$

Hence J is nilpotent of order exactly m. In Jordan normal form,

(i) a_λ is the sum of sizes of blocks with eigenvalue λ. This is the amount of times λ is seen on the diagonal.

(ii) g_λ is the amount of blocks with eigenvalue λ, since each block represents one eigenvector.

(iii) c_λ is the size of the largest block with eigenvalue λ.

Example. Let

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix}$$

We wish to convert this matrix into Jordan normal form; so we seek a basis for which this matrix becomes Jordan normal form.

$$\chi_A(t) = (t-1)^2$$

Hence there exists only one eigenvalue, $\lambda = 1$. $A - I \neq 0$ hence $m_\alpha(t) = (t-1)^2$. Thus, the Jordan normal form of A is of the form

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

Now,

$$\ker(A - I) = \langle v_1 \rangle; \quad v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Further, we seek a v_2 such that

$$(A - I)v_2 = v_1 \implies v_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

Such a v_2 is not unique. Now,

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix}^{-1}$$

9. Properties of bilinear forms

9.1. Changing basis

Let $\phi : V \times V \to \mathbb{F}$ be a bilinear form. Let V be a finite-dimensional F-vector space. Let B be a basis of V and let $[\phi]_B = [\phi]_{BB}$ be the matrix with entries $\phi(e_i, e_j)$.

Lemma. Let ϕ be a bilinear form $V \times V \to F$. Then if B, B' are bases for V, and $P = [I]_{B',B}$ we have

$$[\phi]_{B'} = P^\intercal [\phi]_B P$$

Proof. This is a special case of the general change of basis formula. □

Definition. Let $A, B \in M_n(F)$ be square matrices. We say that A, B are *congruent* if there exists $P \in M_n(F)$ such that $A = P^\intercal B P$.

Remark. Congruence is an equivalence relation.

Definition. A bilinear form ϕ on V is *symmetric* if, for all $u, v \in V$, we have

$$\phi(u, v) = \phi(v, u)$$

Remark. If A is a square matrix, we say A is symmetric if $A = A^\intercal$. Equivalently, $A_{ij} = A_{ji}$ for all i, j. So ϕ is symmetric if and only if $[\phi]_B$ is symmetric for any basis B. Note further that to represent ϕ by a diagonal matrix in some basis B, it must necessarily be symmetric, since

$$P^\intercal A P = D \implies D = D^\intercal = (P^\intercal A P)^\intercal = P^\intercal A^\intercal P \implies A = A^\intercal$$

9.2. Quadratic forms

Definition. A map $Q : V \to F$ is a *quadratic form* if there exists a bilinear form $\phi : V \times V \to F$ such that, for all $u \in V$,

$$Q(u) = \phi(u, u)$$

So a quadratic form is the restriction of a bilinear form to the diagonal.

Remark. Let $B = (e_i)$ be a basis of V. Let $A = [\phi]_B = (\phi(e_i, e_j)) = (a_{ij})$. Then, for $u = \sum_i x_i e_i \in V$,

$$Q(u) = \phi(u, u) = \phi\left(\sum_i x_i e_i, \sum_j x_j e_j\right) = \sum_i \sum_j x_i x_j \phi(e_i, e_j) = \sum_i \sum_j x_i x_j a_{ij}$$

We can check that this is equal to

$$Q(u) = x^\intercal A x$$

where $[u]_B = x$. Note further that

$$x^\intercal A x = \sum_i \sum_j a_{ij} x_i x_j = \sum_i \sum_j a_{ji} x_i x_j = \sum_i \sum_j \frac{a_{ij} + a_{ji}}{2} x_i x_j = x^\intercal \left(\underbrace{\frac{A + A^\intercal}{2}}_{\text{symmetric}} \right) x$$

So we can always express the quadratic form as a symmetric matrix in any basis.

Proposition. If $Q : V \to F$ is a quadratic form, then there exists a unique symmetric bilinear form $\phi : V \times V \to F$ such that $Q(u) = \phi(u, u)$.

Proof. Let ψ be a bilinear form on V such that for all $u \in V$, we have $Q(u) = \psi(u, u)$. Then, let

$$\phi(u, v) = \frac{1}{2}[\psi(u, v) + \psi(v, u)]$$

Certainly ϕ is a bilinear form and symmetric. Further, $\phi(u, u) = \psi(u, u) = Q(u)$. So there exists a symmetric bilinear form ϕ such that $Q(u) = \phi(u, u)$, so it suffices to prove uniqueness. Let ϕ be a symmetric bilinear form such that for all $u \in V$ we have $Q(u) = \phi(u, u)$. Then, we can find

$$Q(u + v) = \phi(u + v, u + v) = \phi(u, u) + \phi(v, v) + 2\phi(u, v)$$

Thus $\phi(u, v)$ is defined uniquely by Q, since

$$2\phi(u, v) = Q(u + v) - Q(u) - Q(v)$$

So ϕ is unique (when 2 is invertible in F). This identity for $\phi(u, v)$ is known as the polarisation identity. □

9.3. Diagonalisation of symmetric bilinear forms

Theorem. Let $\phi : V \times V \to F$ be a symmetric bilinear form, where V is finite-dimensional. Then there exists a basis B of V such that $[\phi]_B$ is diagonal.

Proof. By induction on the dimension, suppose the theorem holds for all dimensions less than n for $n \geq 2$. If $\phi(u, u) = 0$ for all $u \in V$, then $\phi = 0$ by the polarisation identity, which is diagonal. Otherwise $\phi(e_1, e_1) \neq 0$ for some $e_1 \in V$. Let

$$U = (\langle e_1 \rangle)^\perp = \{v \in V : \phi(e_1, v) = 0\}$$

This is a vector subspace of V, which is in particular

$$\ker\{\phi(e_1, \cdot) : V \to F\}$$

By the rank-nullity theorem, $\dim U = n - 1$. We now claim that $U + \langle e_1 \rangle$ is a direct sum. Indeed, for $v = \langle e_1 \rangle \cap U$, we have $v = \lambda e_1$ and $\phi(e_1, v) = 0$. Hence $\lambda = 0$, since by assumption

$\phi(e_1, e_1) \neq 0$. So we find a basis $B' = (e_2, \dots, e_n)$ of U, which we extend by e_1 to $B = (e_1, e_2, \dots, e_n)$. Since $U \oplus \langle e_1 \rangle$ has dimension n, this is a basis of V. Under this basis, we find

$$[\phi]_B = \begin{pmatrix} \phi(e_1, e_1) & 0 \\ 0 & [\phi|_U]_{B'} \end{pmatrix}$$

because

$$\phi(e_1, e_j) = \phi(e_j, e_1) = 0$$

for all $j \geq 2$. By the inductive hypothesis we can take a basis B' such that the restricted ϕ to be diagonal, so $[\phi]_B$ is diagonal in this basis. □

Example. Let $V = \mathbb{R}^3$ and choose the canonical basis (e_i). Let

$$Q(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_1x_3 - 2x_2x_3$$

Then, if $Q(x_1, x_2, x_3) = x^\intercal Ax$, we have

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 2 \end{pmatrix}$$

Note that the off-diagonal terms are halved from their coefficients since in the expansion of $x^\intercal Ax$ they are included twice. Then, we can find a basis in which A is diagonal. We could use the above algorithm to find a basis, or complete the square in each component. We can write

$$Q(x_1, x_2, x_3) = (x_1 + x_2 + x_3)^2 + x_3^2 - 4x_2x_3 = (x_1 + x_2 + x_3)^2 + (x_3 - 2x_2)^2 - (2x_2)^2$$

This yields a new coordinate basis x_1', x_2', x_3'. Then $P^{-1}AP$ is diagonal. P is given by

$$\begin{pmatrix} x_1' \\ x_2' \\ x_3' \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 0 & -2 & 1 \\ 0 & -2 & 0 \end{pmatrix}}_{P^{-1}} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

9.4. Sylvester's law

Corollary. If $F = \mathbb{C}$, for any symmetric bilinear form ϕ there exists a basis of V such that $[\phi]_B$ is

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

Proof. Since any symmetric bilinear form ϕ in a finite-dimensional F-vector space V can be diagonalised, let $E = (e_1, \dots, e_n)$ such that $[\phi]_E$ is diagonal with diagonal entries a_i. Order the a_i such that a_i is nonzero for $1 \leq i \leq r$, and the remaining values (if any) are zero. For $i \leq r$, let $\sqrt{a_i}$ be a choice of a complex root for a_i. Then $v_i = \frac{e_i}{\sqrt{a_i}}$ for $i \leq r$ and $v_i = e_i$ for $i > r$ gives the basis B as required. □

Corollary. Every symmetric matrix of $M_n(\mathbb{C})$ is congruent to a unique matrix of the form

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

where r is the rank of the matrix.

Corollary. Let $F = \mathbb{R}$, and let V be a finite-dimensional $\mathbb{R}$-vector space. Let ϕ be a symmetric bilinear form on V. Then there exists a basis $B = (v_1, \dots, v_n)$ of V such that

$$[\phi]_B = \begin{pmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for some integers p, q.

Proof. Since square roots do not necessarily exist in $\mathbb{R}$, we cannot use the form above. We first diagonalise the bilinear form in some basis E. Then, reorder and group the a_i into a positive group of size p, a negative group of size q, and a zero group. Then,

$$v_i = \begin{cases} \frac{e_i}{\sqrt{a_i}} & i \in \{1, \dots, p\} \\ \frac{e_i}{\sqrt{-a_i}} & i \in \{p+1, \dots, p+q\} \\ e_i & i \in \{p+q+1, \dots, n\} \end{cases}$$

This gives a new basis as required. □

Definition. Let $F = \mathbb{R}$. The *signature* of a bilinear form ϕ is

$$s(\phi) = p - q$$

where p and q are defined as in the corollary above.

Theorem. Let $F = \mathbb{R}$. Let V be a finite-dimensional $\mathbb{R}$-vector space. If a real symmetric bilinear form is represented by some matrix

$$\begin{pmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

in some basis B, and some other matrix

$$\begin{pmatrix} I_{p'} & 0 & 0 \\ 0 & -I_{q'} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

in another basis B', then $p = p'$ and $q = q'$. Thus, the signature of the matrix is well defined.

Definition. Let ϕ be a symmetric bilinear form on a real vector space V. We say that

(i) ϕ is *positive definite* if $\phi(u,u) > 0$ for all nonzero $u \in V$;

(ii) ϕ is *positive semidefinite* if $\phi(u,u) \geq 0$ for all $u \in V$;

(iii) ϕ is *negative definite* or *negative semidefinite* if $\phi(u,u) < 0$ or $\phi(u,u) \leq 0$ respectively for all nonzero $u \in V$.

Example. The matrix

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

is positive definite for $r = n$, and positive semidefinite for $r < n$.

We now prove Sylvester's law.

Proof. In order to prove uniqueness of p, we will characterise the matrix in a way that does not depend on the basis. In particular, we will show that p is the largest dimension of a vector subspace of V such that the restriction of ϕ on this subspace is positive definite. Suppose we have $B = (v_1, \dots, v_n)$ and

$$[\phi]_B = \begin{pmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

We consider

$$X = \langle v_1, \dots, v_p \rangle$$

Then we can easily compute that $\phi|_X$ is positive definite. Let

$$Y = \langle v_{p+1}, \dots, v_n \rangle$$

Then, as above, $\phi|_Y$ is negative semidefinite. Suppose that ϕ is positive definite on another subspace X'. In this case, $Y \cap X' = \{0\}$, since if $y \in Y \cap X'$ we must have $Q(y) \leq 0$, but since $y \in X'$ we have $y = 0$. Thus, $Y + X' = Y \oplus X'$, so $n = \dim V \geq \dim Y + \dim X'$. But $\dim Y = n - p$, so $\dim X' \leq p$. The same argument can be executed for q, hence both p and q are independent of basis. □

9.5. Kernels of bilinear forms

Definition. Let $K = \{v \in V : \forall u \in V, \phi(u,v) = 0\}$. This is the *kernel* of the bilinear form.

Remark. By the rank-nullity theorem,

$$\dim K + \operatorname{rank} \phi = n$$

Using the above notation, we can show that there exists a subspace T of dimension $n - (p + q) + \min\{p, q\}$ such that $\phi|_T = 0$. Indeed, let $B = (v_1, \dots, v_n)$ such that

$$[\phi]_B = \begin{pmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The quadratic form has a zero subspace of dimension $n-(p+q)$ in the bottom right. But by setting

$$T = \{v_1 + v_{p+1}, \dots, v_q + v_{p+q}, v_{p+q+1}, \dots, v_n\}$$

we can combine the positive and negative blocks (assuming here that $p \geq q$) to produce more linearly independent elements of the kernel. In particular, $\dim T$ is the largest possible dimension of a subspace T' of V such that $\phi|_{T'} = 0$.

9.6. Sesquilinear forms

Let $F = \mathbb{C}$. The standard inner product on $\mathbb{C}^n$ is defined to be

$$\left\langle \begin{pmatrix} x_1 \\ \vdots \\ v_n \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right\rangle = \sum_{i=1}^{n} x_i \overline{y}_i$$

This is not a bilinear form on $\mathbb{C}$ due to the complex conjugate, it is linear in the first entry.

Definition. Let V, W be $\mathbb{C}$-vector spaces. A form $\phi : V \times W \to \mathbb{C}$ is called *sesquilinear* if it is linear in the first entry, and

$$\phi(v, \lambda_1 w_1 + \lambda_2 w_2) = \overline{\lambda}_1 \phi(v, w_1) + \overline{\lambda}_2 \phi(v, w_2)$$

so it is antilinear with respect to the second entry.

Lemma. Let $B = (v_1, \dots, v_m)$ be a basis of V and $C = (w_1, \dots, w_n)$ be a basis of W. Let $[\phi]_{B,C} = \big(\phi(v_i, w_j)\big)$. Then,

$$\phi(v, w) = [v]_B^{\intercal} [\phi]_{B,C} \overline{[w]_C}$$

Proof. Let B, B' be bases of V and C, C' be bases of W. Let $P = [I]_{B',B}$ and $Q = [I]_{C',C}$. Then

$$[\phi]_{B',C'} = P^{\intercal} [\phi]_{B,C} \overline{Q}$$

□

9.7. Hermitian forms

Definition. Let V be a finite-dimensional $\mathbb{C}$-vector space. Let ϕ be a sesquilinear form on V. Then ϕ is *Hermitian* if, for all $u, v \in V$,

$$\phi(u, v) = \overline{\phi(v, u)}$$

Remark. If ϕ is Hermitian, then $\phi(u, u) = \overline{\phi(u, u)} \in \mathbb{R}$. Further, $\phi(\lambda u, \lambda u) = |\lambda|^2 \phi(u, u)$. This allows us to define positive and negative definite Hermitian forms.

Lemma. A sesquilinear form $\phi : V \times V \to \mathbb{C}$ is Hermitian if and only if, for any basis B of V,

$$[\phi]_B = [\phi]_B^\dagger$$

Proof. Let $A = [\phi]_B = (a_{ij})$. Then $a_{ij} = \phi(e_i, e_j)$, and $a_{ji} = \phi(e_j, e_i) = \overline{\phi(e_i, e_j)} = \overline{a_{ij}}$. So $\overline{A}^\intercal = A$. Conversely suppose that $[\phi]_B = A = \overline{A}^\intercal$. Now let

$$u = \sum_{i=1}^{n} \lambda_i e_i; \quad v = \sum_{i=1}^{n} \mu_i e_i$$

Then,

$$\phi(u, v) = \phi\left(\sum_{i=1}^{n} \lambda_i e_i, \sum_{i=1}^{n} \mu_i e_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \overline{\mu_j} a_{ij}$$

Further,

$$\overline{\phi(v, u)} = \overline{\phi\left(\sum_{i=1}^{n} \mu_i e_i, \sum_{i=1}^{n} \lambda_i e_i\right)} = \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\mu_j}\overline{\lambda_i}\overline{a_{ij}}$$

which is equivalent. Hence ϕ is Hermitian. □

9.8. Polarisation identity

A Hermitian form ϕ on a complex vector space V is entirely determined by a quadratic form $Q : V \to \mathbb{R}$ such that $v \mapsto \phi(v, v)$ by the formula

$$\phi(u, v) = \frac{1}{4}[Q(u + v) - Q(u - v) + iQ(u + iv) - iQ(u - iv)]$$

9.9. Hermitian formulation of Sylvester's law

Theorem. Let V be a finite-dimensional $\mathbb{C}$-vector space. Let $\phi : V \times V \to \mathbb{C}$ be a Hermitian form on V. Then there exists a basis $B = (v_1, \dots, v_n)$ of V such that

$$[\phi]_B = \begin{pmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where p, q depend only on ϕ and not B.

Proof. The following is a sketch proof; it is nearly identical to the case of real symmetric bilinear forms. If $\phi = 0$, existence is trivial. Otherwise, using the polarisation identity there exists $e_1 \neq 0$ such that $\phi(e_1, e_1) \neq 0$. Let

$$v_1 = \frac{e_1}{\sqrt{|\phi(e_1, e_1)|}} \implies \phi(v_1, v_1) = \pm 1$$

Consider the orthogonal space $W = \{w \in V : \phi(v_1, w) = 0\}$. We can check, arguing analogously to the real case, that $V = \langle v_1 \rangle \oplus W$. Hence, we can inductively diagonalise ϕ.

p, q are unique. Indeed, we can prove that p is the maximal dimension of a subspace on which ϕ is positive definite (which is well-defined since $\phi(u, u) \in \mathbb{R}$). The geometric interpretation of q is similar. □

9.10. Skew-symmetric forms

Definition. Let V be a finite-dimensional $\mathbb{R}$-vector space. Let ϕ be a bilinear form on V. Then ϕ is *skew-symmetric* if, for all $u, v \in V$,

$$\phi(u, v) = -\phi(v, u)$$

Remark. $\phi(u, u) = -\phi(u, u) = 0$. Also, in any basis B of V, we have $[\phi]_B = -[\phi]_B^\intercal$. Any real matrix can be decomposed as the sum

$$A = \frac{1}{2}(A + A^\intercal) + \frac{1}{2}(A - A^\intercal)$$

where the first summand is symmetric and the second is skew-symmetric.

9.11. Skew-symmetric formulation of Sylvester's law

Theorem. Let V be a finite-dimensional $\mathbb{R}$-vector space. Let $\phi : V \times V \to \mathbb{R}$ be a skew-symmetric form on V. Then there exists a basis

$$B = (v_1, w_1, v_2, w_2, \dots, v_m, w_m, v_{2m+1}, v_{2m+2}, \dots, v_n)$$

of V such that

$$[\phi]_B = \begin{pmatrix} 0 & 1 & & & & \\ -1 & 0 & & & & \\ & & 0 & 1 & & \\ & & -1 & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

Corollary. Skew-symmetric matrices have an even rank.

Proof. This is again very similar to the previous case. We will perform an inductive step on the dimension of V. If $\phi \neq 0$, there exist v_1, w_1 such that $\phi_1(v_1, w_1) \neq 0$. After scaling one of the vectors, we can assume $\phi(v_1, w_1) = 1$. Since ϕ is skew-symmetric, $\phi(w_1, v_1) = -1$. Then v_1, w_1 are linearly independent; if they were linearly dependent we would have $\phi(v_1, w_1) = \phi(v_1, \lambda v_1) = 0$. Let $U = \langle v_1, w_1 \rangle$ and let $W = \{v \in V : \phi(v_1, v) = \phi(w_1, v) = 0\}$ and we can show $V = U \oplus W$. Then induction gives the required result. □

10. Inner product spaces

10.1. Definition

Definition. Let V be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A *scalar product* or *inner product* is a positive-definite symmetric (respectively Hermitian) bilinear form ϕ on V. We write

$$\phi(u, v) = \langle u, v \rangle$$

V, when equipped with this inner product, is called a real (respectively complex) *inner product space*.

Example. In $\mathbb{C}^n$, we define

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i \overline{y}_i$$

Example. Let $V = C^0([0, 1], \mathbb{C})$. Then we can define

$$\langle f, g \rangle = \int_0^1 f(t)\overline{g}(t)\,\mathrm{d}t$$

This is the L^2 scalar product.

Example. Let $\omega : [0, 1] : \mathbb{R}_+^\star$ where $\mathbb{R}_+^\star = \mathbb{R}_+ \setminus \{0\}$ and define

$$\langle f, g \rangle = \int_0^1 f(t)\overline{g}(t)w(t)\,\mathrm{d}t$$

Remark. Typically it suffices to check $\langle u, u \rangle = 0 \implies u = 0$ since linearity and positivity are usually trivial.

Definition. Let V be an inner product space. Then for $v \in V$, the *norm* of v induced by the inner product is defined by

$$\|v\| = (\langle v, v \rangle)^{1/2}$$

This is real, and positive if $v \neq 0$.

10.2. Cauchy–Schwarz inequality

Lemma. For an inner product space,

$$|\langle u, v \rangle| \leq \|a\| \cdot \|b\|$$

Proof. Let $t \in F$. Then,

$$0 \leq \|tu - v\| = \langle tu - v, tu - v \rangle = t\overline{t}\,\langle u, u \rangle - u\,\langle u, v \rangle - \overline{t}\,\langle v, u \rangle + \|v\|^2$$

Since the inner product is Hermitian,

$$0 \leq |t|^2 \|u\|^2 + \|v\|^2 - 2\,\mathrm{Re}(t\,\langle u, v\rangle)$$

By choosing

$$t = \frac{\overline{\langle u, v\rangle}}{\|u\|^2}$$

we have

$$0 \leq \frac{|\langle u, v\rangle|^2}{\|u\|^2} + \|v\|^2 - 2\,\mathrm{Re}\left(\frac{|\langle u, v\rangle|^2}{\|u\|^2}\right)$$

Since the term under the real part operator is real, the result holds. □

Note that equality implies collinearity in the Cauchy–Schwarz inequality.

Corollary (triangle inequality)**.** In an inner product space,

$$\|u + v\| \leq \|u\| + \|v\|$$

Proof. We have

$$\|u + v\|^2 = \langle u + v, u + v\rangle = \left\|u^2\right\| + 2\,\mathrm{Re}(\langle u, v\rangle) + \|v\|^2 \leq \left\|u^2\right\| + \|v\|^2 + 2\|u\| \cdot \|v\| = (\|u\| + \|v\|)^2$$

□

Remark. Any inner product induces a norm, but not all norms derive from scalar products.

10.3. Orthogonal and orthonormal sets

Definition. A set $(e_1, \ldots, e_k)$ of vectors of V is said to be *orthogonal* if $\langle e_i, e_j\rangle = 0$ for all $i \neq j$. The set is said to be *orthonormal* if it is orthogonal and $\|e_i\| = 1$ for all i. In this case, $\langle e_i, e_j\rangle = \delta_{ij}$.

Lemma. If $(e_1, \ldots, e_k)$ are orthogonal and nonzero, then they are linearly independent. Further, let $v \in \langle\{e_i\}\rangle$. Then,

$$v = \sum_{j=1}^{k} \lambda_j e_j \implies \lambda_j = \frac{\langle v, e_j\rangle}{\left\|e_j\right\|^2}$$

Proof. Suppose

$$\sum_{i=1}^{k} \lambda_i e_i = 0$$

Then,

$$0 = \left\langle \sum_{i=1}^{k} \lambda_i, e_j \right\rangle \implies 0 = \sum_{i=1}^{k} \lambda_i \langle e_i, e_j\rangle$$

Thus $\lambda_j = 0$ for all j. Further, for v in the span of these vectors,

$$\langle v, e_j \rangle = \sum_{i=1}^{k} \lambda_i \langle e_i, e_j \rangle = \lambda_j \left\| e_j \right\|^2$$

□

10.4. Parseval's identity

Corollary. Let V be a finite-dimensional inner product space. Let $(e_1, \dots, e_n)$ be an orthonormal basis. Then, for any vectors $u, v \in V$, we have

$$\langle u, v \rangle = \sum_{i=1}^{n} \langle u, e_i \rangle \overline{\langle v, e_i \rangle}$$

Hence,

$$\|u\|^2 = \sum_{i=1}^{n} |\langle u, e_i \rangle|^2$$

Proof. By orthonormality,

$$u = \sum_{i=1}^{n} \langle u, e_i \rangle e_i; \quad v = \sum_{i=1}^{n} \langle v, e_i \rangle e_i$$

Hence, by sesquilinearity,

$$\langle u, v \rangle = \sum_{i=1}^{n} \langle u, e_i \rangle \overline{\langle v, e_i \rangle}$$

By taking $u = v$ we find

$$\|u\|^2 = \langle u, u \rangle = \sum_{i=1}^{n} |\langle u, e_i \rangle|^2$$

□

10.5. Gram–Schmidt orthogonalisation process

Theorem. Let V be an inner product space. Let $(v_i)_{i \in I}$ be a linearly independent family of vectors such that I is countable. Then there exists a family $(e_i)_{i \in I}$ of orthonormal vectors such that for all $k \geq 1$,

$$\langle v_1, \dots, v_k \rangle = \langle e_1, \dots, e_k \rangle$$

Proof. This proof is an explicit algorithm to compute the family (e_i), which will be computed by induction on k. For $k = 1$, take $e_1 = \frac{v_1}{\|v_1\|}$. Inductively, suppose $(e_1, \dots, e_k)$ satisfy the conditions as above. Then we will find a valid e_{k+1}. We define

$$e'_{k+1} = v_{k+1} - \sum_{i=1}^{k} \langle v_{k+1}, e_i \rangle e_i$$

This ensures that the inner product between e'_{k+1} and any basis vector e_j is zero, while maintaining the same span. Suppose $e'_{k+1} = 0$. Then, $v_{k+1} \in \langle e_1, \dots, e_k \rangle = \langle v_1, \dots, v_k \rangle$ which contradicts the fact that the family is free. Thus,

$$e_{k+1} = \frac{e'_{k+1}}{\left\| e'_{k+1} \right\|}$$

satisfies the requirements. □

Corollary. In finite-dimensional inner product spaces, there always exists an orthonormal basis. In particular, any orthonormal set of vectors can be extended into an orthonormal basis.

Remark. Let $A \in M_n(\mathbb{R})$ be a real-valued (or complex-valued) matrix. Then, the column vectors of A are orthogonal if $A^\intercal A = I$ (or $A^\intercal \overline{A} = I$ in the complex-valued case).

10.6. Orthogonality of matrices

Definition. A matrix $A \in M_n(\mathbb{R})$ is *orthogonal* if $A^\intercal A = I$, hence $A^\intercal = A^{-1}$. A matrix $A \in M_n(\mathbb{C})$ is *unitary* if $A^\intercal \overline{A} = I$, hence $A^\dagger = A^{-1}$.

Proposition. Let A be a square, non-singular, real-valued (or complex-valued) matrix. Then A can be written as $A = RT$ where T is upper triangular and R is orthogonal (or respectively unitary).

Proof. We apply the Gram–Schmidt process to the column vectors of the matrix. This gives us an orthonormal set of vectors, which gives an upper triangular matrix in this new basis. □

10.7. Orthogonal complement and projection

Definition. Let V be an inner product space. Let $V_1, V_2 \leq V$. Then we say that V is the *orthogonal direct sum* of V_1 and V_2 if $V = V_1 \oplus V_2$ and for all vectors $v_1 \in V_1, v_2 \in V_2$ we have $\langle v_1, v_2 \rangle = 0$. When this holds, we write $V = V_1 \overset{\perp}{\oplus} V_2$.

Remark. If for all vectors v_1, v_2 we have $\langle v_1, v_2 \rangle = 0$, then $v \in V_1 \cap V_2 \implies \|v\|^2 = 0 \implies v = 0$. Hence the sum is always direct if the subspaces are orthogonal.

Definition. Let V be an inner product space and let $W \leq V$. We define the *orthogonal* of W to be

$$W^\perp = \{v \in V : \forall w \in W, \langle v, w\rangle = 0\}$$

Lemma. For any inner product space V and any subspace $W \leq V$, we have $V = W \overset{\perp}{\oplus} W^\perp$.

Proof. First note that $W^\perp \leq V$. Then, if $w \in W$, $w \in W^\perp$, we have

$$\|w\|^2 = \langle w, w\rangle = 0$$

since they are orthogonal, so the vector subspaces intersect only in the zero vector. Now, we need to show $V = W + W^\perp$. Let $(e_1, \dots, e_k)$ be an orthonormal basis of W and extend it into $(e_1, \dots, e_k, e_{k+1}, \dots, e_n)$ which can be made orthonormal. Then, $(e_{k+1}, \dots, e_n)$ are elements of $W^\perp$ and form a basis. □

10.8. Projection maps

Definition. Suppose $V = U \oplus W$, so U is a complement of W in V. Then, we define $\pi : V \to W$ which maps $v = u + w$ to w. This is well defined, since the sum is direct. π is linear, and $\pi^2 = \pi$. We say that π is the *projection* operator onto W.

Remark. The map $\iota - \pi$ is the projection onto U, where ι is the identity map.

Lemma. Let V be an inner product space. Let $W \leq V$ be a finite-dimensional subspace. Let $(e_1, \dots, e_k)$ be an orthonormal basis for W. Then,

(i) $\pi(v) = \sum_{i=1}^k \langle v, e_i\rangle e_i$; and

(ii) for all $v \in V, w \in W$, $\|v - \pi(v)\| \leq \|v - w\|$ with equality if and only if $w = \pi(v)$, hence $\pi(v)$ is the point in W closest to v.

Proof. We define $\pi(v) = \sum_{i=1}^k \langle v, e_i\rangle e_i$. Since $W = \langle\{e_k\}\rangle$, $\pi(v) \in W$ for all $v \in V$. Then, $v = (v - \pi(v)) + \pi(v)$ has a term in W. We claim that the remaining term is in the orthogonal; $v - \pi(v) \in W^\perp$. Indeed, we must show $\langle v - \pi(v), w\rangle = 0$ for all $w \in W$. Equivalently, $\langle v - \pi(v), e_i\rangle = 0$ for all basis vectors e_i of W. We can explicitly compute

$$\langle v - \pi(v), e_j\rangle = \langle v, e_j\rangle - \left\langle \sum_{i=1}^k \langle v, e_i\rangle e_i, e_j \right\rangle = \langle v, e_j\rangle - \sum_{i=1}^k \langle v, e_i\rangle \langle e_i, e_j\rangle = \langle v, e_j\rangle - \langle v, e_j\rangle = 0$$

Hence, $v = (v - \pi(v)) + \pi(v)$ is a decomposition into W and $W^\perp$. Since $W \cap W^\perp = \{0\}$, we have $V = W \overset{\perp}{\oplus} W^\perp$. For the second part, let $v \in V$, $w \in W$, and we compute

$$\|v - w\|^2 = \left\| \underbrace{v - \pi(v)}_{\in W^\perp} + \underbrace{\pi(v) - w}_{\in W} \right\|^2 = \|v - \pi(v)\|^2 + \|\pi(v) - w\|^2 \geq \|v - \pi(v)\|^2$$

with equality if and only if $w = \pi(v)$. □

10.9. Adjoint maps

Definition. Let V, W be finite-dimensional inner product spaces. Let $\alpha \in L(V, W)$. Then there exists a unique linear map $\alpha^\star : W \to V$ such that for all $v, w \in V, W$,

$$\langle \alpha(v), w \rangle = \langle v, \alpha^\star(w) \rangle$$

Moreover, if B is an orthonormal basis of V, and C is an orthonormal basis of W, then

$$[\alpha^\star]_{C,B} = \left(\overline{[\alpha]_{B,C}}\right)^\mathsf{T}$$

Proof. Let $B = (v_1, \dots, v_n)$ and $C = (w_1, \dots, w_m)$ and $A = [\alpha]_{B,C} = (a_{ij})$. To check existence, we define $[\alpha^\star]_{C,B} = \overline{A}^\mathsf{T} = (c_{ij})$ and explicitly check the definition. By orthogonality,

$$\left\langle \alpha\left(\sum \lambda_i v_i\right), \sum \mu_j w_j \right\rangle = \left\langle \sum_{i,k} \lambda_i a_{ki} w_k, \sum_j \mu_j w_j \right\rangle = \sum_{i,j} \lambda_i a_{ji} \overline{\mu_j}$$

Then,

$$\left\langle \sum \lambda_i v_i, \alpha^\star\left(\sum \mu_j w_j\right) \right\rangle = \left\langle \sum_i \lambda_i v_i, \sum_{j,k} \mu_j c_{kj} v_k \right\rangle = \sum_{i,j} \lambda_i \overline{c_{ij} \mu_j}$$

So equality requires $\overline{c_{ij}} = a_{ji}$. Uniqueness follows from the above; the expansions are equivalent for any vector if and only if $\overline{c_{ij}} = a_{ji}$. □

Remark. The same notation, $\alpha^\star$, is used for the adjoint as just defined, and the dual map as defined before. If V, W are real product inner spaces and $\alpha \in L(V, W)$, we define $\psi : V \to V^\star$ such that $\psi(v)(x) = \langle x, v \rangle$ and similarly for W. Then we can check that the adjoint for α is given by the composition of ψ from $V \to V^\star$, then applying the dual, then applying the inverse of ψ for W.

10.10. Self-adjoint and isometric maps

Definition. Let V be a finite-dimensional inner product space, and α be an endomorphism of V. Let $\alpha^\star \in L(V)$ be the adjoint map. Then,

(i) the condition $\langle \alpha v, w \rangle = \langle v, \alpha w \rangle$ is equivalent to the condition $\alpha = \alpha^\star$, and such an α is called *self-adjoint* (for $\mathbb{R}$ we call such endomorphisms *symmetric*, and for $\mathbb{C}$ we call such endomorphisms Hermitian);

(ii) the condition $\langle \alpha v, \alpha w \rangle = \langle v, w \rangle$ is equivalent to the condition $\alpha^\star = \alpha^{-1}$, and such an α is called an *isometry* (for $\mathbb{R}$ it is called *orthogonal*, and for $\mathbb{C}$ it is called *unitary*).

Proposition. The conditions for isometries defined as above are equivalent.

Proof. Suppose $\langle \alpha v, \alpha w\rangle = \langle v, w\rangle$. Then for $v = w$, we find $\|\alpha v\|^2 = \|v\|^2$, so α preserves the norm. In particular, this implies $\ker \alpha = \{0\}$. Since α is an endomorphism and V is finite-dimensional, α is bijective. Then for all $v, w \in V$,

$$\langle v, \alpha^\star(w)\rangle = \langle \alpha v, w\rangle = \big\langle \alpha v, \alpha\big(\alpha^{-1}(w)\big)\big\rangle = \big\langle v, \alpha^{-1}(w)\big\rangle$$

Hence $\alpha^\star = \alpha^{-1}$. Conversely, if $\alpha^\star = \alpha^{-1}$ we have

$$\langle \alpha v, \alpha w\rangle = \langle v, \alpha^\star(\alpha w)\rangle = \langle v, w\rangle$$

as required. □

Remark. Using the polarisation identity, we can show that α is isometric if and only if for all $v \in V$, $\|\alpha(v)\| = \|v\|$.

Lemma. Let V be a finite-dimensional real (or complex) inner product space. Then for $\alpha \in L(V)$,

(i) α is self-adjoint if and only if for all orthonormal bases B of V, we have $[\alpha]_B$ is symmetric (or Hermitian);

(ii) α is an isometry if and only if for all orthonormal bases B of V, we have $[\alpha]_B$ is orthogonal (or unitary).

Proof. Let B be an orthonormal basis for V. Then we know $[\alpha^\star]_B = [\alpha]_B^\dagger$. We can then check that $[\alpha]_B^\dagger = [\alpha]_B$ and $[\alpha]_B^\dagger = [\alpha]_B^{-1}$ respectively. □

Definition. For $F = \mathbb{R}$, we define the *orthogonal group* of V by

$$O(V) = \{\alpha \in L(V) : \alpha \text{ is an isometry}\}$$

Note that $O(V)$ is bijective with the set of orthogonal bases of V. For $F = \mathbb{C}$, we define the *unitary group* of V by

$$U(V) = \{\alpha \in L(V) : \alpha \text{ is an isometry}\}$$

Again, note that $U(V)$ is bijective with the set of orthogonal bases of V.

10.11. Spectral theory for self-adjoint maps

Spectral theory is the study of the spectrum of operators. Recall that in finite-dimensional inner product spaces V, W, $\alpha \in L(V, W)$ yields the adjoint $\alpha^\star \in L(W, V)$ such that for all $v \in V, w \in W$, we have $\langle \alpha(v), w\rangle = \langle v, \alpha^\star(w)\rangle$.

Lemma. Let V be a finite-dimensional inner product space. Let $\alpha \in L(V)$ be a self-adjoint endomorphism. Then α has real eigenvalues, and eigenvectors of α with respect to different eigenvalues are orthogonal.

Proof. Suppose $\lambda \in \mathbb{C}$, $v \in V$ nonzero such that $\alpha(v) = \lambda v$. Then, $\langle \lambda v, v\rangle = \lambda \|v\|^2$ and also

$$\langle \alpha v, v\rangle = \langle v, \alpha v\rangle = \langle v, \lambda v\rangle = \overline{\lambda}\|v\|^2$$

Hence $\lambda = \overline{\lambda}$ since $v \neq 0$. Now, suppose $\mu \neq \lambda$ and $w \in V$ nonzero such that $\alpha(w) = \mu w$. Then,

$$\lambda \langle v, w\rangle = \langle \alpha v, w\rangle = \langle v, \alpha w\rangle = \overline{\mu}\langle v, w\rangle = \mu \langle v, w\rangle$$

So if $\lambda \neq \mu$ we must have $\langle v, w\rangle = 0$. □

Theorem (spectral theorem for self-adjoint maps)**.** Let V be a finite-dimensional inner product space. Let $\alpha \in L(V)$ be self-adjoint. Then V has an orthonormal basis of eigenvectors of α. Hence α is diagonalisable in an orthonormal basis.

Proof. We will consider induction on the dimension of V. Suppose $A = [\alpha]_B$ with respect to the fundamental basis B. By the fundamental theorem of algebra, we know that $\chi_A(\lambda)$ has a (complex) root. But since λ is an eigenvalue of α and α is self-adjoint, $\lambda \in \mathbb{R}$. Now, we choose an eigenvector $v_1 = V \setminus \{0\}$ such that $\alpha(v_1) = \lambda v_1$. We can set $\|v_1\| = 1$ by linearity. Let $U = \langle v_1\rangle^{\perp} \leq V$. We then observe that U is stable by α; $\alpha(U) \leq U$. Indeed, let $u \in U$. Then $\langle \alpha(u), v_1\rangle = \langle u, \alpha(v_1)\rangle = \lambda \langle u, v_1\rangle = 0$ by orthogonality. Hence $\alpha(u) \in U$. We can then restrict α to the domain U, and by induction we can then choose an orthonormal basis of eigenvectors for U. Since $V = \langle v_1\rangle \overset{\perp}{\oplus} U$ we have an orthonormal basis of eigenvectors for V when including v_1. □

Corollary. Let V be a finite-dimensional inner product space. Let $\alpha \in L(V)$ be self-adjoint. Then V is the orthogonal direct sum of the eigenspaces of α.

10.12. Spectral theory for unitary maps

Lemma. Let V be a complex inner product space. Let α be unitary, so $\alpha^\star = \alpha^{-1}$. Then all eigenvalues of α have unit norm. Eigenvectors corresponding to different eigenvalues are orthogonal.

Proof. Let $\lambda \in \mathbb{C}$, $v \in V \setminus \{0\}$ such that $\alpha(v) = \lambda v$. First, $\lambda \neq 0$ since α is invertible, and in particular $\ker \alpha = \{0\}$. Since $v = \lambda\alpha^{-1}(v)$, we can compute

$$\lambda \langle v, v\rangle = \langle \lambda v, v\rangle = \langle \alpha v, v\rangle = \left\langle v, \alpha^{-1}v\right\rangle = \left\langle v, \frac{1}{\lambda}v\right\rangle = \frac{1}{\overline{\lambda}}\langle v, v\rangle$$

Hence $(\lambda\overline{\lambda} - 1)\|v\|^2 = 0$ giving $|\lambda| = 1$. Further, suppose $\mu \in \mathbb{C}$ and $w \in V \setminus \{0\}$ such that $\alpha(w) = \mu w, \lambda \neq \mu$. Then

$$\lambda \langle v, w\rangle = \langle \lambda v, w\rangle = \langle \alpha v, w\rangle = \left\langle v, \alpha^{-1}w\right\rangle = \left\langle v, \frac{1}{\mu}w\right\rangle = \frac{1}{\overline{\mu}}\langle v, w\rangle = \mu \langle v, w\rangle$$

since $\mu\overline{\mu} = 1$. □

Theorem (spectral theorem for unitary maps)**.** Let V be a finite-dimensional complex inner product space. Let $\alpha \in L(V)$ be unitary. Then V has an orthonormal basis of eigenvectors of α. Hence α is diagonalisable in an orthonormal basis.

Proof. Let $A = [\alpha]_B$ where B is an orthonormal basis. Then $\chi_A(\lambda)$ has a complex root λ. As before, let $v_1 \neq 0$ such that $\alpha(v_1) = \lambda v_1$ and $\|v_1\| = 1$. Let $U = \langle v_1 \rangle^\perp$, and we claim that $\alpha(U) = U$. Indeed, let $u \in U$, and we find

$$\langle \alpha(u), v_1 \rangle = \langle u, \alpha^{-1}(v_1) \rangle = \left\langle u, \frac{1}{\lambda} v_1 \right\rangle = \frac{1}{\lambda} \langle u, v_1 \rangle$$

Since $\langle u, v_1 \rangle = 0$, we have $\alpha(u) \in U$. Hence, α restricted to U is a unitary endomorphism of U. By induction we have an orthonormal basis of eigenvectors of α for U and hence for V. □

Remark. We used the fact that the field is complex to find an eigenvalue. In general, a real-valued orthonormal matrix A giving $AA^\intercal = I$ cannot be diagonalised over $\mathbb{R}$. For example, consider

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

This is orthogonal and normalised. However, $\chi_A(\lambda) = 1 + 2\lambda\cos\theta + \lambda^2$ hence $\lambda = e^{\pm i\theta}$ which are complex in the general case.

10.13. Application to bilinear forms

We wish to extend the previous statements about spectral theory into statements about bilinear forms.

Corollary. Let $A \in M_n(\mathbb{R})$ (or $M_n(\mathbb{C})$) be a symmetric (or respectively Hermitian) matrix. Then there exists an orthonormal (respectively unitary) matrix P such that $P^\intercal AP$ (or $P^\dagger AP$) is diagonal with real-valued entries.

Proof. Using the standard inner product, $A \in L(F^n)$ is self-adjoint and hence there exists an orthonormal basis B of F^n such that A is diagonal in this basis. Let $P = (v_1, \dots, v_n)$ be the matrix of this basis. Since B is orthonormal, P is orthogonal (or unitary). The result follows from the fact that $P^{-1}AP$ is diagonal. The eigenvalues are real, hence the diagonal matrix is real. □

Corollary. Let V be a finite-dimensional real (or complex) inner product space. Let $\phi : V \times V \to F$ be a symmetric (or Hermitian) bilinear form. Then, there exists an orthonormal basis B of V such that $[\phi]_B$ is diagonal.

Proof. $A^\intercal = A$ (or respectively $A^\dagger = A$), hence there exists an orthogonal (respectively unitary) matrix P such that $P^{-1}AP$ is diagonal. Let (v_i) be the ith row of $P^{-1} = P^\intercal$ (or $P^\dagger$). Then $(v_1, \dots, v_n)$ is an orthonormal basis B of V such that $[\phi]_V$ is this diagonal matrix. □

Remark. The diagonal entries of $P^{-1}AP$ are the eigenvalues of A. Moreover, we can define the signature $s(\phi)$ to be the difference between the number of positive eigenvalues of A and the number of negative eigenvalues of A.

10.14. Simultaneous diagonalisation

Corollary. Let V be a finite-dimensional real (or complex) vector space. Let ϕ, ψ be symmetric (or Hermitian) bilinear forms on V. Let ϕ be positive definite. Then there exists a basis $(v_1, \dots, v_n)$ of V with respect to which ϕ and ψ are represented with a diagonal matrix.

Proof. Since ϕ is positive definite, V equipped with ϕ is a finite-dimensional inner product space where $\langle u, v\rangle = \phi(u, v)$. Hence, there exists a basis of V in which ψ is represented by a diagonal matrix, which is orthonormal with respect to the inner product defined by ϕ. Then, ϕ in this basis is represented by the identity matrix given by $\phi(v_i, v_j) = \langle v_i, v_j\rangle = \delta_{ij}$, which is diagonal. □

Corollary. Let $A, B \in M_n(\mathbb{R})$ (or $\mathbb{C}$) which are symmetric (or Hermitian). Suppose for all $x \neq 0$ we have $x^\dagger Ax > 0$, so A is positive definite. Then there exists an invertible matrix $Q \in M_n(\mathbb{R})$ (or $\mathbb{C}$) such that $Q^\intercal AQ$ (or $Q^\intercal A\overline{Q}$) and $Q^\intercal BQ$ (or $Q^\intercal B\overline{Q}$) are diagonal.

Proof. A induces a quadratic form $Q(x) = x^\dagger Ax$ which is positive definite by assumption. Similarly, $\widetilde{Q}(x) = x^\dagger Bx$ is induced by B. Then we can apply the previous corollary and change basis. □

VIII. Groups, Rings and Modules

Lectured in Lent 2022 by Dr. R. Zhou

A ring is a an algebraic structure with an addition and multiplication operation. Common examples of rings include $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, the Gaussian integers $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$, the quotient $\mathbb{Z}/n\mathbb{Z}$, and the set of polynomials with complex coefficients. We can study factorisation in a general ring, generalising the idea of factorising integers or polynomials. Certain rings, called unique factorisation domains, have the property like the integers that every nonzero non-invertible element can be expressed as a unique product of irreducibles (in $\mathbb{Z}$, the irreducibles are the prime numbers). This property, and many others, are studied in this course.

Modules are like vector spaces, but instead of being defined over a field, they are defined over an arbitrary ring. In particular, every vector space is a module, because every field is a ring. We use the theory built up over the course to prove that every n-dimensional complex matrix can be written in Jordan normal form.

Contents

1. Review of IA Groups 430
1.1. Definitions 430
1.2. Cosets 430
1.3. Order 431
1.4. Normality and quotients 431
1.5. Homomorphisms 431
1.6. Isomorphisms 431
2. Simple groups 432
2.1. Introduction 432
3. Group actions 433
3.1. Definitions 433
3.2. Cayley's theorem 434
3.3. Conjugation actions 434
4. Alternating groups 436
4.1. Conjugation in alternating groups 436
4.2. Simplicity of alternating groups 436
5. p-groups 438
5.1. p-groups 438
5.2. Sylow theorems 439
6. Matrix groups 441
6.1. Definitions 441
6.2. Möbius maps in modular arithmetic 441
6.3. Properties 443
7. Finite abelian groups 444
7.1. Products of cyclic groups 444
8. Rings 445
8.1. Definitions 445
8.2. Polynomials 445
8.3. Homomorphisms 447
8.4. Ideals 447
8.5. Quotients 448
8.6. Isomorphism theorems 449
8.7. Integral domains 451
8.8. Maximal ideals 453
8.9. Prime ideals 453
9. Factorisation in integral domains 455
9.1. Prime and irreducible elements 455

9.2. Principal ideal domains 456
9.3. Unique factorisation domains 458
9.4. Factorisation in polynomial rings 460
9.5. Eisenstein's criterion 463
10. Algebraic integers 464
10.1. Gaussian integers 464
10.2. Algebraic integers 466
11. Noetherian rings 468
11.1. Definition 468
11.2. Hilbert's basis theorem 468
12. Modules 470
12.1. Definitions 470
12.2. Finitely generated modules 471
12.3. Torsion 472
12.4. Direct sums 473
12.5. Free modules 473
12.6. Row and column operations 475
12.7. Smith normal form 476
12.8. The structure theorem 477
12.9. Primary decomposition theorem 479
12.10. Rational canonical form 480
12.11. Jordan normal form 482
12.12. Modules over principal ideal domains (non-examinable) 483

1. Review of IA Groups

This section contains material covered by IA Groups.

1.1. Definitions

A *group* is a pair $(G, \cdot)$ where G is a set and $\cdot : G \times G \to G$ is a binary operation on G, satisfying

- $a \cdot (b \cdot c) = (a \cdot b) \cdot c$;
- there exists $e \in G$ such that for all $g \in G$, we have $g \cdot e = e \cdot g = g$; and
- for all $g \in G$, there exists an inverse $h \in G$ such that $g \cdot h = h \cdot g = e$.

Remark. (i) Sometimes, such as in IA Groups, a closure axiom is also specified. However, this is implicit in the type definition of $\cdot$. In practice, this must normally be checked explicitly.

(ii) Additive and multiplicative notation will be used interchangeably. For additive notation, the inverse of g is denoted $-g$, and for multiplicative notation, the inverse is instead denoted g^{-1}. The identity element is sometimes denoted 0 in additive notation and 1 in multiplicative notation.

A subset $H \subseteq G$ is a *subgroup* of G, written $H \leq G$, if $h \cdot h' \in H$ for all $h, h' \in H$, and $(H, \cdot)$ is a group. The closure axiom must be checked, since we are restricting the definition of $\cdot$ to a smaller set.

Remark. A non-empty subset $H \subseteq G$ is a subgroup of G if and only if

$$a, b \in H \implies a \cdot b^{-1} \in H$$

An *abelian* group is a group such that $a \cdot b = b \cdot a$ for all a, b in the group. The *direct product* of two groups G, H, written $G \times H$, is the group over the Cartesian product $G \times H$ with operation $\cdot$ defined such that $(g_1, h_1) \cdot (g_2, h_2) = (g_1 \cdot_G g_2, h_1 \cdot_H h_2)$.

1.2. Cosets

Let $H \leq G$. Then, the *left cosets* of H in G are the sets gH for all $g \in G$. The set of left cosets partitions G. Each coset has the same cardinality as H. Lagrange's theorem states that if G is a finite group and $H \leq G$, we have $|G| = |H| \cdot [G : H]$, where $[G : H]$ is the number of left cosets of H in G. $[G : H]$ is known as the *index* of H in G. We can construct Lagrange's theorem analogously using right cosets. Hence, the index of a subgroup is independent of the choice of whether to use left or right cosets; the number of left cosets is equal to the number of right cosets.

1.3. Order

Let $g \in G$. If there exists $n \geq 1$ such that $g^n = 1$, then the least such n is the *order* of G. If no such n exists, we say that g has infinite order. If g has order d, then:

(i) $g^n = 1 \implies d \mid n$;

(ii) $\langle g \rangle = \{1, g, \dots, g^{d-1}\} \leq G$, and by Lagrange's theorem (if G is finite) $d \mid |G|$.

1.4. Normality and quotients

A subgroup $H \leq G$ is *normal*, written $H \trianglelefteq G$, if $g^{-1}Hg = H$ for all $g \in G$. In other words, H is preserved under conjugation over G. If $H \trianglelefteq G$, then the set G/H of left cosets of H in G forms the *quotient group*. The group action is defined by $g_1H \cdot g_2H = (g_1 \cdot g_2)H$. This can be shown to be well-defined.

1.5. Homomorphisms

Let G, H be groups. A function $\phi : G \to H$ is a *group homomorphism* if $\phi(g_1 \cdot_G g_2) = \phi(g_1) \cdot_H \phi(g_2)$ for all $g_1, g_2 \in G$. The *kernel* of ϕ is defined to be $\ker \phi = \{g \in G : \phi(g) = 1\}$, and the *image* of ϕ is $\operatorname{Im} \phi = \{\phi(g) : g \in G\}$. The kernel is a normal subgroup of G, and the image is a subgroup of H.

1.6. Isomorphisms

An *isomorphism* is a homomorphism that is bijective. This yields an inverse function, which is of course also an isomorphism. If $\varphi : G \to H$ is an isomorphism, we say that G and H are isomorphic, written $G \cong H$. Isomorphism is an equivalence relation. The isomorphism theorems are

(i) if $\varphi : G \to H$, then $G/\ker \varphi \cong \operatorname{Im} \varphi$;

(ii) if $H \leq G$ and $N \trianglelefteq G$, then $H \cap N \trianglelefteq H$ and $H/H \cap N \cong HN/N$;

(iii) if $N \leq M \leq G$ such that $N \trianglelefteq G$ and $M \trianglelefteq G$, then $M/N \trianglelefteq G/N$, and $G/N/M/N = G/M$.

2. Simple groups

2.1. Introduction

If $K \trianglelefteq G$, then studying the groups K and G/K give information about G itself. This approach is available only if G has nontrivial normal subgroups. It therefore makes sense to study groups with no normal subgroups, since they cannot be decomposed into simpler structures in this way.

Definition. A group G is *simple* if $\{1\}$ and G are its only normal subgroups.

By convention, we do not consider the trivial group to be a simple group. This is analogous to the fact that we do not consider one to be a prime.

Lemma. Let G be an abelian group. G is simple if and only if $G \cong C_p$ for some prime p.

Proof. Certainly C_p is simple by Lagrange's theorem. Conversely, since G is abelian, all subgroups are normal. Let $1 \neq g \in G$. Then $\langle g \rangle \trianglelefteq G$. Hence $\langle g \rangle = G$ by simplicity. If G is infinite, then $G \cong \mathbb{Z}$, which is not a simple group; $2\mathbb{Z} \triangleleft \mathbb{Z}$. Hence G is finite, so $G \cong C_{o(g)}$. If $o(g) = mn$ for $m, n \neq 1, p$, then $\langle g^m \rangle \leq G$, contradicting simplicity. □

Lemma. If G is a finite group, then G has a *composition series*

$$1 \cong G_0 \triangleleft G_1 \triangleleft \cdots \triangleleft G_n = G$$

where each quotient G_{i+1}/G_i is simple.

Remark. It is not the case that necessarily G_i be normal in G_{i+k} for $k \geq 2$.

Proof. We will consider an inductive step on $|G|$. If $|G| = 1$, then trivially $G = 1$. Conversely, if $|G| > 1$, let G_{n-1} be a normal subgroup of largest possible order not equal to $|G|$. Then, G/G_{n-1} exists, and is simple by the correspondence theorem. □

3. Group actions

3.1. Definitions

Definition. Let X be a set. Then $\mathrm{Sym}(X)$ is the group of permutations of X; that is, the group of all bijections of X to itself under composition. The identity can be written id or id_X.

Definition. A group G is a permutation group of degree n if $G \leq \mathrm{Sym}(X)$ where $|X| = n$.

Example. The symmetric group S_n is exactly equal to $\mathrm{Sym}(\{1, \dots, n\})$, so is a permutation group of order n. A_n is also a permutation group of order n, as it is a subgroup of S_n. D_{2n} is a permutation group of order n.

Definition. A *group action* of a group G on a set X is a function $\alpha : G \times X \to X$ satisfying

$$\alpha(e, x) = x; \quad \alpha(g_1 \cdot g_2, x) = \alpha(g_1, \alpha(g_2, x))$$

for all $g_1, g_2 \in G, x \in X$. The group action may be written $*$, defined by $g * x \equiv \alpha(g, x)$.

Proposition. An action of a group G on a set X is uniquely characterised by a group homomorphism $\varphi : G \to \mathrm{Sym}(X)$.

Proof. For all $g \in G$, we can define $\varphi_g : X \to X$ by $x \mapsto g * x$. Then, for all $x \in X$,

$$\varphi_{g_1 g_2}(x) = (g_1 g_2) * x = g_1 * (g_2 * x) = \varphi_{g_1}(\varphi_{g_2}(x))$$

Thus $\varphi_{g_1 g_2} = \varphi_{g_1} \circ \varphi_{g_2}$. In particular, $\varphi_g \circ \varphi_{g^{-1}} = \varphi_e$. We now define

$$\varphi : G \to \mathrm{Sym}(X); \quad \varphi(g) = \varphi_g \implies \varphi(g)(x) = g * x$$

This is a homomorphism.

Conversely, any group homomorphism $\varphi : G \to \mathrm{Sym}(X)$ induces a group action $*$ by $g * x = \varphi(g)$. This yields $e * x = \varphi(e)(x) = \mathrm{id}\, x = x$ and $(g_1 g_2) * x = \varphi(g_1 g_2)x = \varphi(g_1)\varphi(g_2)x = g_1 * (g_2 * x)$ as required. □

Definition. The homomorphism $\varphi : G \to \mathrm{Sym}(X)$ defined in the above proof is called a *permutation representation* of G.

Definition. Let $G \curvearrowright X$. Then,

(i) the orbit of $x \in X$ is $\mathrm{Orb}_G(x) = \{g * x : g \in G\} \subseteq X$;

(ii) the stabiliser of $x \in X$ is $G_x = \{g \in G : g * x = x\} \leq G$.

Theorem (Orbit-stabiliser theorem)**.** The orbit $\mathrm{Orb}_G(x)$ bijects with the set G/G_x of left cosets of G_x in G (which may not be a quotient group). In particular, if G is finite, we have

$$|G| = |\mathrm{Orb}(x)| \cdot |G_x|$$

Example. If G is the group of symmetries of a cube and we let X be the set of vertices in the cube, $G \curvearrowright X$. Here, for all $x \in X$, $|\mathrm{Orb}(x)| = 8$ and $|G_x| = 6$ (including reflections), hence $|G| = 48$.

Remark. Note that $\ker \varphi = \bigcap_{x \in X} G_x$. The kernel of the permutation representation φ is also referred to as the kernel of the group action itself. If the kernel is trivial the action is said to be *faithful*.

The orbits partition X. In particular, if there is exactly one orbit, the group action is said to be *transitive*.

Note that $G_{g*x} = gG_xg^{-1}$. Hence, if x, y lie in the same orbit, their stabilisers are conjugate.

Example. G acts on itself by left multiplication. This is known as the *left regular action*. The kernel is trivial, hence the action is faithful. The action is transitive, since for all $g_1, g_2 \in G$, the element $g_2g_1^{-1}$ maps g_1 to g_2.

3.2. Cayley's theorem

Theorem (Cayley's theorem)**.** Any finite group G is a permutation group of order $|G|$; it is isomorphic to a subgroup of $S_{|G|}$.

Example. Let $H \leq G$. Then $G \curvearrowright {}^{G}\!/_{H}$ by left multiplication, where ${}^{G}\!/_{H}$ is the set of left cosets of H in G. This is known as the *left coset action*. This action is transitive using the construction above for the left regular action. We have $\ker \varphi = \bigcap_{x \in G} xHx^{-1}$, which is the largest normal subgroup of G contained within H.

Theorem. Let G be a non-abelian simple group, and $H \leq G$ with index $n > 1$. Then $n \geq 5$ and G is isomorphic to a subgroup of A_n.

Proof. Let $G \curvearrowright X = {}^{G}\!/_{H}$ by left multiplication. Let $\varphi \colon G \to \mathrm{Sym}(X)$ be the permutation representation associated to this group action. Since G is simple, $\ker \varphi = 1$ or $\ker \varphi = G$. If $\ker \varphi = G$, then $\operatorname{Im} \varphi = \mathrm{id}$, which is a contradiction since G acts transitively on X, which has index greater than one. Thus $\ker \varphi = 1$, and $G \cong \operatorname{Im} \varphi \leq S_n$. Since $G \leq S_n$ and $A_n \lhd S_n$, the second isomorphism theorem shows that $G \cap A_n \lhd G$, and

$$ {}^{G}\!/_{G \cap A_n} \cong {}^{GA_n}\!/_{A_n} \leq {}^{S_n}\!/_{A_n} \cong C_2 $$

Since G is simple, $G \cap A_n = 1$ or $G \cap A_n = G$. If $G \cap A_n = 1$, then G is isomorphic to a subgroup of C_2, but this is false, since G is non-abelian. Hence $G \cap A_n = G$ so $G \leq A_n$. Finally, if $n \leq 4$ we can check manually that A_n is not simple; A_n has no non-abelian simple subgroups. □

3.3. Conjugation actions

Example. Let $G \curvearrowright G$ by conjugation, so $g * x = gxg^{-1}$. This is known as the *conjugation action*.

Definition. The orbit of the conjugation action is called the *conjugacy class* of a given element $x \in G$, written $\mathrm{ccl}_G(x)$. The stabiliser of the conjugation action is the set C_x of elements which commute with a given element x, called the *centraliser* of x in G. The kernel of φ is the set $Z(G)$ of elements which commute with all elements in x, which is the *centre* of G. This is always a normal subgroup.

Remark. $\varphi : G \to G$ satisfies

$$\varphi(g)(h_1h_2) = gh_1h_2g^{-1} = hh_1g^{-1}gh_2g^{-1} = \varphi(g)(h_1)\varphi(g)(h_2)$$

Hence $\varphi(g)$ is a group homomorphism for all g. It is also a bijection, hence $\varphi(g)$ is an isomorphism from $G \to G$.

Definition. An isomorphism from a group to itself is known as an *automorphism*. We define $\mathrm{Aut}(G)$ to be the set of all group automorphisms of a given group. This set is a group. Note, $\mathrm{Aut}(G) \leq \mathrm{Sym}(G)$, and the $\varphi : G \to \mathrm{Sym}(G)$ above has image in $\mathrm{Aut}(G)$.

Example. Let X be the set of subgroups of G. Then $G \curvearrowright X$ by conjugation: $g * H = gHg^{-1}$. The stabiliser of a subgroup H is $\left\{g \in G : gHg^{-1} = H\right\} = N_G(H)$, called the *normaliser* of H in G. The normaliser of H is the largest subgroup of G that contains H as a normal subgroup. In particular, $H \lhd G$ if and only if $N_G(H) = G$.

4. Alternating groups

4.1. Conjugation in alternating groups

We know that elements in S_n are conjugate if and only if they have the same cycle type. However, elements of A_n that are conjugate in S_n are not necessarily conjugate in A_n. Let $g \in A_n$. Then $C_{A_n}(g) = C_{S_n}(g) \cap A_n$. There are two possible cases.

- If there exists an odd permutation that commutes with g, then $2|C_{A_n}|(g) = |C_{S_n}|(g)$. By the orbit-stabiliser theorem, $|\mathrm{ccl}_{A_n}(g)| = |\mathrm{ccl}_{S_n}(g)|$.
- If there is no odd permutation that commutes with g, we have $|C_{A_n}|(g) = |C_{S_n}|(g)$. Similarly, $2|\mathrm{ccl}_{A_n}(g)| = |\mathrm{ccl}_{S_n}(g)|$.

Example. For $n = 5$, the product $(1\,2)(3\,4)$ commutes with $(1\,2)$, and $(1\,2\,3)$ commutes with $(4\,5)$. Both of these elements are odd. So the conjugacy classes of the above inside S_5 and A_5 are the same. However, $(1\,2\,3\,4\,5)$ does not commute with any odd permutation. Indeed, if that were true for some h, we would have

$$(1\,2\,3\,4\,5) = h(1\,2\,3\,4\,5)h^{-1} = (h(1)\,h(2)\,h(3)\,h(4)\,h(5))$$

Hence h must be a 5-cycle in the subgroup of A_5 generated by $(1\,2\,3\,4\,5)$.

We can then show that A_5 has conjugacy classes of size $1, 15, 20, 12, 12$. If $H \trianglelefteq A_5$, $|H|$ must be a sum of the sizes of the above conjugacy classes. By Lagrange's theorem, $|H|$ must divide 60. We can check explicitly that this is not possible unless $|H| = 1$ or $|H| = 60$. Hence A_5 is simple.

4.2. Simplicity of alternating groups

Lemma. A_n is generated by 3-cycles.

Proof. All elements of A_n are generated by an even number of transpositions. It therefore suffices to show that a product of two transpositions can be written as a product of 3-cycles. Explicitly,

$$(a\,b)(c\,d) = (a\,c\,b)(a\,c\,d); \quad (a\,b)(b\,c) = (a\,b\,c)$$

□

Lemma. If $n \geq 5$, all 3-cycles in A_n are conjugate (in A_n).

Proof. We claim that every 3-cycle is conjugate to $(1\,2\,3)$. If $(a\,b\,c)$ is a 3-cycle, we have $(a\,b\,c) = \sigma(1\,2\,3)\sigma^{-1}$ for some $\sigma \in S_n$. If $\sigma \in A_n$, then the proof is finished. Otherwise, $\sigma \mapsto \sigma(4\,5) \in A_n$ suffices, since $(4\,5)$ commutes with $(1\,2\,3)$. □

Theorem. A_n is simple for $n \geq 5$.

Proof. Suppose $1 \neq N \lhd A_n$. To disprove normality, it suffices to show that N contains a 3-cycle by the lemmas above, since the normality of N would imply N contains all 3-cycles and hence all elements of A_n.

Let $1 \neq \sigma \in N$, writing σ as a product of disjoint cycles.

(i) Suppose σ contains a cycle of length $r \geq 4$. Without loss of generality, let $\sigma = (1\,2\,3\,\ldots\,r)\tau$ where τ fixes $1, \ldots, r$. Now, let $\delta = (1\,2\,3)$. We have

$$\underbrace{\sigma^{-1}}_{\in N}\underbrace{\delta^{-1}\sigma\delta}_{\in N} = (r\,\ldots\,2\,1)(1\,3\,2)(1\,2\,\ldots\,r) = (2\,3\,r)$$

So N contains a 3-cycle.

(ii) Suppose σ contains two 3-cycles, which can be written without loss of generality as $(1\,2\,3)(4\,5\,6)\tau$. Let $\delta = (1\,2\,4)$, and then

$$\sigma^{-1}\delta^{-1}\sigma\delta = (1\,3\,2)(4\,6\,5)(1\,4\,2)(1\,2\,3)(4\,5\,6)(1\,2\,4) = (1\,2\,4\,3\,6)$$

Therefore, there exists an element of N which contains a cycle of length $5 \geq 4$. This reduces the problem to case (i).

(iii) Finally, suppose σ contains two 2-cycles, which will be written $(1\,2)(3\,4)\tau$. Then let $\delta = (1\,2\,3)$ and

$$\sigma^{-1}\delta^{-1}\sigma\delta = (1\,2)(3\,4)(1\,3\,2)(1\,2)(3\,4)(1\,2\,3) = (1\,4)(2\,3) = \pi$$

Let $\varepsilon = (2\,3\,5)$. Then

$$\underbrace{\pi^{-1}}_{\in N}\underbrace{\varepsilon^{-1}\pi\varepsilon}_{\in N} = (1\,4)(2\,3)(2\,5\,3)(1\,4)(2\,3)(2\,3\,5) = (2\,5\,3)$$

Thus N contains a 3-cycle.

There are now three remaining cases, where σ is a transposition, a 3-cycle, or a transposition composed with a 3-cycle. Note that the remaining cases containing transpositions cannot be elements of A_n. If σ is a 3-cycle, we already know A_n contains a 3-cycle, namely σ itself. □

5. p-groups

5.1. p-groups

Definition. Let p be a prime. A finite group G is a *p-group* if $|G| = p^n$ for $n \geq 1$.

Theorem. If G is a p-group, the centre $Z(G)$ is non-trivial.

Proof. For $g \in G$, due to the orbit-stabiliser theorem, $|\mathrm{ccl}(g)||C(g)| = p^n$. In particular, $|\mathrm{ccl}(g)|$ divides p^n, and they partition G. Since G is a disjoint union of conjugacy classes, modulo p we have

$$|G| \equiv \text{number of conjugacy classes of size } 1 \equiv 0 \implies |Z(G)| \equiv 0$$

Hence $Z(G)$ has order zero modulo p so it cannot be trivial. We can check this by noting that $g \in Z(G) \iff x^{-1}gx = g$ for all x, which is true if and only if $\mathrm{ccl}_G(g) = \{g\}$. □

Corollary. The only simple p-groups are the cyclic groups of order p.

Proof. Let G be a simple p-group. Since $Z(G)$ is a normal subgroup of G, we have $Z(G) = 1$ or $Z(G) = G$. But $Z(G)$ may not be trivial, so $Z(G) = G$. This implies G is abelian. The only abelian simple groups are cyclic of prime order, hence $G \cong C_p$. □

Corollary. Let G be a p-group of order p^n. Then G has a subgroup of order p^r for all $r \in \{0, \dots, n\}$.

Proof. Recall that any group G has a composition series $1 = G_1 \lhd \cdots \lhd G_N = G$ where each quotient G_{i+1}/G_i is simple. Since G is a p-group, G_{i+1}/G_i is also a p-group. Each successive quotient is an order p group by the previous corollary, so we have a composition series of nested subgroups of order p^r for all $r \in \{0, \dots, n\}$. □

Lemma. Let G be a group. If $G/Z(G)$ is cyclic, then G is abelian. This then implies that $Z(G) = G$, so in particular $G/Z(G) = 1$.

Proof. Let $gZ(G)$ be a generator for $G/Z(G)$. Then, each coset of $Z(G)$ in G is of the form $g^r Z(G)$ for some $r \in \mathbb{Z}$. Thus, $G = \{g^r z : r \in \mathbb{Z}, z \in Z(G)\}$. Now, we multiply two elements of this group and find

$$g^{r_1} z_1 g^{r_2} z_2 = g^{r_1+r_2} z_1 z_2 = g^{r_1+r_2} z_2 z_1 = z_2 z_1 g^{r_1+r_2} = g^{r_2} z_2 g^{r_1} z_1$$

So any two elements in G commute. □

Corollary. Any group of order p^2 is abelian.

Proof. Let G be a group of order p^2. Then $|Z(G)| \in \{1, p, p^2\}$. The centre cannot be trivial as proven above, since G is a p-group. If $|Z(G)| = p$, we have that $G/Z(G)$ is cyclic as it has order p. Applying the previous lemma, G is abelian. However, this is a contradiction since the centre of an abelian group is the group itself. If $|Z(G)| = p^2$ then $Z(G) = G$ and then G is clearly abelian. □

5.2. Sylow theorems

Theorem. Let G be a finite group of order $p^a m$ where p is a prime and p does not divide m. Then:

(i) The set $\mathrm{Syl}_p(G) = \{P \leq G : |P| = p^a\}$ of Sylow p-subgroups is non-empty.

(ii) All Sylow p-subgroups are conjugate.

(iii) The amount of Sylow p-subgroups $n_p = \left|\mathrm{Syl}_p(G)\right|$ satisfies

$$n_p \equiv 1 \mod p; \quad n_p \mid |G| \implies n_p \mid m$$

Proof. (i) Let Ω be the set of all subsets of G of order p^a. We can directly find

$$|\Omega| = \binom{p^a m}{p^a} = \frac{p^a m}{p^a} \cdot \frac{p^a m - 1}{p^a - 1} \cdots \frac{p^a m - p^a + 1}{1}$$

Note that for $0 \leq k < p^a$, the numbers $p^a m - k$ and $p^a - k$ are divisible by the same power of p. In particular, $|\Omega|$ is coprime to p.

Let $G \curvearrowright \Omega$ by left-multiplication, so $g * X = \{gx : x \in X\}$. For any $X \in \Omega$, the orbit-stabiliser theorem can be applied to show that

$$|G_X||\mathrm{orb}_G(X)| = |G| = p^a m$$

By the above, there must exist an orbit with size coprime to p, since orbits partition Ω. For such an X, $p^a \mid |G_X|$.

Conversely, note that if $g \in G$ and $x \in X$, then $g \in (gx^{-1}) * X$. Hence, we can consider

$$G = \bigcup_{g \in G} g * X = \bigcup_{Y \in \mathrm{orb}_G(X)} Y$$

Thus $|G| \leq |\mathrm{orb}_G(X)| \cdot |X|$, giving $|G_X| = \frac{|G|}{|\mathrm{orb}_G(X)|} \leq |X| = p^a$.

Combining with the above, we must have $|G_X| = p^a$. In other words, the stabiliser G_X is a Sylow p-subgroup of G.

(ii) We will prove a stronger result for this part of the proof. We claim that if P is a Sylow p-subgroup and $Q \leq G$ is a p-subgroup, then $Q \leq gPg^{-1}$ for some $g \in G$. Indeed, let Q act on the set of left cosets of P in G by left multiplication. By the orbit-stabiliser

theorem, each orbit has size which divides $|Q| = p^k$ for some k. Hence each orbit has size p^r for some r.

Since ${}^{G}\!/_{P}$ has size m, which is coprime to p, there must exist an orbit of size 1. Therefore there exists $g \in G$ such that $q * gP = gP$ for all $q \in Q$. Equivalently, $g^{-1}qg \in P$ for all $q \in Q$. This implies that $Q \leq gPg^{-1}$ as required. This then weakens to the second part of the Sylow theorems.

(iii) Let G act on $\mathrm{Syl}_p(G)$ by conjugation. Part (ii) of the Sylow theorems implies that this action is transitive. By the orbit-stabiliser theorem, $n_p = \left|\mathrm{Syl}_p(G)\right| \mid |G|$.

Let $P \in \mathrm{Syl}_p(G)$. Then let P act on $\mathrm{Syl}_p(G)$ by conjugation. Since P is a Sylow p-subgroup, the orbits of this action have size dividing $|P| = p^a$, so the size is some power of p. To show $n_p \equiv 1 \bmod p$, it suffices to show that $\{P\}$ is the unique orbit of size 1. Suppose $\{Q\}$ is another orbit of size 1, so Q is a Sylow p-subgroup which is preserved under conjugation by P. P normalises Q, so $P \leq N_G(Q)$. Notice that P and Q are both Sylow p-subgroups of $N_G(Q)$. By (ii), P and Q are conjugate inside $N_G(Q)$. Hence $P = Q$ since $Q \trianglelefteq N_G(Q)$. Thus, $|P|$ is the unique orbit of size 1, so $n_p \equiv 1 \bmod p$ as required.

□

Corollary. If $n_p = 1$, then there is only one Sylow p-subgroup, and it is normal.

Proof. Let $g \in G$ and $P \in \mathrm{Syl}_p(G)$. Then gPg^{-1} is a Sylow p-subgroup, hence $gPg^{-1} = P$. P is normal in G. □

Example. Let G be a group with $|G| = 1000 = 2^3 \cdot 5^3$. Here, $n_5 \equiv 1 \bmod 5$, and $n_5 \mid 8$, hence $n_5 = 1$. Thus the unique Sylow 5-subgroup is normal. Hence no group of order 1000 is simple.

Example. Let G be a group with $|G| = 132 = 2^2 \cdot 3 \cdot 11$. n_{11} satisfies $n_{11} \equiv 1 \bmod 11$ and $n_{11} \mid 12$, thus $n_{11} \in \{1, 12\}$. Suppose G is simple. Then $n_{11} = 12$. The amount of Sylow 3-subgroups satisfies $n_3 \equiv 1 \bmod 3$ and $n_3 \mid 44$ so $n_3 \in \{1, 4, 22\}$. Since G is simple, $n_3 \in \{4, 22\}$.

Suppose $n_3 = 4$. Then $G \curvearrowright \mathrm{Syl}_3(G)$ by conjugation, and this generates a group homomorphism $\varphi : G \to S_4$. But the kernel of this homomorphism is a normal subgroup of G, so $\ker \varphi$ is trivial or G itself. If $\ker \varphi = G$, then $\operatorname{Im} \varphi$ is trivial, contradicting Sylow's second theorem. If $\ker \varphi = 1$, then $\operatorname{Im} \varphi$ has order 132, which is impossible.

Thus $n_3 = 22$. This means that G has $22 \cdot (3 - 1) = 44$ elements of order 3, and further G has $12 \cdot (11 - 1) = 120$ elements of order 11. However, the sum of these two totals is more than the total of 132 elements, so this is a contradiction. Hence G is not simple.

6. Matrix groups

6.1. Definitions

Definition. Let F be a field, such as $\mathbb{C}$ or $\mathbb{Z}/p\mathbb{Z}$. Let $GL_n(F)$ be set of $n \times n$ invertible matrices over F, which is called the *general linear group*. Let $SL_n(F)$ be set of $n \times n$ matrices with determinant one over F, which is called the *special linear group*. $SL_n(F)$ is the kernel of the determinant homomorphism on $GL_n(F)$, so $SL_n(F) \triangleleft GL_n(F)$.

Let $Z \triangleleft GL_n(F)$ denote the subgroup of *scalar matrices*, the group of nonzero multiples of the identity. The group $PGL_n(F) = GL_n(F)/Z$ is called the *projective general linear group*. Let $PSL_n(F) = SL_n(F)/Z \cap SL_n(F)$. By the second isomorphism theorem, $PSL_n(F)$ is isomorphic to $Z \cdot SL_n(F)/Z$, which is a subgroup of $PGL_n(F)$.

Example. Consider the finite group $G = GL_n\left(\mathbb{Z}/p\mathbb{Z}\right)$. A list of n vectors in $\mathbb{Z}/p\mathbb{Z}$ are the columns of a matrix $A \in G$ if and only if the vectors are linearly independent. Hence, by considering dimensionality of subspaces generated by each column,

$$\begin{aligned} |G| &= (p^n - 1)(p^n - p)(p^n - p^2) \cdots (p^n - p^{n-1}) \\ &= p^{1+2+\cdots+(n-1)}(p^n - 1)(p^{n-1} - 1) \cdots (p - 1) \\ &= p^{\binom{n}{2}} \prod_{i=1}^{n} (p^i - 1) \end{aligned}$$

Hence the Sylow p-subgroups have size $p^{\binom{n}{2}}$. Let U be the set of upper triangular matrices with ones on the diagonal. This forms a Sylow p-subgroup of G, since there are $\binom{n}{2}$ entries in a given upper triangular matrix, and there are p choices for such an entry.

6.2. Möbius maps in modular arithmetic

Recall that $PGL_2(\mathbb{C})$ acts on $\mathbb{C} \cup \{\infty\}$ by Möbius transformations. Likewise, $PGL_2\left(\mathbb{Z}/p\mathbb{Z}\right)$ acts on $\mathbb{Z}/p\mathbb{Z} \cup \{\infty\}$ by Möbius transformations. For a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2\left(\mathbb{Z}/p\mathbb{Z}\right); \quad A : z \mapsto \frac{az + b}{cz + d}$$

Since the scalar matrices act trivially, we obtain an action on the projective general linear group instead of the general linear group. We can represent ∞ as an integer, say, p, for the purposes of constructing a permutation representation.

Lemma. The permutation representation $PGL_2\left(\mathbb{Z}/p\mathbb{Z}\right) \to S_{p+1}$ is injective (and is an isomorphism if $p = 2$ or $p = 3$).

Proof. Suppose that $\frac{az+b}{cz+d} = z$ for all $z \in \mathbb{Z}/p\mathbb{Z} \cup \{\infty\}$. Since $z = 0$, we have $b = 0$. Since $z = \infty$, we find $c = 0$. Thus the matrix is diagonal. Finally, since $z = 1$, $\frac{a}{d} = 1$ hence $a = d$. Thus the matrix is scalar. So the permutation representation from $PGL_2\left(\mathbb{Z}/p\mathbb{Z}\right)$ has trivial kernel, giving injectivity as required.

If $p = 2$ or $p = 3$ we can compute the orders of relevant groups manually and show that the permutation representation is an isomorphism. □

Lemma. Let p be an odd prime. Then

$$\left|PSL_2\left(\mathbb{Z}/p\mathbb{Z}\right)\right| = \frac{(p-1)p(p+1)}{2}$$

Proof. By the example above,

$$\left|GL_2\left(\mathbb{Z}/p\mathbb{Z}\right)\right| = p(p^2-1)(p-1)$$

The homomorphism $GL_2\left(\mathbb{Z}/p\mathbb{Z}\right) \to \left(\mathbb{Z}/p\mathbb{Z}\right)^\times$ given by the determinant is surjective. Since $SL_2\left(\mathbb{Z}/p\mathbb{Z}\right)$ is the kernel of this homomorphism, we have

$$\left|SL_2\left(\mathbb{Z}/p\mathbb{Z}\right)\right| = p(p-1)(p+1)$$

Now, if $\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}$ is an element of the special linear group, then $\lambda^2 \equiv 1 \bmod p$. Then, $p \mid (\lambda-1)(\lambda+1)$ hence $\lambda \equiv \pm 1 \bmod p$. Thus,

$$Z \cap SL_2\left(\mathbb{Z}/p\mathbb{Z}\right) = \{\pm 1\}$$

and the elements are distinct since $p > 2$. Hence the order of the projective special linear group is half the order of the special linear group as required. □

Example. Let $G = PSL_2\left(\mathbb{Z}/5\mathbb{Z}\right)$. Then by the previous lemma, $|G| = 60$. Let $G \curvearrowright \mathbb{Z}/5\mathbb{Z} \cup \{\infty\}$ by Möbius transformations. The permutation representation $\varphi : G \to \mathrm{Sym}(\{0,1,2,3,4,\infty\})$ is injective, since the permutation representation of $PGL_2\left(\mathbb{Z}/p\mathbb{Z}\right)$ is known to be injective by a previous lemma.

We claim that $\mathrm{Im}\,\varphi \subseteq A_6$. Let $\psi = \mathrm{sgn} \circ \varphi$. If we can show ψ is trivial, $\mathrm{Im}\,\varphi \subseteq A_6$. Let $h \in G$, and suppose h has order $2^n m$ for odd m. If $\psi(h^m) = 1$, then since ψ is a group homomorphism we have $\psi(h)^m = 1$ giving $\psi(h) \neq -1 \implies \psi(h) = 1$. So to show ψ is trivial, it suffices to show $\psi(g) = 1$ for all $g \in G$ with order a power of 2. By the second Sylow theorem, if g has order a power of 2, it is contained in a Sylow 2-subgroup. Then it suffices to show that $\psi(H) = 1$ for all Sylow 2-subgroups H. But since $\ker \psi$ is normal and all

Sylow 2-subgroups are conjugate, it suffices to show $\psi(H) = 1$ for a single Sylow 2-subgroup H. The Sylow 2-subgroup must have order 4. Hence consider

$$H = \left\langle \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}\{\pm I\}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\{\pm I\} \right\rangle$$

Both of these elements square to the identity element inside the projective special linear group. This generates a group of order 4 which is necessarily a Sylow 2-subgroup. We can explicitly compute the action of H on $\{0, 1, 2, 3, 4, \infty\}$.

$$\varphi\left(\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}\right) = (1\ 4)(2\ 3); \quad \varphi\left(\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right) = (0\ \infty)(1\ 4)$$

These are products of two transpositions, hence even permutations. Thus $\psi(H) = 1$, proving the claim that $G \leq A_6$. We can prove that for any $G \leq A_6$ of order 60, we have $G \cong A_5$; this is a question from the example sheets.

6.3. Properties

The following properties will not be proven in this course.

- $PSL_n\left(\mathbb{Z}/p\mathbb{Z}\right)$ is simple for all $n \geq 2$ and p prime, except where $n = 2$ and $p = 2, 3$. Such groups are called finite groups of *Lie type*.
- The smallest non-abelian simple groups are $A_5 \cong PSL_2\left(\mathbb{Z}/5\mathbb{Z}\right)$, then $PSL_2\left(\mathbb{Z}/7\mathbb{Z}\right) \cong GL_3\left(\mathbb{Z}/2\mathbb{Z}\right)$ which has order 168.

7. Finite abelian groups

7.1. Products of cyclic groups

Theorem. Every finite abelian group is isomorphic to a product of cyclic groups.

The proof for this theorem will be provided later in the course. Note that the isomorphism provided for by the theorem is not unique. An example of such behaviour is the following lemma.

Lemma. Let m, n be coprime integers. Then $C_m \times C_n \cong C_{mn}$.

Proof. Let g, h be generators of C_m and C_n. Then consider the element $(g,h)^k = (g^k, h^k)$, which has order mn. Thus $\langle (g,h) \rangle$ has order mn. So every element in $C_m \times C_n$ is expressible in this way, giving $\langle (g,h) \rangle = C_m \times C_n$. □

Corollary. Let G be a finite abelian group. Then $G \cong C_{n_1} \times \cdots \times C_{n_k}$ where each n_i is a power of a prime.

Proof. If $n = p_1 a^1 \cdots p^r a^r$ where the p_i are distinct primes, then applying the above lemma inductively gives C_n as a product of cyclic groups which have orders that are powers of primes. We can apply this to the theorem that every finite abelian group is isomorphic to a product of cyclic groups to find the result. □

Later, we will prove the following refinement of this theorem.

Theorem. Let G be a finite abelian group. Then $G \cong C_{d_1} \times \cdots \times C_{d_t}$ where $d_i \mid d_{i+1}$ for all i.

Remark. The integers $n_1, \ldots, n_k$ in the corollary above are unique up to ordering. The integers $d_1, \ldots, d_t$ are also unique, assuming that $d_1 > 1$. The proofs will be omitted.

Example. The abelian groups of order 8 are exactly C_8, $C_2 \times C_4$, and $C_2 \times C_2 \times C_2$. The abelian groups of order 12 are, using the corollary above, $C_2 \times C_2 \times C_3$, $C_4 \times C_3$, and using the above theorem, $C_2 \times C_6$ and C_{12}. However, $C_2 \times C_3 \cong C_6$ and $C_3 \times C_4 \cong C_{12}$, so the groups derived are isomorphic.

Definition. The *exponent* of a group G is the least integer $n \geq 1$ such that $g^n = 1$ for all $g \in G$. Equivalently, the exponent is the lowest common multiple of the orders of elements in G.

Example. The exponent of A_4 is $\operatorname{lcm}\{2, 3\} = 6$.

Corollary. Let G be a finite abelian group. Then G contains an element which has order equal to the exponent of G.

Proof. If $G \cong C_{d_1} \times \cdots \times C_{d_t}$ for $d_i \mid d_{i+1}$, every $g \in G$ has order dividing d_t. Hence the exponent is d_t, and we can choose a generator of C_{d_t} to obtain an element in G of the same order. □

8. Rings

8.1. Definitions

Definition. A *ring* is a triple $(R, +, \cdot)$ where R is a set and $+, \cdot$ are binary operations $R \times R \to R$, satisfying the following axioms.

(i) $(R, +)$ is an abelian group, and we will denote the identity element 0 and the inverse of x as $-x$;

(ii) $(R, \cdot)$ satisfies the group axioms except for the invertibility axiom, and we will denote the identity element 1 and the inverse of x as x^{-1} if it exists;

(iii) for all $x, y, z \in R$ we have $x \cdot (y + z) = x \cdot y + x \cdot z$ and $(y + z) \cdot x = y \cdot x + z \cdot x$.

If multiplication is commutative, we say that R is a *commutative* ring. In this course, we will study only commutative rings.

Remark. For all $x \in R$,

$$0 \cdot x = (0 + 0) \cdot x = 0 \cdot x + 0 \cdot x \implies 0 \cdot x = 0$$

Further,

$$0 = 0 \cdot x = (1 + -1) \cdot x = x + (-1 \cdot x) \implies -1 \cdot x = -x$$

Definition. A subset $S \subseteq R$ is a *subring*, denoted $S \leq R$, if $(S, +, \cdot)$ is a ring with the same identity elements.

Remark. It suffices to check the closure axioms for addition and multiplication; the other properties are inherited.

Example. $\mathbb{Z} \leq \mathbb{Q} \leq \mathbb{R} \leq \mathbb{C}$ are rings. The set $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ is a subring of $\mathbb{C}$. This is known as the ring of Gaussian integers. The set $\mathbb{Q}[\sqrt{2}] = \left\{a + b\sqrt{2} : a, b \in \mathbb{Q}\right\}$ is a subring of $\mathbb{R}$.

Example. The set $\mathbb{Z}/n\mathbb{Z}$ is a ring.

Example. Let R, S be rings. Then the *product* $R \times S$ is a ring under the binary operations

$$(a, b) + (c, d) = (a + c, b + d); \quad (a, b) \cdot (c, d) = (a \cdot c, b \cdot d)$$

The additive identity is $(0_R, 0_S)$ and the multiplicative identity is $(1_R, 1_S)$. Note that the subset $R \times \{0\}$ is preserved under addition and multiplication, so it is a ring, but it is not a subring because the multiplicative identity is different.

8.2. Polynomials

Definition. Let R be a ring. A *polynomial* f over R is an expression

$$f = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$$

for $a_i \in R$. The term X is a formal symbol, no substitution of X for a value will be made. We could alternatively define polynomials as finite sequences of terms in R. The *degree* of a polynomial f is the largest n such that $a_n \neq 0$. A degree-n polynomial is *monic* if $a_n = 1$. We write $R[X]$ for the set of all such polynomials over R. Let $g = b_0 + b_1X + \cdots + b_nX^n$. Then we define

$$f + g = (a_0 + b_0) + (a_1 + b_1)X + \cdots + (a_n + b_n)X^n; \quad f \cdot g = \sum_i \left(\sum_{j=0}^{i} a_j b_{i-j} \right) X^i$$

Then $(R[X], +, \cdot)$ is a ring. The identity elements are the constant polynomials 0 and 1. We can identify the ring R with the subring of $R[X]$ of constant polynomials.

Definition. An element $r \in R$ is a *unit* if r has a multiplicative inverse. The units in a ring, denoted $R^\times$, form an abelian group under multiplication. For instance, $\mathbb{Z}^\times = \{\pm 1\}$ and $\mathbb{Q}^\times = \mathbb{Q} \setminus \{0\}$.

Definition. A *field* is a ring where all nonzero elements are units and $0 \neq 1$.

Example. $\mathbb{Z}/n\mathbb{Z}$ is a field only if n is a prime.

Remark. If R is a ring such that $0 = 1$, then every element in the ring is equal to zero. Indeed, $x = 1 \cdot x = 0 \cdot x = 0$. Thus, the exclusion of rings with $0 = 1$ in the definition of a field simply excludes the trivial ring.

Proposition. Let $f, g \in R[X]$ such that the leading coefficient of g is a unit. Then there exist polynomials $q, r \in R[X]$ such that $f = qg + r$, where the degree of r is less than the degree of g.

Remark. This is the Euclidean algorithm for division, adapted to polynomial rings.

Proof. Let n be the degree of f and m be the degree of g, so

$$f = a_nX^n + \cdots + a_0; \quad g = b_mX^m + \cdots + b_0$$

By assumption, $b_m \in R^\times$. If $n < m$ then let $q = 0$ and $r = f$. Conversely, we have $n \geq m$. Consider the polynomial $f_1 = f - a_n b_m^{-1} g X^{n-m}$. This has degree at most $n - 1$. Hence, we can use induction on n to decompose f_1 as $f_1 = q_1 g + r$. Thus $f = (q_1 + a_n b_m^{-1} X^{n-m})g + r$ as required. □

Remark. If R is a field, then every nonzero element of R is a unit. Therefore, the above algorithm can be applied for all polynomials g unless g is the constant polynomial zero.

Example. Let R be a ring and X be a set. Then the set of functions $X \to R$ is a ring under

$$(f + g)(x) = f(x) + g(x); \quad (f \cdot g)(x) = f(x) \cdot g(x)$$

The set of continuous functions $\mathbb{R} \to \mathbb{R}$ is a subring of the ring of all functions $\mathbb{R} \to \mathbb{R}$, since they are closed under addition and multiplication. The set of polynomial functions $\mathbb{R} \to \mathbb{R}$ is also a subring, and we can identify this with the ring $\mathbb{R}[X]$.

Example. Let R be a ring. Then the *power series ring* $R[\![X]\!]$ is the set of power series on X. This is defined similarly to the polynomial ring, but we permit infinitely many nonzero elements in the expansion. The power series is defined formally; we cannot actually carry out infinitely many additions in an arbitrary ring. We instead consider the power series as a sequence of numbers.

Example. Let R be a ring. Then the ring of *Laurent polynomials* is $R[X, X^{-1}]$ with the restriction that $a_i \neq 0$ for finitely many i.

8.3. Homomorphisms

Definition. Let R and S be rings. A function $\varphi : R \to S$ is a *ring homomorphism* if

(i) $\varphi(r_1 + r_2) = \varphi(r_1) + \varphi(r_2)$;

(ii) $\varphi(r_1 \cdot r_2) = \varphi(r_1) \cdot \varphi(r_2)$;

(iii) $\varphi(1_R) = 1_S$.

We can derive that $\varphi(0_R) = 0_S$ from (i).

A ring homomorphism is an *isomorphism* if it is bijective. The *kernel* of a ring homomorphism is $\ker \varphi = \{r \in R : \varphi(r) = 0\}$.

Lemma. Let R, S be rings. Then a ring homomorphism $\varphi : R \to S$ is injective if and only if $\ker \varphi = \{0\}$.

Proof. Let $\varphi : (R, +) \to (S, +)$ be the induced group homomorphism on addition. The result then follows from the corresponding fact about group homomorphisms. □

8.4. Ideals

Definition. A subset $I \subseteq R$ is an *ideal*, written $I \trianglelefteq R$, if

(i) I is a subgroup of $(R, +)$;

(ii) if $r \in R$ and $x \in I$, then $rx \in I$.

We say that an ideal is *proper* if $I \neq R$.

Lemma. Let $\varphi : R \to S$ be a ring homomorphism. Then $\ker \varphi$ is an ideal of R.

Proof. We know that $\ker \varphi$ is a subgroup by the equivalent fact from groups. If $r \in R$ and $x \in \ker \varphi$, then

$$\varphi(rx) = \varphi(r)\varphi(x) = \varphi(r) \cdot 0 = 0$$

Hence $rx \in \ker \varphi$. □

Remark. If I contains a unit, then the multiplicative identity lies in I. Then all elements lie in I. In particular, if I is a proper ideal, $1 \notin I$. Hence a proper ideal I is not a subring of R.

Lemma. The ideals in $\mathbb{Z}$ are precisely the subsets of the form $n\mathbb{Z}$ for any $n = 0, 1, 2, \ldots$.

Proof. First, we can check directly that any subset of the form $n\mathbb{Z}$ is an ideal. Now, let I be any nonzero ideal of $\mathbb{Z}$ and let n be the smallest positive element. Then $n\mathbb{Z} \subseteq I$. Let $m \in I$. Then by the Euclidean algorithm, $m = qn + r$ for $q, r \in \mathbb{Z}$ and $r \in \{0, 1, \ldots, n-1\}$. Then $r = m - qn$. We know $qn \in I$ since $n \in I$, so $r \in I$. If $r \neq 0$, this contradicts the minimality of n as chosen above. So $I = n\mathbb{Z}$ exactly. □

Definition. For an element $a \in R$, we write (a) to denote the subset of R given by multiples of a; that is, $(a) = \{ra : r \in R\}$. This is an ideal, known as the ideal *generated by* a. More generally, if $a_1, \ldots, a_n \in R$, then $(a_1, \ldots, a_n)$ is the set of elements in R given by linear combinations of the a_i. This is also an ideal.

Definition. Let $I \trianglelefteq R$. Then I is *principal* if there exists some $a \in R$ such that $I = (a)$.

8.5. Quotients

Theorem. Let $I \trianglelefteq R$. Then the set R/I of cosets of I in $(R, +)$ forms the *quotient ring* under the operations

$$(r_1 + I) + (r_2 + I) = (r_1 + r_2) + I; \quad (r_1 + I) \cdot (r_2 + I) = (r_1 \cdot r_2) + I$$

This ring has the identity elements

$$0_{R/I} = 0_R + I; \quad 1_{R/I} = 1_R + I$$

Further, the map $R \to R/I$ defined by $r \mapsto r + I$ is a ring homomorphism called the *quotient map*. The kernel of the quotient map is I. Hence any ideal is the kernel of some homomorphism.

Proof. From the analogous result from groups, the addition defined on the set of cosets yields the group $\left(R/I, +\right)$. If $r_1 + I = r_1' + I$ and $r_2 + I = r_2' + I$, then $r_1' = r_1 + a_1$ and $r_2' = r_2 + a_2$ for some $a_1, a_2 \in I$. Then

$$r_1' r_2' = (r_1 + a_1)(r_2 + a_2) = r_1 r_2 + a_1 r_2 + r_1 a_2 + a_1 a_2$$

Hence $(r_1' r_2') + I = (r_1 r_2) + I$. The remainder of the proof is trivial. □

Example. In the integers $\mathbb{Z}$, the ideals are $n\mathbb{Z}$. Hence we can form the quotient ring $\mathbb{Z}/n\mathbb{Z}$. The ring $\mathbb{Z}/n\mathbb{Z}$ has elements $n\mathbb{Z}, 1 + n\mathbb{Z}, \ldots, (n-1) + n\mathbb{Z}$. Addition and multiplication behave like in modular arithmetic modulo n.

Example. Consider the ideal (X) inside the polynomial ring $\mathbb{C}[X]$. This ideal is the set of polynomials with zero constant term. Let $f(X) = a_n X^n + \cdots + a_0$ be an arbitrary element of $\mathbb{C}[X]$. Then $f(X) + X = a_0 + X$. Thus, there exists a bijection between $\mathbb{C}[X]/(X)$ and $\mathbb{C}$, defined by $f(x)+(X) \mapsto f(0)$, with inverse $a \mapsto a+(X)$. This bijection is a ring homomorphism, hence $\mathbb{C}[X]/(X) \cong \mathbb{C}$.

Example. Consider $(X^2+1) \lhd \mathbb{R}[X]$. For $f(X) = a_n X^n + \cdots + a_0 \in \mathbb{R}[X]$, we can apply the Euclidean algorithm to write $f(X)$ as $q(X)(X^2+1) + r(X)$ where the degree of r is less than two. Hence $r(X) = a + bX$ for some real numbers a and b. Thus, any element of $\mathbb{R}[X]/(X^2+1)$ can be written $a + bX + (X^2+1)$. Suppose a coset can be represented by two representatives: $a + bX + (X^2+1) = a' + b'X + (X^2+1)$. Then,

$$a + bX - a' - b'X = (a - a') - (b - b')X = g(X)(X^2+1)$$

Hence $g(X) = 0$, giving $a - a' = 0$ and $b - b' = 0$. Hence the coset representative is unique. Consider the bijection φ between this quotient ring and the complex numbers given by $a + bX + (X^2+1) \mapsto a + bi$. We can show that φ is a ring homomorphism. Indeed, it preserves addition, and $1 + (X^2+1) \mapsto 1$, so it suffices to check that multiplication is preserved.

$$\begin{aligned}
\varphi((a + bX + (X^2+1)) \cdot (c + dX + (X^2+1))) &= \varphi((a+bX)(c+dX) + (X^2+1)) \\
&= \varphi(ac + (ad+bc)X + bd(X^2+1) - bd + (X^2+1)) \\
&= \varphi(ac - bd + (ad+bc)X + (X^2+1)) \\
&= ac - bd + (ad+bc)i \\
&= (a+bi)(c+di) \\
&= \varphi((a+bX) + (X^2+1))\varphi((c+dX) + (X^2+1))
\end{aligned}$$

Thus $\mathbb{R}[X]/(X^2+1) \cong \mathbb{C}$.

8.6. Isomorphism theorems

Theorem (first isomorphism theorem)**.** Let $\varphi : R \to S$ be a ring homomorphism. Then,

$$\ker\varphi \lhd R; \quad \operatorname{Im}\varphi \leq S; \quad R/\ker\varphi \cong \operatorname{Im}\varphi$$

Proof. We have $\ker\varphi \lhd R$ from above. We know that $\operatorname{Im}\varphi \leq (S, +)$. Now we show that $\operatorname{Im}\varphi$ is closed under multiplication.

$$\varphi(r_1)\varphi(r_2) = \varphi(r_1 r_2) \in \operatorname{Im}\varphi$$

Finally,

$$1_S = \varphi(1_R) \in \operatorname{Im}\varphi$$

Hence $\operatorname{Im}\varphi$ is a subring of S. Let $K = \ker\varphi$. Then, we define $\Phi : {}^{R}\!/_{K} \to \operatorname{Im}\varphi$ by $r + K \mapsto \varphi(r)$. By appealing to the first isomorphism theorem from groups, this is well-defined, a bijection, and a group homomorphism under addition. It therefore suffices to show that Φ preserves multiplication and maps the multiplicative identities to each other.

$$\Phi(1_R + K) = \varphi(1_R) = 1_S; \quad \Phi((r_1 + K)(r_2 + K)) = \Phi(r_1 r_2 + K) = \varphi(r_1 r_2) = \varphi(r_1)\varphi(r_2)$$

The result follows as required. □

Theorem (second isomorphism theorem)**.** Let $R \leq S$ and $J \lhd S$. Then,

$$R \cap J \lhd R; \quad R + J = \{r + a : r \in R, a \in J\} \leq S; \quad {}^{R}\!/_{R \cap J} \cong {}^{R+J}\!/_{J} \leq {}^{S}\!/_{J}$$

Proof. By the second isomorphism theorem for groups, $R+J \leq (S,+)$. Further, $1_S = 1_S + 0_S$, and since R is a subring, $1_S + 0_S \in R + J$ hence $1_S \in R \cap J$. If $r_1, r_2 \in R$ and $a_1, a_2 \in J$, we have

$$(r_1 + a_1)(r_2 + a_2) = \underbrace{r_1 r_2}_{\in R} + \underbrace{r_1 a_2}_{\in J} + \underbrace{r_2 a_1}_{\in J} + \underbrace{r_2 a_2}_{\in J} \in R + J$$

Hence $R + J$ is closed under multiplication, giving $R + J \leq S$.

Let $\varphi : R \to {}^{S}\!/_{J}$ be defined by $r \mapsto r + J$. This is a ring homomorphism, since it is the composite of the inclusion homomorphism $R \subseteq S$ and the quotient map $S \to {}^{S}\!/_{J}$. The kernel of φ is the set $\{r \in R : r + J = J\} = R \cap J$. Since this is the kernel of a ring homomorphism, $R \cap J$ is an ideal in R. The image of φ is $\{r + J \mid r \in R\} = {}^{R+J}\!/_{J} \leq {}^{S}\!/_{J}$. By the first isomorphism theorem, ${}^{R}\!/_{R \cap J} \cong {}^{R+J}\!/_{J}$ as required. □

Remark. If $I \lhd R$, there exists a bijection between ideals in ${}^{R}\!/_{I}$ and the ideals of R containing I. Explicitly,

$$K \mapsto \{r \in R \mid r + I \in K\}; \quad J \mapsto {}^{J}\!/_{I}$$

Theorem (third isomorphism theorem)**.** Let $I \lhd R$ and $J \lhd R$ with $I \subseteq J$. Then,

$${}^{J}\!/_{I} \lhd {}^{R}\!/_{I}; \quad {}^{R/I}\!/_{J/I} \cong {}^{R}\!/_{J}$$

Proof. Let $\varphi : {}^{R}\!/_{I} \to {}^{R}\!/_{J}$ defined by $r + I \mapsto r + J$. We can check that this is a surjective ring homomorphism by considering the third isomorphism theorem for groups. Its kernel is $\{r + I : r \in J\} = {}^{J}\!/_{I}$, which is an ideal in ${}^{R}\!/_{I}$, and we conclude by use of the first isomorphism theorem. □

Remark. ${}^{J}\!/_{I}$ is not a quotient ring, since J is not in general a ring; this notation should be interpreted as a set of cosets.

Example. Consider the surjective ring homomorphism $\varphi : \mathbb{R}[X] \to \mathbb{C}$ which is defined by

$$f = \sum_n a_n X^n \mapsto f(i) = \sum_n a_n i^n$$

Its kernel can be found by the Euclidean algorithm, yielding $\ker \varphi = (X^2 + 1)$. Applying the first isomorphism theorem, we immediately find $\mathbb{R}[X]/(X^2+1) \cong \mathbb{C}$.

Example. Let R be a ring. Then there exists a unique ring homomorphism $i : \mathbb{Z} \to R$. Indeed, we must have

$$0_{\mathbb{Z}} \mapsto 0_R; \quad 1_{\mathbb{Z}} \mapsto 1_R$$

This inductively defines

$$n \mapsto \underbrace{1_R + \cdots + 1_R}_{n \text{ times}}$$

The negative integers are also uniquely defined, since any ring homomorphism is a group homomorphism.

$$-n \mapsto -(\underbrace{1_R + \cdots + 1_R}_{n \text{ times}})$$

We can show that any such construction is a ring homomorphism as required. Then, the kernel of the ring homomorphism is an ideal of $\mathbb{Z}$, hence it is $n\mathbb{Z}$ for some n. Hence, by the first isomorphism theorem, any ring contains a copy of $\mathbb{Z}/n\mathbb{Z}$, since it is isomorphic to the image of i. If $n = 0$, then the ring contains a copy of $\mathbb{Z}$ itself, and if $n = 1$, then the ring is trivial since $0 = 1$. The number n is known as the *characteristic* of R.

For example, $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ have characteristic zero. The rings $\mathbb{Z}/p\mathbb{Z}, \mathbb{Z}/p\mathbb{Z}[X]$ have characteristic p.

8.7. Integral domains

Definition. An *integral domain* is a ring R with $0 \neq 1$ such that for all $a, b \in R$, $ab = 0$ implies $a = 0$ or $b = 0$. A *zero divisor* in a ring R is a nonzero element $a \in R$ such that $ab = 0$ for some nonzero $b \in R$. A ring is an integral domain if and only if it has no zero divisors.

Example. All fields are integral domains. Any subring of an integral domain is an integral domain. For instance, $\mathbb{Z}[i] \leq \mathbb{C}$ is an integral domain.

Example. The ring $\mathbb{Z} \times \mathbb{Z}$ is not an integral domain. Indeed, $(1, 0) \cdot (0, 1) = (0, 0)$.

Lemma. Let R be an integral domain. Then $R[X]$ is an integral domain.

Proof. We will show that any two nonzero elements produce a nonzero element. In particular, let

$$f = \sum_n a_n X^n; \quad g = \sum_n b_n X^n$$

Since these are nonzero, the leading coefficients a_n and b_m are nonzero. Here, the leading term of the product fg has form $a_n b_m X^{n+m}$. Since R is an integral domain, $a_n b_m \neq 0$, so fg is nonzero. Further, the degree of fg is $n+m$, the sum of the degrees of f and g. □

Lemma. Let R be an integral domain, and $f \neq 0$ be a nonzero polynomial in $R[X]$. We define $\text{roots}(f) = \{a \in R : f(a) = 0\}$. Then $|\text{roots}(f)| \leq \deg(f)$.

Proof. Exercise on the example sheets. □

Theorem. Let F be a field. Then any finite subgroup G of $(F^\times, \cdot)$ is cyclic.

Proof. G is a finite abelian group. If G is not cyclic, we can apply a previous structure theorem for finite abelian groups to show that there exists $H \leq G$ such that $H \cong C_{d_1} \times C_{d_1}$ for some integer $d_1 \geq 2$. The polynomial $f(X) = X^{d_1} - 1 \in F[X]$ has degree d_1, but has at least d_1^2 roots, since any element of H is a root. This contradicts the previous lemma. □

Example. $\left(\mathbb{Z}/p\mathbb{Z}\right)^\times$ is cyclic.

Proposition. Any finite integral domain is a field.

Proof. Let $0 \neq a \in R$, where R is an integral domain. Consider the map $\varphi : R \to R$ given by $x \mapsto ax$. If $\varphi(x) = \varphi(y)$, then $a(x-y) = 0$. But $a \neq 0$, hence $x - y = 0$. Hence φ is injective. Since R is finite, φ is a bijection, hence it has an inverse φ^{-1}, which yields the multiplicative inverse of a by considering $\varphi^{-1}(a)$. This may be repeated for all a. □

Theorem. Any integral domain R is a subring of a field F, and every element of F can be written in the form ab^{-1} where $a, b \in R$ and $b \neq 0$. Such a field F is called the *field of fractions* of R.

Proof. Consider the set $S = \{(a,b) \in R : b \neq 0\}$. We can define an equivalence relation

$$(a,b) \sim (c,d) \iff ad = bc$$

This is reflexive and commutative. We can show directly that it is transitive.

$$\begin{aligned}(a,b) \sim (c,d) \sim (e,f) &\implies ad = bc;\ cf = de \\ &\implies adf = bcf = bde \\ &\implies af = be \\ &\implies (a,b) \sim (e,f)\end{aligned}$$

Hence $\sim$ is indeed an equivalence relation. Now, let $F = S/\sim$, and we write $\frac{a}{b}$ for the class $[(a,b)]$. We define the ring operations

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}; \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

These can be shown to be well-defined. Thus, F is a ring with identities $0_F = \frac{0_R}{1_R}$ and $1_F = \frac{1_R}{1_R}$. If $\frac{a}{b} \neq 0_F$, then $a \neq 0$. Thus, $\frac{b}{a}$ exists, and $\frac{a}{b} \cdot \frac{b}{a} = 1$. Hence F is a field.

We can identify R with the subring of F given by $\frac{r}{1}$ for all $r \in R$. This is clearly isomorphic to R. Further, any element of F can be written as $\frac{a}{b} = ab^{-1}$ as required. □

This is analogous to the construction of the rationals using the integers.

Example. Consider $\mathbb{C}[X]$. This has field of fractions $\mathbb{C}(X)$, called the field of *rational functions* in X.

8.8. Maximal ideals

Definition. An ideal $I \lhd R$ is *maximal* if $I \neq R$ and, if $I \subseteq J \lhd R$, we have $J = I$ or $J = R$.

Lemma. A nonzero ring R is a field if and only if its only ideals are zero or R.

Proof. Suppose R is a field. If $0 \neq I \lhd R$, then I contains a nonzero element, which is a unit since R is a field. Hence $I = R$.

Now, suppose a ring R has ideals that are only zero or R. If $0 \neq x \in R$, consider (x). This is nonzero since it contains x. By assumption, $(x) = R$. Thus, the element 1 lies in (x). Hence, there exists $y \in R$ such that $xy = 1$, and hence this y is the multiplicative inverse as required. □

Proposition. Let $I \lhd R$. Then I is maximal if and only if R/I is a field.

Proof. R/I is a field if and only if its ideals are either zero, denoted I/I, or R/I itself. By the correspondence theorem, I and R are the only ideals in R which contain I. Equivalently, $I \lhd R$ is maximal. □

8.9. Prime ideals

Definition. An ideal $I \lhd R$ is *prime* if $I \neq R$ and, for all $a, b \in R$ such that $ab \in I$, we have $a \in I$ or $b \in I$.

Example. The ideals in the integers are (n) for some $n \geq 0$. $n\mathbb{Z}$ is a prime ideal if and only if n is prime or zero. The case for $n = 0$ is trivial. If $n \neq 0$ we can use the property that $p \mid ab$ implies either $p \mid a$ or $p \mid b$. Conversely, if n is composite, we can write $n = uv$ for $u, v > 1$. Then $uv \in n\mathbb{Z}$ but $u, v \notin n\mathbb{Z}$.

Proposition. Let $I \lhd R$. Then I is prime if and only if R/I is an integral domain.

Proof. If I is prime, then for all $ab \in I$ we have $a \in I$ or $b \in I$. Equivalently, for all $a + I, b + I \in {}^{R}\!/_{I}$, we have $(a + I)(b + I) = 0 + I$ if $a + I = 0 + I$ or $b + I = 0 + I$. This is the definition of an integral domain. □

Remark. If I is a maximal ideal, then ${}^{R}\!/_{I}$ is a field. A field is an integral domain. Hence any maximal ideal is prime.

Remark. If the characteristic of a ring is n, then ${}^{\mathbb{Z}}\!/_{n\mathbb{Z}} \leq R$. In particular, if R is an integral domain, then ${}^{\mathbb{Z}}\!/_{n\mathbb{Z}}$ must be an integral domain. Equivalently, $n\mathbb{Z} \lhd \mathbb{Z}$ is a prime ideal. Hence n is zero or prime. Thus, in an integral domain, the characteristic must either be zero or prime. A field always has a characteristic, which is either zero (in which case it contains $\mathbb{Z}$ and hence $\mathbb{Q}$) or prime (in which case it contains ${}^{\mathbb{Z}}\!/_{p\mathbb{Z}} = \mathbb{F}_p$ which is already a field).

9. Factorisation in integral domains

In this section, let R be an integral domain.

9.1. Prime and irreducible elements

Recall that an element $a \in R$ is a unit if it has a multiplicative inverse in R. Equivalently, an element a is a unit if and only if $(a) = R$. Indeed, if $(a) = R$, then $1 \in (a)$ hence there exists a multiple of a equal to 1. We denote the set of units in R by $R^\times$.

Definition. An element $a \in R$ *divides* $b \in R$, written $a \mid b$, if there exists $c \in R$ such that $b = ac$. Equivalently, $(b) \subseteq (a)$.

Two elements $a, b \in R$ are *associates* if $a = bc$ where c is a unit. Informally, the two elements differ by multiplication by a unit. Equivalently, $(a) = (b)$.

Definition. An element $r \in R$ is *irreducible* if r is not zero or a unit, and $r = ab$ implies a is a unit or b is a unit. An element $r \in R$ is *prime* if r is not zero or a unit, and $r \mid ab$ implies $r \mid a$ or $r \mid b$.

Remark. These properties depend on the ambient ring R; for instance, 2 is prime and irreducible in $\mathbb{Z}$, but neither prime nor irreducible in $\mathbb{Q}$. The polynomial $2X$ is irreducible in $\mathbb{Q}[X]$, but not in $\mathbb{Z}[X]$.

Lemma. $(r) \lhd R$ is a prime ideal if and only if $r = 0$ or r is prime.

Proof. Suppose (r) is a prime ideal with $r \neq 0$. Since prime ideals are proper, r cannot be a unit. Suppose $r \mid ab$, or equivalently, $ab \in (r)$. By the definition of a prime ideal, $a \in (r)$ or $b \in (r)$. Hence, $r \mid a$ or $r \mid b$. By definition of a prime element, r is prime.

Conversely, first note that the zero ideal $(0) = \{0\}$ is a prime ideal, since R is an integral domain. Suppose r is prime. We know $(r) \neq R$ since r is not a unit. If $ab \in (r)$, then $r \mid ab$, so $r \mid a$ or $r \mid b$, giving $a \in (r)$ or $b \in (r)$ as required for (r) to be a prime ideal. □

Lemma. Prime elements are irreducible.

Proof. Let r be prime. Then r is nonzero and not a unit. Suppose $r = ab$. Then, in particular, $r \mid ab$, so $r \mid a$ or $r \mid b$ by primality. Let $r \mid a$ without loss of generality. Hence $a = rc$ for some element $c \in R$. Then, $r = ab = rcb$, so $r(1 - cb) = 0$. Since R is an integral domain, and $r \neq 0$, we have $cb = 1$, so b is a unit. □

Example. The converse does not hold in general. Let

$$R = \mathbb{Z}[\sqrt{-5}] = \left\{a + b\sqrt{-5} : \ a, b \in \mathbb{Z}\right\} \leq \mathbb{C}; \quad R \cong {}^{\mathbb{Z}[X]}\!/_{(X^2+5)}$$

Since R is a subring of the field $\mathbb{C}$, it is an integral domain. We can define the *norm* $N : R \to \mathbb{Z}$ by $N(a + b\sqrt{-5}) = a^2 + 5b^2 \geq 0$. Note that this norm is multiplicative: $N(z_1 z_2) = N(z_1)N(z_2)$.

We claim that the units are exactly ± 1. Indeed, if $r \in R^\times$, then $rs = 1$ for some element $s \in R$. Then, $N(r)N(s) = N(1) = 1$, so $N(r) = N(s) = 1$. But the only elements $r \in R$ with $N(r) = 1$ are $r = \pm 1$.

We will now show that the element $2 \in R$ is irreducible. Suppose $2 = rs$ for $r, s \in R$. By the multiplicative property of N, $N(2) = 4 = N(r)N(s)$ can only be satisfied by $N(r), N(s) \in \{1, 2, 4\}$. Since $a^2 + 5b^2 = 2$ has no integer solutions, R has no elements of norm 2. Hence, either r or s has unit norm and is thus a unit by the above discussion. We can show similarly that $3, 1 + \sqrt{-5}, 1 - \sqrt{-5}$ are irreducible, as there exist no elements of norm 3.

We can now compute directly that $(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6 = 2 \cdot 3$, hence $2 \mid (1 + \sqrt{-5})(1 - \sqrt{-5})$. But $2 \nmid (1 + \sqrt{-5})$ and $2 \nmid (1 - \sqrt{-5})$, which can be checked by taking norms. Hence, 2 is irreducible but not a prime.

In order to construct this example, we have exhibited two factorisations of 6 into irreducibles: $(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6 = 2 \cdot 3$. Since $R^\times = \{\pm 1\}$, these irreducibles in the factorisations are not associates.

9.2. Principal ideal domains

Definition. An integral domain R is a *principal ideal domain* if all ideals are principal ideals. In other words, for all ideals I, there exists an element r such that $I = (r)$.

Example. $\mathbb{Z}$ is a principal ideal domain.

Proposition. In a principal ideal domain, all irreducible elements are prime.

Proof. Let $r \in R$ be irreducible, and suppose $r \mid ab$. If $r \mid a$, the proof is complete, so suppose $r \nmid a$. Since R is a principal ideal domain, the ideal (a, r) is generated by a single element $d \in R$. In particular, since $r \in (d)$, we have $d \mid r$ so $r = cd$ for some $c \in R$.

Since r is irreducible, either c or d is a unit. If c is a unit, $(a, r) = (r)$, so in particular $r \mid a$, which contradicts the assumption that $r \nmid a$, so c cannot be a unit. Thus, d is a unit. In this case, $(a, r) = R$. By definition of (a, r), there exist $s, t \in R$ such that $1 = sa + tr$. Then, $b = sab + trb$. We have $r \mid sab$ since $r \mid ab$, and we know $r \mid trb$. Hence $r \mid b$ as required. □

Lemma. Let R be a principal ideal domain. Then an element r is irreducible if and only if (r) is maximal.

Proof. Suppose r is irreducible. Since r is not a unit, $(r) \neq R$. Suppose $(r) \subseteq J \subseteq R$ where J is an ideal in R. Since R is a principal ideal domain, $J = (a)$ for some $a \in R$. In particular, $r = ab$ for some $b \in R$, since $(r) \subseteq J$. Since r is irreducible, either a or b is a unit. But if a

is a unit, we have $J = R$. If b is a unit, then a and r are associates so they generate the same ideal. Hence, (r) is maximal.

Conversely, suppose (r) is maximal. Note that r is not a unit, since $(r) \neq R$. Suppose $r = ab$. Then $(r) \subseteq (a) \subseteq R$. But since (r) is maximal, either $(a) = (r)$ or $(a) = R$. If $(a) = (r)$, then b is a unit. If $(a) = R$, then a is a unit. Hence r is irreducible. Note that this direction of the proof did not require that R was a principal ideal domain, however R must still be an integral domain. □

Remark. Let R be a principal ideal domain, and $0 \neq r \in R$. Then, (r) is maximal if and only if r is irreducible, which is true if and only if r is prime, which is equivalent to the fact that (r) is prime. Hence, the maximal ideals are the nonzero prime ideals.

Definition. An integral domain is a *Euclidean domain* if there exists a function $\varphi : R\backslash\{0\} \to \mathbb{Z}_{\geq 0}$ such that, for all $a, b \in R$.

(i) if $a \mid b$ then $\varphi(a) \leq \varphi(b)$;

(ii) if $b \neq 0$ then $\exists q, r \in R$ such that $a = bq + r$ and either $r = 0$ or $\varphi(r) < \varphi(b)$.

Such a φ is called a *Euclidean function*.

Example. $\mathbb{Z}$ is a Euclidean domain, where the Euclidean function φ is the absolute value function.

Proposition. Euclidean domains are principal ideal domains.

Proof. Let R have Euclidean function φ. Let $I \lhd R$ be a nonzero ideal. Let $b \in I \setminus \{0\}$ that minimises $\varphi(b)$. Then $(b) \subseteq I$. For any element $a \in I$, we can use the Euclidean algorithm to show $a = bq + r$ where $r = 0$ or $\varphi(r) < \varphi(b)$. But since $r = a - bq \in I$, $\varphi(r)$ cannot be lower than the minimal element $\varphi(b)$. Thus $r = 0$, so $a = bq$. Hence, $I = (b)$, so all ideals are principal. □

Remark. In the above proof, only the second property of the Euclidean function was used. The first property is included in the definition since it will allow us to easily describe the units in the ring.

$$R^\times = \{u \in R : u \neq 0, \varphi(u) = \varphi(1)\}$$

It can be shown that, if there exists a function φ satisfying (ii), there exists a (possibly not unique) function φ' satisfying (i) and (ii).

Example. Let F be a field. Then $F[X]$ is a Euclidean domain with Euclidean function $\varphi(f) = \deg(f)$. We have already proven the requisite properties of Euclidean functions.

The ring $R = \mathbb{Z}[i]$ is a Euclidean domain with $\varphi(u + iv) = N(u + iv) = u^2 + v^2$. Since the norm is multiplicative, $N(zw) = N(z)N(w)$ which immediately gives property (i) in the definition. Consider $z, w \in \mathbb{Z}[i]$ where $w \neq 0$. Consider $\frac{z}{w} \in \mathbb{C}$. This has distance less than

1 from the nearest element q of R. Let $r = z - wq \in R$. Then $z = wq + r$ where

$$\varphi(r) = |r|^2 = |z - wq|^2 < |w|^2 = \varphi(w)$$

So property (ii) is satisfied.

Hence $F[X]$ and $\mathbb{Z}[i]$ are principal ideal domains.

Example. Let A be a nonzero $n \times n$ matrix over a field F. Let $I = \{f \in F[X] : f(A) = 0\}$. I is an ideal. Indeed, if $f, g \in I$, then $(f - g)(A) = f(A) - g(A) = 0$, and for $f \in I$ and $g \in F[X]$, we have $(f \cdot g)(A) = f(A) \cdot g(A) = 0$ as required. Since $F[X]$ is a principal ideal domain, $I = (f)$ for some polynomial $f \in F[X]$. All units in $F[X]$ are the nonzero constant polynomials. Hence, the polynomial of smallest degree in I is unique up to multiplication by a unit, so without loss of generality we may assume f is monic. This yields the minimal polynomial of A.

Example. Let $\mathbb{F}_2$ be the finite field of order 2, which is isomorphic to $\mathbb{Z}/2\mathbb{Z}$. Let $f(X)$ be the polynomial $X^3 + X + 1 \in \mathbb{F}_2[X]$.

We claim that f is irreducible. Suppose $f = gh$ where the degrees of g, h are positive. Since the degree of f is 3, one of g, h must have degree 1. Hence f has a root. But we can check that $f(0) = f(1) = 1$ so f has no root in $\mathbb{F}_2$. Hence f is irreducible as required.

Since $\mathbb{F}_2[X]$ is a principal ideal domain, we have that $(f) \lhd \mathbb{F}_2[X]$ is a maximal ideal. Hence, $\mathbb{F}_2[X]/(f)$ is a field. We can verify that this field has order 8, using the Euclidean algorithm. Any element in this quotient has coset representative $aX^2 + bX + c$ for $a, b, c \in \mathbb{F}_2$. We can show that all 8 of these possibilities yields different polynomials. So we have constructed a field of order 8. This technique will be explored further in Part II Galois Theory.

Example. The ring $\mathbb{Z}[X]$ is not a principal ideal domain. Consider the ideal $I = (2, X) \lhd \mathbb{Z}[X]$. We can write

$$I = \{2f_1(X) + Xf_2(X) : f_1, f_2 \in \mathbb{Z}[X]\} = \{f \in \mathbb{Z}[X] : 2 \mid f(0)\}$$

Suppose $I = (f)$ for some element f. Since $2 \in I$, we must have $2 = fg$ for some polynomial g. By comparing degrees, the degrees of f and g must be zero, since $\mathbb{Z}$ is an integral domain. Hence f is an integer, so $f = \pm 1$ or $f = \pm 2$. If $f = \pm 1$ then $I = \mathbb{Z}[X]$, and if $f = \pm 2$ then $I = 2\mathbb{Z}[X]$. These both lead to contradictions, since $1 \notin I$ and $X \in I$.

9.3. Unique factorisation domains

Definition. An integral domain is a *unique factorisation domain* if

(i) every nonzero, non-unit element is a product of irreducibles;

(ii) if $p_1 \cdots p_m = q_1 \cdots q_n$ where p_i, q_i are irreducible, then $m = n$, and p_i, q_i are associates, up to reordering.

Proposition. Let R be an integral domain satisfying property (i) above (every nonzero, non-unit element is a product of irreducibles). Then R is a unique factorisation domain if and only if every irreducible is prime.

Proof. Suppose R is a unique factorisation domain. Let $p \in R$ be irreducible, and $p \mid ab$. Then $ab = pc$ for some $c \in R$. Writing a, b, c as products of irreducibles, it follows from uniqueness of factorisation that $p \mid a$ or $p \mid b$. Hence p is prime.

Conversely, suppose every irreducible is prime. Suppose $p_1 \cdots p_m = q_1 \cdots q_n$ where p_i, q_i are irreducible and hence prime. Since $p_1 \mid q_1 \cdots q_n$, we have $p_1 \mid q_i$ for some i. After reordering, we may assume that $p_1 \mid q_1$, so $p_1 u = q_1$ for $u \in R$. Since q_1 is irreducible, u is a unit since p_1 cannot be a unit. Hence p_1, q_1 are associates. Cancelling p_1 from both sides, we find $p_2 \cdots p_m = u q_2 \cdots q_n$. We may absorb this unit into q_2 without loss of generality. Inductively, all p_i and q_i are associates, for each i. Hence R is a unique factorisation domain. □

Definition. Let R be a ring. Suppose, for all nested sequences of ideals in R written $I_1 \subseteq I_2 \subseteq \cdots$, there exists N such that $I_n = I_{n+1}$ for all $n \geq N$. Then, we say that R is a *Noetherian* ring.

This condition is known as the 'ascending chain condition'. In other words, we cannot infinitely nest distinct ideals in a Noetherian ring.

Lemma. Principal ideal domains are Noetherian rings.

Proof. Let $I = \bigcup_{i=1}^{\infty} I_i$. Then, I is an ideal in R. Since R is a principal ideal domain, $I = (a)$ for some $a \in R$. Then $a \in \bigcup_{i=1}^{\infty} I_i$, so in particular $a \in I_N$ for some N. But then for all $n \geq N$, $(a) \subseteq I_N \subseteq I_n \subseteq I_{n+1} \subseteq I = (a)$. So all inclusions are equalities, so in particular $I_n = I_{n+1}$. □

Theorem. If R is a principal ideal domain, then it is a unique factorisation domain.

Proof. First, we verify property (i), that every nonzero, non-unit element is a product of irreducibles. Let $x \neq 0$ be an element of R which is not a unit. Suppose x does not factor as a product of irreducibles. This implies in particular that x is not irreducible. By definition, we can write x as the product of two elements x_1, y_1 where x_1, y_1 are not units. Then either x_1 or y_1 is not a product of irreducibles, so without loss of generality we can suppose x_1 is not a product of irreducibles. We have $(x) \subset (x_1)$. This inclusion is strict, since y_1 is not a unit. Now, we can write $x_1 = x_2 y_2$ where x_2 is not a unit, and inductively we can create $(x) \subset (x_1) \subset (x_2) \subset \cdots$. But R is Noetherian, so this is a contradiction. So every nonzero, non-unit element is indeed a product of irreducibles.

By the proposition above, it suffices to show that every irreducible is prime. This has already been shown previously. Hence R is a unique factorisation domain. □

Example. We have shown that all Euclidean domains are principal ideal domains, and all principal ideal domains are unique factorisation domains, and all unique factorisation

domains are integral domains. We now provide examples for counterexamples to the converses.

The ring $\mathbb{Z}/4\mathbb{Z}$ is not an integral domain since 2 is a zero divisor.

The ring $\mathbb{Z}[\sqrt{-5}] \leq \mathbb{C}$ is integral, but not a unique factorisation domain.

The ring $\mathbb{Z}[X]$ has been shown to be not a principal ideal domain. We can show using later results that this is a unique factorisation domain.

We can construct the ring $\mathbb{Z}\left[\frac{1+\sqrt{-19}}{2}\right]$, which can be shown to be not a Euclidean domain, but is a principal ideal domain. This proof is beyond the scope of Part IB Groups, Rings and Modules, but will be proved in Part II Number Fields.

Finally, $\mathbb{Z}[i]$ is a Euclidean domain, and is hence a principal ideal domain, a unique factorisation domain, and an integral domain.

Definition. Let R be an integral domain.

(i) $d \in R$ is a *common divisor* of $a_1, \dots, a_n \in R$ if $d \mid a_i$ for all i;

(ii) $d \in R$ is a *greatest common divisor* of $a_1, \dots, a_n$ if for all common divisors d', we have $d' \mid d$;

(iii) $m \in R$ is a *common multiple* of $a_1, \dots, a_n$ if $a_i \mid m$ for all i;

(iv) $m \in R$ is a *least common multiple* of $a_1, \dots, a_n$ if for all common multiples m', we have $m \mid m'$.

Remark. Greatest common divisors and lowest common multiples are unique up to associates, if they exist.

Proposition. In unique factorisation domains, greatest common divisors and least common multiples always exist.

Proof. Let $a_i = u_i \prod_j p_j^{n_{ij}}$ where the p_j are irreducible and pairwise non-associate, u_i is a unit, and $n_{ij} \in \mathbb{Z}_{\geq 0}$. We claim that $d = \prod_j p_j^{m_j}$, where $m_j = \min_{1 \leq i \leq n} n_{ij}$, is the greatest common divisor. Certainly d is a common divisor. If d' is a common divisor, then d' can be written as a product of irreducibles, which will be denoted $d' = w \prod_j p_i^{t_j}$. We can see that $t_j \leq n_{ij}$ for all i, so in particular, $t_j \leq m_j$. This implies $d' \mid d$. Hence d is a greatest common divisor. The argument for the least common multiple is similar, replacing minima with maxima. □

9.4. Factorisation in polynomial rings

Theorem. Let R be a unique factorisation domain. Then $R[X]$ is also a unique factorisation domain.

The proof for this theorem will require a number of key lemmas. In this subsection, R will denote a unique factorisation domain, with field of fractions F. We have $R[X] \leq F[X]$. Since polynomial rings over fields are Euclidean domains, $F[X]$ is a principal ideal domain, and hence a unique factorisation domain. This does not immediately imply that $R[X]$ is a unique factorisation domain, however.

Definition. The *content* of a polynomial $f = \sum_{i=0}^{n} a_i X^i \in R[X]$ is $c(f) = \gcd\{a_0, \dots, a_n\}$. This is well-defined up to multiplication by a unit.

We say that f is *primitive* if $c(f)$ is a unit.

Lemma. The product of primitive polynomials is primitive. Further, for $f, g \in R[X]$, $c(fg)$ and $c(f)c(g)$ are associates.

Proof. Let $f = \sum_{i=0}^{n} a_i X^i$ and $g = \sum_{i=0}^{m} b_i X^i$. Suppose fg is not primitive, so $c(fg)$ is not a unit. This implies that there exists a prime p such that $p \mid c(fg)$. Since f, g are primitive, $p \nmid c(f)$ and $p \nmid c(g)$.

Suppose p does not divide all of the a_k or the b_ℓ. Let k, ℓ be the smallest values such that $p \nmid a_k$ and $p \nmid b_\ell$. Then, the coefficient of $X^{k+\ell}$ in fg is given by

$$\sum_{i+j=k+\ell} a_i b_j = \underbrace{\cdots + a_{k-1}b_{\ell+1}}_{\text{divisible by } p} + a_k b_\ell + \underbrace{a_{k+1}b_{\ell-1} + \cdots}_{\text{divisible by } p}$$

Thus $p \mid a_k b_\ell$. This is a contradiction as we have $p \mid a_k$ or $p \mid b_\ell$.

To prove the second part, let $f = c(f)f_0$ for some $f_0 \in R[X]$. Here, f_0 is primitive. Similarly, $g = c(g)g_0$ for a primitive g_0. Thus $fg = c(f)c(g)f_0g_0$. The expression f_0g_0 is a primitive polynomial by the first part, so $c(fg)$ is equal to $c(f)c(g)$ up to associates. □

Corollary. If $p \in R$ is prime in R, then p is prime in $R[X]$.

Proof. Since R is an integral domain, we have $R[X]^\times = R^\times$, so p is not a unit. Let $f \in R[X]$. Then $p \mid f$ in $R[X]$ if and only if $p \mid c(f)$ in R. Thus, if $p \mid gh$ in $R[X]$, we have $p \mid c(gh) = c(g)c(h)$. In particular, since p is prime in R, we have $p \mid c(g)$ or $p \mid c(h)$, so $p \mid g$ or $p \mid h$. So p is prime in $R[X]$. □

Lemma. Let $f, g \in R[X]$, where g is primitive. Then if $g \mid f$ in $F[X]$, then $g \mid f$ in $R[X]$.

Proof. Let $f = gh$, where $h \in F[X]$. We can find a nonzero $a \in R$, such that $ah \in R[X]$. In particular, we can multiply the denominators of the coefficients of h to form a. Now, $ah = c(ah)h_0$ where h_0 is primitive. Then $af = c(ah)h_0g$. Since h_0 and g are primitive, so is h_0g. Thus, taking contents, $a \mid c(ah)$. This implies $h \in R[X]$. Hence $g \mid f$ in $R[X]$. □

Lemma (Gauss' lemma)**.** Let $f \in R[X]$ be primitive. Then if f is irreducible in $R[X]$, we have that f is irreducible in $F[X]$.

Proof. Since $f \in R[X]$ is irreducible and primitive, its degree must be larger than zero. Hence f is not a unit in $F[X]$. Suppose f is not irreducible in $F[X]$, so $f = gh$ for $g, h \in F[X]$ with degree larger than zero. Let $\lambda \in F^\times$ such that $\lambda^{-1}g \in R[X]$ is primitive. For example, let $b \in R$ such that $bg \in R[X]$ to clear denominators, then $bg = c(bg)g_0$, giving $\lambda = c(bg)b^{-1}$. Replacing g by $\lambda^{-1}g$ and h by λh, we still have a factorisation of f. Hence, we may assume without loss of generality that $g \in R[X]$ and is primitive. By the previous lemma, we have that $h \in R[X]$, with degrees larger than zero. This contradicts irreducibility. □

Remark. We will see that the reverse implication in Gauss' lemma also holds.

Lemma. Let $g \in R[X]$ be primitive. If g is prime in $F[X]$, then g is prime in $R[X]$.

Proof. It suffices to show that if $f_1, f_2 \in R[X]$, then $g \mid f_1 f_2$ implies $g \mid f_1$ or $g \mid f_2$. Since g is prime in $F[X]$, $g \mid f_1$ or $g \mid f_2$ in $F[X]$. By the previous lemma, $g \mid f_1$ or $g \mid f_2$ in $R[X]$ as required. □

We can now prove the first theorem of this subsection, that polynomial rings over unique factorisation domains are unique factorisation domains.

Proof. Let $f \in R[X]$. Then, $f = c(f)f_0$ for f_0 primitive in $R[X]$. Since R is a unique factorisation domain, $c(f)$ is a product of irreducibles in R. If an element of R is irreducible, it is irreducible as an element of $R[X]$. Hence, it suffices to find a factorisation of f_0.

Suppose f_0 is not irreducible, so $f_0 = gh$ for $g, h \in R[X]$. Since f_0 is primitive, g and h are primitive, and the degrees of g, h are larger than zero. By induction on the degree, we can factor f_0 as a product of primitive irreducibles in $R[X]$.

It now suffices to show uniqueness of the factorisation. By a previous proposition, it in fact suffices to show that every irreducible element of $R[X]$ is prime. Let f be irreducible. Write $f = c(f)f_0$, where f_0 is primitive. Since f is irreducible, f must be constant or primitive.

Suppose f is constant. Since f is irreducible in $R[X]$, it must be irreducible in R. As R is a unique factorisation domain, f is prime in R. By a previous corollary, f is prime in $R[X]$.

Now, suppose f is primitive. Since f is irreducible in $R[X]$, we can use Gauss' lemma to show that f is irreducible in $F[X]$. Thus, f is prime in $F[X]$, as $F[X]$ is a unique factorisation domain. Finally, we can see that f is prime in $R[X]$ by the previous lemma. □

Remark. We know that the prime elements in an integral domain are irreducible. This implies that the implications in the last paragraph above are in fact equivalences. In particular, in Gauss' lemma, the implication is an equivalence.

Example. The above theorem implies that $\mathbb{Z}[X]$ is a unique factorisation domain.

Let $R[X_1, \dots, X_n]$ be the ring of polynomials in n variables. We can rewrite this as $R[X_1] \dots [X_n]$, so by induction this is a unique factorisation domain if R is.

9.5. Eisenstein's criterion

Proposition. Let R be a unique factorisation domain, and $f(X) = \sum_{i=0}^{n} a_i X^i \in R[X]$ be a primitive polynomial. Let $p \in R$ be irreducible (or, equivalently, prime) such that

(i) $p \nmid a_n$;

(ii) $p \mid a_i$ for all $i < n$; and

(iii) $p^2 \nmid a_0$.

Then f is irreducible in $R[X]$.

Proof. Suppose $f = gh$ for $g, h \in R[X]$ not units. Since f is primitive, g, h must have positive degree. Let $g(X) = \sum_{i=0}^{k} r_i X^i$ and $h(X) = \sum_{i=0}^{\ell} s_i X^i$, so $k + \ell = n$. Then $p \nmid a_n = r_k s_\ell$, so $p \nmid r_k$ and $p \nmid s_\ell$. Further, $p \mid a_0 = r_0 s_0$ so $p \mid r_0$ or $p \mid s_0$. Without loss of generality, we may assume $p \mid r_0$. There exists a minimal $j \leq k$ such that $p \mid r_i$ for all $i < j$ but $p \nmid r_j$.

$$a_j = r_0 s_j + r_1 s_{j-1} + \cdots + r_{j-1} s_1 + r_j s_0$$

By assumption, a_j is divisible by p since $j < n$. Further, the first j terms in the expansion are divisible by p. Thus, $p \mid r_j s_0$. By assumption, $p \nmid r_j$, so $p \mid s_0$. In particular, $p^2 \mid r_0 s_0 = a_0$, contradicting the third criterion. □

Example. Let $f(X) = X^3 + 2X + 5 \in \mathbb{Z}[X]$. We will show this is irreducible as a polynomial over $\mathbb{Q}$. If f is not irreducible in $\mathbb{Z}[X]$, then it factorises as $f(X) = (X + a)(X^2 + bX + c)$ up to multiplication by units. Here, $ac = 5$. But $\pm 1, \pm 5$ are not roots of f, so this is irreducible in $\mathbb{Z}[X]$. By Gauss' lemma, f is irreducible in $\mathbb{Q}[X]$, since $\mathbb{Q}$ is the field of fractions of $\mathbb{Z}$. In particular, $\mathbb{Q}[X]/(f)$ is a field, since the ideal (f) is maximal.

Example. Let $p \in \mathbb{Z}$ be a prime, and let $f(X) = X^n - p$. By Eisenstein's criterion, f is irreducible in $\mathbb{Z}[X]$. It is then irreducible in $\mathbb{Q}[X]$ by Gauss' lemma.

Example. Consider $f(X) = X^{p-1} + X^{p-2} + \cdots + X + 1 \in \mathbb{Z}[X]$, where p is prime. Eisenstein's criterion does not apply directly. Consider

$$f(X) = \frac{X^p - 1}{X - 1}; \quad Y = X - 1$$

By using this substitution of Y,

$$f(Y + 1) = \frac{(Y+1)^p - 1}{Y - 1 + 1} = Y^{p-1} + \binom{p}{1} Y^{p-2} + \cdots + \binom{p}{p-2} Y + \binom{p}{p-1}$$

We can apply Eisenstein's criterion to this new polynomial, since $p \mid \binom{p}{i}$ for all $1 \leq i \leq p-1$, and $p^2 \nmid \binom{p}{p-1} = p$. Thus, $f(Y + 1)$ is irreducible in $\mathbb{Z}[Y]$, so $f(X)$ is irreducible in $\mathbb{Z}[X]$. Of course, $f(X)$ is therefore irreducible in $\mathbb{Q}[X]$ as before.

10. Algebraic integers

10.1. Gaussian integers

Recall the ring of Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\} \leq \mathbb{C}$. There is a norm function $N : \mathbb{Z}[i] \to \mathbb{Z}_{\geq 0}$ given by $a + bi \mapsto a^2 + b^2$, and $N(xy) = N(x)N(y)$. This norm is a Euclidean function, giving the Gaussian integers the structure of a Euclidean domain and hence a principal ideal domain and a unique factorisation domain. In particular, the primes are the irreducibles. The units in $\mathbb{Z}[i]$ are $\pm 1, \pm i$, since they are the only elements of unit norm.

Example. 2 is not irreducible in $\mathbb{Z}[i]$, since it factors as $(1 + i)(1 - i)$. 5 is not irreducible, since it factors as $(2 + i)(2 - i)$. These are nontrivial factorisations since the norms of the factors are not unit length.

3 is a prime, since it is irreducible. Indeed, $N(3) = 9$, so if 3 were reducible it would factor as ab where $N(a) = N(b) = 3$. But $\mathbb{Z}[i]$ has no elements of norm 3. Similarly, 7 is a prime.

Proposition. Let $p \in \mathbb{Z}$ be a prime. Then, the following are equivalent.

(i) p is not prime in $\mathbb{Z}[i]$;

(ii) $p = a^2 + b^2$ for $a, b \in \mathbb{Z}$;

(iii) $p = 2$ or $p \equiv 1 \bmod 4$.

Proof. Suppose p is not prime in $\mathbb{Z}[i]$. So let $p = xy$ for $x, y \in \mathbb{Z}[i]$ not units. Then, $p^2 = N(p) = N(x)N(y)$. Since x, y are not units, $N(x), N(y) > 1$ and in particular $N(x) = N(y) = p$. Writing $x = a + bi$ for $a, b \in \mathbb{Z}$, we have $p = N(x) = a^2 + b^2$, which is the condition in (ii).

Now, suppose $p = a^2 + b^2$. The only squares modulo 4 are 0 and 1. Since $p \equiv a^2 + b^2 \bmod 4$, we have that p cannot be congruent to 3, modulo 4.

Finally, suppose $p = 2$ or $p \equiv 1 \bmod 4$. We have already observed above that 2 is not prime. It hence suffices to consider the case where $p \equiv 1 \bmod 4$. We have that $\left(\mathbb{Z}/p\mathbb{Z}\right)^\times$ is cyclic of order $p - 1$ by a previous theorem. Hence, if $p \equiv 1 \bmod 4$, we have that $4 \mid p - 1$, and hence $\left(\mathbb{Z}/p\mathbb{Z}\right)^\times$ contains an element of order 4. In particular, there exists $x \in \mathbb{Z}$ with $x^4 \equiv 1 \bmod p$, but $x^2 \not\equiv 1 \bmod p$. Then $x^2 \equiv -1 \bmod p$, or in other words, $p \mid (x^2 + 1)$. But this factorises as $p \mid (x + i)(x - i)$. We can see that $p \nmid x + i$, $p \nmid x - i$, so p cannot be prime. □

Remark. The proof that (iii) implies (ii) is entirely nontrivial. It required lots of theory in order to reach the result, even though its statement did not require even the notion of a complex number.

Theorem. The primes in $\mathbb{Z}[i]$ are, up to associates,

(i) $a + bi$, where $a, b \in \mathbb{Z}$ and $a^2 + b^2 = p$ is a prime in $\mathbb{Z}$ with $p = 2$ or $p \equiv 1 \bmod 4$; and

(ii) the primes p in $\mathbb{Z}$ satisfying $p \equiv 3 \bmod 4$.

Proof. First, we must check that all such elements are prime. For (i), note that $N(a+bi) = p$ is prime, so $a + bi$ is irreducible. We can use the above proof to deduce that primes in $\mathbb{Z}$ of form (ii) are primes in $\mathbb{Z}[i]$.

It now suffices to show that any prime in the Gaussian integers satisfies one of the two above conditions. Let z be prime in $\mathbb{Z}[i]$. We note that $\overline{z}$ is also irreducible. Now, $N(z) = z\overline{z}$, which is a factorisation of the norm into irreducibles.

Let p be a prime in $\mathbb{Z}$ dividing $N(z)$. If $p \equiv 3 \bmod 4$, p is prime in $\mathbb{Z}[i]$. So $p \mid z$ or $p \mid \overline{z}$ so p is associate to z or $\overline{z}$.

Otherwise, $p = a^2 + b^2 = (a + bi)(a - bi)$ where $a \pm bi$ are prime in $\mathbb{Z}[i]$ as they have norm p. So we have $p = (a + bi)(a - bi) \mid z\overline{z}$, so z is an associate of $a + bi$ or $a - bi$ by uniqueness of factorisation. □

Remark. In the above theorem, if $p = a^2 + b^2$, $a + bi$ and $a - bi$ are not associate unless $p = 2$.

Corollary. An integer $n \geq 1$ is the sum of two squares if and only if every prime factor p of n with $p \equiv 3 \bmod 4$ divides n to an even power.

Proof. Suppose $n = a^2 + b^2$. So $n = N(a + bi)$. Hence n is a product of norms of primes in the Gaussian integers. By the classification above, those norms are

(i) the primes $p \in \mathbb{Z}$ with $p \not\equiv 3 \bmod 4$; and

(ii) squares of primes $p \in \mathbb{Z}$ with $p \equiv 3 \bmod 4$.

The result follows. □

Example. We can write $65 = 5 \cdot 13$ as the sum of two primes since $5, 13 \equiv 1 \bmod 4$. We first factorise 5 and 13 into primes in the Gaussian integers.

$$5 = (2 + i)(2 - i); \quad 13 = (2 + 3i)(2 - 3i)$$

Thus, the factorisation of 65 into irreducibles in $\mathbb{Z}[i]$ is

$$\begin{aligned} 65 &= (2 + 3i)(2 + i)(2 - 3i)(2 - i) \\ &= [(2 + 3i)(2 + i)]\overline{[(2 + 3i)(2 + i)]} \\ &= N((2 + 3i)(2 - i)) \\ &= N(1 + 8i) = 1^2 + 8^2 \end{aligned}$$

This was dependent on the choice of grouping of terms. Alternatively,

$$65 = N((2 + i)(2 - 3i)) = N(7 + 4i) = 7^2 + 4^2$$

10.2. Algebraic integers

Definition. A number $\alpha \in \mathbb{C}$ is *algebraic* if α is a root of some nonzero polynomial $f \in \mathbb{Q}[X]$. α is an *algebraic integer* if it is a root of some monic polynomial $f \in \mathbb{Z}[X]$.

Let $R \leq S$, and $\alpha \in S$. We write $R[\alpha]$ to denote the smallest subring of S containing R and α. Alternatively, $R[\alpha]$ is the intersection of all subrings of S containing R and α. Further, $R[\alpha] = \operatorname{Im}\varphi$ where $\varphi\colon R[X] \to S$ is the homomorphism $g(X) \mapsto g(\alpha)$.

Definition. Let α be an algebraic number. Consider the homomorphism $\varphi\colon \mathbb{Q}[X] \to \mathbb{C}$ where $g(X) \mapsto g(\alpha)$. Since $\mathbb{Q}[X]$ is a a principal ideal domain, $\ker\varphi = (f)$ for some $f \in \mathbb{Q}[X]$. This ideal contains a nonzero element since α is an algebraic number, hence f is nonzero. Multiplying f by a unit, we may assume f is monic without loss of generality. This unique f is known as the *minimal polynomial* of α.

Corollary. All minimal polynomials are irreducible. By the first isomorphism theorem, $\mathbb{Q}[X]/(f) \cong \mathbb{Q}[\alpha] \leq \mathbb{C}$. Any subring of a field is an integral domain. Hence (f) is a prime ideal in $\mathbb{Q}[X]$, and hence f is irreducible. In particular, this implies that $\mathbb{Q}[\alpha]$ is a field.

Proposition. Let α be an algebraic integer, and $f \in \mathbb{Q}[X]$ be its minimal polynomial. Then $f \in \mathbb{Z}[X]$, and $(f) = \ker\theta \lhd \mathbb{Z}[X]$ where $\theta\colon \mathbb{Z}[X] \to \mathbb{C}$ is given by $g(X) \mapsto g(\alpha)$.

Remark. If α is an algebraic integer, then the polynomial in the definition can be taken to be minimal without loss of generality. $\mathbb{Z}[X]$ is not a principal ideal domain, so the above argument cannot work verbatim.

Proof. Let f be the minimal polynomial of α. Let $\lambda \in \mathbb{Q}^\times$ such that λf has coefficients in $\mathbb{Z}$ and is primitive. Then $\lambda f(\alpha) = 0$, so $\lambda f \in \ker\theta$.

Let $g \in \ker\theta$, so in particular $g \in \mathbb{Z}[X]$. Then $g \in \ker\varphi$, and hence $\lambda f \mid g$ in $\mathbb{Q}[X]$. By a previous lemma, $\lambda f \mid g$ in $\mathbb{Z}[X]$. Thus, $\ker\theta = (\lambda f)$.

Now, since α is an algebraic integer, we know that there exists a monic polynomial $g \in \ker\theta$ such that $g(\alpha) = 0$. Then $\lambda f \mid g$ in $\mathbb{Z}[X]$, so $\lambda = \pm 1$ as both f, g are monic. Hence, $f \in \mathbb{Z}[X]$, and $(\lambda f) = (f) = \ker\theta$. □

Let $\alpha \in \mathbb{C}$ be an algebraic integer. Then, applying the isomorphism theorem to θ, $\mathbb{Z}[X]/(f) \cong \mathbb{Z}[\alpha]$. For example:

$$\begin{aligned}
\mathbb{Z}[X]/(X^2+1) &\cong \mathbb{Z}[i] \\
\mathbb{Z}[X]/(X^2-2) &\cong \mathbb{Z}\left[\sqrt{2}\right] \\
\mathbb{Z}[X]/(X^2+X+1) &\cong \mathbb{Z}\left[\frac{-1+\sqrt{-3}}{2}\right] \\
\mathbb{Z}[X]/(X^n-p) &\cong \mathbb{Z}\left[\sqrt[n]{p}\right]
\end{aligned}$$

Corollary. If α is an algebraic integer, and $\alpha \in \mathbb{Q}$, then $\alpha \in \mathbb{Z}$.

Proof. Let $\alpha \neq 0$, since the case where $\alpha = 0$ is trivial. Then the minimal polynomial of α has coefficients in $\mathbb{Z}$. Since α is rational, the minimal polynomial is $X - \alpha$. Hence $\alpha \in \mathbb{Z}$ as it is a coefficient of the minimal polynomial. □

11. Noetherian rings

11.1. Definition

Recall the definition of a Noetherian ring.

Definition. A ring R is *Noetherian* if, for all sequences of nested ideals $I_1 \subseteq I_2 \subseteq \cdots$, there exists $N \in \mathbb{N}$ such that for all $n > N$, $I_n = I_{n+1}$.

Lemma. Let R be a ring. Then R satisfies the ascending chain condition (so R is Noetherian) if and only if all ideals in R are finitely generated.

We have already shown that principal ideal domains are Noetherian, since they satisfy this 'ascending chain' condition. This now will immediately follow from the lemma.

Proof. First, suppose that all ideals in R are finitely generated. Let $I_1 \subseteq I_2 \subseteq \cdots$ be an ascending chain of ideals. Consider $I = \bigcup_{i=1}^{\infty} I_i$, which is an ideal. I is finitely generated, so $I = (a_1, \dots, a_n)$. These elements belong to a nested union of ideals. In particular, we can choose $N \in \mathbb{N}$ such that all a_i are contained within I_N. Then, for $n \geq N$, we find

$$(a_1, \dots, a_n) \subseteq I_N \subseteq I_n \subseteq I = (a_1, \dots, a_n)$$

So the inclusions are all equalities, so $I_N = I_n$.

Conversely, suppose that R is Noetherian. Suppose that there exists an ideal $J \lhd R$ which is not finitely generated. Let $a_1 \in J$. Then since J is not finitely generated, $(a_1) \subset J$. We can therefore choose $a_2 \in J \setminus (a_1)$, and then $(a_1) \subset (a_1, a_2) \subset J$. Continuing inductively, we contradict the ascending chain condition. □

11.2. Hilbert's basis theorem

Theorem. Let R be a Noetherian ring. Then $R[X]$ is Noetherian.

Proof. Suppose there exists an ideal J that is not finitely generated. Let $f_1 \in J$ be an element of minimal degree. Then $(f_1) \subset J$. So we can choose $f_2 \in J \setminus (f_1)$, which is also of minimal degree. Inductively we can construct a sequence $f_1, f_2, \dots$, where the degrees are non-decreasing. Let a_i be the leading coefficient of f_i, for all i. We then obtain a sequence of ideals $(a_1) \subseteq (a_1, a_2) \subseteq (a_1, a_2, a_3) \subseteq \cdots$ in R. Since R is Noetherian, there exists $m \in \mathbb{N}$ such that for all $n \geq m$, we have $a_n \in (a_1, \dots, a_m)$. Let $a_{m+1} = \sum_{i=1}^{m} \lambda_i a_i$, since a_{m+1} lies in the ideal $(a_1, \dots, a_m)$. Now we define

$$g(X) = \sum_{i=1}^{m} \lambda_i X^{\deg(f_{m+1} - f_i)} f_i$$

The degree of g is equal to the degree of f_{m+1}, and they have the same leading coefficient a_{m+1}. Then, consider $f_{m+1} - g \in J$ and $\deg(f_{m+1} - g) < \deg f_{m+1}$. By minimality of the degree of f_{m+1}, $f_{m+1} - g \in (f_1, \dots, f_m)$, hence $f_{m+1} \in (f_1, \dots, f_m)$. This contradicts the choice of f_{m+1}, so J is in fact finitely generated. □

Corollary. $\mathbb{Z}[X_1, \dots, X_n]$ is Noetherian. Similarly, $F[X_1, \dots, X_n]$ is Noetherian for any field F, since fields satisfy the ascending chain condition.

Example. Let $R = \mathbb{C}[X_1, \dots, X_n]$. Let $V \subseteq \mathbb{C}^n$ be a subset of the form

$$V = \{(a_1, \dots, a_n) \in \mathbb{C}^n : f(a_1, \dots, a_n) = 0, \forall f \in \mathcal{F}\}$$

where $\mathcal{F} \subseteq R$ is a (possibly infinite) set of polynomials. Such a set is referred to as an *algebraic variety*. Let

$$I = \left\{ \sum_{i=1}^{m} \lambda_i f_i : m \in \mathbb{N}, \lambda_i \in R_i, f_i \in \mathcal{F} \right\}$$

We can check that $I \lhd R$. Since R is Noetherian, $I = (g_1, \dots, g_r)$. Hence

$$V = \{(a_1, \dots, a_n) \in \mathbb{C}^n : g(a_1, \dots, a_n) = 0, \forall g \in I\}$$

Lemma. Let R be a Noetherian ring, and $I \lhd R$. Then R/I is Noetherian.

Proof. Let $J_1' \subseteq J_2' \subseteq \cdots$ be a chain of ideals in R/I. By the ideal correspondence, J_i' corresponds to an ideal J_i that contains I, so $J_i' = J_i/I$. So $J_1 \subseteq J_2 \subseteq \cdots$ is a chain of ideals in R. Since R is Noetherian, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, we have $J_N = J_n$, and so $J_N' = J_n'$. Hence R/I satisfies the ascending chain condition. □

Example. The ring of Gaussian integers $\mathbb{Z}/(X^2 + 1)$ is Noetherian. If $R[X]$ is Noetherian, then $R[X]/(X) \cong R$ is Noetherian. This is a converse to the Hilbert basis theorem.

The ring of polynomials in countably many variables is not Noetherian.

$$\mathbb{Z}[X_1, X_2, \dots] = \bigcup_{n \in \mathbb{N}} \mathbb{Z}[X_1, \dots, X_n]$$

In particular, consider the ascending chain $(X_1) \subset (X_1, X_2) \subset (X_1, X_2, X_3) \subset \cdots$.

Let $R = \{f \in \mathbb{Q}[X] : f(0) \in \mathbb{Z}\} \leq \mathbb{Q}[X]$. Even though $\mathbb{Q}[X]$ is Noetherian, R is not. Indeed, consider $(X) \subset \left(\frac{1}{2}X\right) \subset \left(\frac{1}{4}X\right) \subset \left(\frac{1}{8}X\right) \subset \cdots$. These inclusions are strict, since $2 \in R$ is not a unit.

12. Modules

12.1. Definitions

Definition. Let R be a ring. A *module over R* is a triple $(M, +, \cdot)$ consisting of a set M and two operations $+ : M \times M \to M$ and $\cdot : R \times M \to M$, that satisfy

(i) $(M, +)$ is an abelian group with identity $0 = 0_M$;

(ii) $(r_1 + r_2) \cdot m = r_1 \cdot m + r_2 \cdot m$;

(iii) $r \cdot (m_1 + m_2) = r \cdot m_1 + r \cdot m_2$;

(iv) $r_1 \cdot (r_2 \cdot m) = (r_1 \cdot r_2) \cdot m$;

(v) $1_R \cdot m = m$;

Remark. Closure is implicitly required by the types of the $+$ and $\cdot$ operations.

Example. A module over a field is precisely a vector space.

A $\mathbb{Z}$-module is precisely the same as an abelian group, since

$$\cdot : \mathbb{Z} \times A \to A; \quad n \cdot a = \begin{cases} \underbrace{a + \cdots + a}_{n \text{ times}} & \text{if } n > 0 \\ 0 & \text{if } n = 0 \\ -\left(\underbrace{a + \cdots + a}_{-n \text{ times}}\right) & \text{if } n < 0 \end{cases}$$

Let F be a field, and V be a vector space over F. Let $\alpha : V \to V$ be an endomorphism. We can turn V into an $F[X]$-module by

$$\cdot : F[X] \times V \to V; \quad f \cdot v = (f(\alpha))(v)$$

Note that the structure of the $F[X]$-module depends on the choice of α. We can write $V = V_\alpha$ to disambiguate.

For any ring R, we can consider R^n as an R-module via

$$r \cdot (r_1, \ldots, r_n) = (r \cdot r_1, \ldots, r \cdot r_n)$$

In particular, the case $n = 1$ shows that any ring R can be considered an R-module where the scalar multiplication in the ring and the module agree.

For an ideal $I \lhd R$, we can regard I as an R-module, since I is preserved under multiplication by elements in R. The quotient ring R/I is also an R-module, defining multiplication as $r \cdot (s + I) = rs + I$.

Let $\varphi : R \to S$ be a ring homomorphism. Then any S-module can be regarded as an R-module. We define $r \cdot m = \varphi(r) \cdot m$. In particular, this applies when R is a subring of S, and φ is the inclusion map. So any module over a ring can be viewed as a module over any subring.

Definition. Let M be an R-module. Then $N \subseteq M$ is an *R-submodule of M*, written $N \leq M$, if $(N, +) \leq (M, +)$, and for all $rn \in N$ for all $r \in R$ and $n \in N$.

Example. By considering R as an R-module, a subset of R is an R-submodule if and only if it is an ideal. If $R = F$ is a field, this definition corresponds to the definition of a vector subspace.

Definition. Let $N \leq M$ be R-modules. Then, the *quotient* M/N is defined as the quotient of groups under addition, and with scalar multiplication defined as $r \cdot (m + N) = rm + N$. This is well-defined, since N is preserved under scalar multiplication. This makes M/N an R-module.

Remark. Submodules are analogous both to subrings and to ideals.

Definition. Let M, N be R-modules. Then $f : M \to N$ is a *R-module homomorphism* if it is a homomorphism of $(M, +)$ and $(N, +)$, and scalar multiplication is preserved: $f(r \cdot m) = r \cdot f(m)$. An *$R$-module isomorphism* is an R-module homomorphism that is a bijection.

Example. If $R = F$ is a field, F-module homomorphisms are exactly linear maps.

Theorem. Let $f : M \to N$ be an R-module homomorphism. Then

(i) $\ker f = \{m \in M : f(m) = 0\} \leq M$;

(ii) $\operatorname{Im} f = \{f(m) \in N : m \in M\} \leq N$;

(iii) $M/\ker f \cong \operatorname{Im} f$.

Theorem. Let $A, B \leq M$ be R-submodules. Then

(i) $A + B = \{a + b : a \in A, b \in B\} \leq M$;

(ii) $A \cap B \leq M$;

(iii) $A/A \cap B \cong A + B/B$.

Theorem. For $N \leq L \leq M$ are R-submodules, then

$$M/N/L/N \cong M/L$$

For $N \leq M$, there is a correspondence between submodules of M/N and submodules of M containing N. These isomorphism theorems can be proved exactly as before. Note that these results apply to vector spaces; for example, the first isomorphism theorem immediately gives the rank-nullity theorem.

12.2. Finitely generated modules

Definition. Let M be an R-module. If $m \in M$, then we write $Rm = \{rm : r \in R\}$. This is an R-submodule of M, known as the submodule *generated by m*.

If $A, B \leq M$, we can define $A + B = \{a + b : a \in A, b \in B\}$, known as the *sum of submodules*. In particular, this sum is commutative.

Definition. A module M is *finitely generated* if it is the sum of finitely many submodules generated by a single element. In other words, $M = Rm_1 + \cdots + Rm_n$.

This is the analogue of finite dimensionality in linear algebra.

Lemma. An R-module M is finitely generated if and only if there exists a surjective R-module homomorphism $f : R^n \to M$ for some n.

Proof. If M is finitely generated, we have $M = Rm_1 + \cdots + Rm_n$. We define $f : R^n \to M$ by $(r_1, \ldots, r_n) \mapsto r_1m_1 + \cdots + r_nm_n$. This is surjective.

Conversely, suppose such a surjective homomorphism f exists. Let $e_i = (0, \ldots, 1, \ldots, 0)$ be the element of R^n with all entries zero except for 1 in the ith place. Let $m_i = f(e_i)$. Then, since f is surjective, any element $m \in M$ is contained in the image of f, so is of the form $f(r_1, \ldots, r_n) = r_1m_1 + \cdots + r_nm_n$. □

Corollary. Any quotient by a submodule of a finitely generated module is finitely generated.

Proof. Let $N \leq M$, where M is finitely generated. Then there exists a surjective R-module homomorphism $f : R^n \to M$. Then $q \circ f$, where q is the quotient map, is also a surjective homomorphism. So M/N is finitely generated. □

Example. It is not always the case that a submodule of a finitely generated module is finitely generated. Let R be a non-Noetherian ring, and I an ideal in R that is not finitely generated (in the ring sense). R is a finitely generated R-module, since $R1 = R$. I is a submodule of R, which is not finitely generated (in the module sense).

Remark. If R is Noetherian, it is always the case that submodules of finitely generated R-modules are finitely generated. This will be shown on the example sheets.

12.3. Torsion

Definition. Let M be an R-module.

(i) $m \in M$ is *torsion* if there exists $0 \neq r \in R$ such that $rm = 0$;

(ii) M is a *torsion module* if every element is torsion;

(iii) M is a *torsion-free module* if 0 is the only torsion element.

Example. The torsion elements in a $\mathbb{Z}$-module (which is an abelian group) are precisely the elements of finite order. If F is a field, any F-module is torsion-free.

12.4. Direct sums

Definition. Let $M_1, \dots, M_n$ be R-modules. Then the *direct sum* of $M_1, \dots, M_n$, written $M_1 \oplus \cdots \oplus M_n$, is the set $M_1 \times \cdots \times M_n$, with the operations of addition and scalar multiplication defined componentwise. We can show that the direct sum of (finitely many) R-modules is an R-module.

Example. $R^n = R \oplus \cdots \oplus R$, where we take the direct sum of n copies of R.

Lemma. Let $M = \bigoplus_{i=1}^{n} M_i$, and for each M_i, let $N_i \leq M_i$. Then $N = \bigoplus_{i=1}^{n} N_i$ is a submodule of M. Further,

$$M/N = \bigoplus_{i=1}^{n} M_i \Big/ \bigoplus_{i=1}^{n} N_i \cong \bigoplus_{i=1}^{n} M_i/N_i$$

Proof. First, we can see that this N is a submodule. Applying the first isomorphism theorem to the surjective R-module homomorphism $M \to \bigoplus_{i=1}^{n} M_i/N_i$ given by $(m_1, \dots, m_n) \mapsto (m_1 + N_1, \dots, m_n + N_n)$, the result follows as required, since the kernel is N. □

12.5. Free modules

Definition. Let $m_1, \dots, m_n \in M$. The set $\{m_1, \dots, m_n\}$ is *independent* if $\sum_{i=1}^{n} r_i m_i = 0$ implies that the r_i are all zero.

Definition. A subset $S \subseteq M$ *generates M freely* if:

(i) S generates M, so for all $m \in M$, we can find finitely many entries s_i and coefficients r_i such that $m = \sum_{i=1}^{k} r_i s_i$;

(ii) any function $\psi : S \to N$, where N is an R-module, extends to an R-module homomorphism $\theta : M \to N$.

Remark. In (ii), such an extension θ is always unique if it exists, by (i).

Definition. An R-module M freely generated by some subset $S \subseteq M$ is called *free*. We say that S is a *free basis* for M.

Remark. Free bases in the study of modules are analogous to bases in linear algebra. All vector spaces are free modules, but not all modules are free.

Proposition. For a finite subset $S = \{m_1, \dots, m_n\} \subseteq M$, the following are equivalent.

(i) S generates M freely;

(ii) S generates M, and S is independent;

(iii) every element of M can be written uniquely as $r_1 m_1 + \cdots + r_n m_n$ for some $r_i \in R$;

(iv) the R-module homomorphism $R^n \to M$ given by $(r_1, \dots, r_n) \mapsto r_1 m_1 + \cdots + r_n m_n$ is bijective, so is an isomorphism.

Proof. Not all implications are shown, but they are similar to arguments found in Part IB Linear Algebra. We show (i) implies (ii). Let S generate M freely. Suppose S is not independent. Then there exist r_i such that $\sum_{i=1}^n r_i m_i = 0$ but not all r_i are zero. Let $r_j \neq 0$. Since S generates M freely, consider the module homomorphism $\psi : S \to R$ given by

$$\psi(m_i) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Then

$$0 = \theta(0) = \theta\left(\sum_{i=1}^n r_i m_i\right) = \sum_{i=1}^n r_i \theta(m_i) = r_j \neq 0$$

This is a contradiction, so S is independent.

To show (ii) implies (iii), it suffices to show uniqueness. If there exist two ways to write an element as a linear combination, consider their difference to find a contradiction from (ii).

We can show (iii) implies (i). Then it remains to show (iii) and (iv) are equivalent. □

Example. A non-trivial finite abelian group is not a free $\mathbb{Z}$-module.

The set $\{2, 3\}$ generates $\mathbb{Z}$ as a $\mathbb{Z}$-module. This is not a free basis, since they are not independent: $2 \cdot 3 - 3 \cdot 2 = 0$. However, it contains no subset that is a free basis. This is different to vector spaces, where we can always construct a basis from a subset of a spanning set.

Proposition (invariance of dimension)**.** Let R be a nonzero ring. If $R^m \cong R^n$ as R-modules, then $m = n$.

Proof. Let $I \lhd R$, and M an R-module. We define $IM = \{\sum a_i m_i : a_i \in I, m_i \in M\}$. Since I is an ideal, we can show that IM is a submodule of M. The quotient module M/IM is an R-module, but we can also show that it is an R/I-module, by defining scalar multiplication as

$$(r + I) \cdot (m + IM) = (r \cdot m + IM)$$

We can check that this is well-defined; this follows from the fact that for $b \in I$, $b \cdot (m + IM) = bm + IM$, but $b \in I$ so $bm \in IM$.

Now, suppose that $R^m \cong R^n$. Then let $I \lhd R$ be a maximal ideal in R. We can prove the existence of such an ideal under the assumption of the axiom of choice, and in particular using Zorn's lemma. By the above discussion, we find an isomorphism of R/I-modules

$$\left(R/I\right)^m \cong R^m/IR^m \cong R^n/IR^n \cong \left(R/I\right)^n$$

This is an isomorphism of vector spaces over R/I which is a field, since I is maximal. Hence, using the corresponding result from linear algebra, $n = m$. □

12.6. Row and column operations

We will assume that R is a Euclidean domain in this subsection, and let φ be a Euclidean function for R. We will consider an $m \times n$ matrix with entries in R.

Definition. The *elementary row operations* on a matrix are

(i) add $\lambda \in R$ multiplied by the jth row to the ith row, where $i \neq j$;

(ii) swap the ith row and the jth row;

(iii) multiply the ith row by $u \in R^\times$.

Each of these operations can be realised by left-multiplication by some $m \times m$ matrix. These operations are all invertible, so their matrices are all invertible.

We can define elementary column operations in an analogous way, using right-multiplication by an $n \times n$ matrix instead.

Definition. Two $m \times n$ matrices A, B are *equivalent* if there exists a sequence of elementary row and column operations that transforms one matrix into the other. If they are equivalent, then there exist invertible matrices P, Q such that $B = QAP$.

Definition. A $k \times k$ *minor* of an $m \times n$ matrix A is the determinant of a $k \times k$ submatrix of A, which is a matrix of A produced by removing $m - k$ rows and $n - k$ columns.

The kth Fitting ideal $\mathrm{Fit}_k(A) \lhd R$ is the ideal generated by the $k \times k$ minors of A.

Lemma. The kth Fitting ideal of a matrix is invariant under elementary row and column operations.

Proof. It suffices by symmetry to show that the elementary row operations do not change the Fitting ideal. For the first elementary row operation on a matrix A, suppose we add $\lambda \in R$ multiplied by the jth row to the ith row, yielding a matrix A'. In particular, $a_{ik} \mapsto a_{ik} + \lambda a_{jk}$ for all k. Let C be a $k \times k$ submatrix of A and C' the corresponding submatrix of A'.

If row i was not chosen in C, then C and C' are the same matrix. Hence the corresponding minors are equal. If row i and row j were both chosen in C, we have that C, C' differ by a row operation. Since the determinant is invariant under this elementary row operations, the corresponding minors are equal.

If row i was chosen but row j was not chosen, by expanding the determinant along the ith row, we find

$$\det C' = \det C + \lambda \det D$$

where we can show that D is a $k \times k$ submatrix of A that includes row j but not row i. By definition, $\det D \in \mathrm{Fit}_k(A)$ and $\det C \in \mathrm{Fit}_k(A)$, so certainly $\det C' \in \mathrm{Fit}_k(A)$. Hence $\mathrm{Fit}_k(A') \subseteq \mathrm{Fit}_k(A)$. By the invertibility of the elementary row operations, $\mathrm{Fit}_k(A') \supseteq \mathrm{Fit}_k(A)$.

The proofs for the other elementary row operations are left as an exercise. □

12.7. Smith normal form

Theorem. An $m \times n$ matrix $A = (a_{ij})$ over a Euclidean domain R is equivalent to a matrix of the form

$$\begin{pmatrix} d_1 & & & & & \\ & \ddots & & & & \\ & & d_t & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & \end{pmatrix}; \quad d_1 \mid d_2 \mid \cdots \mid d_t$$

The d_i are known as *invariant factors*, and they are unique up to associates.

Proof. If $A = 0$, the matrix is already in Smith normal form. Otherwise, we can swap columns and rows such that $a_{11} \neq 0$. We will reduce $\varphi(a_{11})$ as much as possible until it divides every other element in the matrix, using the following algorithm.

If $a_{11} \nmid a_{1j}$ for some $j \geq 2$, then $a_{1j} = qa_{11} + r$ where $q, r \in R$ and $\varphi(r) < \varphi(a_{11})$. We can subtract q multiplied by column 1 from column j. Swapping such columns leaves $a_{11} = r$. If $a_{11} \nmid a_{i1}$ for some $i \geq 2$, then repeat the above process using row operations. Now, $a_{11} \mid a_{ij}$ for all i, j. These steps are repeated until a_{11} divides all entries of the first row and first column. This algorithm will always terminate, for example because the Euclidean function takes values in $\mathbb{Z}_{\geq 0}$ and $\varphi(a_{11})$ strictly decreases in each iteration.

Now, we can subtract multiples of the first row and column from the others to give

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & A' & \\ 0 & & & \end{pmatrix}$$

If $a_{11} \nmid a_{ij}$ for $i, j \geq 2$, then add the ith row to the first row. There is now an element in the first row that does a_{11} not divide. We can then perform column operations as above to decrease $\varphi(a_{11})$. We will then restart the algorithm. After finitely many steps, this algorithm will terminate and a_{11} will divide all elements a_{ij} of the matrix.

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & A' & \\ 0 & & & \end{pmatrix}; \quad a_{11} \equiv d_1 \mid a_{ij}$$

We can now apply the algorithm to A', since column and row operations not including the first row or column do not change whether $a_{11} \mid a_{ij}$.

We now demonstrate uniqueness of the invariant factors. Suppose A has Smith normal form with invariant factors d_i where $d_1 \mid \cdots \mid d_t$. Then, for all k, $\mathrm{Fit}_k(A)$ can be evaluated in Smith normal form by invariance of the Fitting ideal under row and column operations. Hence $\mathrm{Fit}_k(A) = (d_1 d_2 \cdots d_k) \lhd R$. Thus, the product $d_1 \cdots d_k$ depends only on A, and is unique up to associates. Cancelling, we can see that each d_i depends only on A, up to associates. □

Example. Consider the matrix over $\mathbb{Z}$ given by

$$A = \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$$

Using elementary row and column operations,

$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} \xrightarrow{c_1 \mapsto c_1 + c_2} \begin{pmatrix} 1 & -1 \\ 3 & 2 \end{pmatrix} \xrightarrow{c_2 \mapsto c_1 + c_2} \begin{pmatrix} 1 & 0 \\ 3 & 5 \end{pmatrix} \xrightarrow{r_2 \mapsto -3r_1 + r_2} \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$$

This is in Smith normal form as $1 \mid 5$.

Alternatively, $(d_1) = (2, -1, 1, 2) = (1)$. So $d_1 = \pm 1$. Further, $(d_1 d_2) = (\det A) = (5)$. So $d_1 d_2 = \pm 5$ and hence $d_2 = \pm 5$.

12.8. The structure theorem

Lemma. Let R be a Euclidean domain with Euclidean function φ (or, indeed, a principal ideal domain). Any submodule of the free module R^m is generated by at most m elements.

Proof. Let $N \leq R^m$. Consider

$$I = \{r \in R : \exists r_2, \dots, r_m \in R, (r, r_2, \dots, r_m) \in N\}$$

Since N is a submodule, this is an ideal. Since R is a principal ideal domain, $I = (a)$ for some $a \in R$. Let $n = (a, a_2, \dots, a_m) \in N$. For $(r_1, \dots, r_m) \in N$, we have $r_1 = ra$ for some r. Hence $(r_1, \dots, r_m) - rn = (0, r_2 - ra_2, \dots, r_m - ra_m)$, which lies in $N' = N \cap \{0\} \times R^{m-1} \leq R^{m-1}$, hence $N = Rn + N'$. By induction, N' is generated by $n_2, \dots, n_m$, hence $(n, n_2, \dots, n_m)$ generate N. □

Theorem. Let R be a Euclidean domain, and $N \leq R^m$. Then there is a free basis $x_1, \dots, x_m$ for R^m such that N is generated by $d_1 x_1, \dots, d_t x_t$ for some $d_i \in R$ and $t \leq m$, and such that $d_1 \mid \cdots \mid d_t$.

Proof. By the above lemma, we have $N = Ry_1 + \cdots + Ry_n$ for some $y_i \in R^m$ for some $n \leq m$. Each y_i belongs to R^m so we can form the $m \times n$ matrix A which has columns y_i. A is equivalent to a matrix A' in Smith normal form with invariant factors $d_1 \mid \cdots \mid d_t$.

A' is obtained from A by elementary row and column operations. Switching row i and row j in A corresponds to reassigning the standard basis elements e_i and e_j to each other. Adding a multiple of row i to row j corresponds to replacing $e_1, \dots, e_m$ with a linear combination of these basis elements which is a free basis. In general, each row operation simply changes the choice of free basis used for R^m. Analogously, each column operation changes the set of generators y_i for N.

Hence, after applying these row and column operations, the free basis e_i of R^m is converted into $x_1, \dots, x_m$, and N is generated by $d_1 x_1, \dots, d_t x_t$. □

Theorem (structure theorem for finitely generated modules over Euclidean domains)**.** Let R be a Euclidean domain, and M a finitely generated module over R. Then

$$M \cong {}^{R}\!/_{(d_1)} \oplus \cdots \oplus {}^{R}\!/_{(d_t)} \oplus \underbrace{R \oplus \cdots \oplus R}_{k \text{ copies}} \cong {}^{R}\!/_{(d_1)} \oplus \cdots \oplus {}^{R}\!/_{(d_t)} \oplus R^k$$

for some $0 \neq d_i \in R$ and $d_1 \mid \cdots \mid d_t$, and where $k \geq 0$. The d_i are called invariant factors.

Proof. Since M is a finitely generated module, there exists a surjective R-module homomorphism $\varphi \colon R^m \to M$ for some m. By the first isomorphism theorem, $M \cong {}^{R^m}\!/_{\ker \varphi}$. By the previous theorem, there exists a free basis $x_1, \ldots, x_m$ for R^m such that $\ker \varphi \leq R^m$ is generated by $d_1 x_1, \ldots, d_t x_t$ and where $d_1 \mid \cdots \mid d_t$. Then,

$$\begin{aligned} M &\cong \frac{\underbrace{R \oplus \ldots R}_{k \text{ copies}}}{d_1 R \oplus \cdots \oplus d_t R \oplus \underbrace{0 \oplus \cdots \oplus 0}_{m-t \text{ copies}}} \\ &\cong {}^{R}\!/_{(d_1)} \oplus \cdots \oplus {}^{R}\!/_{(d_t)} \oplus \underbrace{R \oplus \cdots \oplus R}_{m-t \text{ copies}} \end{aligned}$$

□

Remark. After deleting those d_i which are units, the invariant factors of M are unique up to associates. The proof is omitted.

Corollary. Let R be a Euclidean domain. Then any finitely generated torsion-free module is free.

Proof. Since M is torsion-free, there are no submodules of the form ${}^{R}\!/_{(d)}$ with d nonzero, since then multiplying an element of M by d would give zero. Hence, by the structure theorem, $M \cong R^m$ for some m. □

Example. Consider $R = \mathbb{Z}$, and the abelian group $G = \langle a, b \rangle$ subject to the relations $2a + b = 0$ and $-a + 2b = 0$, so $G \cong {}^{\mathbb{Z}^2}\!/_{N}$ where N is the $\mathbb{Z}$-submodule of $\mathbb{Z}^2$ generated by $(2, 1)$ and $(-1, 2)$. Consider

$$A = \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$$

which has Smith normal form $d_1 = 1$ and $d_2 = 5$. Hence, by changing basis for $\mathbb{Z}^2$, we can let N be generated by $(1, 0)$ and $(0, 5)$. Hence,

$$G \cong {}^{\mathbb{Z} \oplus \mathbb{Z}}\!/_{\mathbb{Z} \oplus 5\mathbb{Z}} \cong {}^{\mathbb{Z}}\!/_{5\mathbb{Z}}$$

12.9. Primary decomposition theorem

More generally, applying the structure theorem to $\mathbb{Z}$-modules, we obtain the structure theorem for finitely generated abelian groups:

Theorem. Let G be a finitely generated abelian group. Then

$$G \cong C_{d_1} \times \cdots \times C_{d_t} \times \mathbb{Z}^r$$

where $d_1 \mid \cdots \mid d_t$ in $\mathbb{Z}$, and $r \geq 0$.

We have replaced the submodule notation $\mathbb{Z}/n\mathbb{Z}$ and $\oplus$ with the group notation C_n and $\times$. The previous theorem for the structure of finite abelian groups is a special case of this theorem, where $r = 0$. We have also seen that any finite abelian group can be written as a product of cyclic groups of prime power order. This also has a generalisation for modules. The previous result relied on the lemma $C_{mn} \cong C_m \times C_n$ where m and n are coprime. There is an analogous result for principal ideal domains.

Lemma. Let R be a principal ideal domain, and $a, b \in R$ with unit greatest common divisor. Then, treating these quotients as R-modules,

$$R/(ab) \cong R/(a) \oplus R/(b)$$

Proof. Since R is a principal ideal domain, $(a, b) = (d)$ for some $d \in R$. The greatest common divisor of a, b is a unit, so d is a unit, giving $(a, b) = R$. Hence, there exist $r, s \in R$ such that $ra + sb = 1$. This is a generalisation of Bézout's theorem.

Now, we define an R-module homomorphism $\psi : R \to R/(a) + R/(b)$ by $\psi(x) = (x + (a), x + (b))$. Then $\psi(sb) = (sb + (a), sb + (b)) = (1 - ra + (a), sb + (b)) = (1 + (a), (b))$, and similarly $\psi(ra) = ((a), 1 + (b))$. Hence, $\psi(sbx + rby) = (x + (a), y + (b))$ so ψ is surjective.

Clearly we have $(ab) \subset \ker \psi$, so it suffices to show the converse. If $x \in \ker \psi$, then $x \in (a)$ and $x \in (b)$, so $x \in (a) \cap (b)$. Since $x = x(ra + sb) = r(ax) + s(bx)$, we must have that $s(bx) \in (a)$ and $r(ax) \in (b)$, so $x \in (ab)$. Hence $\ker \psi = (ab)$, and the result follows from the first isomorphism theorem for modules. □

Lemma (primary decomposition theorem)**.** Let R be a Euclidean domain and M a finitely generated R-module. Then

$$M \cong R/(p_1^{n_1}) \oplus \cdots \oplus R/(p_k^{n_k}) \oplus R^m$$

where the quotients are considered as R-modules, where p_i are primes in R, which are not necessarily distinct, and where $m \geq 0$.

Proof. By the structure theorem,

$$M \cong R/(d_1) \oplus \cdots \oplus R/(d_t) \oplus \underbrace{R \oplus \cdots \oplus R}_{k \text{ copies}} \cong R/(d_1) \oplus \cdots \oplus R/(d_t) \oplus R^m$$

where $d_1 \mid \cdots \mid d_t$. So it suffices to show that each $R/(d_i)$ can be written as a product of factors of the form $R/(p_j^{n_j})$. Since R is a unique factorisation domain and a principal ideal domain, d_i can be written as a product $up_1^{\alpha_1} \cdots p_r^{\alpha_r}$ where u is a unit and the p_j are pairwise non-associate primes. By the previous lemma,

$$R/(d_i) \cong R/(p_1^{\alpha_1}) \oplus \ldots R/(p_r^{\alpha_r})$$

□

12.10. Rational canonical form

Let V be a vector space over a field F, and $\alpha : V \to V$ be a linear map. Let V_α denote the $F[X]$-module V where scalar multiplication is defined by $f(X) \cdot v = f(\alpha)(v)$.

Lemma. If V is finite-dimensional as a vector space, then V_α is finitely generated as an $F[X]$-module.

Proof. Consider a basis $v_1, \ldots, v_n$ of V, so $v_1, \ldots, v_n$ generate V as an F-vector space. Then, these vectors generate V_α as an $F[X]$-module, since $F \leq F[X]$. □

Example. Suppose $V_\alpha \cong F[X]/(X^n)$ as an $F[X]$-module. Then, $1, X, X^2, \ldots, X^{n-1}$ is a basis for $F[X]/(X^n)$ as an F-vector space. With respect to this basis, α has the matrix form

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \qquad (*)$$

Example. Suppose $V_\alpha \cong F[X]/(X-\lambda)^n$ as an $F[X]$-module. Consider the basis $1, X-\lambda, (X-\lambda)^2, \ldots, (X-\lambda)^{n-1}$ for $F[X]/(X-\lambda)^n$ as an F-vector space. Here, $\alpha - \lambda\,\mathrm{id}$ has matrix $(*)$ from the previous example. Hence, α has matrix $(*) + \lambda I$.

Example. Suppose $V_\alpha \cong F[X]/(f)$ where $f \in F[X]$ as an $F[X]$-module, such that f is monic. Let

$$f(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_0$$

With respect to basis $1, X, \dots, X^{n-1}$, α has matrix

$$C(f) = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ 0 & 0 & 1 & \cdots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}$$

since f is monic and the last column represents X^n. The above matrix is known as the *companion matrix* of the monic polynomial.

Theorem (Rational canonical form)**.** Let F be a field, V be a finite-dimensional F-vector space, and $\alpha : V \to V$ be a linear map. Then the $F[X]$-module V_α decomposes as

$$V_\alpha \cong {F[X]}/{(f_1)} \oplus \cdots \oplus {F[X]}/{(f_t)}$$

for some monic polynomials $f_i \in F[X]$, and $f_1 \mid \cdots \mid f_t$. Moreover, with respect to a suitable basis, α has matrix

$$\begin{pmatrix} C(f_1) & & & \\ & C(f_2) & & \\ & & \ddots & \\ & & & C(f_t) \end{pmatrix} \tag{$**$}$$

Proof. We know that V_α is finitely generated as an $F[X]$-module, since V is finite-dimensional. Since $F[X]$ is a Euclidean domain, the structure theorem applies, and

$$V_\alpha \cong {F[X]}/{(f_1)} \oplus \cdots \oplus {F[X]}/{(f_t)} \oplus F[X]^m$$

for some m, where $f_1 \mid \cdots \mid f_t$. Since V is finite-dimensional, $m = 0$. As F is a field, without loss of generality we may multiply each f_i by a unit to ensure that they are monic. Then, using the previous example, we can construct the companion matrices for each polynomial and obtain the matrix as required. □

Remark. If α is represented by an $n \times n$ matrix A, there exists a change of basis matrix P such that PAP^{-1} has form $(*)$ as stated in the theorem, so A is similar to such a block diagonal matrix of companion matrices. Note further that $(**)$ can be used to find the minimal and characteristic polynomials of α; the minimal polynomial is f_t, and the characteristic polynomial is $f_1 \cdots f_t$. In particular, the minimal polynomial divides the characteristic polynomial, and this implies the Cayley–Hamilton theorem.

Example. Consider $\dim V = 2$. Here, $\sum \deg f_i = 2$, so there are two cases: one polynomial of degree two, or two polynomials of degree one. Consider $V_\alpha \cong {F[X]}/{(X-\lambda)} \oplus {F[X]}/{(X-\mu)}$. Since one of the f_i must divide the other, we have $\lambda = \mu$. If we have one polynomial of degree two, we have $V_\alpha \cong {F[X]}/{(f)}$, where f is the characteristic polynomial of α.

Corollary. Let A, B be invertible 2×2 non-scalar matrices over a field F. Then A, B are similar if and only if their characteristic polynomials are equal.

Proof. Certainly if A, B are similar they have the same characteristic polynomial, which is proven in Part IB Linear Algebra. Conversely, if the matrices are non-scalar, the modules V_α, V_β are of the form $F[X]/(f)$ by the previous example, so they are both similar to the companion matrix of f, where f is the characteristic polynomial of A or B. □

Definition. The *annihilator* of an R-module M is

$$\mathrm{Ann}_R(M) = \{r \in R : \forall m \in M,\ rm = 0\} \lhd R$$

Example. Let $I \lhd R$. Then the annihilator of R/I is $\mathrm{Ann}_R\left(R/I\right) = I$.

Let A be a finite abelian group. Then, considering A as a $\mathbb{Z}$-module, $\mathrm{Ann}_{\mathbb{Z}}(A) = (e)$ where e is the *exponent* of the group, which is the lowest common multiple of the orders of elements in the group.

Let V_α be as above. Then $\mathrm{Ann}_{F[X]}(V_\alpha) = (f)$ where f is the minimal polynomial of α.

12.11. Jordan normal form

Jordan normal form concerns matrix similarity in $\mathbb{C}$. The following results are therefore restricted to this particular field.

Lemma. The primes (or equivalently, irreducibles) in $\mathbb{C}[X]$ are the polynomials $X - \lambda$ for $\lambda \in \mathbb{C}$, up to associates.

Proof. By the fundamental theorem of algebra, any non-constant polynomial with complex coefficients has a complex root. By the Euclidean algorithm, we can show that having a root λ is equivalent to having a linear factor $X - \lambda$. Hence the irreducibles have degree one, and thus are $X - \lambda$ exactly, up to associates. □

Theorem. Let $\alpha : V \to V$ be an endomorphism of a finite-dimensional $\mathbb{C}$-vector space V. Let V_α be the set V as a $\mathbb{C}[X]$-module, where scalar multiplication is defined by $f \cdot v = f(\alpha)(v)$. Then, there exists an isomorphism of $\mathbb{C}[X]$-modules

$$V_\alpha \cong \mathbb{C}[X]/((X-\lambda_1)^{n_1}) \oplus \cdots \oplus \mathbb{C}[X]/((X-\lambda_t)^{n_t})$$

where $\lambda_i \in \mathbb{C}$ are not necessarily distinct. In particular, there exists a basis for this vector space such that α has matrix in block diagonal form

$$\begin{pmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{n_t}(\lambda_t) \end{pmatrix}$$

where each *Jordan block* $J_{n_i}(\lambda_i)$ is an $n_i \times n_i$ matrix of the form

$$J_{n_i}(\lambda_i) = \begin{pmatrix} \lambda_i & 0 & 0 & \cdots & 0 \\ 1 & \lambda_i & 0 & \cdots & 0 \\ 0 & 1 & \lambda_i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i \end{pmatrix}$$

Proof. Note $\mathbb{C}[X]$ is a Euclidean domain using the degree function, and V_α is finitely generated as a $\mathbb{C}[X]$-module. These are the assumptions of the primary decomposition theorem. Applying this, we find the module decomposition as required, noting that the primes in $\mathbb{C}[X]$ are the linear polynomials. Note that the free factor $\mathbb{C}[X]$ cannot appear in the decomposition since V is finite-dimensional.

We have already seen that for a module $W_\alpha \cong {}^{F[X]}\!/_{((X-\lambda)^n)}$, multiplication by X is represented by the matrix $J_n(\lambda)$ with respect to the basis $1, (X-\lambda), \dots, (X-\lambda)^{n-1}$. Hence the result follows by considering the union of these bases. □

Remark. If α is represented by a matrix A, then A is similar to a matrix in Jordan normal form. This is the form of the result often used in linear algebra.

The Jordan blocks are uniquely determined up to reordering. This can be proven by considering the dimensions of the *generalised eigenspaces*, which are $\ker((\alpha - \lambda \operatorname{id})^m)$ for some $m \in \mathbb{N}$.

The minimal polynomial of α is $\prod_\lambda (X-\lambda)^{c_\lambda}$ where c_λ is the size of the largest λ-block. The characteristic polynomial of α is $\prod_\lambda (X-\lambda)^{a_\lambda}$ where a_λ is the sum of the sizes of the λ-blocks.

The number of λ-blocks is the dimension of the eigenspace of λ.

12.12. Modules over principal ideal domains (non-examinable)

The structure theorem above was proven for Euclidean domains. This also holds for principal ideal domains. Some of the ideas relevant to this proof are illustrated in this subsection.

Theorem. Let R be a principal ideal domain. Then any finitely generated torsion-free R-module is free.

If R is a Euclidean domain, this was proven as a corollary to the structure theorem.

Lemma. Let R be a principal ideal domain and M be an R-module. Let $r_1, r_2 \in R$ be not both zero, and let d be their greatest common divisor. Then,

(i) there exists $A \in SL_2(R)$ such that

$$A \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}$$

(ii) if $x_1, x_2 \in M$, then there exist $x_1', x_2' \in M$ such that $Rx_1 + Rx_2 = Rx_1' + Rx_2'$, and $r_1x_1 + r_2x_2 = dx_1' + 0 \cdot x_2'$.

Proof. Since R is a principal ideal domain, $(r_1, r_2) = (d)$. Hence, by definition, $d = \alpha r_1 + \beta r_2$ for some $\alpha, \beta \in R$. Let $r_1 = s_1 d$ and $r_2 = s_2 d$. Then $\alpha s_1 + \beta s_2 = 1$. Now, let

$$A = \begin{pmatrix} \alpha & \beta \\ -s_2 & s_1 \end{pmatrix} \implies \det A = 1; \quad A \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}$$

as required.

For the second part, let $x_1' = s_1x_1 + s_2x_2$ and $x_2' = -\beta x_1 + \alpha x_2$. Then $Rx_1' + Rx_2' \subseteq Rx_1 + Rx_2$. The matrix defining x_1', x_2' in terms of x_1, x_2 is invertible since its determinant is a unit; we can solve for x_1, x_2 in terms of x_1', x_2'. So $Rx_1' + Rx_2' = Rx_1 + Rx_2$. Then by direct computation we can see that $r_1x_2 + r_2x_2 = dx_1' + 0 \cdot x_2'$. □

The structure theorem for principal ideal domains follows the same method; it is deduced for Smith normal form. That theorem also holds for principal ideal domains. The above lemma allows one to prove Smith normal form for principal ideal domains. In a Euclidean domain, we used the Euclidean function for a notion of size in order to perform induction; in a principal ideal domain we can count the irreducibles in a factorisation.

Proof of theorem. Let $M = Rx_1 + \cdots + Rx_n$ where n is minimal. If $x_1, \ldots, x_n$ are independent, then M is free as required. Suppose that the x_i are not independent, so there exists r_i such that $\sum r_i x_i = 0$ but not all of the r_i are zero. By reordering, we can suppose that $r_1 \neq 0$. By using part (ii) of the previous lemma, after replacing x_1 and x_2 by suitable x_1', x_2', we may assume that $r_1 \neq 0$ and $r_2 = 0$. By repeating this process with x_1 and x_i for all $i \geq 2$, we obtain $r_1 \neq 0$ and $r_2 = \cdots = r_n = 0$, so $r_1x_1'' = 0$ for some nonzero $x_1'' \in M$. But M is torsion-free, so r_1 must be zero, and this is a contradiction. □

IX. Complex Analysis

Lectured in Lent 2022 by Prof. N. Wickramasekera

Complex differentiation is a stronger notion than real differentiation. Many functions that are differentiable as a function of two real variables are not complex differentiable, for example the complex conjugate function. This stronger notion allows us to prove some surprising results. It turns out that if a function is complex differentiable once in a neighbourhood of a point, then it is given by a convergent power series in some neighbourhood of that point.

Another interesting result is Cauchy's integral formula: if a function is complex differentiable in a neighbourhood around a point, one can evaluate the function at that point using a certain integral over any loop around that point. A similar result can be used to obtain an arbitrary derivative of a function at a point by using a single integral.

Contents

1. Differentiation 488
1.1. Basic notions 488
1.2. Continuity and differentiability 488
1.3. Cauchy–Riemann equations 489
1.4. Curves and path-connectedness 490
1.5. Power series 491
1.6. Exponentials 492
1.7. Logarithms 493
1.8. Conformality 494
2. Integration 496
2.1. Introduction 496
2.2. Integrating along curves 497
2.3. Fundamental theorem of calculus 499
2.4. Star-shaped domains 500
2.5. Cauchy's integral formula 503
2.6. Liouville's theorem 505
2.7. Taylor series 506
2.8. Zeroes of holomorphic functions 508
2.9. Analytic continuation 509
2.10. Uniform limits of holomorphic functions 512
3. More integration 515
3.1. Winding numbers 515
3.2. Continuity of derivative function 517
3.3. Cauchy's theorem and Cauchy's integral formula 518
3.4. Homotopy 521
3.5. Simply connected domains 523
4. Singularities 524
4.1. Motivation 524
4.2. Removable singularities 524
4.3. Poles 525
4.4. Essential singularities 527
4.5. Laurent series 527
4.6. Coefficients of Laurent series 530
4.7. Residues 531
4.8. Jordan's lemma 533
5. The argument principle, local degree, and Rouché's theorem . . . 537
5.1. The argument principle 537

5.2. Local degree theorem 539
5.3. Open mapping theorem 540
5.4. Rouché's theorem 540

1. Differentiation

1.1. Basic notions

We use the following definitions.

- The complex plane is denoted $\mathbb{C}$.
- The complex conjugate of a complex number z is denoted $\overline{z}$.
- The modulus is denoted $|z|$.
- The function $d(z, w) = |z - w|$ is a metric on $\mathbb{C}$. All topological notions will be with respect to this metric.
- We define the disc $D(a, r) = \{z \in \mathbb{C} : |z - a| < r\}$ to be the open ball with centre a and radius r.
- A subset $U \subset \mathbb{C}$ is said to be open if it is open with respect to the above metric. In particular, by identifying $\mathbb{C}$ with $\mathbb{R}^2$, we can see that $U \subset \mathbb{C}$ is open if and only if $U \subset \mathbb{R}^2$ is open with respect to the Euclidean metric.

The course concerns itself with complex-valued functions of a single complex variable. Identifying $\mathbb{C}$ with $\mathbb{R}^2$ allows us to construct $f(z) = u(x, y) + iv(x, y)$, where u, v are real-valued functions. We can denote these parts by $u = \mathrm{Re}(f)$ and $v = \mathrm{Im}(f)$.

1.2. Continuity and differentiability

The definition of continuity is carried over from metric spaces. That is, $f : A \to \mathbb{C}$ is continuous at a point $w \in A$ if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall z \in A, |z - w| < \delta \implies |f(z) - f(w)| < \varepsilon$$

Equivalently, the limit $\lim_{z \to w} f(z)$ exists and takes the value $f(w)$. We can easily check that f is continuous at $w = c + id \in A$ if and only if u, v are continuous at (c, d) with respect to the Euclidean metric on $A \subset \mathbb{R}^2$.

Definition. Let $f : U \to \mathbb{C}$, where U is open in $\mathbb{C}$.

(i) f is *differentiable* at $w \in U$ if the limit

$$f'(w) = \lim_{z \to w} \frac{f(z) - f(w)}{z - w}$$

exists, and its value is complex. We say that $f'(w)$ is the derivative of f at w.

(ii) f is *holomorphic* at $w \in U$ if there exists $\varepsilon > 0$ such that $D(w, \varepsilon) \subset U$ and f is differentiable at every point in $D(w, \varepsilon)$.

(iii) f is holomorphic in U if f is holomorphic at every point in U, or equivalently, f is differentiable everywhere.

Differentiation of composite functions, sums, products and quotients can be computed in the complex case exactly as they are in the real case.

Example. Polynomials $p(z) \sum_{j=0}^{n} a_j z^j$ for complex coefficients a_j are holomorphic on $\mathbb{C}$. Further, if p, q are polynomials, $\frac{p}{q}$ is holomorphic on $\mathbb{C} \setminus \{z : q(z) = 0\}$.

Remark. The differentiability of f at a point $c + id$ is not equivalent to the differentiability of u, v at (c, d). $u : U \to \mathbb{R}$ is differentiable at $(c, d) \in U$ if there is a 'good' affine approximation of u at (c, d); there exists a linear transformation $L : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\lim_{(x,y)\to(c,d)} \frac{u(x, y) - (u(c, d) - L(x - c, y - d))}{\sqrt{(x - c)^2 + (y - d)^2}} = 0$$

If u is differentiable at (c, d), then L is uniquely defined, and can be denoted $L = Du(c, d)$. L is given by the partial derivatives of u, which are

$$L(x, y) = \left(\frac{\partial u}{\partial x}(c, d)\right)x + \left(\frac{\partial u}{\partial y}(c, d)\right)y$$

This seems to imply that the differentiability of f requires more than the differentiability of u, v.

1.3. Cauchy–Riemann equations

Theorem. $f = u + iv : U \to \mathbb{C}$ is differentiable at $w = c + id \in U$ if and only if $u, v : U \to \mathbb{R}$ are differentiable at $(c, d) \in U$ and u, v satisfy the Cauchy–Riemann equations at (c, d), which are

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}; \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

If f is differentiable at $w = c + id$, then

$$f'(w) = \frac{\partial u}{\partial x}(c, d) + i\frac{\partial v}{\partial x}(c, d)$$

and other expressions, which follow directly from the Cauchy–Riemann equations.

Proof. All of the following statements will be bi-implications. Suppose f is differentiable at w with $f'(w) = p + iq$, so

$$\lim_{z\to w} \frac{f(z) - f(w)}{z - w} = p + iq$$

$$\lim_{z\to w} \frac{f(z) - f(w) - (z - w)(p + iq)}{|z - w|} = 0$$

By separating real and imaginary parts, writing $w = c + id$ we have

$$\lim_{(x,y)\to(c,d)} \frac{u(x, y) - u(c, d) - p(x - c) + q(y - d)}{\sqrt{(x - c)^2 + (y - d)^2}} = 0$$

$$\lim_{(x,y)\to(c,d)} \frac{v(x, y) - v(c, d) - q(x - c) - p(y - d)}{\sqrt{(x - c)^2 + (y - d)^2}} = 0$$

Thus, u is differentiable at (c, d) with $Du(c, d)(x, y) = px - qy$ and v is differentiable at (c, d) with $Dv(c, d)(x, y) = qx + py$.

$$u_x(c, d) = v_y(c, d) = p; \quad -u_y(c, d) = v_x(c, d) = q$$

Hence the Cauchy–Riemann equations hold at (c, d). We also find that if f is differentiable at w, we have $f'(w) = u_x(c, d) + iv_x(c, d)$. □

Remark. If u, v simply satisfy the Cauchy–Riemann equations alone, that does not imply differentiability of f. u, v must also be differentiable.

Remark. If we simply want to show that the differentiability of f implies that the Cauchy–Riemann equations hold, we can proceed in a simpler way. For $t \in \mathbb{R}$,

$$f'(w) = \lim_{t \to 0} \left(\frac{u(c + t, d) - u(c, d)}{t} + i \frac{v(c + t, d) - v(c, d)}{t} \right)$$

Hence the real part and the complex part both exist, so $u_x(c, d)$ and $v_x(c, d)$ exist, and $f'(w) = u_x(c, d) + iv_x(c, d)$. If we instead considered a perturbation along the imaginary axis, we find $f'(w) = v_y(c, d) - iu_y(c, d)$, giving the Cauchy–Riemann equations.

Example. The complex conjugate function $z \mapsto \overline{z}$ is not differentiable. Here, $u(x, y) = x$, and $v(x, y) = -y$, so the Cauchy–Riemann equations do not hold.

Corollary. If u, v have continuous partial derivatives at (c, d) and satisfy the Cauchy–Riemann equations at this point, then f is differentiable at $c + id$. In particular, if u, v are C^1 functions on U (i.e. have continuous partial derivatives in U) satisfying the Cauchy–Riemann equations everywhere, then f is holomorphic (in U).

Proof. If u, v have continuous partial derivatives then u, v are differentiable at (c, d) by Analysis and Topology. □

1.4. Curves and path-connectedness

Definition. A *curve* is a continuous function $\gamma : [a, b] \to \mathbb{C}$, where $a, b \in \mathbb{R}$. γ is a C^1 curve if γ' exists and is continuous on $[a, b]$. An open set $U \subset \mathbb{C}$ is *path-connected* if for any two points $z, w \in U$, there exists $\gamma : [0, 1] \to U$ such that $\gamma(0) = z$ and $\gamma(1) = w$. A *domain* is a non-empty, open, path-connected subset of $\mathbb{C}$.

Corollary. Let U be a domain. Let $f : U \to \mathbb{C}$ be a holomorphic function with derivative zero everywhere. Then f is constant on U.

Proof. By the Cauchy–Riemann equations, $f' = 0$ implies that $Du = Dv = 0$ in U. By Analysis and Topology, the path-connectedness of U implies that u and v are constant functions. □

1.5. Power series

Recall the following theorem from IA Analysis.

Theorem. Let $(c_n)_{n=0}^{\infty}$ be a sequence of complex numbers. Then, the power series

$$\sum_{n=0}^{\infty} c_n(z-a)^n$$

has a unique *radius of convergence* $R \in [0,\infty]$ such that the power series converges absolutely for $|z-a| < R$ and diverges if $|z-a| > R$. Further, if $0 < r < R$, the series converges uniformly with respect to z on the compact disc $D(a,r)$.

Note that

$$R = \sup\left\{r \geq 0 : \lim_{n\to\infty} |c_n| r^n = 0\right\}; \quad \frac{1}{R} = \limsup_{n\to\infty} |c_n|^{\frac{1}{n}}$$

Theorem. Let the sequence (c_n) define a power series f centred around a with positive radius of convergence R. Then, the function $f : D(a,R) \to \mathbb{C}$ satisfies

(i) f is holomorphic on $D(a,R)$;

(ii) the term-by-term differentiated series $\sum_{n=1}^{\infty} nc_n(z-a)^{n-1}$ also has radius of convergence equal to R, and this series is exactly the value of f';

(iii) f has derivatives of all orders on $D(a,R)$ and $c_n = \frac{f^{(n)}(a)}{n!}$;

(iv) if f vanishes on $D(a,\varepsilon)$ for any $\varepsilon > 0$, then $f \equiv 0$ on $D(a,R)$.

Proof. (i) Without loss of generality, let $a = 0$. $\sum_{n=1}^{\infty} nc_n(z-a)^{n-1}$ has some radius of convergence R_1.

Let $z \in D(0,R)$ and choose ρ such that $|z| < \rho < R$. Then,

$$n|c_n||z|^{n-1} = n|c_n|\left|\frac{z}{\rho}\right|^{n-1}\rho^{n-1} \leq |c_n|\rho^{n-1}$$

for sufficiently large n, since $n\left|\frac{z}{\rho}\right|^{n-1} \to 0$ as $n \to \infty$. Since $\sum |c_n|\rho^n$ converges, we must have that $n|c_n||z|^{n-1}$ converges. Hence $R_1 \geq R$.

Now, since

$$|c_n||z|^n \leq n|c_n||z|^n = |z|\left(n|c_n||z|^{n-1}\right)$$

If $\sum n|c_n|z^{n-1}$ converges then so does $\sum |c_n||z|^n$. Hence $R_1 \leq R$. This leads us to conclude $R_1 = R$.

(ii) Let $z \in D(0,R)$. The statement that f' is the above differentiated power series at z is equivalent to continuity at z of the function

$$g : D(0,R) \to \mathbb{C}; \quad g(w) = \begin{cases} \frac{f(w)-f(z)}{w-z} & w \neq z \\ \sum_{n=1}^{\infty} n c_n z^{n-1} & w = z \end{cases}$$

Substituting for f, we have $g(w) = \sum_{n=1}^{\infty} h_n(w)$ for $w \in D(0,R)$ where

$$h_n(w) = \begin{cases} \frac{c_n(w^n - z^n)}{w-z} & w \neq z \\ n c_n z^{n-1} & w = z \end{cases}$$

Note that h_n is continuous on $D(0,R)$. Further, note that

$$\frac{w^n - z^n}{w - z} = \sum_{j=0}^{n-1} z^j w^{n-1-j}$$

We have that for all r with $|z| < r < R$ and all $w \in D(0,r)$, $|h_n|(w) \leq n|c_n|r^{n-1} \equiv M_n$. Since $\sum M_n < \infty$, the Weierstrass M test shows that $\sum h_n$ converges uniformly on $D(0,r)$. A uniform limit of continuous functions is continuous, hence $g = \sum h_n$ is continuous in $D(0,r)$ and in particular at z.

(iii) Part (ii) can be applied inductively. The equation $c_n = \frac{f^{(n)}(a)}{n!}$ can be found by differentiating the series n times.

(iv) If $f \equiv 0$ in some disc $D(a,\varepsilon)$, then $f^{(n)}(a) = 0$ for all n. Thus the power series is identically zero.

□

1.6. Exponentials

Definition. If $f : \mathbb{C} \to \mathbb{C}$ is holomorphic on $\mathbb{C}$, we say that f is *entire*.

Definition. The *complex exponential function* is defined by

$$e^z = \exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

Proposition. (i) e^z is entire, and $(e^z)' = e^z$;

(ii) $e^z \neq 0$ and $e^{z+w} = e^z e^w$ for all complex z, w;

(iii) $e^{x+iy} = e^x(\cos y + i \sin y)$ for real x, y;

(iv) $e^z = 1$ if and only if $z = 2\pi n i$ for an integer n;

(v) if $z \in \mathbb{C}$, then there exists w such that $e^w = z$ if and only if $z \neq 0$.

Proof. (i) We can show that the radius of convergence is infinite. We can thus differentiate term by term and find $(e^z)' = e^z$.

(ii) Let $w \in \mathbb{C}$, and $F(z) = e^{z+w}e^{-w}$. Then we have

$$F'(z) = -e^{z+w}e^{-z} + e^{z+w}e^{-z} = 0$$

Hence $F(z)$ is constant. But $F(0) = e^w$, so $F(z) = e^w$. Taking $w = 0$, we have $e^z e^{-z} = 1$, so $e^z \neq 0$. Further, $e^{z+w} = e^z e^w$.

(iii) By part (ii), $e^{x+iy} = e^x e^{iy}$. Then, the series expansions of the sine and cosine functions can be used to finish the proof.

The rest of the proof is left as an exercise, which follows from (iii). □

1.7. Logarithms

Definition. Let $z \in \mathbb{C}$. Then, $w \in \mathbb{C}$ is a *logarithm* of z if $e^w = z$.

By part (v) above, z has a logarithm if and only if $z \neq 0$. In particular, $z \neq 0$ has infinitely many logarithms of the form $w + 2\pi i n$ for $n \in \mathbb{Z}$. If w is a logarithm of z, then $e^{\operatorname{Re} w} = |z|$, and hence $\operatorname{Re}(w) = \ln|z|$, where ln here is the unique real logarithm. In particular, $\operatorname{Re}(w)$ is uniquely determined by z.

Definition. Let $U \subset \mathbb{C} \setminus \{0\}$ be an open set. A *branch of logarithm* on U is a continuous function $\lambda : U \to \mathbb{C}$ such that $e^{\lambda(z)} = z$ for all $z \in U$.

Remark. Note that if λ is a branch of logarithm on U then λ is holomorphic in U with $\lambda'(z) = \frac{1}{z}$.

Proof. If $w \in U$ we have

$$\begin{aligned}
\lim_{z\to w} \frac{\lambda(z)-\lambda(w)}{z-w} &= \lim_{z\to w} \frac{\lambda(z)-\lambda(w)}{e^{\lambda(z)}-e^{\lambda(w)}} \\
&= \lim_{z\to w} \frac{1}{\left(\frac{e^{\lambda(z)}-e^{\lambda(w)}}{\lambda(z)-\lambda(w)}\right)} \\
&= \frac{1}{e^{\lambda(w)}} \lim_{z\to w} \frac{1}{\left(\frac{e^{\lambda(z)-\lambda(w)}-1}{\lambda(z)-\lambda(w)}\right)} \\
&= \frac{1}{e^{\lambda(w)}} \lim_{h\to 0} \frac{1}{\left(\frac{e^h-1}{h}\right)} \\
&= \frac{1}{e^{\lambda(w)}} \\
&= \frac{1}{w}
\end{aligned}$$

□

Definition. The *principal branch of logarithm* is the function

$$\mathrm{Log}:\ U_1 = \mathbb{C} \setminus \{x \in \mathbb{R} :\ x \leq 0\} \to \mathbb{C}; \quad \mathrm{Log}(z) = \ln|z| + i\arg(z)$$

where $\arg(z)$ is the unique argument of $z \in U_1$ in $(-\pi, \pi)$.

This is a branch of logarithm. Indeed, to check continuity, note that $z \mapsto \log|z|$ is continuous on $\mathbb{C} \setminus \{0\}$, and $z \mapsto \arg(z)$ is continuous since $\theta \mapsto e^{i\theta}$ is a homeomorphism $(-\pi, \pi) \to \mathbb{S}^1 \setminus \{-1\}$, and $z \mapsto \frac{z}{|z|}$ is continuous on $\mathbb{C} \setminus \{0\}$. Further,

$$e^{\mathrm{Log}(z)} = e^{\ln|z|}e^{i\arg(z)} = |z|(\cos\arg z + i\sin\arg z) = z$$

Note that Log cannot be continuously extended to $\mathbb{C} \setminus \{0\}$, since $\arg z \to \pi$ as $z \to -1$ with $\mathrm{Im}(z) > 0$, and $\arg z \to -\pi$ as $z \to -1$ with $\mathrm{Im}(z) < 0$. We will later prove that no branch of logarithm can exist on all of $\mathbb{C} \setminus \{0\}$.

Proposition. (i) Log is holomorphic on U_1 with $(\mathrm{Log}\, z)' = \frac{1}{z}$; and

(ii) for $|z| < 1$, we have

$$\mathrm{Log}(1+z) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}z^n}{n}$$

Proof. Part (i) follows from the above. The radius of convergence of the given series is one, and $1 + z \in U_1$, so both sides of the equation are defined on the unit disc. Then,

$$F(z) = \mathrm{Log}(1+z) - \sum_{n=1}^{\infty} \frac{(-1)^{n-1}z^n}{n} \implies F'(z) = \frac{1}{1+z} - \sum_{n=1}^{\infty}(-z)^{n-1} = 0 \implies F(z) = F(0) = 0$$

□

We can now define the *principal branch of* z^α by

$$z^\alpha = e^{\alpha\,\mathrm{Log}(z)}$$

Note that z^α is holomorphic on U_1 with $(z^\alpha)' = \alpha z^{\alpha-1}$. We can use exponentials to define the trigonometric and hyperbolic functions, which are all entire functions with derivatives matching those of the real definitions of these functions.

1.8. Conformality

Let $f:\ U \to \mathbb{C}$ be holomorphic, where U is an open set. Let $w \in U$ and suppose that $f'(w) \neq 0$. Let $\gamma_1, \gamma_2:\ [-1, 1] \to U$ be C^1 curves, such that $\gamma_i(0) = w$ and $\gamma_i'(0) \neq 0$. Then $f \circ \gamma_i$ are C^1 curves passing through $f(w)$. Further, $(f \circ \gamma_i)'(0) = f'(w)\gamma_i'(0) \neq 0$. Thus

$$\frac{(f \circ \gamma_1)'(0)}{(f \circ \gamma_2)'(0)} = \frac{\gamma_1'(0)}{\gamma_2'(0)}$$

Hence,

$$\arg(f \circ \gamma_1)'(0) - \arg(f \circ \gamma_2)'(0) = \arg \gamma_1'(0) - \arg \gamma_2'(0)$$

In other words, the angle that the curves make when they intersect at w is the same angle that their images $f \circ \gamma_i$ make when they intersect at $f(w)$, and the orientation also is preserved (clockwise or anticlockwise). Hence, f is angle-preserving at w whenever $f'(w) \neq 0$. In particular, if γ_i are tangential at w, the curves $f \circ \gamma_i$ are tangential at $f(w)$.

Remark. If f is C^1, then the converse holds. If $w \in U$ and $(f \circ \gamma)'(0) \neq 0$ for any C^1 curve γ with $\gamma(0) = w$ and $\gamma'(0) \neq 0$, and if f is angle-preserving at w in the above sense, then $f'(w)$ exists and is nonzero.

Definition. A holomorphic function $f : U \to \mathbb{C}$ on an open set U is *conformal* at $w \in U$ if $f'(w) \neq 0$.

Definition. Let $U, \widetilde{U}$ be domains in $\mathbb{C}$. A map $f : U \to \widetilde{U}$ is a *conformal equivalence* between $U, \widetilde{U}$ if f is a bijective holomorphic map with $f'(z) \neq 0$ for all $z \in U$.

Remark. We will prove later that if f is holomorphic and injective, then $f'(z) \neq 0$ for all z. Thus, in the above definition, the condition $f'(z) \neq 0$ is redundant.

Remark. It is automatic that $f^{-1} : \widetilde{U} \to U$ is holomorphic, which will follow from the holomorphic inverse function theorem.

Example. Möbius maps

$$f(z) = \frac{az + b}{cz + d}$$

are conformal on $\mathbb{C} \setminus \{-d/c\}$ if $c \neq 0$, and conformal on $\mathbb{C}$ if $c = 0$. Möbius maps are sometimes used as explicit conformal equivalences between subdomains of $\mathbb{C}$. For instance, let $\mathbb{H}$ be the open upper half plane in $\mathbb{C}$. Then

$$z \in \mathbb{H} \iff |z - i| < |z + i| \iff \left|\frac{z - i}{z + i}\right| < 1$$

Thus the map $z \mapsto \frac{z-i}{z+i}$ maps $\mathbb{H}$ onto $D(0, 1)$, so g is a conformal equivalence.

Example. Let $f : z \mapsto z^n$ for $n \geq 1$. Then

$$f : \left\{z \in \mathbb{C} \setminus \{0\} : 0 < \arg z < \frac{\pi}{n}\right\} \to \mathbb{H}$$

is the restricted map on a sector. The restricted f is a conformal equivalence with $f^{-1}(z) = z^{1/n}$, the principal branch of $z^{1/n}$.

Example. The function

$$\exp : \{z \in \mathbb{C} : -\pi < \operatorname{Im} z < \pi\} \to \mathbb{C} \setminus \{x \in \mathbb{R} : x \leq 0\}$$

is a conformal equivalence, with inverse Log.

Theorem (Riemann mapping theorem)**.** *This theorem is non-examinable.*

Any simply connected domain $U \subset \mathbb{C}$ with $U \neq \mathbb{C}$ is conformally equivalent to $D(0, 1)$.

2. Integration

2.1. Introduction

Definition. If $f : [a, b] \subset \mathbb{R} \to \mathbb{C}$ is a complex function, and the real and imaginary parts of f are Riemann integrable, then we define

$$\int_a^b f(t)\,\mathrm{d}t = \int_a^b \mathrm{Re}(f(t))\,\mathrm{d}t + i\int_a^b \mathrm{Im}(f(t))\,\mathrm{d}t$$

In particular, for $g : [a, b] \to \mathbb{R}$, we have

$$\int_a^b ig(t)\,\mathrm{d}t = i\int_a^b g(t)\,\mathrm{d}t$$

Thus, for a complex constant $w \in \mathbb{C}$, we can find

$$\int_a^b wf(t)\,\mathrm{d}t = w\int_a^b f(t)\,\mathrm{d}t$$

Proposition (basic estimate)**.** If $f : [a, b] \to \mathbb{C}$ is continuous, then

$$\left|\int_a^b f(t)\,\mathrm{d}t\right| \le \int_a^b |f(t)|\,\mathrm{d}t \le (b-a)\sup_{t\in[a,b]} |f(t)|$$

Equality holds if and only if f is constant.

Proof. If $\int_a^b f(t)\,\mathrm{d}t = 0$ then the proof is complete. Otherwise, we can write the value of the integral as $re^{i\theta}$ for $\theta \in [0, 2\pi)$. Let $M = \sup_{t\in[a,b]} |f(t)|$. Then we have

$$\begin{aligned}\left|\int_a^b f(t)\,\mathrm{d}t\right| &= r\\ &= e^{-i\theta}\int_a^b f(t)\,\mathrm{d}t\\ &= \int_a^b e^{-i\theta}f(t)\,\mathrm{d}t\\ &= \int_a^b \mathrm{Re}(e^{-i\theta}f(t))\,\mathrm{d}t + i\int_a^b \mathrm{Im}(e^{-i\theta}f(t))\,\mathrm{d}t\end{aligned}$$

Since the left hand side is real, the imaginary integral vanishes.

$$\begin{aligned}\left|\int_a^b f(t)\,\mathrm{d}t\right| &= \int_a^b \mathrm{Re}(e^{-i\theta}f(t))\,\mathrm{d}t\\ &\le \int_a^b \left|e^{-i\theta}f(t)\right|\mathrm{d}t = \int_a^b |f(t)|\,\mathrm{d}t\\ &\le (b-a)M\end{aligned}$$

Equality holds if and only if $|f(t)| = M$ and $\mathrm{Re}(e^{-i\theta} f(t)) = M$ for all $t \in [a, b]$, which is true only if $|f(t)| = M$ and $\arg(f(t)) = \theta$ hence $f = Me^{i\theta}$ for all t. □

2.2. Integrating along curves

Definition. Let $U \subset \mathbb{C}$ be an open set and let $f : U \to \mathbb{C}$ be continuous. Let $\gamma : [a, b] \to U$ be a C^1 curve. Then the *integral of f along γ* is

$$\int_\gamma f(z)\,\mathrm{d}z = \int_a^b f(\gamma(t))\gamma'(t)\,\mathrm{d}t$$

This definition is consistent with the previous definition of the integral of a function f along the interval $[a, b]$. The integral along a curve has various convenient properties.

(i) It is invariant under the choice of parametrisation. Let $\varphi : [a_1, b_1] \to [a, b]$ be C^1 and injective with $\varphi(a_1) = a$ and $\varphi(b_1) = b$. Let $\delta = \gamma \circ \varphi : [a_1, b_1] \to U$. Then

$$\int_\delta f(z)\,\mathrm{d}z = \int_\gamma f(z)\,\mathrm{d}z$$

Indeed,

$$\begin{aligned}\int_\delta f(z)\,\mathrm{d}z &= \int_{a_1}^{b_1} f(\gamma(\varphi(t)))\gamma'(\varphi(t))\varphi'(t)\,\mathrm{d}t \\ &= \int_a^b f(\gamma(s))\gamma'(s)\,\mathrm{d}s \\ &= \int_\gamma f(z)\,\mathrm{d}z\end{aligned}$$

(ii) The integral is linear. It is easy to check that

$$\int_\gamma (\lambda f(z) + \mu g(z))\,\mathrm{d}z = \lambda \int_\gamma f(z)\,\mathrm{d}z + \mu \int_\gamma g(z)\,\mathrm{d}z$$

for complex constants $\lambda, \mu \in \mathbb{C}$.

(iii) The additivity property states that if $\gamma : [a, b] \to U$ is C^1 and $a < c < b$, then

$$\int_\gamma f(z)\,\mathrm{d}z = \int_{\gamma|_{[a,c]}} f(z)\,\mathrm{d}z + \int_{\gamma|_{c,b}} f(z)\,\mathrm{d}z$$

(iv) We define the *inverse path* $(-\gamma) : [-b, -a] \to U$ by $(-\gamma)(t) = \gamma(-t)$. Then

$$\int_{(-\gamma)} f(z)\,\mathrm{d}z = -\int_\gamma f(z)\,\mathrm{d}z$$

Definition. Let $\gamma : [a, b] \to \mathbb{C}$ be a C^1 curve. Then the *length* of γ is

$$\text{length}(\gamma) = \int_a^b |\gamma'(t)| \, \mathrm{d}t$$

Definition. A *piecewise C^1 curve* is a continuous map $\gamma : [a, b] \to \mathbb{C}$ such that there exists a finite subdivision

$$a = a_0 < a_1 < \cdots < a_n = b$$

such that each $\gamma_j = \gamma|_{[a_{j-1}, a_j]}$ is C^1 for $1 \leq j \leq n$. Then, for such a piecewise C^1 curve, we define

$$\int_\gamma f(z) \, \mathrm{d}z = \sum_{j=1}^n \int_{\gamma_j} f(z) \, \mathrm{d}z$$

and

$$\text{length}(\gamma) = \sum_{j=1}^n \text{length}(\gamma_j) = \sum_{j=1}^n \int_{a_{j-1}}^{a_j} |\gamma'(t)| \, \mathrm{d}t$$

By the additivity property, both definitions are invariant under changing the subdivision. From here, we will use 'curve' to refer to 'piecewise C^1 curve', unless stated otherwise.

Definition. If $\gamma_1 : [a, b] \to \mathbb{C}$ and $\gamma_2 : [c, d]$ are curves with $\gamma_1(b) = \gamma_2(c)$, we define the *sum* of γ_1 and γ_2 to be the curve

$$(\gamma_1 + \gamma_2) : [a, b + d - c] \to \mathbb{C}; \quad (\gamma_1 + \gamma_2)(t) = \begin{cases} \gamma_1(t) & a \leq t \leq b \\ \gamma_2(t - b + c) & b \leq t \leq b + d - c \end{cases}$$

Proposition. Let $f : U \to \mathbb{C}$ be continuous and $\gamma : [a, b] \to \mathbb{C}$, we have

$$\left| \int_\gamma f(z) \, \mathrm{d}z \right| \leq \text{length}(\gamma) \sup_\gamma |f|$$

where $\sup_\gamma g \equiv \sup_{t \in [a,b]} g(\gamma(t))$.

Proof. If γ is C^1, then

$$\left| \int_\gamma f(z) \, \mathrm{d}z \right| = \left| \int_a^b f(\gamma(t)) \gamma'(t) \, \mathrm{d}t \right| \leq \int_a^b |f(\gamma(t))| \cdot |\gamma'(t)| \, \mathrm{d}t \leq \sup_{t \in [a,b]} |f(\gamma(t))| \text{length}(\gamma)$$

If γ is piecewise C^1, then the result follows from the definition of a piecewise C^1 function and the above. □

2.3. Fundamental theorem of calculus

Theorem (fundamental theorem of calculus)**.** Let $f : U \to \mathbb{C}$ be continuous on an open set $U \subset \mathbb{C}$. Let $F : U \to \mathbb{C}$ be a function such that $F'(z) = f(z)$ for all $z \in U$. Then, for any curve $\gamma : [a, b] \to U$, we have

$$\int_\gamma f(z)\,\mathrm{d}z = F(\gamma(b)) - F(\gamma(a))$$

If γ is a closed curve, then $\int_\gamma f(z) = 0$. Such a function F is known as an *antiderivative* of f.

Proof.

$$\int_\gamma f(z)\,\mathrm{d}z = \int_a^b f(\gamma(t))\gamma'(t)\,\mathrm{d}t = \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}F(\gamma(t))\,\mathrm{d}t = F(\gamma(b)) - F(\gamma(a))$$

□

Remark. Note that we assume that F exists such that $F'(z) = f(z)$; such an F is not provided for by the theorem.

Example. For an integer n and the curve $\gamma(t) = Re^{2\pi it}$ for $t = [0, 1]$, consider the integral $\int_\gamma z^n\,\mathrm{d}z$. For $n \neq -1$, the function $\frac{z^{n+1}}{n+1}$ is an antiderivative of z^n. Hence, $\int_\gamma z^n\,\mathrm{d}z = 0$ since γ is a closed curve. If $n = -1$, we can use the definition of the integral to find

$$\int_\gamma \frac{1}{z}\,\mathrm{d}z = \int_0^1 \frac{1}{\gamma(t)}\gamma'(t)\,\mathrm{d}t = \int_0^1 \frac{1}{Re^{2\pi it}}2\pi iRe^{2\pi it}\,\mathrm{d}t = 2\pi i$$

This is not zero, hence for all $R > 0$, $\frac{1}{z}$ has no antiderivative in any open set containing the circle $\{|z| = R\}$. In particular, for any branch of logarithm λ, it has derivative $\frac{1}{z}$, hence there exists no branch of logarithm on $\mathbb{C}^\star = \mathbb{C} \setminus \{0\}$.

Theorem (converse to fundamental theorem of calculus)**.** Let $U \subset \mathbb{C}$ be a domain. If $f : U \to \mathbb{C}$ is continuous and if $\int_\gamma f(z)\,\mathrm{d}z = 0$ for every closed curve γ in U, then f has an antiderivative. In other words, there exists a holomorphic function $F : U \to \mathbb{C}$ such that $F' = f$ in U.

Proof. Let $a_0 \in U$. Then for $w \in U$, we can define

$$F(w) = \int_{\gamma_w} f(z)\,\mathrm{d}z$$

where $\gamma_w : [0, 1] \to \mathbb{C}$ is a curve with $\gamma_w(0) = a_0$ and $\gamma_w(1) = w$.

The definition of F is independent of the choice of γ_w. Indeed, suppose two paths γ_w, γ'_w exist. Then the curve $\gamma_w + (-\gamma'_w)$ is a closed path, and by assumption the integral along this

curve is zero. Thus F is independent of the choice of path as claimed. So F is a well-defined function.

Now, let $w \in U$. Since U is an open set, there exists $r > 0$ such that $D(w, r) \subset U$. For $h \in \mathbb{C}$ with $0 < |h| < r$, let δ_h be the radial path $t \mapsto w + th$ for $t \in [0, 1]$. Now we define

$$\gamma = \gamma_w + \delta_h + (-\gamma_{w+h})$$

This is a closed curve contained within U, hence $\int_\gamma f(z)\,\mathrm{d}z = 0$. Thus

$$\int_{\gamma_{w+h}} f(z)\,\mathrm{d}z = \int_{\gamma_w} f(z)\,\mathrm{d}z + \int_{\delta_h} f(z)\,\mathrm{d}z$$

Informally, the integral has an additivity property which is independent of the path taken. Rewriting this using F,

$$\begin{aligned} F(w+h) &= F(w) + \int_{\delta_h} f(z)\,\mathrm{d}z \\ &= F(w) + \int_{\delta_h} (f(z) + f(w) - f(w))\,\mathrm{d}z \\ &= F(w) + hf(w) + \int_{\delta_h} (f(z) - f(w))\,\mathrm{d}z \end{aligned}$$

Hence, by continuity of f,

$$\begin{aligned} \left|\frac{F(w+h) - F(w)}{h} - f(w)\right| &= \frac{1}{|h|}\left|\int_{\delta_h} (f(z) - f(w))\,\mathrm{d}z\right| \\ &\leq \frac{1}{|h|}\operatorname{length}(\delta_h) \sup_{z \in \operatorname{Im}\delta_h} |f(z) - f(w)| \\ &= \sup_{z \in \operatorname{Im}\delta_h} |f(z) - f(w)| \\ \therefore \lim_{h\to 0}\left|\frac{F(w+h) - F(w)}{h} - f(w)\right| &= \lim_{h\to 0} \sup_{z \in \operatorname{Im}\delta_h} |f(z) - f(w)| = 0 \end{aligned}$$

Thus, F is differentiable at w with $F'(w) = f(w)$. □

2.4. Star-shaped domains

Definition. A domain U is *star-shaped*, or a *star domain*, if there exists a (not necessarily unique) centre $a_0 \in U$ such that for all $w \in U$, the straight line segment $[a_0, w]$ is contained within U.

Remark. Any disc is convex; any convex domain is star-shaped; any star-shaped domain is path-connected. The reverse implications are not true in general.

Definition. A *triangle* in $\mathbb{C}$ is the *convex hull* of three points in $\mathbb{C}$. The (closed) convex hull of a set S is the smallest (closed) convex set C such that $S \subseteq C$. In this case, if $z_1, z_2, z_3 \in \mathbb{C}$, we have

$$T = \{az_1 + bz_2 + cz_3 : 0 \leq a, b, c \leq 1, a + b + c = 1\}$$

When used as a curve, the boundary ∂T represents the piecewise affine closed curve $\gamma = \gamma_1 + \gamma_2 + \gamma_3$ where γ_i are affine functions parametrising the three line segments on the boundary of T.

Corollary. Let U be a star-shaped domain. Let $f : U \to \mathbb{C}$ be continuous and $\int_{\partial T} f(z)\,\mathrm{d}z = 0$ for any triangle $T \subset U$. Then f has an antiderivative in U.

Remark. This is a relaxation of the conditions from the previous theorem.

Proof. Let a_0 be a centre for the domain U. Let w be an arbitrary point in U. Then let γ_w be the affine function parametrising the directed line segment $[a_0, w]$, and let $F(w) = \int_{\gamma_w} f(z)\,\mathrm{d}z$. Using h and δ_h as above, by letting $\gamma = \gamma_w + \delta_h + (-\gamma_{w+h})$ we then have $\int f(z)\,\mathrm{d}z = \pm \int_{\partial T} f(z)\,\mathrm{d}z$ for a triangle $T \subset U$. Since the integral around a triangle is zero by hypothesis, $\int_\gamma f(z)\,\mathrm{d}z = 0$. We then complete the proof in analogous way to the general case. □

Theorem (Cauchy's theorem for triangles)**.** Let $U \subset \mathbb{C}$ be an open set and $f : U \to \mathbb{C}$ be a holomorphic function. Then $\int_{\partial T} f(z)\,\mathrm{d}z = 0$ for all triangles $T \subset U$.

Proof. Let $\eta(t) = \int_{\partial T} f(z)\,\mathrm{d}z$. We will subdivide the triangle T into four smaller triangles $T^{(1)}, T^{(2)}, T^{(3)}, T^{(4)}$. The vertices of the inner triangle are the midpoints of the sides of T, and the three other triangles are constructed to fill the remaining area of T. Thus,

$$\eta(T) = \int_{\partial T^{(1)}} f(z)\,\mathrm{d}z + \int_{\partial T^{(2)}} f(z)\,\mathrm{d}z + \int_{\partial T^{(3)}} f(z)\,\mathrm{d}z + \int_{\partial T^{(4)}} f(z)\,\mathrm{d}z$$

Then, by the triangle inequality, there exists a triangle $T^{(j)}$ such that

$$\left|\int_{\partial T^{(j)}} f(z)\,\mathrm{d}z\right| \geq \frac{|\eta(T)|}{4}$$

Let $T_0 = T$, and $T_1 = T^{(j)}$, so $|\eta(T_1)| \geq \frac{1}{4}|\eta(T_0)|$. We can show geometrically that for any choice of T_i, $\mathrm{length}(\partial T_1) = \frac{1}{2}\mathrm{length}(\partial T_0)$. Inductively, we can subdivide T_i and produce T_{i+1}, such that

$$T_0 \supset T_1 \supset \cdots; \quad |\eta(T_n)| \geq \frac{1}{4}|\eta(T_{n-1})|; \quad \mathrm{length}(\partial T_n) = \frac{1}{2}\mathrm{length}(\partial T_{n-1})$$

Hence,

$$|\eta(T_n)| \geq \frac{1}{4^n}|\eta(T_0)|; \quad \mathrm{length}(\partial T_n) = \frac{1}{2^n}\mathrm{length}(\partial T_0)$$

Since T_n are non-empty, nested closed subsets with diameter converging to zero, we can show that $\bigcap_{n=1}^{\infty} T_n = \{z_0\}$ for some $z_0 \in \mathbb{C}$. Let $\varepsilon > 0$. Since f is differentiable at z_0, there exists $\delta > 0$ such that

$$\begin{aligned} z \in U, |z - z_0| < \delta &\implies \left| \frac{f(z) - f(z_0)}{z - z_0} - f'(z_0) \right| \leq \varepsilon \\ &\implies |f(z) - f(z_0) - f'(z_0)(z - z_0)| \leq \varepsilon |z - z_0| \end{aligned}$$

Now, observe that for all n,

$$\int_{\partial T_n} f(z)\,\mathrm{d}z = \int_{\partial T_n} (f(z) - f(z_0) - f'(z_0)(z - z_0))\,\mathrm{d}z$$

since $\int_{\partial T_n} \mathrm{d}z = \int_{\partial T_n} z\,\mathrm{d}z = 0$ by the fundamental theorem of calculus. Let n such that $T_n \subset D(z_0, \delta)$. Hence,

$$\begin{aligned} \frac{1}{4^n}|\eta(T_0)| &\leq |\eta(T_n)| \\ &= \left| \int_{\partial T_n} f(z)\,\mathrm{d}z \right| \\ &= \left| \int_{\partial T_n} (f(z) - f(z_0) - f'(z_0)(z - z_0))\,\mathrm{d}z \right| \\ &\leq \left(\sup_{z \in \partial T_n} |f(z) - f(z_0) - f'(z_0)(z - z_0)| \right) \mathrm{length}(\partial T_n) \\ &\leq \varepsilon \left(\sup_{z \in \partial T_n} |z - z_0| \right) \mathrm{length}(\partial T_n) \\ &\leq \varepsilon \cdot \mathrm{length}(\partial T_n)^2 \\ &= \frac{\varepsilon}{4^n} \mathrm{length}(\partial T_0)^2 \\ \therefore |\eta(T_0)| &\leq \varepsilon \cdot \mathrm{length}(\partial T_0)^2 \end{aligned}$$

ε was arbitrary, hence $\eta(T_0)$ must be zero. □

We can generalise the above theorem for functions that are holomorphic except at a finite number of points.

Theorem. Let $U \subset \mathbb{C}$ be an open set and $f : U \to \mathbb{C}$ be a continuous function. Let $S \subset U$ be a finite set and suppose that f is holomorphic on $U \setminus S$. Then $\int_{\partial T} f(z)\,\mathrm{d}z = 0$ for all triangles $T \subset U$.

Proof. By the procedure above, we can subdivide T into a total of 4^n smaller triangles; at each step we join the midpoints of the sides of the triangles of the previous step. We will keep all of the smaller triangles, and let the sequence of such smaller triangles be denoted

$T_1, \dots, T_N$. Then, since the integrals along the sides of the smaller triangles that are interior to T cancel, we have

$$\int_{\partial T} f(z)\,\mathrm{d}z = \sum_{j=1}^{N} \int_{\partial T_j} f(z)\,\mathrm{d}z$$

By the previous theorem, $\int_{\partial T_j} f(z)\,\mathrm{d}z = 0$ unless T_j intersects with S. So by letting $I = \{j : T_j \cap S \neq \varnothing\}$, we have

$$\int_{\partial T} f(z)\,\mathrm{d}z = \sum_{j\in I} \int_{\partial T_j} f(z)\,\mathrm{d}z$$

Since any point may be in at most six of the smaller triangles, and $\mathrm{length}(\partial T_j) = \frac{1}{2^n}\mathrm{length}(\partial T)$, we find

$$\left|\int_{\partial T} f(z)\,\mathrm{d}z\right| \leq 6|S|\left(\sup_{z\in T}|f(z)|\right)\frac{\mathrm{length}(\partial T)}{2^n}$$

Then let $n \to \infty$ and the result then holds as required. □

We can now prove the 'convex Cauchy' theorem.

Corollary (Cauchy's theorem for convex sets)**.** Let $U \subset \mathbb{C}$ be convex, or more generally, a star domain. Let $f : U \to \mathbb{C}$ be continuous on U and holomorphic in $U \setminus S$ where S is a finite set. Then $\int_\gamma f(z)\,\mathrm{d}z = 0$ for any closed curve γ in U.

Proof. By the theorems above, $\int_{\partial T} f(z)\,\mathrm{d}z = 0$ for any triangle $T \subset U$. Since U is a star domain and f is continuous, this means that f has an antiderivative in U. The result then follows from the fundamental theorem of calculus. □

Remark. We will soon show that if $f : U \to \mathbb{C}$ is continuous and holomorphic in $U \setminus S$ where S is finite, then f is holomorphic in U.

2.5. Cauchy's integral formula

For a disc $D(a,\rho)$ we will write $\int_{\partial D(a,\rho)} f(z)\,\mathrm{d}z$ to mean $\int_\gamma f(z)\,\mathrm{d}z$ where $\gamma : [0,1] \to \mathbb{C}$ is the curve $\gamma(t) = a + \rho e^{2\pi i t}$.

Theorem (Cauchy's integral formula for a disc)**.** Let $D = D(a,r)$ and let $f : D \to \mathbb{C}$ be holomorphic. Then, for any ρ with $0 < \rho < r$ and any $w \in D(a,\rho)$, we have

$$f(w) = \frac{1}{2\pi i}\int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{z-w}$$

In particular, taking $w = a$,

$$f(a) = \frac{1}{2\pi i}\int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{z-a} = \int_0^1 f(a+\rho e^{2\pi i t})\,\mathrm{d}t$$

This final equation is called the *mean value property* for holomorphic functions.

We first need the following lemma.

Lemma. If $\gamma : [a, b] \to \mathbb{C}$ is a curve and (f_n) is a sequence of continuous complex functions on $\operatorname{Im}\gamma$ converging uniformly to f on $\operatorname{Im}\gamma$, then $\int_\gamma f_n(z)\,\mathrm{d}z \to \int_\gamma f(z)\,\mathrm{d}z$.

Proof. We have

$$\left|\int_\gamma f_n(z)\,\mathrm{d}z - \int_\gamma f(z)\,\mathrm{d}z\right| = \left|\int_\gamma (f_n(z) - f(z))\,\mathrm{d}z\right| \leq \sup_{z\in\operatorname{Im}\gamma} |f_n(z) - f(z)|\operatorname{length}(\gamma)$$

□

We can now prove Cauchy's integral formula for a disc.

Proof. Let $w \in D(a, \rho)$ be fixed, and define $h : D \to \mathbb{C}$ by

$$h(z) = \begin{cases} \frac{f(z)-f(w)}{z-w} & \text{if } z \neq w \\ f'(w) & \text{if } z = w \end{cases}$$

Then h is continuous on D and holomorphic in $D \setminus \{w\}$. By Cauchy's theorem for convex sets,

$$\int_{\partial D(a,\rho)} h(z)\,\mathrm{d}z = 0$$

Substituting for h, we find

$$f(w)\int_{\partial D(a,\rho)} \frac{\mathrm{d}z}{z-w} = \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{z-w}$$

It now suffices to prove that

$$\int_{\partial D(a,\rho)} \frac{\mathrm{d}z}{z-w} = 2\pi i$$

Note that

$$\frac{1}{z-w} = \frac{1}{z-a+a-w} = \frac{1}{(z-a)\left(1-\frac{w-a}{z-a}\right)} = \sum_{j=0}^{\infty} \frac{(w-a)^j}{(z-a)^{j+1}}$$

where the convergence is uniform for $z \in \partial D(a, \rho)$ by the Weierstrass M-test. Therefore, by the above lemma, we interchange summation and integration to find

$$\int_{\partial D(a,\rho)} \frac{\mathrm{d}z}{z-w} = \sum_{j=0}^{\infty} (w-a)^j \int_{\partial D(a,\rho)} \frac{\mathrm{d}z}{(z-a)^{j+1}}$$

For $j \geq 1$, the function $\frac{1}{(z-a)^{j+1}}$ has an antiderivative in a neighbourhood of $\partial D(a, \rho)$, hence all integrals on the right hand side for $j \geq 1$ vanish. For $j = 0$, we can compute directly that $\int_{\partial D(a,\rho)} \frac{\mathrm{d}z}{z-a} = 2\pi i$. Hence, $\int_{\partial D(a,\rho)} \frac{\mathrm{d}z}{z-w} = 2\pi i$, proving Cauchy's integral formula.

Now, taking $w = a$ in Cauchy's integral formula, we find

$$f(a) = \frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{z-a}$$

By direct computation using the parametrisation $t \mapsto a + \rho e^{2\pi i t}$ for $t \in [0,1]$, we find

$$f(a) = \int_0^1 f(a + \rho e^{2\pi i t})\,\mathrm{d}t$$

as required. □

2.6. Liouville's theorem

Theorem. Let $f : \mathbb{C} \to \mathbb{C}$ be entire and bounded. Then f is constant. More generally, if f is entire with sublinear growth (there exist $K \geq 0$ and $\alpha < 1$ such that $|f(z)| \leq K(1 + |z|^{\alpha})$ for all $z \in \mathbb{C}$) then f is constant.

Proof. Let $w \in \mathbb{C}$ and $\rho > |w|$. By Cauchy's integral formula, we have

$$f(w) = \frac{1}{2\pi i} \int_{\partial D(0,\rho)} \frac{f(z)\,\mathrm{d}z}{z-w}; \quad f(0) = \frac{1}{2\pi i} \int_{\partial D(0,\rho)} \frac{f(z)\,\mathrm{d}z}{z}$$

Thus,

$$\begin{aligned} |f(w) - f(0)| &= \frac{1}{2\pi} \left| \int_{\partial D(0,\rho)} \frac{w f(z)\,\mathrm{d}z}{z(z-w)} \right| \\ &\leq \frac{|w|}{2\pi} \sup_{z \in \partial D(0,\rho)} \frac{|f(z)|}{|z| \cdot ||z| - |w||} \mathrm{length}(\partial D(0,\rho)) \\ &\leq \frac{|w| K(1+\rho^{\alpha})}{2\pi\rho(\rho - |w|)} 2\pi\rho \\ &= \frac{|w| K(1+\rho^{\alpha})}{\rho - |w|} \end{aligned}$$

By letting $\rho \to \infty$, we can conclude $f(w) = f(0)$. □

Theorem (fundamental theorem of algebra)**.** Every non-constant polynomial with complex coefficients has a complex root.

Proof. Let $p(z) = a_n z^n + \cdots + a_0$ be a complex polynomial of degree $n \geq 1$. Then $a_n \neq 0$, and for $z \neq 0$ we can write

$$p(z) = z^n \Big(a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n} \Big)$$

By the triangle inequality,

$$|p(z)| \geq |z|^n \left(|a_n| - \frac{|a_{n-1}|}{|z|} - \cdots - \frac{|a_0|}{|z|^n} \right)$$

Hence, there exists $R > 0$ such that $|p(z)| \geq 1$ for $|z| > R$.

Now, if $p(z) \neq 0$ for all z, then $g(z) = \frac{1}{p(z)}$ is entire. By the above, $|g(z)| \leq 1$ for $|z| > R$. By continuity of g, we have further that $|g(z)|$ is bounded from above on the compact set $\{|z| \leq R\}$. Hence, g is a bounded entire function. By Liouville's theorem, g is constant. Since p is non-constant, this is a contradiction. Hence p has a zero. □

Theorem (local maximum modulus principle)**.** Let $f : D(a, R) \to \mathbb{C}$ be holomorphic, and $|f(z)| \leq |f(a)|$ for all $z \in D(a, R)$. Then f is constant.

Proof. By the mean value property,

$$f(a) = \int_0^1 f(a + \rho e^{2\pi i t}) \, \mathrm{d}t$$

Therefore,

$$|f(a)| = \left| \int_0^1 f(a + \rho e^{2\pi i t}) \, \mathrm{d}t \right| \leq \sup_{t \in [0,1]} \left| f(a + \rho e^{2\pi i t}) \right| \leq |f(a)|$$

where the last inequality is by hypothesis. Therefore, both inequalities must be equalities. Equality in the first inequality implies that $f(a + \rho e^{2\pi i t}) = c_\rho$ for some constant c_ρ and all $t \in [0, 1]$. Then, by the first equality, $|c_\rho| = |f(a)|$ for all $\rho \in (0, R)$. Thus, $\left| f(a + \rho e^{2\pi i t}) \right|$ is constant for all $\rho \in (0, R)$ and $t \in [0, 1]$. Hence $|f(z)|$ is constant on $D(a, R)$. By the Cauchy–Riemann equations, f must be constant. □

2.7. Taylor series

Theorem. Let $f : D(a, R) \to \mathbb{C}$ be holomorphic. Then f has a convergent power series representation on $D(a, R)$. More precisely, there exists a sequence of complex numbers $c_0, c_1, \ldots$ such that

$$f(w) = \sum_{n=0}^{\infty} c_n (w - a)^n$$

The coefficient c_n is given by

$$c_n = \frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z) \, \mathrm{d}z}{(z - a)^{n+1}}$$

for any $\rho \in (0, R)$.

Proof. Let $0 < \rho < R$. Then, for any $w \in D(0, \rho)$, we have by Cauchy's integral formula that

$$\begin{aligned} f(w) &= \frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{z-w} \\ &= \frac{1}{2\pi i} \int_{\partial D(a,\rho)} f(z) \sum_{n=0}^{\infty} \frac{(w-a)^n}{(z-a)^{n+1}}\,\mathrm{d}z \\ &= \sum_{n=0}^{\infty} \left(\frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{n+1}} \right) (w-a)^n \end{aligned}$$

The last equality holds since the series under the integral converges uniformly for all $z \in \partial D(a, \rho)$. Let

$$c_n(\rho) = \frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{n+1}}$$

Then we have shown that $f(w) = \sum_{n=0}^{\infty} c_n(\rho)(w-a)^{n+1}$ for all $w \in D(a, \rho)$. By a previous theorem, the function f has derivatives of all orders in $D(a, \rho)$ and hence $c_n(\rho) = \frac{f^{(n)}(a)}{n!}$, which is independent of ρ, so we can let $c_n = c_n(\rho)$ for an arbitrary ρ. □

Corollary. If f is holomorphic on an open set $U \subset \mathbb{C}$, then f has derivatives of all orders in U, and those derivatives are holomorphic on U.

Proof. We have a power series representation for f near every points, so its derivatives of all orders exist everywhere. Hence, the derivatives of all orders are holomorphic. □

Remark. We can explicitly compute from the c_n above that

$$f^{(n)}(a) = \frac{n!}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{n+1}}$$

This is a special case of a Cauchy integral formula for derivatives.

Note also that by taking $n = 1$, we can apply the estimate for the integral to find

$$|f'(a)| \leq \frac{1}{\rho} \left(\sup_{z \in \partial D(a,\rho)} |f(z)| \right)$$

This can be thought of as a localised version of Liouville's theorem, and it directly implies Liouville's theorem. Indeed, if f is entire and bounded, let $a \in \mathbb{C}$ and by applying the estimate and letting $\rho \to \infty$ we can conclude $f' = 0$ on $\mathbb{C}$, giving that f is constant.

Definition. A function f is *analytic* at a point $a \in \mathbb{C}$ (or $\mathbb{R}$) if there exists a neighbourhood of a such that f is given by a convergent power series about a.

Remark. If f is analytic at a, we must have derivatives of all orders of f near a. The above corollary implies that if f is complex, the following are equivalent.

(i) f is analytic at a

(ii) f has complex derivatives of all orders in a neighbourhood of a

(iii) f is complex differentiable once in a neighbourhood in a neighbourhood of a (so f is holomorphic at a)

For real functions, this is not the case. For example, consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \exp(-x^{-2})$. This has $f^{(n)}(0) = 0$ for all n, so f is not given by a convergent power series near zero.

Let $U \subset \mathbb{C}$ be an open set. Now, we have that $f = u + iv$ is holomorphic in U if and only if u and v have continuous partial derivatives in U, and that $u.v$ satisfy the Cauchy–Riemann equations. Further, the corollary above implies that u, v are C^2 functions. This shows that u and v are harmonic.

Theorem (Morera's theorem)**.** Let $U \subseteq \mathbb{C}$ be open, and $f : U \to \mathbb{C}$ be a continuous function such that $\int_\gamma f(z)\,\mathrm{d}z = 0$ for all closed curves γ in U. Then f is holomorphic in U.

Remark. This can be thought of as a converse to Cauchy's theorem.

Proof. We know that f has a holomorphic antiderivative F on U. Then, we know that F is twice differentiable in U. Since $F' = f$, f is holomorphic. □

Corollary. Let $U \subseteq \mathbb{C}$ be an open set. Let $f : U \to \mathbb{C}$ be a continuous function and holomorphic in $U \setminus S$, where S is a finite set. Then f is holomorphic in U.

Proof. For all $a \in U$, there exists $r > 0$ such that $D = D(a, r) \subset U$. Since D is convex, we can apply Cauchy's formula for convex sets to observe that $\int_\gamma f(z)\,\mathrm{d}z = 0$ for all closed curves in D. Then by Morera's theorem, f is holomorphic. □

2.8. Zeroes of holomorphic functions

Definition. Let f be a holomorphic function on a disc $D = D(a, R)$. By the Taylor series theorem, there exist constants c_n such that

$$f(z) = \sum_{n=0}^{\infty} c_n (z-a)^n$$

for all $z \in D$. Then if f is not identically zero, there exists n such that $c_n \neq 0$. Let $m = \min\{n : c_n \neq 0\}$. Then,

$$f(z) = (z-a)^m g(z); \quad g(z) = \sum_{n=m}^{\infty} c_n (z-a)^{n-m}$$

Note that g is holomorphic on D, and $g(a) = c_m \neq 0$.

If $m \neq 0$, we say that f has a *zero of order m at $z = a$*. Hence m is the smallest natural number n such that $f^{(n)}(a) \neq 0$. If $S \subseteq \mathbb{C}$, then a point $w \in S$ is an *isolated point* of S if there exists $r > 0$ such that $S \cap D(w, r) = \{w\}$.

Theorem (principle of isolated zeroes)**.** Let $f : D(a, R) \to \mathbb{C}$ be holomorphic and not identically zero. Then there exists $r \in (0, R)$ such that $f(z) \neq 0$ whenever $0 < |z - a| < r$.

Remark. If $f(a) = 0$, then $\{z : f(z) = 0\}$ intersects $D(a, r)$ only at a. Hence, a is an isolated point of the set of zeroes. For instance, there exists no nonzero holomorphic function that vanishes on a line segment or a disc.

We can show that certain identities from real analysis hold for complex functions. For instance, consider the function $g(z) = \sin^2 z + \cos^2 z - 1$. Since this g is holomorphic and vanishes on the real line, g must be identically zero in the complex plane.

The zero set may have an *accumulation point* on the boundary of the domain of f. Consider $f(z) = \sin \frac{1}{z}$ for $z \in D(1, 1)$. Here, if $a_n = \frac{1}{2n\pi}$, then $a_n \in D(1, 1)$ and $f(a_n) = 0$ and $a_n \to 0 \in \partial D(1, 1)$.

Proof. If $f(a) \neq 0$, then by continuity of f there exists $r > 0$ such that $f(z) \neq 0$ for all $z \in D(a, r)$. If $f(a) = 0$, then there exists an integer $m \geq 1$ such that $f(z) = (z - a)^m g(z)$ for $z \in D(a, R)$, where g is holomorphic with $g(a) \neq 0$. In this case, we find that there exists $r > 0$ such that $g(z) \neq 0$ for $z \in D(a, r)$ and hence $f(z) \neq 0$ for $z \in D(a, r) \setminus \{a\}$. □

2.9. Analytic continuation

Theorem. Let $U \subset V$ be domains. If $g_1, g_2 : V \to \mathbb{C}$ are analytic and $g_1 = g_2$ on U, then $g_1 = g_2$ on V. Equivalently, if $f : U \to \mathbb{C}$ is analytic, then there is at most one analytic function $g : V \to \mathbb{C}$ such that $g = f$ on U. We say that g is the *analytic continuation* of f to V, if it exists.

Proof. Let $g_1, g_2 : V \to \mathbb{C}$ be analytic with $g_1|_U = g_2|_U$. Then, $h = g_1 - g_2 : V \to \mathbb{C}$ is analytic, and $h|_U \equiv 0$. We want to show that $h \equiv 0$. Let

$$V_0 = \left\{ z \in V : \exists r > 0, h\Big|_{D(z,r)} \equiv 0 \right\}$$

and

$$V_1 = \left\{ z \in V : \exists n \geq 0, h^{(n)}(z) \neq 0 \right\}$$

Let $z \in V$ and suppose that $z \notin V_0$. Then for any disc $D = D(z, r) \subset V$, we have $h \not\equiv 0$ in D. Hence, by Taylor series, $h^{(n)}(z) \neq 0$ for some n, so $z \in V_1$. Thus, $V = V_0 \cup V_1$. We also know that $V_0 \cap V_1 = \varnothing$.

Note that V_0 is open by definition, and V_1 is by continuity of the derivatives $h^{(n)}$. By connectedness of the domain V, either V_0 or V_1 is empty. Since $U \subset V_0$, we must have $V_1 = \varnothing$. Thus, $V = V_0$ so $h \equiv 0$. □

Remark. The above proof does not rely on holomorphicity but on analyticity. Thus, the theorem holds for real analytic functions. For example, due to *elliptic regularity* (see Part II Analysis of Functions), we can show that harmonic functions are real analytic, and hence have a unique analytic continuation if one exists.

Given a holomorphic function f defined on a disc, we can compute the largest domain containing the disc to which there exists an analytic continuation of f. This is nontrivial to answer in general.

Example. Let $f(z) = \sum_{n=0}^{\infty} z^n$. The radius of convergence of this series is 1, so f is analytic in $D(0,1)$, and there is no larger *disc* $D(0,r) \supset D(0,1)$ such that g has an analytic continuation to $D(0,r)$. However, since $f(z) = \frac{1}{1-z}$ for $z \in D(0,1)$ and the function $\frac{1}{1-z}$ is analytic in $\mathbb{C} \setminus \{1\}$, f indeed has an analytic continuation to the larger domain $\mathbb{C} \setminus \{1\}$.

Example. Let $f(z) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} z^n}{n}$. This function also has a radius of convergence of 1, so f is analytic on $D(0,1)$. It has analytic continuation $\operatorname{Log}(1+z)$ to the domain $\mathbb{C} \setminus \{x \in \mathbb{R} : x \leq -1\}$ containing $D(0,1)$.

Example. Let $f(z) = \sum_{n=0}^{\infty} z^{n!}$. This has radius of convergence 1, so f is analytic in $D(0,1)$. However, f has no analytic continuation to any larger domain containing $D(0,1)$. The boundary $\partial D(0,1)$ is known as the *natural boundary* of f.

We can find in fact that for any given domain $U \subset \mathbb{C}$, there exists a holomorphic function $f : U \to \mathbb{C}$ which has no analytic continuation to a domain properly containing U.

The failure of analytic continuation in some cases can be explained as the result of loss of a regularity condition, such as boundedness, continuity, differentiability, or so on. However, this is not always the reason, and analytic continuation may remain impossible even when regularity conditions are all satisfied.

Example. Let $f(z) = \sum_{n=0}^{\infty} \exp\left(-2^{n/2}\right) z^{2^n}$, which has unit radius of convergence. f, and its derivatives of any order, are uniformly continuous on the closed disc $\overline{D(0,1)}$. However, we can prove that it has natural boundary $\partial D(0,1)$, using the following theorem which will not be proven.

Theorem (Ostrowski–Hadamard gap theorem)**.** Let (p_n) be a sequence of positive integers such that $p_{n+1} > (1+\delta)p_n$ for all n and some fixed $\delta > 0$. If (c_n) is a sequence of complex numbers such that $f(z) = \sum_{n=0}^{\infty} c_n z^{p_n}$ has unit radius of convergence, then $\partial D(0,1)$ is the natural boundary of f.

Corollary (identity principle)**.** Let $f, g : U \to \mathbb{C}$ be holomorphic functions in a domain U. If the set $S = \{z \in U : f(z) = g(z)\}$ contains a non-isolated point, then $f = g$ in U.

Proof. Let $h = f - g$, so $S = \{z \in U : h(z) = 0\}$. Suppose that S has a non-isolated point w. If there exists $r > 0$ such that $h \not\equiv 0$ in $D(w, r)$, then by the principle of isolated zeroes, we can find $\varepsilon > 0$ such that $f(z) \neq 0$ whenever $0 < |z - w| < \varepsilon$. However, this contradicts the assumption that w is a non-isolated point of S. Thus, $h \equiv 0$ on $D(w, r)$ for all $D(w, r) \subset U$. Thus, $h \equiv 0$ on U, so $f = g$ on U. □

Corollary (global maximum principle)**.** Let U be a bounded open set. Suppose $f : \overline{U} \to \mathbb{C}$ is a continuous function such that f is holomorphic in U. Then $|f|$ attains its maximum on ∂U.

Proof. $\overline{U}$ is compact, and $|f|$ is continuous on $\overline{U}$. Hence, it attains its maximum; there exists $w \in \overline{U}$ such that $|f(w)| = \max_{z \in \overline{U}} |f(z)|$. If $w \notin U$, then $w \in \partial U$ as required. Otherwise, let $D = D(w, r) \subset U$. Since $|f(z)| \leq |f(w)|$ for all $z \in D$, the local maximum principle implies that f is constant on D. Hence, by the identity principle, f is constant on the connected component of U containing D, which will be written U'. By continuity, f is constant on the closure of this connected component $\overline{U'}$. In particular, $|f(z)| = |f(w)|$ for all $z \in \partial U' \subseteq \partial U$ as required. □

Theorem (Cauchy's integral formula for derivatives)**.** Let $f : D(a, R) \to \mathbb{C}$ be holomorphic. For any $\rho \in (0, R)$ and $w \in D(a, \rho)$, we have

$$f^{(k)}(w) = \frac{k!}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z - w)^{k+1}}$$

Further,

$$\sup_{z \in D(a,R/2)} \left|f^{(k)}(z)\right| \leq \frac{C}{R^k} \sup_{z \in D(a,R)} |f(z)|$$

where $C = k!2^{k+1}$ is a constant which depends only on k. This final result is called a Cauchy estimate for the kth derivative.

Remark. Directly applying Cauchy's integral formula to $f^{(n)}$, we find a formula for $f^{(n)}(w)$ in terms of an integral involving $f^{(n)}$. The significance of the above theorem is that the integral involves f alone, and not its derivatives.

Note that we have already observed the special case $w = a$. This was proven during the discussion on Taylor series.

Proof. If $k = 0$, we have the usual Cauchy integral formula. For $k = 1$, let $g(z) = \frac{f(z)}{z-w}$, which is holomorphic in $D(a, R) \setminus \{w\}$, with derivative

$$g'(z) = \frac{f'(z)}{z - w} - \frac{f(z)}{(z - w)^2}$$

Since $\partial D(a,\rho) \subset D(a,R) \setminus \{w\}$, we know that $\int_{\partial D(a,\rho)} g'(z)\,\mathrm{d}z = 0$ by the fundamental theorem of calculus. Applying the usual Cauchy integral formula to f',

$$f'(w) = \frac{1}{2\pi i}\int_{\partial D(a,\rho)} \frac{f'(z)\,\mathrm{d}z}{z-w}$$

Combining these results give the result for $k = 1$. For higher derivatives, we can proceed by induction. Let $k \geq 2$, and then suppose the formula holds for this value of k, for all holomorphic functions $D(a,R) \to \mathbb{C}$. For any holomorphic function $f : D(a,R) \to \mathbb{C}$, consider

$$g(z) = \frac{f(z)}{(z-w)^{k+1}} \implies g'(z) = \frac{f'(z)}{(z-w)^{k+1}} - \frac{(k+1)f(z)}{(z-w)^{k+2}}$$

which is defined in $D(a,R) \setminus \{w\}$. Then, since $\int_{\partial D(a,\rho)} g'(z)\,\mathrm{d}z = 0$, we find

$$\int_{\partial D(a,\rho)} \frac{f'(z)\,\mathrm{d}z}{(z-w)^{k+1}} = (k+1)\int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-w)^{k+2}}$$

By substituting f' into the induction hypothesis,

$$f^{(k+1)}(w) = \frac{k!}{2\pi i}\int_{\partial D(a,\rho)} \frac{f'(z)\,\mathrm{d}z}{(z-w)^{k+1}}$$

We can then combine the previous two expressions to find

$$f^{(k+1)}(w) = \frac{(k+1)!}{2\pi i}\int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-w)^{k+2}}$$

as required.

For the second part, let $\sup_{z\in D(a,R)} |f(z)| < \infty$ without loss of generality. Let $\rho \in (R/2, R)$. Then, by the first part, for all $w \in D(a, R/2)$ we have

$$\left|f^{(k)}(w)\right| \leq \frac{k!}{2\pi}\left(\sup_{z\in\partial D(a,\rho)} \frac{|f(z)|}{|z-w|^{k+1}}\right)\mathrm{length}(\partial D(a,\rho))$$

As $|z-w| \geq \rho - R/2$ for all $z \in \partial D(a,\rho)$ and all $w \in D(a,R/2)$, this implies

$$\sup_{w\in D(a,R/2)} \left|f^{(k)}(w)\right| \leq \frac{k!\rho}{(\rho - R/2)^{k+1}} \sup_{z\in D(a,R)} |f(z)|$$

Now, as $\rho \to R$, the result follows. □

2.10. Uniform limits of holomorphic functions

Definition. Let $U \subseteq \mathbb{C}$ be open, and let $f_n : U \to \mathbb{C}$ be a sequence of functions. We say that (f_n) converges *locally uniformly* on U if, for all $a \in U$, there exists $r > 0$ such that (f_n) converges uniformly on $D(a,r)$.

Example. Let $f_n(z) = z^n$. Then $f_n \to 0$ locally uniformly, but not uniformly.

Proposition. (f_n) converges locally uniformly on an open set $U \subseteq \mathbb{C}$ if and only if (f_n) converges uniformly on each compact subset $K \subseteq U$.

Proof. The forward implication is simple, due to the definition of compactness. The converse follows since for all $a \in U$, there exists a compact disc $\overline{D(a,r)} \subset U$. □

Theorem (uniform limits of holomorphic functions)**.** Let $U \subseteq \mathbb{C}$ be open, and $f_n : U \to \mathbb{C}$ be holomorphic for each $n \in \mathbb{N}$. If (f_n) converges locally uniformly on U to some function $f : U \to \mathbb{C}$, then f is holomorphic.

Further, $f_n' \to f'$ locally uniformly on U, and by induction, for each k we have $f_n^{(k)} \to f^{(k)}$ locally uniformly on U as $n \to \infty$.

Remark. This is not true for real analytic functions. The *Weierstrass approximation theorem* states the following. Let $f : [a,b] \to \mathbb{R}$ be a continuous function on a compact interval $[a,b] \subset \mathbb{R}$. Then, there exists a sequence of polynomials (p_n) converging uniformly to f on $[a,b]$.

There exist continuous, nowhere differentiable functions $f : [a,b] \to \mathbb{R}$. Applying the Weierstrass approximation theorem to such functions f shows that the uniform limit of real analytic functions need not have a single point of differentiability.

Proof. Let $a \in U$ and choose $r > 0$ such that $\overline{D(a,r)} \subset U$ and $f_n \to f$ uniformly on $\overline{D(a,r)}$. Since the f_n are continuous, by a result from Analysis and Topology we have that f is continuous in $\overline{D(a,r)}$.

Let γ be a closed curve in $D(a,r)$. Since $D(a,r)$ is convex, by the convex Cauchy theorem we have $\int_\gamma f_n(z)\,\mathrm{d}z = 0$. Since $f_n \to f$ uniformly on $D(a,r)$, it follows that

$$\int_\gamma f(z)\,\mathrm{d}z = \lim_{z\to\infty} \int_\gamma f_n(z)\,\mathrm{d}z = 0$$

By Morera's theorem, f is holomorphic in $D(a,r)$. Since a is arbitrary, f is holomorphic on all of U.

Now, let $a \in U$ be arbitrary and let $D(a,r)$ be as above. We can apply the Cauchy estimate for $k = 1, R = r$, applied to the function $f_n - f$. This gives

$$\sup_{z\in D(a,r/2)} |f_n'(z) - f'(z)| \le \frac{4}{r} \sup_{z\in D(a,r)} |f_n(z) - f(z)|$$

Since the right hand side converges to zero as $n \to \infty$, the claim follows. □

Remark. Many of the key results proven for holomorphic functions have analogues for real harmonic functions on domains not just in $\mathbb{R}^2$ but in $\mathbb{R}^n$ for any n. For instance:

(i) (Liouville's theorem) if $u : \mathbb{R}^n \to \mathbb{R}$ is a bounded harmonic function then u is constant;

(ii) (local maximum principle) if $u : D(a,r)$ is a C^2 harmonic function on an open ball $D(a,r)$ in $\mathbb{R}^n$, and if $u(x) \leq u(a)$ for all $x \in D(a,r)$, then u is constant;

(iii) (global maximum principle) a harmonic function on a bounded open set U that is continuous on $\overline{U}$ attains its maximum on ∂U;

(iv) harmonic functions are real analytic;

(v) the unique analytic continuation principle holds;

(vi) uniform limits of harmonic functions are harmonic;

(vii) derivative estimates hold: if $u : D(a,R) \subseteq \mathbb{R}^n \to \mathbb{R}$ is harmonic, then

$$\sup_{D(a,R/2)} |D^k u| \leq CR^{-k} \sup_{D(a,R)} |u|; \quad C = C(n,k)$$

For the case $n = 2$, the result for harmonic functions can often be deduced directly from the corresponding results for holomorphic functions. For instance, for Liouville's theorem, given that u is a harmonic function on $\mathbb{R}^2$, we find a function v such that $u+iv$ is holomorphic on $\mathbb{C}$ (which is always possible in a simply connected domain). Then $g = e^f$ is holomorphic with $|g| = e^u$, so if u is bounded then g is bounded. By Liouville's theorem for holomorphic functions, g and hence f is constant.

3. More integration

3.1. Winding numbers

Let $\gamma : [a,b] \to \mathbb{C}$ be a closed, piecewise C^1 curve, and let $w \notin \operatorname{Im}\gamma$. For all t, there exists $r(t) > 0$ and $\theta(t) \in \mathbb{R}$ such that $\gamma(t) = w + r(t)e^{i\theta(t)}$. Then, the function $r : [a,b] \to \mathbb{R}$ is given by $r(t) = |\gamma(t) - w|$, so it is uniquely determined and piecewise C^1.

Definition. If we have a continuous choice of $\theta : [a,b] \to \mathbb{R}$ such that $\gamma(t) = w + r(t)e^{i\theta(t)}$, then we define the *winding number* or the *index* of γ about w as

$$I(\gamma; w) = \frac{\theta(b) - \theta(a)}{2\pi}$$

If γ is a closed curve, $I(\gamma; w)$ is an integer. This is because

$$\gamma(a) = \gamma(b) \implies \exp(i\theta(b) - i\theta(a)) = 1$$

If $\theta_1 : [a,b] \to \mathbb{C}$ is also continuous such that $\gamma(t) = w + re^{i\theta_1(t)}$, then $\exp(i\theta(t) - i\theta_1(t)) = 1$, so

$$\frac{\theta_1(t) - \theta(t)}{2\pi} \in \mathbb{Z}$$

Since $\theta_1 - \theta$ is continuous, this quotient must be a constant. Hence, $I(\gamma; w)$ is well-defined and independent of the (continuous) choice of θ.

Lemma. Let $w \in \mathbb{C}$ and $\gamma : [a,b] \to \mathbb{C} \setminus \{w\}$, where γ is piecewise C^1. Then, there exists a piecewise C^1 function $\theta : [a,b] \to \mathbb{R}$ such that $\gamma(t) = w + r(t)e^{i\theta(t)}$, where $r(t) = |\gamma(t) - w|$. If γ is closed, then we also have

$$I(\gamma; w) = \frac{1}{2\pi i} \int_\gamma \frac{\mathrm{d}z}{z - w}$$

Remark. If γ is C^1, and there is a C^1 function θ such that $\gamma(t) = w + r(t)e^{i\theta(t)}$, then

$$\gamma'(t) = (r'(t) + ir(t)\theta'(t))e^{i\theta(t)} = \left(\frac{r'(t)}{r(t)} + i\theta'(t)\right)r(t)e^{i\theta(t)} = \left(\frac{r'(t)}{r(t)} + i\theta'(t)\right)(\gamma(t) - w)$$

Hence,

$$\theta'(t) = \operatorname{Im}\frac{\gamma'(t)}{\gamma(t) - w} \implies \theta(t) = \theta(a) + \operatorname{Im}\int_a^t \frac{\gamma'(s)\,\mathrm{d}s}{\gamma(s) - w}$$

Proof. Let $h(t) = \int_a^t \frac{\gamma'(s)}{\gamma(s)-w}\,\mathrm{d}s$. The integrand is bounded on $[a,b]$, and is continuous except at the finite number of points at which γ' may be discontinuous. Hence, $h : [a,b] \to \mathbb{C}$ is continuous. Further, h is differentiable with $h'(t) = \frac{\gamma'(t)}{\gamma(t)-w}$ at each t where γ' is continuous. Hence, h is piecewise C^1. This induces an ordinary differential equation for $\gamma(t) - w$.

$$(\gamma(t) - w)' - (\gamma(t) - w)h'(t) = 0$$

which is true for all $t \in [a, b]$ except possibly for a finite set. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\big((\gamma(t) - w)e^{-h(t)}\big) = \gamma'(t)e^{-h(t)} - (\gamma(t) - w)e^{-h(t)}h'(t) = 0$$

except for finitely many t. Since $(\gamma(t)-w)e^{-h(t)}$ is continuous, it must be constant, and equal to its value at $t = a$. Hence

$$\gamma(t) - w = (\gamma(a) - w)e^{h(t)} = (\gamma(a) - w)e^{\operatorname{Re} h(t)}e^{i \operatorname{Im} h(t)} = |\gamma(a) - w|e^{\operatorname{Re} h(t)}e^{i(\alpha + \operatorname{Im} h(t)}$$

for α such that $e^{i\alpha} = \frac{\gamma(a)-w}{|\gamma(a)-w|}$. Hence, we can set $\theta(t) = \alpha + \operatorname{Im} h(t)$.

For the second part, note that

$$I(\gamma; w) = \frac{\theta(b) - \theta(a)}{2\pi} = \frac{\operatorname{Im}(h(b) - h(a))}{2\pi} = \frac{\operatorname{Im} h(b)}{2\pi}$$

Since $\gamma(t) - w = (\gamma(a) - w)e^{h(t)}$ and $\gamma(b) = \gamma(a)$, we have $e^{h(b)} = 1$, so $\operatorname{Re} h(b) = 0$ and $\operatorname{Im} h(b) = -ih(b)$. Thus,

$$I(\gamma; w) = \frac{1}{2\pi i}h(b) = \frac{1}{2\pi i}\int_a^b \frac{\gamma'(s)}{\gamma(s) - w}\,\mathrm{d}s = \frac{1}{2\pi i}\int_\gamma \frac{\mathrm{d}z}{z - w}$$

□

Remark. It is also true that θ exists and is continuous if γ is merely continuous, but the formula for the winding number is not useful, so we omit this proof.

Proposition. If $\gamma\colon [a, b] \to \mathbb{C}$ is a closed curve, then the function $w \mapsto I(\gamma; w)$ is continuous on $\mathbb{C} \setminus \operatorname{Im}\gamma$. Since $I(\gamma; w)$ is integer-valued, $I(\gamma; w)$ is locally constant. So $I(\gamma; w)$ is constant for each connected component of the open set $\mathbb{C} \setminus \operatorname{Im}\gamma$.

Proof. Exercise. □

Proposition. If $\gamma\colon [a, b] \to D(z_0, R)$ is a closed curve, then $I(\gamma; w) = 0$ for all $w \in \mathbb{C} \setminus D(z_0, R)$.

If $\gamma\colon [a, b] \to \mathbb{C}$ is a closed curve, then there exists a unique unbounded connected component Ω of $\mathbb{C} \setminus \gamma([a, b])$, and $I(\gamma; w) = 0$ for all $w \in \Omega$.

Proof. For the first part, if $w \in \mathbb{C} \setminus D(z_0, R)$, then the function $f(z) = \frac{1}{z-w}$ is holomorphic in $D(z_0, R)$. Hence $I(\gamma; w) = 0$ by the convex version of Cauchy's theorem.

For the second part, since $\gamma([a, b])$ is compact (by continuity of γ), there exists $R > 0$ such that $\gamma([a, b]) \subset D(0, R)$. Since $\mathbb{C} \setminus D(0, R)$ is a connected subset of $\mathbb{C} \setminus \gamma([a, b])$, there exists a connected component Ω of $\mathbb{C} \setminus \gamma([a, b])$ such that $\mathbb{C} \setminus D(0, R) \subseteq \Omega$. This component is unbounded. Any other component is disjoint from $\mathbb{C} \setminus D(0, R)$, so is contained within $D(0, R)$ and is hence bounded. So the unbounded component is unique. Since $I(\gamma; w)$ is locally constant and zero on $\mathbb{C} \setminus D(0, R)$, it is zero on Ω. □

3.2. Continuity of derivative function

Lemma. Let $f : U \to \mathbb{C}$ be holomorphic, and define $g : U \times U \to \mathbb{C}$ by

$$g(z,w) = \begin{cases} \frac{f(z)-f(w)}{z-w} & \text{if } z \neq w \\ f'(w) & \text{if } z = w \end{cases}$$

Then g is continuous. Moreover, if γ is a closed curve in U, then the function $h(w) = \int_\gamma g(z,w)\,\mathrm{d}z$ is holomorphic on U.

Proof. It is clear that g is continuous at (z,w) if $z \neq w$. To check continuity at a point $(a,a) \in U \times U$, let $\varepsilon > 0$ and choose $\delta > 0$ such that $D(a,\delta) \subseteq U$ and $|f'(z) - f'(a)| < \varepsilon$ for all $z \in D(a,\delta)$. This is always possible since f' is continuous.

Let $z, w \in D(a,\delta)$. If $z = w$, then $|g(z,w) - g(a,a)| = |f'(z) - f'(a)| < \varepsilon$. If $z \neq w$, we have $tz + (1-t)w \in D(a,\delta)$ for $t \in [0,1]$. Hence,

$$\begin{aligned} f(z) - f(w) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f(tz + (1-t)w)\,\mathrm{d}t \\ &= \int_0^1 f'(tz + (1-t)w)(z-w)\,\mathrm{d}t \\ &= (z-w)\int_0^1 f'(tz + (1-t)w)\,\mathrm{d}t \end{aligned}$$

Thus,

$$\begin{aligned} |g(z,w) - g(a,a)| &= \left|\frac{f(z)-f(w)}{z-w} - f'(a)\right| \\ &= \left|\int_0^1 [f'(tz + (1-t)w) - f'(a)]\,\mathrm{d}t\right| \\ &\leq \sup_{t\in[0,1]} |f'(tz + (1-t)w) - f'(a)| < \varepsilon \end{aligned}$$

Hence $|(z,w) - (a,a)| < \delta$ implies $|g(z,w) - g(a,a)| < \varepsilon$, so g is continuous at (a,a).

To show h is holomorphic, we must first check that h is continuous. Let $w_0 \in W$, and suppose $w_n \to w_0$. Let $\delta > 0$ such that $\overline{D(w_0,\delta)} \subset U$. The function g is continuous on $U \times U$, so it is uniformly continuous on the compact subset $\operatorname{Im}\gamma \times \overline{D(w_0,\delta)} \subset U \times U$. Thus, if we let $g_n(z) = g(z,w_n)$ and $g_0(z) = g(z,w_0)$ for $z \in \operatorname{Im}\gamma$, then $g_n \to g_0$ uniformly on $\operatorname{Im}\gamma$. Hence $\int_\gamma g_n(z)\,\mathrm{d}z \to \int_\gamma g_0(z)\,\mathrm{d}z$. In other words, $h(w_n) \to h(w_0)$. Thus, h is continuous.

Now, we can use the convex Cauchy's theorem and Morera's theorem to show h is holomorphic on U. For $w_0 \in U$, we can choose a disc $D(w_0,\delta) \subset U$. Suppose that γ is parametrised over $[a,b]$, and let $\beta : [c,d] \to D(w_0,\delta)$ be any closed curve. Then $h(w) =$

$\int_\gamma g(z,w)\,\mathrm{d}z = \int_a^b g(\gamma(t),w)\gamma'(t)\,\mathrm{d}t$, hence

$$\begin{aligned}\int_\beta h(w)\,\mathrm{d}w &= \int_c^d \left(\int_a^b g(\gamma(t),\beta(s))\gamma'(t)\beta'(s)\,\mathrm{d}t\right)\mathrm{d}s\\ &= \int_a^b \left(\int_c^d g(\gamma(t),\beta(s))\gamma'(t)\beta'(s)\,\mathrm{d}s\right)\mathrm{d}t\\ &= \int_\gamma \left(\int_\beta g(z,w)\,\mathrm{d}w\right)\mathrm{d}z\end{aligned}$$

by Fubini's theorem, which will be proven below. By a previous theorem, for all $z \in U$, the function $w \mapsto g(z,w)$ is holomorphic in $D(w_0,\delta)$ (and hence in U), since it is continuous in U and holomorphic except at a single point z. Hence, by the convex version of Cauchy's theorem, $\int_\beta g(z,w)\,\mathrm{d}w = 0$. Hence, $\int_\beta h(w)\,\mathrm{d}w = 0$. By Morera's theorem, h is holomorphic in $D(w_0,\delta)$ and hence on U. □

Lemma (Fubini's theorem)**.** If $\varphi : [a,b]\times[c,d] \to \mathbb{R}$ is continuous, then $f_1 : s \mapsto \int_c^d \varphi(s,t)\,\mathrm{d}t$ is continuous on $[a,b]$, the function $f_2 : t \mapsto \int_a^b \varphi(s,t)\,\mathrm{d}t$ is continuous on $[c,d]$, and

$$\int_a^b \left(\int_c^d \varphi(s,t)\,\mathrm{d}t\right)\mathrm{d}s = \int_c^d \left(\int_a^b \varphi(s,t)\,\mathrm{d}s\right)\mathrm{d}t$$

Proof. Since φ is continuous on the compact set $[a,b]\times[c,d]$, it is uniformly continuous. Hence, given $\varepsilon > 0$, there exists $\delta > 0$ such that $|s_1 - s_2| < \delta \implies |\varphi(s_1,t) - \varphi(s_2,t)| < \varepsilon$ for all $t \in [c,d]$, so $|f_1(s_1) - f_1(s_2)| < (d-c)\varepsilon$, so f_1 is continuous. Similarly, f_2 is continuous. Note that since φ is uniformly continuous, it is the uniform limit of a sequence of step functions of the form $g(x,y) = \sum_{j=1}^N \alpha_j \chi_{R_j}(x,y)$ where α_j are constants, and R_j are sub-rectangles of the form $R_j = [a_j,b_j)\times[c_j,d_j)$ such that $\bigcup R_j$ is a finite partition of $[a,b)\times[c,d)$, and χ_{R_j} is the characteristic function of R_j For such step functions, we can easily check the interchangability of the integrals. □

3.3. Cauchy's theorem and Cauchy's integral formula

Definition. Let $U \subseteq \mathbb{C}$ be open. A closed curve $\gamma : [a,b] \to U$ is said to be *homologous to zero* in U if $I(\gamma;w) = 0$ for all $w \in \mathbb{C} \setminus U$.

Theorem. Let U be a non-empty open subset of $\mathbb{C}$, and γ be a closed curve in U homologous to zero in U. Then,

$$I(\gamma;w)f(w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)\,\mathrm{d}z}{z-w}$$

for every holomorphic function $f : U \to \mathbb{C}$ and every $w \in U \setminus \operatorname{Im}\gamma$. Further,

$$\int_\gamma f(z)\,\mathrm{d}z = 0$$

for every holomorphic $f : U \to \mathbb{C}$.

Remark. Cauchy's theorem states that if $\int_\gamma f(z)\,\mathrm{d}z = 0$ for a specific family of holomorphic functions on U, namely for $f_w(z) = \frac{1}{z-w}$ where $w \in \mathbb{C} \setminus U$, then $\int_\gamma f(z)\,\mathrm{d}z = 0$ for any holomorphic function $f : U \to \mathbb{C}$.

The first and second parts as statements are equivalent. Indeed, if we assume the Cauchy integral formula holds, simply apply the formula with $F(z) = (z-w)f(z)$. Since $F(w) = 0$, we have $\int_\gamma f(z)\,\mathrm{d}z = 0$. If we assume Cauchy's theorem, for any $w \in U$, the function

$$g(z) = \begin{cases} \frac{f(z)-f(w)}{z-w} & \text{if } z \neq w \\ f'(w) & \text{if } z = w \end{cases}$$

is holomorphic in U as seen above. Hence $\int_\gamma g(z)\,\mathrm{d}z = 0$, so $\frac{1}{2\pi i}\int_\gamma \frac{f(z)\mathrm{d}z}{z-w} = I(\gamma; w)f(w)$ for all $w \notin \operatorname{Im}\gamma$.

Note that the statement that γ is homologous to zero is equivalent to Cauchy's theorem being valid for all f. For example, given $w \in \mathbb{C} \setminus U$, we can apply Cauchy's theorem to $f(z) = \frac{1}{z-w}$ to get $I(\gamma; w) = 0$. The converse is proven in the theorem following this proof. This is also equivalent to Cauchy's integral formula being valid for all f.

Proof. It suffices to prove part (i). Equivalently, for all $w \in U \setminus \operatorname{Im}\gamma$,

$$\int_\gamma \frac{f(z)-f(w)}{z-w}\,\mathrm{d}z = 0 \iff \int_\gamma g(z,w)\,\mathrm{d}z = 0$$

where

$$g(z,w) = \begin{cases} \frac{f(z)-f(w)}{z-w} & \text{if } z \neq w \\ f'(w) & \text{if } z = w \end{cases}$$

Now, define

$$h : U \to \mathbb{C}; \quad h(w) = \int_\gamma g(z,w)\,\mathrm{d}z$$

By the above lemma, this is holomorphic on U. We will show that $h = 0$. We will extend h to a holomorphic function $H : \mathbb{C} \to \mathbb{C}$ and prove that $H(w) \to 0$ as $w \to \infty$, then we can apply Liouville's theorem.

To extend h into an entire function H, by definition of γ being homologous to zero in U, we have $\mathbb{C} \setminus U \subseteq V \equiv \{w \in \mathbb{C} \setminus \operatorname{Im}\gamma : I(\gamma; w) = 0\}$. So $\mathbb{C} = U \cup V$, and V is open since $I(\gamma; \cdot)$ is locally constant. For $w \in U \cap V$, we have

$$h(w) = \int_\gamma \frac{f(z)-f(w)}{z-w}\,\mathrm{d}z = \int_\gamma \frac{f(z)\,\mathrm{d}z}{z-w}$$

since $\int_\gamma \frac{\mathrm{d}z}{z-w} = 2\pi i \cdot I(\gamma; w) = 0$ as $w \in V$. Hence, on $U \cap V$, the function h agrees with

$$h_1 : V \to \mathbb{C}; \quad h_1(w) = \int_\gamma \frac{f(z)\,\mathrm{d}z}{z-w}$$

We know that h_1 is holomorphic on V. Hence, the function $H : \mathbb{C} \to \mathbb{C}$ defined by

$$H(w) = \begin{cases} h(w) & w \in U \\ h_1(w) & w \in V \end{cases}$$

is well-defined and holomorphic.

Now, we will show $H(w) \to 0$ as $|w| \to \infty$. Let $R > 0$ such that $\operatorname{Im}\gamma \subset D(0,R)$, which is possible since $\operatorname{Im}\gamma$ is compact. Hence, $\mathbb{C} \setminus D(0,R) \subseteq V$. If $|w| > R$,

$$|H(w)| = |h_1(w)| = \left| \int_\gamma \frac{f(z)\,\mathrm{d}z}{z-w} \right| \leq \frac{1}{|w| - R} \left(\sup_{z \in \operatorname{Im}\gamma} |f(z)| \right) \operatorname{length}(\gamma)$$

Hence, $H(w) \to 0$ as $|w| \to \infty$, as claimed. Hence H is bounded, since H is continuous, and $|H(w)| \leq 1$ outside some closed disc $\overline{D(0,R_1)}$. By Liouville's theorem, H is constant, and by the claim, $H = 0$. In particular, $h = 0$. □

Corollary. Let $U \subset \mathbb{C}$ be open and $\gamma_1, \dots, \gamma_n$ be closed curves in U such that $\sum_{j=1}^n I(\gamma_j; w) = 0$ for all $w \in \mathbb{C} \setminus U$. Then, for any holomorphic $f : U \to \mathbb{C}$, we have

$$f(w) \sum_{j=1}^n I(\gamma_j; w) = \sum_{j=1}^n \frac{1}{2\pi i} \int_{\gamma_j} \frac{f(z)\,\mathrm{d}z}{z-w}$$

for all $w \in U \setminus \bigcup_{j=1}^n \operatorname{Im}\gamma_j$, and

$$\sum_{j=1}^n \int_{\gamma_j} f(z)\,\mathrm{d}z = 0$$

Proof. For the first part, define $g(z,w)$ as before, but let

$$V = \left\{ w \in \mathbb{C} \setminus \bigcup_{j=1}^n \operatorname{Im}\gamma_j : \sum_{j=1}^n I(\gamma_j; w) = 0 \right\}$$

In the definitions of h and h_1, use the sum of the integrals over γ_j. Then we can proceed as above. The second part follows from the first as before. □

Corollary. Let $U \subset \mathbb{C}$ be open and let β_1, β_2 be closed curves in U such that $I(\beta_1; w) = I(\beta_2; w)$ for all $w \in \mathbb{C} \setminus U$. Then

$$\int_{\beta_1} f(z)\,\mathrm{d}z = \int_{\beta_2} f(z)\,\mathrm{d}z$$

for all holomorphic functions $f : U \to \mathbb{C}$.

Proof. We can apply the second part of the previous corollary with $\gamma_1 = \beta_1$ and $\gamma_2 = -\beta_2$, noting that $I(-\beta_2; w) = -I(\beta_2; w)$ for any $w \notin \operatorname{Im} \beta_2$. □

3.4. Homotopy

The set of closed curves in U such that Cauchy's theorem is valid is the set of holomorphic functions homologous to zero. We will now construct a more restrictive condition, the condition of being *null-homotopic*.

Definition. Let $U \subseteq \mathbb{C}$ be a domain, and let $\gamma_0, \gamma_1 : [a, b] \to U$ be closed curves. We say that γ_0 is *homotopic to* γ_1 *in* U if there exists a continuous map $H : [0, 1] \times [a, b] \to U$ such that for all $s \in [0, 1]$, $t \in [a, b]$,

$$H(0, t) = \gamma_0(t); \quad H(1, t) = \gamma_1(t); \quad H(s, a) = H(s, b)$$

Such a map is called a *homotopy* between γ_0, γ_1.

For $0 \leq s \leq 1$, if we let $\gamma_s : [a, b] \to U$ be defined by $\gamma_s(t) = H(s, t)$ for $t \in [a, b]$, then the above conditions imply that $\{\gamma_s : s \in [0, 1]\}$ is a family of continuous closed curves in U which deform γ_0 to γ_1 continuously without leaving U.

Definition. A closed curve is *null-homotopic* in a certain domain if it is homotopic to a constant curve in the domain, such as $\gamma(t) = z$ for z fixed.

Theorem. If $\gamma_0, \gamma_1 : [a, b] \to U$ are homotopic closed curves in U, then $I(\gamma_0; w) = I(\gamma_1; w)$ for all $w \in \mathbb{C} \setminus U$. In particular, if a closed curve γ is null-homotopic in U, it is homologous to zero in U.

Proof. Let $H : [0, 1] \times [a, b] \to U$ be a homotopy between γ_0 and γ_1. Since H is continuous and $[0, 1] \times [a, b]$ is compact, the image $K = H([0, 1] \times [a, b])$ is a compact subset of the open set U. Therefore, there exists $\varepsilon > 0$ such that for all $w \in \mathbb{C} \setminus U$, $|w - H(s, t)| > 2\varepsilon$ for all $(s, t) \in [0, 1] \times [a, b]$. Since H is uniformly continuous on $[0, 1] \times [a, b]$, there exists $n \in \mathbb{N}$ such that

$$\forall (s, t), (s', t') \in [0, 1] \times [a, b], \ |s - s'| + |t - t'| \leq \frac{1}{n} \implies |H(s, t) - H(s', t')| < \varepsilon$$

For $k = 0, 1, 2, \dots, n$, we let $\Gamma_k(t) = H(k/n, t)$ for $a \leq t \leq b$. Then the Γ_k are closed continuous curves with $\Gamma_0 = \gamma_0$ and $\Gamma_n = \gamma_1$. Hence, for all $t \in [a, b]$,

$$\underbrace{|\Gamma_{k-1}(t) - \Gamma_k(t)|}_{<\varepsilon} < \underbrace{|w - \Gamma_{k-1}(t)|}_{>2\varepsilon}$$

On the example sheets we have shown that for piecewise C^1 closed curves $\gamma, \tilde{\gamma}$, if we have $|\gamma(t) - \tilde{\gamma}(t)| < |w - \gamma(t)|$ for all t, then $I(\gamma; w) = I(\tilde{\gamma}; w)$. Hence, if Γ_k are piecewise C^1, we can see that $I(\Gamma_{k-1}; w) = I(\Gamma_k; w)$ for all k, and hence $I(\gamma_0; w) = I(\gamma_1; w)$ as required.

We have only assumed that H is continuous, so Γ_k need not be piecewise C^1. We can fix this problem by approximating each Γ_k by a polygonal curve. We can take

$$\tilde{\Gamma}_k(t) = \left(1 - \frac{n(t - a_{j-1})}{b-a}\right) H\left(\frac{k}{n}, a_{j-1}\right) + \frac{n(t-a_{j-1})}{b-a} H\left(\frac{k}{n}, a_j\right)$$

for $a_{j-1} \le t \le a_j$, where

$$a_j = a + \frac{(b-a)j}{n}$$

If we choose n so that

$$|s - s'| + |t - t'| \le \frac{\min\{1, b-a\}}{n} \implies |H(s,t) - H(s',t')| < \varepsilon$$

the curves $\tilde{\Gamma}_k$ satisfy

$$|\tilde{\Gamma}_{k-1}(t) - \tilde{\Gamma}_k(t)| < |w - \tilde{\Gamma}_{k-1}(t)|$$

for all $t \in [a, b]$. This is because for $t \in [a_{j-1}, a_j]$,

$$\begin{aligned} |\tilde{\Gamma}_{k-1}(t) - \tilde{\Gamma}_k(t)| &\le \left(1 - \frac{n(t-a_{j-1})}{b-a}\right)\left|H\left(\frac{k-1}{n}, a_{j-1}\right) - H\left(\frac{k}{n}, a_{j-1}\right)\right| \\ &\quad + \frac{n(t-a_{j-1})}{b-a}\left|H\left(\frac{k-1}{n}, a_j\right) - H\left(\frac{k}{n}, a_j\right)\right| \\ &< \varepsilon \end{aligned}$$

and

$$|w - \tilde{\Gamma}_{k-1}(t)| \ge |w - \Gamma_{k-1}(t)| - |\Gamma_{k-1}(t) - \tilde{\Gamma}_{k-1}(t)| > 2\varepsilon - \varepsilon = \varepsilon$$

We also have, for all $t \in [a, b]$,

$$|\tilde{\Gamma}_0(t) - \gamma_0(t)|; \quad |\tilde{\Gamma}_n - \gamma_1(t)| < |w - \gamma_1(t)|$$

Hence the result follows from the same example sheet question. □

Remark. If γ is homologous to zero in U, it is not necessarily the case that γ is null-homotopic. For insance, let $U = \mathbb{C} \setminus \{w_1, w_2\}$ for $w_1 \ne w_2$, and let $U_1 = U \cup \{w_1\} = \mathbb{C} \setminus \{w_2\}$ and $U_2 = U \cup \{w_2\} = \mathbb{C} \setminus \{w_1\}$. Then, consider a curve γ which is not null-homotopic in U, but null-homotopic in each of the larger domains U_1, U_2. Then γ is homologous to zero in U_1 and U_2. Hence $I(\gamma; w_1) = I(\gamma; w_2) = 0$, so γ is homologous to zero in U.

Corollary. If $\gamma_0, \gamma_1 : [a, b] \to U$ are homotopic closed curves in U, then

$$\int_{\gamma_0} f(z)\,\mathrm{d}z = \int_{\gamma_1} f(z)\,\mathrm{d}z$$

for all holomorphic $f : U \to \mathbb{C}$.

This is immediate from previous results. However, we can make a direct proof that does not require the most general form of Cauchy's theorem.

Proof. With $\tilde{\Gamma}_k$ as above, consider the closed curve comprised of

(i) the curve $\tilde{\Gamma}_{k-1}$ on $[a_{j-1}, a_j]$;

(ii) the line segment $\left[\tilde{\Gamma}_{k-1}(a_j), \tilde{\Gamma}_k(a_j)\right]$;

(iii) the curve $-\tilde{\Gamma}_k$ on $[a_j, a_{j-1}]$;

(iv) the line segment $\left[\tilde{\Gamma}_k(a_{j-1}), \tilde{\Gamma}_{k-1}(a_{j-1})\right]$.

This curve is contained in the disc $D(\tilde{\Gamma}_{k-1}(a_{j-1}), \varepsilon) \subseteq U$. We can apply the convex version of Cauchy's theorem and sum over j to find

$$\int_{\tilde{\Gamma}_{k-1}} f(z)\,\mathrm{d}z = \int_{\tilde{\Gamma}_k} f(z)\,\mathrm{d}z$$

Similarly we can find

$$\int_{\tilde{\Gamma}_0} f(z)\,\mathrm{d}z = \int_{\gamma_0} f(z)\,\mathrm{d}z\,; \quad \int_{\tilde{\Gamma}_n} f(z)\,\mathrm{d}z = \int_{\gamma_1} f(z)\,\mathrm{d}z$$

□

3.5. Simply connected domains

Definition. A domain U is *simply connected* if every closed curve in u is null-homotopic in U.

Star domains U are simply connected. Indeed, there exists a centre $a \in U$ such that $[a, z] \subset U$ for all $z \in U$. If $\gamma : [a, b] \to U$ is a closed curve, let $H(z, t) = (1 - s)a + s\gamma(t) \in U$ for $(s, t) \in [0, 1] \times [a, b]$. Then $H(s, t) \in U$, and H is a homotopy between γ and the constant curve $\gamma_0(t) = a$.

Theorem (Cauchy's theorem for simply connected domains)**.** If U is simply connected, then

$$\int_\gamma f(z)\,\mathrm{d}z = 0$$

for all holomorphic $f : U \to \mathbb{C}$, and every closed curve γ in U.

This is an immediate application of the above. The converse is also true, but is harder to prove.

Hence, U is simply connected if and only if $\int_\gamma f(z)\,\mathrm{d}z = 0$ for all holomorphic f and all closed γ in U. In particular, U is simply connected if and only if every closed curve in U is homologous to zero in U. Contrast this to the previous remark that if a curve is homologous to zero it is not necessarily null-homotopic.

4. Singularities

4.1. Motivation

Let U be open, and γ be a closed curve in U homologous to zero in U. Then, if $f : U \to \mathbb{C}$ is holomorphic, we have Cauchy's integral formula

$$\int_\gamma \underbrace{\frac{f(z)\,\mathrm{d}z}{z-a}}_{g(z)\mathrm{d}z} = 2\pi i \cdot I(\gamma; a) f(a)$$

for all $a \in U \setminus \operatorname{Im}\gamma$. This allows us to compute $\int_\gamma g(z)\,\mathrm{d}z$ for a holomorphic function $g : U \setminus \{a\} \to \mathbb{C}$ where γ does not pass through the point a, provided that g satisfies a particular condition: $(z-a)g(z)$ is the restriction to $U \setminus \{a\}$ of a holomorphic function $f : U \to \mathbb{C}$. We wish to drop this restriction and observe the consequences; that is, we wish to compute $\int_\gamma g(z)\,\mathrm{d}z$ for arbitrary holomorphic functions $g : U \setminus \{a\} \to \mathbb{C}$ for $a \in U$ and $a \notin \operatorname{Im}\gamma$. For example, consider $g(z) = e^{z^{-1}}$ for $U = \mathbb{C}$ and $a = 0$, $\gamma = \partial D(0,1)$. Note that $zg(z) = ze^{z^{-1}}$ is not continuous at $z = 0$, so it is certainly not holomorphic. This leads us to the study of singularities, and to eventually prove the residue theorem.

4.2. Removable singularities

Definition. Let $U \subseteq \mathbb{C}$ be open. If $a \in U$ and $f : U \setminus \{a\} \to \mathbb{C}$ is holomorphic, we say that f has an *isolated singularity* at a.

Definition. An isolated singularity a of f is a *removable singularity* if f can be defined at a such that the extended function is holomorphic on U.

Proposition. Let U be open, $a \in U$, and $f : U \setminus \{a\} \to \mathbb{C}$ be holomorphic. Then, the following are equivalent.

(i) f has a removable singularity at a;

(ii) $\lim_{z\to a} f(z)$ exists in $\mathbb{C}$;

(iii) there exists $D(a,\varepsilon) \subseteq U$ such that $|f(z)|$ is bounded in $D(a,\varepsilon) \setminus \{a\}$;

(iv) $\lim_{z\to a}(z-a)f(z) = 0$.

Proof. We can see that (i) implies (ii). If a is a removable singularity of f, then by definition there is a holomorphic function $g : U \to \mathbb{C}$ such that $f(z) = g(z)$ for all $z \in U \setminus \{a\}$. Then $\lim_{z\to a} f(z) = \lim_{z\to a} g(z) = g(a) \in \mathbb{C}$. Similarly, (ii) implies (iii) and (iii) implies (iv) are clear.

It suffices to check (iv) implies (i). Consider the function

$$h(z) = \begin{cases} (z-a)^2 f(z) & \text{if } z \neq a \\ 0 & \text{if } z = a \end{cases}$$

We have

$$\lim_{z\to a}\frac{h(z)-h(a)}{z-a}=\lim_{z\to a}(z-a)f(z)=0$$

Hence h is differentiable at a with $h'(a)=0$. Since h is differentiable in $U\setminus\{a\}$, we must have that h is holomorphic in U. Since $h(a)=h'(a)=0$, we can find $r>0$ and a holomorphic $g\colon D(a,r)\to\mathbb{C}$ such that $h(z)=(z-a)^2g(z)$ for $z\in D(a,r)$. Comparing this to the definition of h, we have that $f(z)=g(z)$ for $D(a,r)\setminus\{a\}$. By defining $f(a)=g(a)$, we have that f is differentiable at a with $f'(a)=g'(a)$. So a is a removable singularity of f. □

Example. Consider $f(z)=\frac{e^z-1}{z}$. Certainly f is holomorphic on $\mathbb{C}\setminus\{0\}$, and $\lim_{z\to 0} zf(z)=0$. So $z=0$ is a removable singularity. We can also see directly by the Taylor series of e^z at $z=0$ that $f(z)=\sum_{k=1}^{\infty}\frac{z^{k-1}}{k!}$ for $z\neq 0$, and the series on the right hand side defines an entire function.

Remark. If $u\colon D(0,1)\setminus\{0\}\to\mathbb{R}$ is a C^2 harmonic function, when can we say that $z=0$ is a removable singularity, i.e. that u extends to $z=0$ as a harmonic function? We can relate this to the study of holomorphic functions. However, unlike with previous cases, the analogy is more subtle in this case. We cannot necessarily construct a harmonic conjugate v such that $u+iv$ is holomorphic in $D(0,1)\setminus\{0\}$, because U is not simply connected.

There is a similar result, however. If $\lim_{z\to 0}u(z)$ exists, then the extended function is in fact C^2 and harmonic. More generally, if u is bounded near $z=0$, there exists a harmonic extension. We can also consider the case $\lim_{z\to 0}|z||u(z)|=0$; this is explored on the example sheets.

4.3. Poles

Note, if a holomorphic function f has a non-removable singularity, f is not bounded in $D(a,r)\setminus\{a\}$ for any $r>0$.

Definition. If $a\in U$ is an isolated singularity of f, then a is a *pole* of f if

$$\lim_{z\to a}|f(z)|=\infty$$

Example. $f(z)=(z-a)^{-k}$ for $k\in\mathbb{N}$ has a pole at a.

Definition. If $a\in U$ is an isolated singularity of f that is not removable or a pole, it is an *essential singularity*.

Remark. An equivalent characterisation for a to be an essential singularity is that the limit $\lim_{z\to a}|f(z)|$ does not exist. This follows from the previous proposition and the definition of a pole.

Example. $f(z)=e^{\frac{1}{z}}$ has $|f(iy)|=1$ for all $y\in\mathbb{R}\setminus\{0\}$ and $\lim_{x\to 0^+}f(x)=\infty$. So $z=0$ is an essential singularity of f.

Proposition. Let $f : U \setminus \{a\} \to \mathbb{C}$ be holomorphic. The following are equivalent.

(i) f has a pole at a;

(ii) there exists $\varepsilon > 0$ and a holomorphic function $h : D(a, \varepsilon) \to \mathbb{C}$ with $h(a) = 0$ and $h(z) \neq 0$ for all $z \neq a$ such that $f(z) = \frac{1}{h(z)}$ for $z \in D(a, \varepsilon) \setminus \{a\}$;

(iii) there exists a unique integer $k \geq 1$ and a unique holomorphic function $g : U \to \mathbb{C}$ with $g(a) \neq 0$ such that $f(z) = (z - a)^{-k} g(z)$ for $z \in U \setminus \{a\}$.

Remark. Since (i) implies (iii), there exists no holomorphic function on a punctured disc $f : D(a, R) \setminus \{a\} \to \mathbb{C}$ such that $|f(z)| \to \infty$ as $z \to a$ at the rate of a negative non-integer power of $|z - a|$, i.e. with $c|z - a|^{-s} \leq |f(z)| \leq C|z - a|^{-s}$ for some constants $s \in (0, \infty) \setminus \mathbb{N}$, $c > 0$, $C > 0$, and all $z \in D(a, R) \setminus \{a\}$.

Proof. We show (i) implies (ii). Since $\lim_{z \to a} |f(z)| = \infty$, there exists $\varepsilon > 0$ such that $|f(z)| \geq 1$ for all $0 < |z - a| < \varepsilon$. Hence $\frac{1}{f(z)}$ is holomorphic and bounded in $D(a, \varepsilon) \setminus \{a\}$. By the above proposition, $\frac{1}{f}$ has a removable singularity at a, so there exists a holomorphic function $h : D(a, \varepsilon) \to \mathbb{C}$ such that $\frac{1}{f} = h$, or equivalently, $f = \frac{1}{h}$, for $z \in D(a, \varepsilon) \setminus \{a\}$. Since $|f(z)| \to \infty$ as $z \to a$, we have that $h(a) = 0$.

Now we show (ii) implies (iii). Let ε and h be as in the definition of (ii). By Taylor series, there exists $k \geq 1$ and a holomorphic function $h_1 : D(a, \varepsilon) \to \mathbb{C}$ with $h_1(z) \neq 0$ for all $z \in D(a, \varepsilon)$ such that $h(z) = (z - a)^k h_1(z)$. If $g_1 = \frac{1}{h_1}$, then g_1 is holomorphic in $D(a, \varepsilon)$, $g_1 \neq 0$ in $D(a, \varepsilon)$, and $f(z) = (z - a)^{-k} g_1(z)$ in $D(a, \varepsilon) \setminus \{a\}$.

We can now define $g : U \to \mathbb{C}$ by $g(z) = g_1(z)$ for $z \in D(a, \varepsilon)$, and $g(z) = (z - a)^k f(z)$ for $z \in U \setminus \{a\}$. Since $f(z) = (z - a)^{-k} g_1(z)$, the definitions agree on $D(a, \varepsilon) \setminus \{a\}$, so g is well-defined and holomorphic in U, and $g(a) = g_1(a) \neq 0$. This proves the existence of an integer $k \geq 1$ and a holomorphic $g : U \to \mathbb{C}$ with $g(a) \neq 0$ such that $f(z) = (z - a)^{-k} g(z)$ for all $z \in U \setminus \{a\}$.

To prove uniqueness of k and g, suppose there exists $\tilde{k} \geq 1$ and a holomorphic $\tilde{g} : U \to \mathbb{C}$ with $\tilde{g}(a) \neq 0$ such that $f(z) = (z - a)^{-\tilde{k}} \tilde{g}(z)$ for all $z \in U \setminus \{a\}$. Then we have $g(z) = (z - a)^{k - \tilde{k}} \tilde{g}(z)$ for $z \in U \setminus \{a\}$. Since $g, \tilde{g}$ are holomorphic with $g(a) \neq 0$ and $\tilde{g}(a) \neq 0$, this can only be true if $k = \tilde{k}$, and hence $g = \tilde{g}$ on $U \setminus \{a\}$, and then at a by continuity.

It is clear that (iii) implies (i). □

Definition. If f has a pole at $z = a$, then the unique positive integer k given by the above proposition is the *order* of the pole at a. If $k = 1$, we say that f has a *simple pole* at a.

Let U be open and $S \setminus U$ be a discrete subset of U, so all points of S are isolated. If $f : U \setminus S \to \mathbb{C}$ is holomorphic and each $a \in S$ is either a removable singularity or a pole of f, then f is a *meromorphic* function on U. In particular, if $S = \varnothing$, f is holomorphic.

Remark. If $f : U \setminus \{a\} \to \mathbb{C}$ is holomorphic and the singularity $z = a$ is a pole of f, we can regard f as a continuous mapping onto the Riemann sphere $f : U \to \mathbb{C} \cup \{\infty\}$, by setting $f(a) = \infty$. Here, f is holomorphic on U. Holomorphicity of the extended map near the pole a follows from the fact that in a punctured disc about a, $\frac{1}{f}$ has the form $\frac{(z-a)^k}{g(z)}$ for some holomorphic g with $g(z) \neq 0$ near a; and the fact that any function h defined in a neighbourhood of ∞ in the Riemann sphere is holomorphic, by definition, if the function $\widetilde{h}(z) = h\left(\frac{1}{z}\right)$ if $z \neq 0$, $\widetilde{h}(0) = h(\infty)$ is holomorphic near zero. Hence $h \circ f = \widetilde{h} \circ \left(\frac{1}{f}\right)$ is holomorphic near a for all holomorphic h in a neighbourhood of ∞ in the Riemann sphere.

Hence, any meromorphic function $f : U \setminus S \to \mathbb{C}$ can be viewed as a holomorphic function $U \to \mathbb{C} \cup \{\infty\}$. Geometrically, therefore, poles are not 'real' singularities, and the only true isolated singularities are the essential singularities. This is explored further in Part II Riemann Surfaces.

4.4. Essential singularities

Remark. Suppose $z = a$ is an essential singularity of a holomorphic $f : U \setminus \{a\} \to \mathbb{C}$. Then there exists a sequence of points $a_n \in U \setminus \{a\}$, $a_n \to a$, such that $f(a_n) \to \infty$. This is because a is not removable. There is also another sequence of points $b_n \in U \setminus \{a\}$, $b_n \to a$ such that $(f(b_n))$ is bounded. This is because a is not a pole. We can generalise this further.

Theorem (Casorati–Weierstrass theorem)**.** If $f : U \setminus \{a\} \to \mathbb{C}$ is holomorphic and $a \in U$ is an essential singularity of f, then for any $\varepsilon > 0$, the set $f(D(a, \varepsilon) \setminus \{a\})$ is dense in $\mathbb{C}$.

The proof is an exercise on the second example sheet.

Theorem (Picard's theorem)**.** If $f : U \setminus \{a\} \to \mathbb{C}$ is holomorphic and $a \in U$ is an essential singularity of f, then there exists $w \in \mathbb{C}$ such that for any $\varepsilon > 0$, $\mathbb{C} \setminus \{w\} \subseteq f(D(a, \varepsilon) \setminus \{a\})$. In other words, in any neighbourhood $D(a, \varepsilon) \setminus \{a\}$, f attains all complex numbers except possibly one.

The proof is omitted.

4.5. Laurent series

If $z = a$ is a removable singularity of f, then for some $R > 0$, f is given by a power series $\sum_{n=0}^{\infty} c_n (z - a)^n$, which is the Taylor series of the holomorphic extension of f to $D(a, R)$, for all $z \in D(a, R) \setminus \{a\}$. If a is a pole of some order $k \geq 1$, then for some $R > 0$ we have $f(z) = (z - a)^{-k} g(z)$ for some holomorphic $g : D(a, R) \to \mathbb{C}$ and all $z \in D(a, R) \setminus \{a\}$, so using the Taylor seires of g, we find a series of the form $f(z) = \sum_{n=-k}^{\infty} c_n (z - a)^n$, for $z \in D(a, R) \setminus \{a\}$. When a is an essential singularity, we can still obtain an analogous series expansion with infinitely many terms with negative powers. More generally, we have the following.

Theorem (Laurent expansion)**.** Let f be holomorphic on an annulus

$$A = \{z \in \mathbb{C} : r < |z - a| < R\}$$

for $0 \leq r < R \leq \infty$. Then:

(i) f has a unique convergent series expansion

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n \equiv \sum_{n=1}^{\infty} c_{-n}(z-a)^{-n} + \sum_{n=0}^{\infty} c_n (z-a)^n$$

where the c_n are constants;

(ii) for any $\rho \in (r, R)$, the coefficient c_n is given by

$$c_n = \frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{n+1}}$$

(iii) if $r < \rho' \leq \rho < R$, then the two series in (i) separately converge uniformly on the set

$$\{z \in \mathbb{C} : \rho' \leq |z - a| \leq \rho\}$$

Remark. If f is the restriction of A of a holomorphic function g on the full disc $D(a, R)$, then by the formula in part (ii), we have for any negative $n = -m$, $m \geq 1$, the coefficient c_{-m} is zero by Cauchy's theorem. In this case, the Laurent series of f is the Taylor series of g restricted to A. The new content of the theorem is simply when f has no holomorphic extension to $D(a, R)$.

Proof. Let $w \in A$ and consider the function

$$g(z) = \begin{cases} \frac{f(z)-f(w)}{z-w} & \text{if } z \neq w \\ f'(w) & \text{if } z = w \end{cases}$$

This g is continuous in A and holomorphic in $A \setminus \{w\}$. Hence, this is holomorphic in A since this is a removable singularity. Let ρ_1, ρ_2 such that $r < \rho_1 < |w - a| < \rho_2 < R$. The two positively oriented curves $\partial D(a, \rho_1)$ and $\partial D(a, \rho_2)$ are homotopic in A. Hence,

$$\int_{\partial D(a,\rho_1)} g(z)\,\mathrm{d}z = \int_{\partial D(a,\rho_2)} g(z)\,\mathrm{d}z$$

Substituting for g,

$$\int_{\partial D(a,\rho_1)} \frac{f(z)\,\mathrm{d}z}{z-w} - 2\pi i \cdot I(\partial D(a,\rho_1); w) f(w) = \int_{\partial D(a,\rho_2)} \frac{f(z)\,\mathrm{d}z}{z-w} - 2\pi i \cdot I(\partial D(a,\rho_2); w) f(w)$$

We have

$$I(\partial D(a,\rho_1); w) = 0; \quad I(\partial D(a,\rho_2); w) = I(\partial D(a,\rho_2); a) = 1$$

Hence,

$$f(w) = \frac{1}{2\pi i}\int_{\partial D(a,\rho_2)} \frac{f(z)\,\mathrm{d}z}{z-w} - \frac{1}{2\pi i}\int_{\partial D(a,\rho_1)} \frac{f(z)\,\mathrm{d}z}{z-w}$$

This is an analogue of Cauchy's integral formula for annular domains. We can now proceed as before when proving the Taylor series expansion for holomorphic functions.

For the first integral, consider the expansion

$$\frac{1}{z-w} = \frac{1}{z-a-(w-a)} = \sum_{n=0}^{\infty} \frac{(w-a)^n}{(z-a)^{n+1}}$$

This series converges uniformly over $z \in \partial D(a,\rho_2)$, since $\left|\frac{w-a}{z-a}\right| < 1$. For the second integral, consider

$$\frac{1}{z-w} = \frac{1}{z-a-(w-a)} = -\frac{1}{(w-a)\left(1-\frac{z-a}{w-a}\right)} = -\sum_{n=0}^{\infty} \frac{(z-a)^n}{(w-a)^{n+1}}$$

Likewise, this series converges uniformly over $z \in \partial D(a,\rho_1)$, since $\left|\frac{z-a}{w-a}\right| < 1$ in this disc. Substituting these into the representation formula, we can switch integration and summation due to uniform convergence. This gives

$$f(w) = \sum_{n=0}^{\infty} c_n (w-a)^n + \sum_{n=1}^{\infty} c_{-n}(w-a)^{-n}$$

where

$$c_n = \frac{1}{2\pi i}\int_{\partial D(a,\rho_2)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{n+1}}$$

for $n \geq 0$, and

$$c_n = \frac{1}{2\pi i}\int_{\partial D(a,\rho_1)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{n+1}}$$

for $n \leq -1$. Since $\partial D(a,\rho_1)$ and $\partial D(a,\rho_2)$ are homotopic in A to $\partial D(a,\rho)$ for any $\rho \in (r,R)$, we have that

$$c_n = \frac{1}{2\pi i}\int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{z-a}$$

for any $\rho \in (r,R)$ and $n \in \mathbb{Z}$, so (i) and the formula (ii) both hold.

To show (iii) and uniqueness, suppose there exist constants c_n such that, for all $z \in A$, we have

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n \qquad (*)$$

Let $r < \rho' \leq \rho < R$. Then the power series $\sum_{n=0}^{\infty} c_n (z-a)^n$ converges for $z \in A$, so it has radius of convergence at least R, and converges uniformly for $|z-a| \leq \rho$. Further, the series

$\sum_{n=1}^{\infty} c_{-n}(z-a)^{-n}$ converges on A. Let $\zeta = (z-a)^{-1}$. Then the power series $\sum_{n=1}^{\infty} c_{-n}\zeta^n$ converges for $\frac{1}{R} < |\zeta| < \frac{1}{r}$ so it has radius of convergence at least $\frac{1}{r}$ and converges uniformly for $|\zeta| \le \frac{1}{\rho'}$. Thus, the series $\sum_{n=1}^{\infty} c_{-n}(z-a)^{-n}$ converges uniformly for $|z-a| \ge \rho'$. Hence $(*)$ converges uniformly in $\rho' \le |z-a| \le \rho$. Hence, for any $m \in \mathbb{Z}$, we have

$$\int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{m+1}} = \sum_{n=-\infty}^{\infty} c_n \int_{\partial D(a,\rho)} (z-a)^{n-m-1}\,\mathrm{d}z$$

By the fundamental theorem of calculus, the only nonzero integral on the right hand side occurs when $n-m-1=-1$, which occurs for $n=m$ only. This integral gives

$$c_m = \frac{1}{2\pi i} \int_{\partial D(a,\rho)} \frac{f(z)\,\mathrm{d}z}{(z-a)^{m+1}}$$

for all $\rho \in (r,R)$. This formula also implies the uniqueness of the c_n for which the series expansion is valid. □

Remark. The above proof shows that if $f : A \equiv D(a,R) \setminus \overline{D(a,r)} \to \mathbb{C}$ is holomorphic, then there is a holomorphic function $f_1 : D(a,R) \to \mathbb{C}$ and a holomorphic function $f_2 : \mathbb{C} \setminus \overline{D(a,r)} \to \mathbb{C}$ such that $f = f_1 + f_2$ on A. This decomposition is not unique, since we can take $f_1 \mapsto f_1 + g$ and $f_2 \mapsto f_2 - g$ for an entire function g. However, if we also require $f_2(z) \to 0$ as $z \to \infty$, the decomposition into two series given in (ii) above is unique.

4.6. Coefficients of Laurent series

Let $f : D(a,R) \setminus \{a\} \to \mathbb{C}$ be holomorphic, so $z=a$ is an isolated singularity of f. Then, by the Laurent series with $r=0$, we have a unique set of complex numbers c_n such that

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n$$

Then,

(i) If $c_n = 0$ for all $n < 0$, we have $f(z) = \sum_{n=0}^{\infty} c_n(z-a)^n \equiv g(z)$ on $D(a,R) \setminus \{a\}$. Since g is holomorphic on $D(a,R)$, $z=a$ is a removable singularity.

(ii) If $c_{-k} \neq 0$ for some $k \ge 1$ and $c_{-n} = 0$ for all $n \ge k+1$, we have

$$f(z) = \frac{c_{-k}}{(z-a)^k} + \frac{c_{-k+1}}{(z-a)^{k+1}} + \cdots + \frac{c_{-1}}{z-a} + \sum_{n=0}^{\infty} c_n (z-a)^n$$

Hence, $f(z) = (z-a)^{-k} g(z)$ for a function g which is holomorphic on $D(a,R)$, and where $g(a) = c_{-k} \neq 0$. Equivalently, $z=a$ is a pole of order k.

(iii) If $c_n \neq 0$ for infinitely many $n < 0$, $z=a$ is an essential singularity. This holds since the above two parts were all bidirectional implications.

4.7. Residues

Definition. Let $f : D(a,R) \setminus \{a\} \to \mathbb{C}$ be holomorphic. The coefficient c_{-1} of the Laurent series of f in $D(a,R) \setminus \{a\}$ is called the *residue of f at a*, denoted $\operatorname{Res}_f(a)$. The series

$$f_P = \sum_{n=1}^{\infty} c_{-n}(z-a)^{-n}$$

is known as the *principal part of f at a*.

We know that f_P is holomorphic on $\mathbb{C}\setminus\{a\}$, with the series defining f_P converging uniformly on compact subsets of $\mathbb{C} \setminus \{a\}$. By the Laurent series, $f = f_P + h$ on $D(a,R) \setminus \{a\}$, where h is holomorphic on $D(a,R)$. Let γ be a closed curve in $D(a,R)$ with $a \notin \operatorname{Im}\gamma$. Then $\int_\gamma h(z)\,\mathrm{d}z = 0$ by Cauchy's theorem, and hence $\int_\gamma f(z)\,\mathrm{d}z = \int_\gamma f_P(z)\,\mathrm{d}z = 2\pi i \cdot I(\gamma;a)\operatorname{Res}_f(a)$, where the last inequality holds by uniform convergence of the series for f_P and the fundamental theorem of calculus. This reasoning can be extended to the case of more then one isolated singularity.

Theorem (residue theorem)**.** Let U be an open set, $\{a_1,\dots,a_k\} \subset U$ be finite, and $f : U \setminus \{a_1,\dots,a_k\} \to \mathbb{C}$ be holomorphic. If γ is a closed curve in U homologous to zero in U, and if $a_j \notin \operatorname{Im}\gamma$ for each j, then

$$\frac{1}{2\pi i}\int_\gamma f(z)\,\mathrm{d}z = \sum_{j=1}^{k} I(\gamma;a_j)\operatorname{Res}_f(a_j)$$

This is a generalisation of Cauchy's integral formula.

Proof. Let $f_P^{(j)} = \sum_{n=1}^{\infty} c_{-n}^{(j)}(z-a_j)^{-n}$ be the principal part of f at a_j. Then $f_P^{(j)}$ is holomorphic in $\mathbb{C}\setminus\{a_j\}$, and hence is holomorphic in $\mathbb{C}\setminus\{a_1,\dots,a_k\}$. Let

$$h \equiv f - \left(f_P^{(1)} + \cdots + f_P^{(k)}\right)$$

This h is holomorphic in $U \setminus \{a_1,\dots,a_k\}$. Let j be fixed. Then $f - f_P^{(j)}$ has a removable singularity at $z = a_j$. For all $\ell \neq j$, $f_P^{(\ell)}$ is holomorphic at a_j. Hence h has a removable singularity at a_j. This is true for all j, so h extends to all of U as a holomorphic function. By Cauchy's theorem, $\int_\gamma h(z)\,\mathrm{d}z = 0$. Hence

$$\frac{1}{2\pi i}\int_\gamma f(z)\,\mathrm{d}z = \sum_{j=1}^{k}\frac{1}{2\pi i}\int_\gamma f_P^{(j)}(z)\,\mathrm{d}z$$

By termwise integration of the series for $f_P^{(j)}$, which converges uniformly on compact subsets of $\mathbb{C}\setminus\{a_j\}$, we have

$$\frac{1}{2\pi i}\int_\gamma f_P^{(j)}(z)\,\mathrm{d}z = I(\gamma;a_j)\operatorname{Res}_f(a_j)$$

as required. □

There are simple ways to calculate residues if we know information about the singularity in question.

(i) If f has a simple pole at $z = a$, then

$$\mathrm{Res}_f(a) = \lim_{z\to a}(z-a)f(z)$$

Indeed, near a, we have $f(z) = (z-a)^{-1}g(z)$ where g is holomorphic and $g(a) \neq 0$. Hence, by the Taylor expansion of g, we have that $\mathrm{Res}_f(a) = g(a)$.

(ii) If f has a pole of order k at a, then near a we have that $f(z) = (z-a)^{-k}g(z)$ where g is holomorphic and $g(a) \neq 0$. In this case, the residue $\mathrm{Res}_f(a)$ is the coefficient of the $(z-a)^{k-1}$ term of the Taylor series of g at a, which is

$$\mathrm{Res}_f(a) = \frac{g^{(k-1)}(a)}{(k-1)!}$$

(iii) If $f = \frac{g}{h}$ where g and h are holomorphic at $z = a$, such that $g(a) \neq 0$ and h has a simple zero at $z = a$, then from (i) we have

$$\mathrm{Res}_f(a) = \lim_{z\to a}\frac{(z-a)g(z)}{h(z)} = \lim_{z\to a}\frac{g(z)}{\frac{h(z)-h(a)}{z-a}} = \frac{g(a)}{h'(a)}$$

Example. For $0 < \alpha < 1$, we will show that

$$\int_0^\infty \frac{x^{-\alpha}}{1+x}\,\mathrm{d}x = \frac{\pi}{\sin\pi\alpha}$$

Let $g(z) = z^{-\alpha}$ be the branch of $z^{-\alpha}$ defined by $g(z) = e^{-\alpha\ell(z)}$, where $\ell(z)$ is the holomorphic branch of logarithm on $U = \mathbb{C}\backslash\{x \in \mathbb{R} : x \geq 0\}$. given by $\ell(z) = \log|z| + i\arg z$ where $\arg(z)$ takes values in $(0, 2\pi)$. Let $f(z) = \frac{g(z)}{1+z}$. Then

$$f(z) = \frac{|z|^{-\alpha}e^{-i\alpha\arg z}}{1+z}$$

and f is holomorphic in $U\backslash\{-1\}$ where $z = -1$ is a simple pole with $\mathrm{Res}_f(-1) = \lim_{z\to-1}(z+1)f(z) = e^{-i\pi\alpha}$.

Let ε, R be such that $0 < \varepsilon < 1 < R$ and $\theta > 0$ be small. Let γ be the positively-oriented 'keyhole countour' determined by the two circular arcs $\gamma_R : [\theta, 2\pi-\theta] \to U$ and the two line segments $\gamma_1, \gamma_2 : [\varepsilon, R] \to U$ given by

$$\gamma_R(t) = Re^{it}; \quad \gamma_\varepsilon(t) = \varepsilon e^{i(2\pi-t)}; \quad \gamma_1(t) = te^{i\theta}; \quad \gamma_2(t) = te^{i(2\pi-\theta)}$$

The domain U is star shaped and hence simply connected, and so γ is homologous to zero in U. Directly from the definitions of γ and the winding number, we can show that $I(\gamma; -1) = 1$.

By the residue theorem, we find $\int_\gamma f(z)\,\mathrm{d}z = 2\pi i e^{-i\pi\alpha}$. Now,

$$\int_{\gamma_1} f(z)\,\mathrm{d}z = \int_\varepsilon^R f(te^{i\theta})e^{i\theta}\,\mathrm{d}t = \int_\varepsilon^R \frac{t^{-\alpha}e^{i(1-\alpha)\theta}}{1+te^{i\theta}}\,\mathrm{d}t$$

and

$$\int_{\gamma_2} f(z)\,\mathrm{d}z = \int_\varepsilon^R f(te^{i(2\pi-\theta)})e^{i(2\pi-\theta)}\,\mathrm{d}t = \int_\varepsilon^R \frac{t^{-\alpha}e^{i(1-\alpha)(2\pi-\theta)}}{1+te^{i(2\pi-\theta)}}\,\mathrm{d}t$$

As $\theta \to 0^+$, we can show that the integrands converge uniformly on $[\varepsilon, R]$ to $\frac{t^{-\alpha}}{1+t}$ and $\frac{e^{-2i\pi\alpha}t^{-\alpha}}{1+t}$ respectively. Hence,

$$\lim_{\theta\to 0^+}\left[\int_{\gamma_1} f(z)\,\mathrm{d}z + \int_{(-\gamma_2)} f(z)\,\mathrm{d}z\right] = \left(1-e^{-2i\pi\alpha}\right)\int_\varepsilon^R \frac{t^{-\alpha}}{1+t}\,\mathrm{d}t$$

For all $z \in \operatorname{Im}\gamma_R$, we have $|f(z)| \le \frac{R^{-\alpha}}{R-1}$; and for all $z \in \operatorname{Im}\gamma_\varepsilon$, we have $|f(z)| \le \frac{\varepsilon^{-\alpha}}{1-\varepsilon}$. Hence,

$$\left|\int_{\gamma_R} f(z)\,\mathrm{d}z + \int_{\gamma_\varepsilon} f(z)\,\mathrm{d}z\right| \le \frac{2\pi R^{1-\alpha}}{R-1} + \frac{2\pi\varepsilon^{1-\alpha}}{1-\varepsilon}$$

Note that the right hand side is independent of θ, even though γ_R and γ_ε depend on θ. Since

$$\int_\gamma f(z)\,\mathrm{d}z - \left(\int_{\gamma_1} f(z)\,\mathrm{d}z + \int_{(-\gamma_2)} f(z)\,\mathrm{d}z\right) = \int_{\gamma_R} f(z)\,\mathrm{d}z + \int_{\gamma_\varepsilon} f(z)\,\mathrm{d}z$$

we then have that

$$\left|2\pi i e^{-i\pi\alpha} - \left(\int_{\gamma_1} f(z)\,\mathrm{d}z + \int_{(-\gamma_2)} f(z)\,\mathrm{d}z\right)\right| \le \frac{2\pi R^{1-\alpha}}{R-1} + \frac{2\pi\varepsilon^{1-\alpha}}{1-\varepsilon}$$

First letting $\theta \to 0^+$ in this, and then letting $\varepsilon \to 0^+$ and $R \to \infty$, we conclude

$$\left(1-e^{-2\pi i\alpha}\right)\int_0^\infty \frac{t^{-\alpha}}{1+t}\,\mathrm{d}t = 2\pi i e^{-i\pi\alpha}$$

or,

$$\int_0^\infty \frac{t^{-\alpha}}{1+t}\,\mathrm{d}t = \frac{\pi}{\sin\pi\alpha}$$

4.8. Jordan's lemma

Lemma. Let f be a continuous complex-valued function on the semicircle $C_R^+ = \operatorname{Im}\gamma_R^+$ in the upper half-plane, where $R > 0$ and $\gamma_R^+(t) = Re^{it}$ for $0 \le t \le \pi$. Then, for $\alpha > 0$,

$$\left|\int_{\gamma_R^+} f(z)e^{i\alpha z}\,\mathrm{d}z\right| \le \frac{\pi}{\alpha}\sup_{z\in C_R^+}|f(z)|$$

In particular, if f is continuous in $H^+ \setminus D(0,R_0)$ for $R_0 > 0$ where $H^+ = \{z : \operatorname{Im} z \geq 0\}$ and if $\sup_{z\in C_R^+} |f(z)| \to 0$ as $R \to \infty$, then for each $\alpha > 0$, we have

$$\int_{\gamma_R^+} f(z)e^{i\alpha z}\,\mathrm{d}z \to 0$$

as $R \to \infty$.

A similar statement holds for $\alpha < 0$ and the semicircle $C_R^- = \operatorname{Im}\gamma_R^-$ in the lower half-plane where $\gamma_R^-(t) = -Re^{it}$ for $R > 0$ and $0 \leq t \leq \pi$.

Proof. Let $M_R = \sup_{z\in C_R^+} |f(z)|$. Then,

$$\begin{aligned}
\left|\int_{\gamma_R^+} f(z)e^{i\alpha z}\,\mathrm{d}z\right| &= \left|\int_0^\pi f(Re^{it})e^{-\alpha R\sin t + i\alpha R\cos t} iRe^{it}\,\mathrm{d}t\right| \\
&\leq RM_R\int_0^\pi e^{-\alpha R\sin t}\,\mathrm{d}t \\
&= RM_R\left(\int_0^{\frac{\pi}{2}} e^{-\alpha R\sin t}\,\mathrm{d}t + \int_{\frac{\pi}{2}}^\pi e^{-\alpha R\sin t}\,\mathrm{d}t\right) \\
&= 2RM_R\int_0^{\frac{\pi}{2}} e^{-\alpha R\sin t}\,\mathrm{d}t \\
&\leq 2RM_R\int_0^{\frac{\pi}{2}} e^{\frac{-2\alpha Rt}{\pi}}\,\mathrm{d}t \\
&= \frac{\pi M_R}{\alpha}\left(1 - e^{-\alpha R}\right) \leq \frac{\pi M_R}{\alpha}
\end{aligned}$$

where we have used the fact that for $t \in (0, \frac{\pi}{2}]$, $\varphi(t) \equiv \frac{\sin t}{t} \geq \frac{2}{\pi}$ since $\varphi(\frac{\pi}{2}) = \frac{2}{\pi}$ and $\varphi'(t) \leq 0$ on $[0, \frac{\pi}{2}]$. □

Lemma (integrals on small circular arcs)**.** Let f be holomorphic in $D(a,R) \setminus \{a\}$ with a simple pole at $z = a$. Let $\gamma_\varepsilon : [\alpha,\beta] \to \mathbb{C}$ be the circular arc $\gamma_\varepsilon(t) = a + \varepsilon e^{it}$. Then

$$\lim_{\varepsilon\to 0^+}\int_{\gamma_\varepsilon} f(z)\,\mathrm{d}z = (\beta - \alpha)i\operatorname{Res}_f(a)$$

Proof. Let $f(z) = \frac{c}{z-a} + g(z)$ where g is holomorphic in $D(a,R)$ and $c = \operatorname{Res}_f(a)$. Then

$$\left|\int_{\gamma_\varepsilon} g(z)\,\mathrm{d}z\right| = \left|\int_\alpha^\beta g(a+\varepsilon e^{it})\varepsilon ie^{it}\right| \leq \varepsilon(\beta-\alpha)\sup_{t\in[\alpha,\beta]}|g(a+\varepsilon e^{it})| \to 0$$

as $\varepsilon \to 0^+$. By direct calculation,

$$\int_{\gamma_\varepsilon} \frac{c}{z-a}\,\mathrm{d}z = (\beta-\alpha)i\,\mathrm{Res}_f(a)$$

Hence the claim follows. □

Example. Consider $\int_0^\infty \frac{\sin x}{x}\,\mathrm{d}x$. Let $f(z) = \frac{e^{iz}}{z}$. Consider the integral $\int_\gamma f(z)\,\mathrm{d}z$ over the curve $\gamma = \gamma_R + \gamma_1 + \gamma_\varepsilon + \gamma_2$ where

(i) $\gamma_R(t) = Re^{it}$ for $0 \le t \le \pi$;

(ii) $\gamma_1(t) = t$ for $-R \le t \le -\varepsilon$;

(iii) $\gamma_\varepsilon(t) = \varepsilon e^{-it}$ for $-\pi \le t \le 0$;

(iv) $\gamma_2(t) = t$ for $\varepsilon \le t \le R$.

By Jordan's lemma, $\int_{\gamma_R} f(z)\,\mathrm{d}z \to 0$ as $R \to \infty$. f has a simple pole at $z = 0$ with $\mathrm{Res}_f(0) = \lim_{z\to 0} zf(z) = 1$. By the above lemma, $\int_{-\gamma_\varepsilon} f(z)\,\mathrm{d}z \to \pi i$ as $\varepsilon \to 0^+$.

Since f is holomorphic in $U = \mathbb{C} \setminus \{0\}$ and γ is homologous to zero in U, Cauchy's theorem gives that

$$\int_\gamma f(z)\,\mathrm{d}z = 0 \implies \int_{\gamma_R} f(z)\,\mathrm{d}z + \int_{-R}^{-\varepsilon} \frac{e^{it}}{t}\,\mathrm{d}t + \int_{\gamma_\varepsilon} f(z)\,\mathrm{d}z + \int_\varepsilon^R \frac{e^{it}}{t}\,\mathrm{d}t = 0$$

Combining the two integrals on the real axis under a change of variables,

$$\int_\varepsilon^R \frac{e^{it} - e^{-it}}{t}\,\mathrm{d}t + \int_{\gamma_R} f(z)\,\mathrm{d}z + \int_{\gamma_\varepsilon} f(z)\,\mathrm{d}z = 0$$

Letting $R \to \infty$ and $\varepsilon \to 0^+$, we have

$$\int_0^\infty \frac{\sin t}{t}\,\mathrm{d}t = \frac{\pi}{2}$$

Example. We prove that $\sum_{n=1}^\infty \frac{1}{n^2} = \frac{\pi^2}{6}$. Consider the function

$$f(z) = \frac{\pi\cot(\pi z)}{z^2} = \frac{\pi\cos(\pi z)}{z^2\sin(\pi z)}$$

This is holomorphic in $\mathbb{C}$ except for simple poles at each point in $\mathbb{Z} \setminus \{0\}$, and an order 3 pole at zero. Near $n \in \mathbb{Z} \setminus \{0\}$, we have $f(z) = \frac{g(z)}{h(z)}$ where $g(n) \neq 0$ and h has a simple zero at n, and so

$$\mathrm{Res}_f(n) = \frac{g(n)}{h'(n)} = \frac{1}{n^2}$$

To compute the residue at zero, consider

$$\cot z = \frac{\cos z}{\sin z} = \left(1 - \frac{z^2}{2} + O(z^4)\right)\left(z - \frac{z^3}{6} + O(z^5)\right)^{-1} = \frac{1}{z} - \frac{z}{3} + O(z^2)$$

Hence,

$$\frac{\pi \cot(\pi z)}{z^2} = \frac{1}{z^3} - \frac{\pi^2}{3z} + \dots$$

This shows that $\operatorname{Res}_f(0) = -\frac{\pi^2}{3}$. For $N \in \mathbb{N}$, let γ_N be the positively oriented boundary of the square defined by the lines $x = \pm(N + \frac{1}{2})$ and $y = \pm(N + \frac{1}{2})$. By the residue theorem,

$$\int_{\gamma_N} f(z)\,\mathrm{d}z = 2\pi i\left[2\left(\sum_{n=1}^{N} \frac{1}{n^2}\right) - \frac{\pi^2}{3}\right] \tag{$*$}$$

Since $\operatorname{length}(\gamma_N) = 4(2N+1)$, we have

$$\begin{aligned}\left|\int_{\gamma_N} f(z)\,\mathrm{d}z\right| &\le \sup_{\gamma_N}\left|\frac{\pi\cot(\pi z)}{z^2}\right| \cdot 4(2N+1)\\ &\le \sup_{\gamma_N}|\cot(\pi z)| \cdot \frac{4(2N+1)\pi}{(N+\frac{1}{2})^2}\\ &= \frac{16\pi}{2N+1} \cdot \sup_{\gamma_N}|\cot(\pi z)|\end{aligned}$$

On γ_N, it is possible to show that $\cot(\pi z)$ is bounded independently of N. Hence,

$$\int_{\gamma_N} f(z)\,\mathrm{d}z \to 0$$

as $N \to \infty$. Letting $N \to \infty$ in $(*)$, we find

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

5. The argument principle, local degree, and Rouché's theorem

5.1. The argument principle

Proposition. If f has a zero (or pole) of order $k \geq 1$ at $z = a$, then $\frac{f'}{f}$ has a simple pole at $z = a$ with residue k (or $-k$, respectively).

Proof. If $z = a$ is a zero of order k, there is a disc $D(a, r)$ such that $f(z) = (z-a)^k g(z)$ for $z \in D(a, r)$ where $g : D(a, r) \to \mathbb{C}$ is holomorphic with $g(z) \neq 0$ for all $z \in D(a, r)$. Hence,

$$f'(z) = k(z-a)^{k-1} g(z) + (z-a)^k g'(z)$$

and

$$\frac{f'(z)}{f(z)} = \frac{k}{z-a} + \frac{g'(z)}{g(z)}$$

for all $z \in D(a, r) \setminus \{a\}$. Since $\frac{g'}{g}$ is holomorphic in $D(a, R)$, the claim follows. A similar argument holds for poles. □

Definition. The order of a zero or pole a of a holomorphic function f is denoted $\operatorname{ord}_f(a)$.

Theorem (the argument principle)**.** Let f be a meromorphic function on a domain U with finitely many zeroes $a_1, \dots, a_k$ and finitely many poles $b_1, \dots, b_\ell$. If γ is a closed curve in U homologous to zero in U, and if $a_i, b_j \notin \operatorname{Im} \gamma$ for all i, j, then

$$\frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)} \, \mathrm{d}z = \sum_{i=1}^{k} I(\gamma; a_i) \operatorname{ord}_f(a_i) - \sum_{j=1}^{\ell} I(\gamma; b_j) \operatorname{ord}_f(b_j)$$

Proof. Apply the residue theorem to $g = \frac{f'}{f}$. If $z_0 \in U$ is not a pole of f, then f and hence f' are holomorphic near z_0. If additionally z_0 is not a zero of f, g is holomorphic near z_0. So the set of singularities of g is precisely $\{a_1, \dots, a_k\} \cup \{b_1, \dots, b_\ell\}$. By the previous proposition, their residues are known, and the result follows. □

Remark. Let f, γ be as in the theorem, and let $\Gamma(t) = f(\gamma(t))$. Then $\Gamma(t)$ is a closed curve with image $\operatorname{Im} \Gamma \subset \mathbb{C} \setminus \{0\}$, since no zeroes or poles of f are in $\operatorname{Im} \gamma$. Moreover, if $[a, b]$ is the domain of γ, we have

$$I(\Gamma; 0) = \frac{1}{2\pi i} \int_\Gamma \frac{\mathrm{d}z}{z} = \frac{1}{2\pi i} \int_a^b \frac{\Gamma'(t)}{\Gamma(t)} \, \mathrm{d}t = \frac{1}{2\pi i} \int_a^b \frac{f'(\gamma(t))\gamma'(t)}{f(\gamma(t))} \, \mathrm{d}t = \frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)} \, \mathrm{d}z$$

Thus, $\frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)}$ is the number of times the image curve $f \circ \gamma$ winds around zero as we move along γ.

Definition. Let Ω be a domain, and let γ be a closed curve in $\mathbb{C}$. We say that γ *bounds* Ω if $I(\gamma; w) = 1$ for all $w \in \Omega$, and $I(\gamma; w) = 0$ for all $w \in \mathbb{C} \setminus (\Omega \cup \operatorname{Im}\gamma)$.

Example. $\partial D(0,1)$ bounds $D(0,1)$, but does not bound $D(0,1) \setminus \{0\}$.

Remark. If γ bounds Ω, then

(i) Ω is bounded. Indeed, let $D(a,R)$ such that $\operatorname{Im}\gamma \subseteq D(a,R)$. Then $I(\gamma; w) = 0$ for $w \in \mathbb{C} \setminus D(a,R)$. Since $I(\gamma; w) = 1$ for all $w \in \Omega$, we must have $\Omega \subset D(a,R)$.

(ii) the topological boundary $\partial\Omega$ is contained within $\operatorname{Im}\gamma$, but it need not be the case that $\partial\Omega = \operatorname{Im}\gamma$.

There is a large class of closed curves that bound domains, namely, *simple closed curves*, which are curves $\gamma : [a,b] \to \mathbb{C}$ with $\gamma(a) = \gamma(b)$, and such that $\gamma(t_1) = \gamma(t_2)$ implies $t_1 = t_2$ or $t_1, t_2 \in \{a,b\}$. That a simple closed curve bounds a domain is a highly non-trivial fact guaranteed by the Jordan curve theorem: if γ is a simple closed curve, then $\mathbb{C} \setminus \operatorname{Im}\gamma$ consists precisely of two connected components, one of which is bounded and the other unbounded, and moreover, γ (or $-\gamma$) bounds the bounded component, and $\operatorname{Im}\gamma$ is the boundary of each of the two components. Thus, if Ω_1 is the bounded component and Ω_2 is the unbounded component, then after possibly changing the orientation of γ, we have $I(\gamma; w) = 1$ for $w \in \Omega_1$, and $I(\gamma; w) = 0$ for $w \in \Omega_2$. This last assertion is simply that for any disc $D(a,R) \supset \operatorname{Im}\gamma$, we have $I(\gamma; w) = 0$ for all $w \in \mathbb{C} \setminus D(a,R)$.

For a domain bounded by a closed curve, the argument principle gives the following.

Corollary. Let γ be a closed curve bounding a domain Ω, and let f be meromorphic in an open set U with $\Omega \cup \operatorname{Im}\gamma \subseteq U$. Suppose that f has no zeroes or poles on $\operatorname{Im}\gamma$. Then f has finitely many zeroes and finitely many poles in Ω.

Let the number of zeroes in Ω be N, and the number of poles in Ω be P, both counted with multiplicity. Then in addition we have that

$$N - P = \frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z = I(\Gamma; 0)$$

where $\Gamma = f \circ \gamma$.

Proof. Since f is meromorphic in U, its singularities form a discrete set $S \subset U$ consisting of poles or removable singularities. Since γ bounds Ω, we have that Ω is bounded and hence $\overline{\Omega}$ is compact. Also, $\overline{\Omega} \subseteq \Omega \cup \operatorname{Im}\gamma \subseteq U$. If $\overline{\Omega} \cap S$ is infinite, then by compactness of $\overline{\Omega}$, there exists a point $w \in \overline{\Omega}$ and distinct points $w_j \in \overline{\Omega} \cap S$ such that $w_j \to w$. If $w \notin S$, then f is defined and holomorphic near w which is impossible since $w_j \in S$ and $w_j \to w$. So $w \in S$, but this is impossible since S is a discrete set. So $\overline{\Omega} \cap S$ is finite, and in particular P is finite.

If f has infinitely many zeroes in Ω, then by compactness there exists $z \in \overline{\Omega} \subset U$ and distinct zeroes $z_j \in \Omega$ such that $z_j \to z$. Then either $z \in U \setminus S$, or (if $z \in S$) z is a removable singularity, since otherwise z would be a pole and hence $|f(\zeta)| \to \infty$ as $\zeta \to z$ which is

impossible since $z_j \to z$ and $f(z_j) = 0$. In either case, by the principle of isolated zeroes, f must be identically zero in $D(z, \rho) \setminus \{z\}$ for some $\rho > 0$. Since f is holomorphic in $\Omega \setminus S$ which is connected (since $\Omega \cap S$ is finite and Ω) is connected, it follows from the unique continuation principle that $f \equiv 0$ in Ω. This is impossible since f has no zeroes in $\operatorname{Im}\gamma$, so N must be finite.

By the definition of γ bounding Ω, we have that $I(\gamma; w) = 1$ for all $w \in \Omega$, and $I(\gamma; w) = 0$ for all $w \in \mathbb{C} \setminus (\Omega \cup \operatorname{Im}\gamma)$. In particular, γ is homologous to zero in U. The final conclusion then follows from the fact that Γ is a closed curve in $\mathbb{C} \setminus \{0\}$ and $I(\gamma; 0) = \frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z$ as proven above. □

5.2. Local degree theorem

Definition. Let f be a holomorphic function on a disc $D(a, R)$ that is not constant. Then the *local degree of f at a*, denoted $\deg_f(a)$, is the order of the zero of $f(z) - f(a)$ at $z = a$. This is a finite positive integer.

Example. If $f(z) = (z-1)^4 + 1$ has $\deg_f(1) = 4$.

Theorem. Let $f : D(a, R) \to \mathbb{C}$ be holomorphic and non-constant, with $\deg_f(a) = d$. Then there exists $r_0 > 0$ such that for any $r \in (0, r_0]$, there exists $\varepsilon > 0$ such that for all w with $0 < |f(a) - w| < \varepsilon$, the equation $f(z) = w$ has precisely d distinct roots in $D(a, r) \setminus \{a\}$.

Proof. Let $g(z) = f(z) - f(a)$. Since g is non-constant, $g' \not\equiv 0$ in $D(a, R)$. Applying the principle of isolated zeroes to g and g', there exists $r_0 \in (0, R)$ such that $g(z) \neq 0$ and $g'(z) \neq 0$ for $z \in \overline{D(a, r_0)} \setminus \{a\}$.

We will show that the conclusion holds for this choice of r_0. Let $r \in (0, r_0]$, and for $t \in [0, 1]$, let $\gamma(t) = a + re^{2\pi i t}$ and $\Gamma(t) = g(\gamma(t))$. Note that $\operatorname{Im}\Gamma$ is compact and hence closed in $\mathbb{C}$, and $0 \notin \operatorname{Im}\Gamma$ since $g \neq 0$ on $\partial D(a, r)$. Hence there exists $\varepsilon > 0$ such that $D(0, \varepsilon) \subseteq \mathbb{C} \setminus \operatorname{Im}\Gamma$.

We now show that this ε satisfies the conditions in the theorem for this r. Let w such that $0 < |w - f(a)| < \varepsilon$. Then $w - f(a) \in D(0, \varepsilon) \subseteq \mathbb{C} \setminus \operatorname{Im}\Gamma$. Since $z \mapsto I(\Gamma; z)$ is locally constant, it is constant on $D(0, \varepsilon)$, so in particular $I(\Gamma; w - f(a)) = I(\Gamma; 0)$.

By direct calculation,

$$I(\Gamma; w - f(a)) = \frac{1}{2\pi i} \int_0^1 \frac{g'(\gamma(t))\gamma'(t)}{g(\gamma(t)) - (w - f(a))}\,\mathrm{d}t = \frac{1}{2\pi i} \int_{\partial D(a,r)} \frac{f'(z)}{f(z) - w}\,\mathrm{d}z$$

By the argument principle, $I(\Gamma; 0) = d$, since $I(\Gamma; 0)$ is the number of zeroes of g in $D(a, r)$ counted with multiplicity; the zero of g at $z = a$ has order d, and it is the only zero in $D(a, r)$. Hence,

$$\frac{1}{2\pi i} \int_{\partial D(a,r)} \frac{f'(z)}{f(z) - w}\,\mathrm{d}z = d$$

Again, the argument principle shows that the number of zeroes of $f(z) - w$ in $D(a, r)$ is d, counted with multiplicity. None of these zeroes is equal to a since $w \neq f(a)$. Since $f'(z) = g'(z) \neq 0$ in $D(a, r) \setminus \{a\}$, it follows from the Taylor series that these zeroes are simple. Thus $f(z) - w$ has d distinct zeroes in $D(a, r) \setminus \{a\}$. □

5.3. Open mapping theorem

Corollary. A non-constant holomorphic function maps open sets to open sets. That is, non-constant holomorphic functions are open maps.

Proof. Let $f : U \to \mathbb{C}$ be holomorphic and non-constant, and let $V \subseteq U$ be an open set. Let $b \in f(V)$. Then $b = f(a)$ for some $a \in V$. Since V is open, there exists $r > 0$ such that $D(a, r) \subseteq V$. By the local degree theorem, if r is sufficiently small, there exists $\varepsilon > 0$ such that $w \in D(f(a), \varepsilon) \setminus \{f(a)\} \implies w = f(z)$ for some $z \in D(a, r) \setminus \{a\}$, hence $D(f(a), \varepsilon) \setminus \{f(a)\} \subseteq f(D(a, r) \setminus \{a\})$. Hence $D(b, \varepsilon) = D(f(a), \varepsilon) \subseteq f(D(a, r)) \subseteq f(V)$. Thus, for all $b \in f(V)$, there exists a disc $D(b, \varepsilon) \subseteq f(V)$, so $f(V)$ is open. □

5.4. Rouché's theorem

Theorem. Let γ be a closed curve bounding a domain Ω, and let f, g be holomorphic functions on an open set U containing $\Omega \cup \operatorname{Im}\gamma$. If $|f(z) - g(z)| < |g(z)|$ for all $z \in \operatorname{Im}\gamma$, then f and g have the same number of zeroes in Ω, counted with multiplicity.

Proof. The strict inequality $|f - g| < |g|$ on $\operatorname{Im}\gamma$ implies that f, g are never zero on $\operatorname{Im}\gamma$ and hence never zero on some open set V containing $\operatorname{Im}\gamma$. So $h = \frac{f}{g}$ is holomorphic and nonzero in V. In particular, g is not identically zero in Ω, and hence the zeroes of g in $\Omega \cup V$ are isolated. Hence h is meromorphic in $\Omega \cup V$, and h has no zeroes or poles on $\operatorname{Im}\gamma$. Also, f, g have finitely many zeroes in Ω.

Now, $|h(z) - 1| < 1$ for all $z \in \operatorname{Im}\gamma$. Hence, the curve $\Gamma = h \circ \gamma$ has image contained within $D(1, 1)$. Since zero is outside this disc, $I(\Gamma; 0) = 0$, and so by the argument principle,

$$\sum_{w \in \mathcal{P}} \operatorname{ord}_h(w) = \sum_{w \in \mathcal{Z}} \operatorname{ord}_h(w)$$

where $\mathcal{P}$ and $\mathcal{Z}$ denote the sets of distinct poles and zeroes of h respectively, and the sums are finite. Now, $\mathcal{P} = \mathcal{P}_1 + \mathcal{P}_2$ and $\mathcal{Z} = \mathcal{Z}_1 \cup \mathcal{Z}_2$, where

$$\begin{aligned}
\mathcal{P}_1 &= \{w \in \Omega : g(w) = 0; f(w) \neq 0\};\\
\mathcal{P}_2 &= \{w \in \Omega : g(w) = f(w) = 0; \operatorname{ord}_g(w) > \operatorname{ord}_f(w)\};\\
\mathcal{Z}_1 &= \{w \in \Omega : f(w) = 0; g(w) \neq 0\};\\
\mathcal{Z}_2 &= \{w \in \Omega : f(w) = g(w) = 0; \operatorname{ord}_f(w) > \operatorname{ord}_g(w)\}
\end{aligned}$$

Hence,

$$\sum_{w\in\mathcal{P}_1} \operatorname{ord}_g(w) + \sum_{w\in\mathcal{P}_2} \big(\operatorname{ord}_g(w) - \operatorname{ord}_f(w)\big) = \sum_{w\in\mathcal{Z}_1} \operatorname{ord}_f(w) + \sum_{w\in\mathcal{Z}_2} \big(\operatorname{ord}_f(w) - \operatorname{ord}_g(w)\big)$$

Equivalently,

$$\sum_{w\in\mathcal{P}_1} \operatorname{ord}_g(w) + \sum_{w\in\mathcal{P}_2} \operatorname{ord}_g(w) + \sum_{w\in\mathcal{Z}_2} \operatorname{ord}_g(w) = \sum_{w\in\mathcal{Z}_1} \operatorname{ord}_f(w) + \sum_{w\in\mathcal{Z}_2} \operatorname{ord}_f(w) + \sum_{w\in\mathcal{P}_2} \operatorname{ord}_f(w)$$

Adding $\sum_{w\in\mathcal{R}} \operatorname{ord}_g(w)$ to the left hand side and the equal number $\sum_{w\in\mathcal{R}} \operatorname{ord}_f(w)$ to the right hand side, where

$$\mathcal{R} = \big\{w \in \Omega : \ f(w) = g(w) = 0; \operatorname{ord}_f(w) = \operatorname{ord}_g(w)\big\}$$

we have

$$\sum_{w\in\Omega\,:\ g(w)=0} \operatorname{ord}_g(w) = \sum_{w\in\Omega\,:\ f(w)=0} \operatorname{ord}_f(w)$$

as required. □

Example. $z^4 + 6z + 3$ has three roots counted with multiplicity in $\{1 < |z| < 2\}$. Let $f(z) = z^4 + 6z + 3$.

On $|z| = 2$ we have $|z^4| = 16$ and $|6z + 3| \le 6|z| + 3 = 15$, so $|z|^4 > |6z + 3|$. By Rouché's theorem, f has the same number of roots inside $\{|z| < 2\}$ as z^4, counting with multiplicity. Thus, all roots of $z^4 + 6z + 3$ lie inside $\{|z| < 2\}$; this is all of the roots since f is a polynomial with degree 4.

On $|z| = 1$, we have $|6z| = 6$ and $|z^4 + 3| \le |z|^4 + 3 \le 4$. Again by Rouché's theorem, f has one root inside $\{|z| < 1\}$, as $6z$ has one root in this region. From the strict inequalities, no roots lie on $\{|z| = 2\}$ or $\{|z| = 1\}$. Hence three roots of f lie in $|z \in \mathbb{C} : \ 1 < |z| < 2|$.

X. Geometry

Lectured in Lent 2022 by Prof. I. Smith

This course serves as an introduction to the modern study of surfaces in geometry. A surface is a topological space that locally looks like the plane. The notions of length and area on a surface are governed by mathematical objects called the fundamental forms of the surface at particular points. We can use integrals to work out exact lengths and areas. We study various spaces, including spaces of constant curvature, such as the plane, spheres, and hyperbolic space.

Contents

1. Surfaces 546
1.1. Basic definitions 546
1.2. Subdivisions 551
1.3. Euler classification 552
2. Smooth surfaces 554
2.1. Charts and atlases 554
3. Smooth surfaces in $\mathbb{R}^3$ 556
3.1. Definitions and equivalent characterisations 556
3.2. Inverse and implicit function theorems 557
3.3. Conditions for smoothness 558
3.4. Orientability 559
3.5. Tangent planes 561
4. Geometry of surfaces in $\mathbb{R}^3$ 564
4.1. First fundamental form 564
4.2. Conformality 567
4.3. Area 568
4.4. Second fundamental form 569
4.5. Gauss maps 571
4.6. Gauss curvature 573
4.7. Elliptic, hyperbolic, and parabolic points 574
5. Geodesics 578
5.1. Definitions 578
5.2. The geodesic equations 578
5.3. Geodesics on the plane 580
5.4. Geodesics on the sphere 580
5.5. Geodesics on the torus 581
5.6. Equivalent characterisation of geodesics 581
5.7. Planes of symmetry 583
5.8. Surfaces of revolution 583
5.9. Local existence of geodesics 585
5.10. Surfaces of constant curvature 586
6. Riemannian metrics 588
6.1. Definitions 588
6.2. The length metric 589
6.3. The hyperbolic metric 591
6.4. The hyperbolic upper half-plane 593
6.5. Isometries of hyperbolic space 594

6.6. Hyperbolic triangles 596
6.7. Area of triangles 597
6.8. Surfaces of constant negative curvature 600
6.9. Gauss–Bonnet theorem 602
6.10. Green's theorem (non-examinable) 604
6.11. Alternate flat toruses 605
6.12. Further courses 606

1. Surfaces

1.1. Basic definitions

Definition. A *topological surface* is a topological space Σ such that

(i) for all points $p \in \Sigma$, there exists an open neighbourhood $p \in U \subset \Sigma$ such that U is homeomorphic to $\mathbb{R}^2$, or a disc $D^2 \subset \mathbb{R}^2$, with its usual Euclidean topology;

(ii) Σ is Hausdorff and second countable.

Remark. $\mathbb{R}^2$ is homeomorphic to the open disc $D(0,1) = \{x \in \mathbb{R}^2 : \|x\| < 1\}$. Recall that a space X is Hausdorff if two points $p \neq q \in X$ have open neighbourhoods U, V such that $U \cap V = \varnothing$. A space X is *second countable* if it has a countable base; there exists a countable family of open sets U_i, such that every open set is a union of some of the U_i.

Note that subspaces of Hausdorff and second countable spaces are also Hausdorff and second countable. In particular, Euclidean space $\mathbb{R}^n$ is Hausdorff (as $\mathbb{R}^n$ is a metric space) and second countable (consider the set of balls $D(p,q)$ for points p with rational coordinates, and rational radii q). Hence, any subspace of $\mathbb{R}^n$ is implicitly Hausdorff and second countable. These topological requirements are typically not the purpose of considering topological spaces, but they are occasionally technical requirements to prove interesting theorems.

Example. $\mathbb{R}^2$ is a topological surface. Any open subset of $\mathbb{R}^2$ is also a topological surface. For example, $\mathbb{R}^2 \setminus \{0\}$ and $\mathbb{R}^2 \setminus \{(0,0)\} \cup \left\{\left(0, \frac{1}{n}\right) : n = 1, 2, \dots\right\}$ are topological surfaces.

Example. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. The graph of f, denoted Γ_f, is defined by

$$\Gamma_f = \{(x, y, f(x,y)) : (x,y) \in \mathbb{R}^2\}$$

with the subspace topology when embedded in $\mathbb{R}^3$. Recall that a product topology $X \times Y$ has the feature that $f : Z \to X \times Y$ is continuous if and only if $\pi_x \circ f : Z \to X$ and $\pi_y \circ f : Z \to Y$ are continuous. Hence, any graph $\Gamma \subseteq X \times Y$ is homeomorphic to X if f is continuous. Indeed, the projection π_x projects each point in the graph onto the domain. The function $s : x \mapsto (x, f(x))$ is continuous by the above. In particular, in our case, the graph Γ_f is homeomorphic to $\mathbb{R}^2$, which we know is a surface.

Remark. As a topological surface, Γ_f is independent of the function f. However, we will later introduce more ways to describe topological spaces that will ascribe new properties to Γ_f which do depend on f.

Example. The sphere:

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$$

is a topological surface, when using the subspace topology in $\mathbb{R}^3$. Consider the stereographic projection of S^2 onto $\mathbb{R}^2$ from the north pole $(0,0,1)$. The projection satisfies $\pi_+ : S^2 \setminus \{(0,0,1)\}$ and

$$(x, y, z) \mapsto \left(\frac{x}{1-z}, \frac{y}{1-z}\right)$$

Certainly, π_+ is continuous, since we do not consider the point $(0,0,1)$ to be in its domain. The inverse map is given by

$$(u,v) \mapsto \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1} \right)$$

This is also a continuous function. Hence π_+ is a homeomorphism. Similarly, we can construct the stereographic projection from the south pole, π_-. This is a homeomorphism. Hence, every point in S^2 lies either in the domain of π_+ or π_-, and hence sits in an open set $S^2 \setminus \{(0,0,1)\}$ or $S^2 \setminus \{(0,0,-1)\}$ which is homeomorphic to $\mathbb{R}^2$.

Remark. S^2 is compact by the Heine–Borel theorem; it is a closed bounded set in $\mathbb{R}^3$.

Example. The real projective plane is a topological surface. The group $\mathbb{Z}/2$ acts on S^2 by homeomorphisms via the *antipodal map* $a\colon S^2 \to S^2$, mapping $x \mapsto -x$. There exists a homeomorphism $\mathbb{Z}/2$ to the group $\mathrm{Homeo}(\mathbb{S}^2)$ of homeomorphisms of S^2, by mapping $1 + \mathbb{Z} \mapsto a$. We now define the real projective plane to be the quotient of S^2 given by identifying every point x with its image $-x$ under a.

$$\mathbb{RP}^2 = {}^{S^2}\!/_{\mathbb{Z}/2} = {}^{S^2}\!/_{\sim}; \quad x \sim a(x)$$

Lemma. $\mathbb{RP}^2$ bijects with the set of straight lines in $\mathbb{R}^3$ through the origin.

Proof. Any line through the origin intersects S^2 exactly in a pair of antipodal points $x, -x$. Similarly, pairs of antipodal points uniquely define a line through the origin. □

Lemma. $\mathbb{RP}^2$ is a topological surface.

Proof. We must check that $\mathbb{RP}^2$ is Hausdorff since it is constructed by a quotient, not a subspace. If X is a space and $q\colon X \to Y$ is a quotient map, then by definition $V \subset Y$ is open if and only if $q^{-1}(V) \subset X$ is open. If $[p] \neq [q] \in \mathbb{RP}^2$, then $\pm p, \pm q \in S^2$ are distinct antipodal pairs. We can therefore construct distinct open discs around p, q in S^2, and their antipodal images. These uniquely define open neighbourhoods of $[p], [q]$, which are disjoint.

Similarly, we can check that $\mathbb{RP}^2$ is second countable. We know that S^2 is second countable, so let $\mathcal{U}$ be a countable base for the topology on S^2. Without loss of generality, we can assert that for all sets $U \in \mathcal{U}$, we have $-U \in \mathcal{U}$. Let $\overline{\mathcal{U}}$ be the family of open sets in $\mathbb{RP}^2$ of the form $q(U) \cup q(-U)$ for $U \in \mathcal{U}$, where q is the quotient map. Now, if $V \subseteq \mathbb{RP}^2$ is open, then by definition $q^{-1}(V)$ is open in S^2 hence $q^{-1}(V)$ contains some $U \in \mathcal{U}$ and hence contains $U \cup (-U)$. Hence $\overline{\mathcal{U}}$ is a countable base for the quotient topology on $\mathbb{RP}^2$.

Finally, let $p \in S^2$ and $[p] \in \mathbb{RP}^2$ its image. Let $\overline{D}$ be a small (contained in an open hemisphere) closed disc, which is a neighbourhood of $p \in S^2$. The quotient map restricted to $\overline{D}$, written $q|_{\overline{D}} : \overline{D} \to q(\overline{D}) \subset \mathbb{RP}^2$, is a continuous function from a compact space to a Hausdorff space. Further, q is injective on $\overline{D}$ since the disc was contained entirely in a single hemisphere. The topological inverse function theorem states that a continuous bijection

from a compact space to a Hausdorff space is a homeomorphism. So $q|_{\overline{D}}$ is a homeomorphism from $\overline{D}$ to $q(\overline{D})$. This then induces the homeomorphism $q|_D$ from the open disc $D = \overline{D}^\circ$ to $q(D)$. So by construction, $[p] \in q(D)$; it has an open neighbourhood in $\mathbb{RP}^2$ which is homeomorphic to an open disc, concluding the proof. □

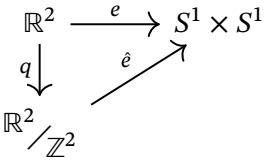
Example. Let S^1 be the unit circle in $\mathbb{C}$, and then we define the torus to be the product space $S^1 \times S^1$, with the subspace topology from $\mathbb{C}^2$ (which is identical to the product topology).

Lemma. The torus is a topological surface.

Proof. Consider the map $e : \mathbb{R}^2 \to S^1 \times S^1$ defined by

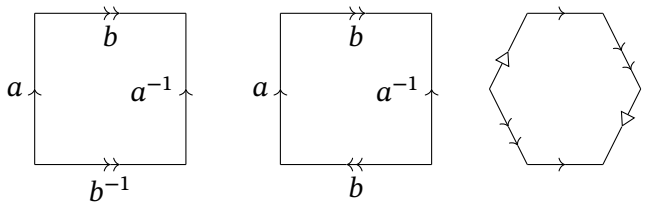
$$(s,t) \mapsto \left(e^{2\pi i s}, e^{2\pi i t}\right)$$

Note that this induces a map $\hat{e}$ from $\mathbb{R}^2/\mathbb{Z}^2$, since e is constant under translations by $\mathbb{Z}^2$.



Under the quotient topology given by the quotient map q, $\mathbb{R}^2/\mathbb{Z}^2$ is a topological space. The map $[0,1]^2 \to \mathbb{R}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ is surjective, so $\mathbb{R}^2/\mathbb{Z}^2$ is compact. So $\hat{e}$ is a continuous map from a compact space to a Hausdorff space, and $\hat{e}$ is bijective, so $\hat{e}$ is a homeomorphism. We already have that $S^1 \times S^1$ is compact and Hausdorff (as a closed and bounded set in $\mathbb{C}^2$), so it suffices to show it is locally homeomorphic to $\mathbb{R}^2$. Let $[p] = q(p) \in S^1 \times S^1$, then we can choose a small disc $\overline{D}(p)$ such that $\overline{D}(p) \cap \left(\overline{D}(p) + (n,m)\right) = \varnothing$ for nonzero $(n,m) \in \mathbb{Z}^2$. Hence $e|_{\overline{D}(p)}$ is injective and $q|_{\overline{D}(p)}$ is injective. Now, restricting to the open disc as before, we can find an open disc neighbourhood of $[p]$. Since $[p]$ was chosen arbitrarily, $S^1 \times S^1$ is a topological surface. □

Example. Let P be a planar Euclidean polygon, with oriented edges. We will pair the edges, and without loss of generality we will assume that paired edges have the same Euclidean length.



We can assign letter names to each edge pair, and denote a polygon by the sequence of edges found when traversing in a clockwise orientation. The edge pair name is inverted if the edge is traversed in the reverse direction. Note the difference between the annotations on the first two shapes above, due to the reversed direction of the edge. If two edges $e, \hat{e}$ are paired, this defines a unique Euclidean isometry from e to $\hat{e}$ respecting the orientation, which will be written $f_{e\hat{e}} : e \to \hat{e}$. The set of all such functions generate an equivalence relation on the polygon, identifying paired edges with each other.

Lemma. $P/\sim$, with the quotient topology, is a topological surface.

Example. Consider the torus, defined here as $T^2 = [0,1]^2/\sim$. Let P be the polygon $[0,1]^2$. If p is in the interior of P, then construct a sufficiently small disc that lies entirely within the interior. The quotient map is injective on the closure of the disc and is a homeomorphism on its interior.

Let p be on an edge, but not a vertex. Let us say without loss of generality that $p = (0, y_0) \sim (1, y_0)$. Let δ be sufficiently small that the closed half-discs U, V centred on p with radius δ do not intersect any vertices. Then we define a map from the union of the two half-discs to the disc $B(0,\delta) \subseteq \mathbb{R}^2$ via $(x, y) \mapsto (x, y - y_0)$ or $(x, y) \mapsto (x - 1, y - y_0)$, which will be a bijective map. Recall the gluing lemma from Analysis and Topology: that if $X = A \cup B$ is a union of closed subspaces, and $f : A \to Y$, $g : B \to Y$ are continuous and $f|_{A\cap B} = g|_{A\cap B}$, they define a continuous map on X. Let f_U, f_V be the maps on the half-discs U, V. By the definition of the quotient topology, $q \circ f_U$ and $q \circ f_V$ are also continuous. On the overlapping area, the functions $q \circ f_U$ and $q \circ f_V$ agree. Hence, by the gluing lemma, we can construct a function $f : U \cup V \to B(0,\delta)$. We can show that this is a homeomorphism using the usual process: pass to the closed disc, apply the topological inverse function theorem, then apply the result to the interior. If $[p] \in T^2$ lies on the image of an edge in $[0,1]^2$, it has indeed a neighbourhood homeomorphic to a disc.

Now it suffices to consider points p on a vertex. All four vertices of the square are identified to the same point in the torus. A neighbourhood of each vertex can be identified with a quarter-disc in $\mathbb{R}^2$. We can repeatedly apply the gluing lemma to construct the whole disc $B(0,\delta) \subseteq \mathbb{R}^2$ and complete the argument as before.

Thus, $[0,1]^2/\sim$ is a topological surface.

We can generalise this proof to an arbitrary planar Euclidean polygon P, such as the hexagon above. The equivalence relation $x \sim f_{e\hat{e}}(x)$ induces an equivalence relation on the vertices of P, by considering the images of the vertices under all $f_{e\hat{e}}$. However, it is not necessarily the case that an equivalence class of vertices contains exactly four vertices, so quarter-discs are not necessarily applicable. Again, there are three types of point:

- interior points, for which a neighbourhood not intersecting the boundary is chosen;
- points on edges, for which a corresponding point exists and two half-discs can be glued to form the neighbourhood; and

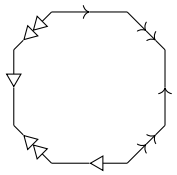
- points on vertices. For this case, all vertices of the polygon have a neighbourhood which is a sector of a circle. Let there be r vertices in a given equivalence class. Let α be the sum of the angles of the sectors in a given class. Any sector can be identified with a given sector in the disc $B(0,\delta) \subseteq \mathbb{R}^2$, which we will choose to have angle α/r. Then, we can glue each sector together in $\mathbb{R}^2$, compatibly with the orientations of the edges and arrows, inducing a neighbourhood which is locally homeomorphic to a disc. If $r = 1$, we have an equivalence class comprising a single vertex, which gives a single sector. For r to be one, the two edges attached to this vertex must be paired and have the same direction (either both inwards or outwards from the vertex). This quotient space is simply a cone, which is homeomorphic to a disc as required.

We can also show that the quotient space is Hausdorff and second countable. By construction, two distinct points in the quotient space can be separated by open neighbourhoods by selecting a sufficiently small radius such that the discs considered in the derivation above are disjoint. For second countability, consider

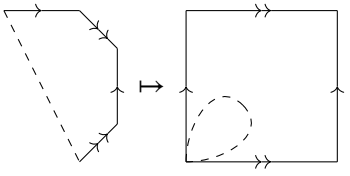
- discs in the interior of P with rational centres and radii;
- for each edge of P, consider an isometry $e \to [0, \ell]$ where ℓ is the length of e, taking discs on e which are centred at rational values in $[0, \ell]$; and
- for each vertex, consider discs centred at these vertices with rational radii.

Example. Given topological surfaces Σ_1, Σ_2 we can remove an open disc from each and glue the resulting circles. Explicitly, we form a quotient relation on the disjoint union of the surfaces with the discs removed. This process is known as forming the *connect sum* of the surfaces, written $\Sigma_1 \# \Sigma_2$. Typically, the information about where the discs were removed from is discarded when considering the connect sum. The connect sum of two topological surfaces is a topological surface.

Example. Consider the following octagon.

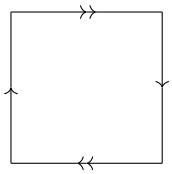


The associated quotient space $P/\sim$ can be seen to be homeomorphic to a surface with two holes, known as a double torus. All vertices are identified as the same vertex in the quotient space. We can cut the octagon along a diagonal, leaving two topological surfaces which are homeomorphic to a torus.



Thus, the connect sum of the two half-octagons are the connect sum of two toruses.

Example. Consider the following square.



This is homeomorphic to the real projective plane $\mathbb{RP}^2$. This is because we identify points on the boundary with their antipodes, when interpreting the square as the closed disc $B(0,1)$. The real projective plane was constructed by identifying points on the unit sphere with their antipodes. Thus, we can construct a homeomorphism by considering only points in the upper hemisphere (taking antipodes as required), and then orthographically projecting onto the xy plane. Under this transformation, points on the boundary are identified with their antipodes as required.

1.2. Subdivisions

Definition. A *subdivision* of a compact topological surface Σ comprises

(i) a finite subset $V \subseteq \Sigma$ of vertices;

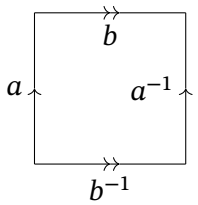
(ii) a finite subset $E = \{e_i : [0,1] \to \Sigma\}$ which are continuous injections and pairwise disjoint except perhaps at the endpoints;

(iii) such that each connected component of the complement of $V \cup E$ in Σ is homeomorphic to an open disc, and each such component will be called a face. In particular, the boundary of each face has boundary inside the union of the edges and the vertices.

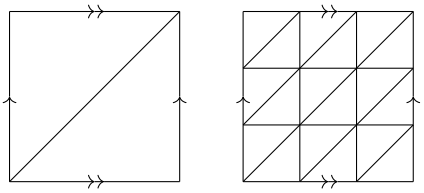
We say that a subdivision is a *triangulation* if each closed face (closure of a face) contains exactly three edges, and two closed faces meet either at exactly one edge or not at any edges.

Example. A cube displays a subdivision of S^2. A tetrahedron displays a triangulation of S^2.

Example. We can display subdivisions of surfaces constructed from polygons.

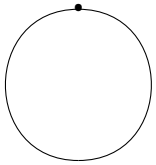


This is a subdivision of a torus with one edge, two edges, and one face. We can construct additional subdivisions of a torus, for example:



The first of these examples is not a triangulation, since the two faces meet in more than one edge. The second is a triangulation.

Remark. The following is a very degenerate subdivision of S^2.



This has one vertex, no edges, and one face.

1.3. Euler classification

Definition. The *Euler characteristic* of a subdivision is

$$\#V - \#E + \#F$$

Theorem. (i) Every compact topological surface has a subdivision (and indeed triangulations).

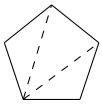
(ii) The Euler characteristic is invariant under choice of subdivision, and is topologically invariant.

Hence, we might say that a surface has a particular Euler characteristic, without referring to subdivisions. We write this $\chi(\Sigma)$.

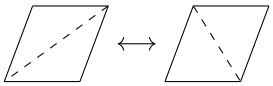
No proof will be given.

Example. The Euler characteristic of S^2 is $\chi(S^2) = 2$. For the torus, $\chi(T^2) = 0$. If Σ_1, Σ_2 are compact surfaces, then the connect sum $\Sigma_1 \# \Sigma_2$ can be constructed by removing a face of a triangulation, then gluing together the boundary circles (three edges) in a way that matches the edges. Then the connect sum inherits a subdivision, and we can find that it has Euler characteristic $\chi(\Sigma_1 \# \Sigma_2) = \chi(\Sigma_1) + \chi(\Sigma_2) - 2$, where the remaining term corresponds to the two faces that were removed; the changes of three vertices and three edges cancel each other. In particular, a surface Σ_g with g holes can be written $\#_{i=1}^{g} T^2$, so $\chi(\Sigma_g) = 2 - 2g$. We call g the *genus* of Σ.

Remark. It is not trivial to prove part (i). For part (ii), note that subdivisions can be converted into triangulations by constructing triangle fans.



Triangulations can be related by local moves, such as



It is easy to check that both of these moves do not change the Euler characteristic. However, it is hard to make this argument rigorous, and it does not give much explanation for why the result is true. In Part II Algebraic Topology, a more advanced definition of the Euler characteristic is given, which admits a more elegant proof.

2. Smooth surfaces

2.1. Charts and atlases

Recall that if Σ is a topological surface, any point lies in an open neighbourhood homeomorphic to a disc.

Definition. A pair (U, φ), where U is an open set in Σ and $\varphi : U \to V$ is a homeomorphism to an open set $V \subseteq \mathbb{R}^2$, is called a *chart* for Σ. If $p \in U$, we might say that (U, φ) is a chart for Σ *at* p. A collection of charts whose domains cover Σ is known as an *atlas* for Σ. The inverse $\sigma = \varphi^{-1} : V \to U$ is known as a *local parametrisation* for the surface.

Example. If $Z \subseteq \mathbb{R}^2$ is closed, $\mathbb{R}^2 \setminus Z$ is a topological surface with an atlas containing one chart, $(\mathbb{R}^2 \setminus Z, \phi = \mathrm{id})$.

For S^2, there is an atlas with two charts, which are the two stereographic projections from the poles. We could consider alternative charts, for instance the projection to the yz plane, but this would be insufficient for describing the poles.

Definition. Let (U_i, φ_i) be charts containing the point $p \in \Sigma$, for $i = 1, 2$. Then the map

$$* : \varphi_1(U_1 \cap U_2) \to \varphi_2(U_1 \cap U_2); \quad * = \varphi_2 \circ \varphi_1^{-1}\Big|_{\varphi_1(U_1 \cap U_2)}$$

converts between the corresponding charts, and is called a *transition map*. This is a homeomorphism of open sets in $\mathbb{R}^2$.

Recall from Analysis and Topology that if $V \subseteq \mathbb{R}^n$ and $V' \subseteq \mathbb{R}^m$ are open, then a continuous map $f : V \to V'$ is called *smooth* if it is infinitely differentiable. Equivalently, it is smooth if partial derivatives of all orders in all variables exist at all points. If $n = m$, then in particular the homeomorphism $f : V \to V'$ is called a *diffeomorphism* if it is smooth and has smooth inverse.

Definition. An *abstract smooth surface* is a topological space Σ together with an atlas of charts (U_i, φ_i) such that all transition maps $\varphi_i \circ \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$ are diffeomorphisms.

Remark. We could not simply consider a smoothness condition for Σ itself without appealing to atlases, since Σ is an arbitrary topological space and could have almost any topology.

Example. The atlas of two charts with stereographic projections gives S^2 the structure of an abstract smooth surface.

Example. For the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$, we can find charts of all points by choosing sufficiently small discs in $\mathbb{R}^2$ such that they do not intersect any of their non-trivial integer translates. The transition maps for this atlas are all translations of $\mathbb{R}^2$. Hence T^2 inherits the structure of an abstract smooth surface. Explicitly, let us define $e : \mathbb{R}^2 \to T^2$ by $(t, s) \mapsto \left(e^{2\pi i t}, e^{2\pi i s}\right)$, then consider the atlas

$$\left\{(e(D_\varepsilon(x, y)), e^{-1} \text{ on this image})\right\}$$

for $\varepsilon < \frac{1}{3}$. These are charts on T^2, and the transition maps are (restricted to appropriate domains) translations in $\mathbb{R}^2$. Hence T^2, via this atlas, has the structure of an abstract smooth surface.

Remark. The definition of a topological surface is a notion of structure. One can observe a topological space and determine whether it is a topological surface. Conversely, to be an abstract smooth surface is to have a specific set of data; that is, we must provide charts for the surface in order to see that it is indeed an abstract smooth surface.

Definition. Let Σ be an abstract smooth surface, and $f : \Sigma \to \mathbb{R}^n$ be a continuous map. We say that f is *smooth* at $p \in \Sigma$ if, for all charts (U, φ) of p belonging to the smooth atlas for Σ, the map

$$f \circ \varphi^{-1} : \varphi(U) \to \mathbb{R}^n$$

is smooth at $\varphi(p) \in \mathbb{R}^2$.

Remark. Note that the choice of chart and atlas was arbitrary, but smoothness of f at p is independent of the choice of chart, since the transition maps between two such charts are diffeomorphisms.

Definition. Let Σ_1, Σ_2 be abstract smooth surfaces. Then a map $f : \Sigma_1 \to \Sigma_2$ is *smooth* if it is 'smooth in the local charts'. Given a chart (U, φ) at p and a chart (U', ψ) at $f(p)$, both mapping to open subsets of $\mathbb{R}^2$, the map $\psi \circ f \circ \varphi^{-1}$ is smooth at $\varphi(p)$. Smoothness of f does not depend on the choice of chart, provided that the charts all belong to the same atlas.

Definition. Two surfaces Σ_1, Σ_2 are *diffeomorphic* if there exists a homeomorphism $f : \Sigma_1 \to \Sigma_2$ which is smooth and has smooth inverse.

Remark. Often, we convert from a given smooth atlas for an abstract smooth surface Σ to the *maximal compatible* smooth atlas. That is, we consider the atlas with the maximal possible set of charts, all of which have transition maps that are diffeomorphisms. This can be accomplished formally by use of Zorn's lemma.

3. Smooth surfaces in $\mathbb{R}^3$

3.1. Definitions and equivalent characterisations

Recall that if $V \subseteq \mathbb{R}^n$ and $V' \subseteq \mathbb{R}^m$, then $f : V \to V'$ is smooth if it is infinitely differentiable.

Definition. If Z is an arbitrary subset of $\mathbb{R}^n$, we say that a continuous function $f : Z \to \mathbb{R}^m$ is smooth at $p \in Z$ if there exists an open ball $p \in B \subseteq \mathbb{R}^n$ and a smooth map $F : B \to \mathbb{R}^m$ which extends f such that they agree on $B \cap Z$. In other words, f is locally the restriction of a smooth map defined on an open set.

Definition. Let $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$. We say that X and Y are *diffeomorphic* if there exists a continuous function $f : X \to Y$ such that f is a smooth homeomorphism with smooth inverse.

Definition. A *smooth surface in* $\mathbb{R}^3$ is a subspace of $\mathbb{R}^3$ such that for all points $p \in \Sigma$, there exists an open subset $p \in U \subseteq \Sigma$ that is diffeomorphic to an open set in $\mathbb{R}^2$. In other words, for all $p \in \Sigma$, there exists an open ball $p \in B \subseteq \mathbb{R}^3$ such that if $U = B \cap \Sigma$ and there exists a map $f : B \to V \subseteq \mathbb{R}^2$ such that $f|_U : U \to V$ is a homeomorphism, and the inverse map $V \to U \subseteq \Sigma \subseteq \mathbb{R}^3$ is smooth.

Definition. Let $\sigma : V \to U$ where $V \subseteq \mathbb{R}^2$ is open and $U \subseteq \Sigma \subseteq \mathbb{R}^3$ is open in Σ, such that σ is a smooth homeomorphism and $D\,\sigma|_x$ has rank 2 for all $x \in V$. Then σ is called an *allowable parametrisation*. If $\sigma(0) = p$, we say that σ is an allowable parametrisation *near* p.

Theorem. For a subset $\Sigma \subseteq \mathbb{R}^3$, the following are equivalent.

(a) Σ is a smooth surface in $\mathbb{R}^3$;

(b) Σ is locally the graph of a smooth function, over one of the three coordinate planes: for all $p \in \Sigma$ there exists an open ball $p \in B \subseteq \mathbb{R}^3$ and an open set $V \subseteq \mathbb{R}^2$ such that

$$\Sigma \cap B = \{(x, y, g(x, y)) : g : V \to \mathbb{R} \text{ smooth}\}$$

or one of the other coordinate planes;

(c) Σ is locally cut out by a smooth function: for all $p \in \Sigma$ there exists an open ball $p \in B \subseteq \mathbb{R}^3$ and a smooth function $f : B \to \mathbb{R}$ such that

$$\Sigma \cap B = f^{-1}(0); \quad D\,f\Big|_x \neq 0$$

for all $x \in B$;

(d) Σ is locally the image of an allowable parametrisation near all points.

Remark. Part (b) implies that if Σ is a smooth surface in $\mathbb{R}^3$, each $p \in \Sigma$ belongs to a chart (U, φ) where φ is (the restriction of) one of the three coordinate plane projections

$\pi_{xy}, \pi_{yz}, \pi_{xz}$ from $\mathbb{R}^3$ to $\mathbb{R}^2$. Consider the transition map between two such charts. If the two charts are based on the same projection such as π_{xy}, then the transition map is the identity. If they are based on different projections π_{xy} and π_{yz}, then the transition map is

$$(x, y) \mapsto (x, y, g(x, y)) \mapsto (y, g(x, y))$$

which has inverse

$$(y, z) \mapsto (h(y, z), y, z) \mapsto (h(y, z), y)$$

Hence all of the transition maps between such charts involve projection maps and the smooth maps involved in defining Σ as a graph. This gives Σ the structure of an abstract smooth surface.

Some of the relations given in the above theorem are easy to prove, but others come as a result of the inverse function theorem.

3.2. Inverse and implicit function theorems

Theorem (inverse function theorem)**.** Let $U \subseteq \mathbb{R}^n$ be open, and $f : U \to \mathbb{R}^n$ be continuously differentiable. Let $p \in U$ and $f(p) = q$. Suppose $D\,f|_p$ is invertible. Then there is an open neighbourhood V of q and a differentiable map $g : V \to \mathbb{R}^n$ and $g(q) = p$ with image an open neighbourhood $U' \subseteq U$ of p such that $f \circ g = \mathrm{id}_V$. If f is smooth, then g is also.

Remark. The chain rule then implies that $D\,g|_q = \left(D\,f|_p\right)^{-1}$. The inverse function theorem concerns functions $\mathbb{R}^n \to \mathbb{R}^n$, where $D\,f|_p$ is an isomorphism. If we have a map $\mathbb{R}^n \to \mathbb{R}^m$ for $n > m$, then we can discuss the behaviour when $D\,f|_p$ is surjective. The derivative $D\,f|_p$ is an $n \times m$ matrix, so if it has full rank, up to the permutation of coordinates we have that the last m columns are linearly independent.

Theorem (implicit function theorem)**.** Let $p = (x_0, y_0)$ be a point in an open set $U \subset \mathbb{R}^k \times \mathbb{R}^\ell$. Let $f : U \to \mathbb{R}^\ell$ such that $p \mapsto 0$ and $\left(\frac{\partial f_i}{\partial y_j}\right)_{\ell \times \ell}$ is an isomorphism. Then there is an open neighbourhood V of x_0 in $\mathbb{R}^k$ and a continuously differentiable map $g : V \to \mathbb{R}^\ell$ with $x_0 \mapsto y_0$ such that if $(x, y) \in U \cap (V \times \mathbb{R}^\ell)$, then $f(x, y) = 0 \iff y = g(x)$. If f is smooth, so is g.

Proof. Let $F : U \to \mathbb{R}^k \times \mathbb{R}^\ell$ be defined by $(x, y) \mapsto (x, f(x, y))$. Then note that

$$DF = \begin{pmatrix} I & * \\ 0 & \frac{\partial f_i}{\partial y_j} \end{pmatrix}$$

hence DF is an isomorphism at (x_0, y_0). By the inverse function theorem, F is locally invertible near $F(x_0, y_0) = (x_0, f(x_0, y_0)) = (x_0, 0)$. Consider an open neighbourhood $V \times V' \subseteq \mathbb{R}^k \times \mathbb{R}^\ell$ on which this continuously differentiable inverse $G : V \times V' \to U' \subseteq U \subseteq \mathbb{R}^k \times \mathbb{R}^\ell$ exists, such that $F \circ G = \mathrm{id}_{V \times V'}$. Then,

$$G(x, y) = (\varphi(x, y), \psi(x, y)) \implies F \circ G(x, y) = (\varphi(x, y), f(\varphi(x, y), \psi(x, y))) = (x, y)$$

Hence $\varphi(x,y) = x$. We have $f(x,\psi(x,y)) = y$ when $(x,y) \in V \times V'$. This gives $f(x,y) = 0 \iff y = \psi(x,0)$. We then define $g : V \to \mathbb{R}^\ell$ by $x \mapsto \psi(x,0)$. □

Example. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be smooth and $f(x_0,y_0) = 0$, and suppose $\frac{\partial f}{\partial y} \neq 0$ at (x_0,y_0). Then there exists a smooth map $g : (x_0-\varepsilon, x_0+\varepsilon) \to \mathbb{R}$ with $g(x_0) = y_0$ and $f(x,y) = 0 \iff y = g(x)$ for (x,y) in some open neighbourhood of (x_0,y_0). Since $f(x,g(x)) = 0$ in this open neighbourhood, we can differentiate that expression to find

$$g'(x) = \frac{-f_x}{f_y}$$

noting that $f_y \neq 0$ in some neighbourhood near (x_0,y_0). Note that the level set $f(x,y) = 0$ is implicitly defined by g, which is a function for which we have an integral expression.

Example. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a smooth map with $f(x_0,y_0,z_0) = 0$. Consider the level set $\Sigma = f^{-1}(0)$, assuming that $Df \neq 0$ at (x_0,y_0,z_0). Permuting coordinates if necessary, we can assume $\frac{\partial f}{\partial z} \neq 0$ at this point. Then there exists an open neighbourhood V of (x_0,y_0) and a smooth function $g : V \to \mathbb{R}$ such that $(x_0,y_0) \mapsto z_0$ with the property that for an open set $(x_0,y_0,z_0) \in U$, the set $f^{-1}(0) \cap U = \Sigma \cap U$ is the graph of the function g, which is $\{(x,y,g(x,y)) : (x,y) \in V\}$.

3.3. Conditions for smoothness

We now prove the theorem stated above, relating equivalent conditions for smoothness of a surface Σ.

Proof. First, we show that (b) implies all of the other conditions. If Σ is locally a graph $\{(x,y,g(x,y))\}$, we find a chart from the coordinate plane projection π_{xy} of that graph. Since this projection is smooth and defined on an open neighbourhood of points of Σ in its domains, this shows that Σ is a smooth surface in $\mathbb{R}^3$ (a). Further, since Σ is locally the given graph, it is cut out by the function $f(x,y,z) = z - g(x,y)$ and $\frac{\partial f}{\partial z} \neq 0$ (c). Finally, the local parametrisation $\sigma(x,y) = (x,y,g(x,y))$ is allowable; g is smooth, the partial derivatives of σ are linearly independent by considering the x and y components, and σ is injective where required (d).

Now, we show (a) implies (d). This is simply part of the definition of being a smooth surface in $\mathbb{R}^3$, being locally diffeomorphic to $\mathbb{R}^2$. In particular, at $p \in \Sigma$, Σ is locally diffeomorphic to $\mathbb{R}^2$ and the inverse of such a local diffeomorphism is an allowable parametrisation.

We have already shown (c) implies (b); this was the example of the implicit function theorem provided above.

Finally, we must prove (d) implies (a) and (b), and then the result will hold. Let $p \in \Sigma$ and V be an open set in $\mathbb{R}^2$ with an allowable parametrisation to Σ such that $\sigma(0) = p$. If

$\sigma = (\sigma_1(u,v), \sigma_2(u,v), \sigma_3(u,v))$, we have

$$D\sigma = \begin{pmatrix} \frac{\partial \sigma_1}{\partial u} & \frac{\partial \sigma_1}{\partial v} \\ \frac{\partial \sigma_2}{\partial u} & \frac{\partial \sigma_2}{\partial v} \\ \frac{\partial \sigma_3}{\partial u} & \frac{\partial \sigma_3}{\partial v} \end{pmatrix}$$

This has rank 2, hence there exist two rows defining an invertible matrix. Suppose those are the first two rows, and let $\mathrm{pr} = \pi_{xy}$ be the projection map. Consider $\mathrm{pr} \circ \sigma : V \to \mathbb{R}^2$. This has isomorphic derivative at zero, so we can apply the inverse function theorem. Hence Σ is locally a graph over the xy coordinate plane, so (b) holds. Moreover, let $\varphi = \mathrm{pr} \circ \sigma$, and consider the open ball $B(p,\delta) \subseteq \mathbb{R}^3$ and a map such that $(x,y,z) \mapsto \varphi^{-1}(x,y)$ in this ball. Here, $\varphi : W \to \Sigma$ where W is an open set in $\mathrm{pr}(B(p,\delta))$. This is a locally defined map, which is smooth on an open set in $\mathbb{R}^3$, which is a smooth inverse to σ. Hence Σ is a smooth surface in $\mathbb{R}^3$, so (a) holds. □

Example. The unit sphere S^2 in $\mathbb{R}^3$ is $f^{-1}(0)$ for $f(x,y,z) = x^2 + y^2 + z^2 - 1$. For any point on S^2, $Df \neq 0$, so S^2 is a smooth surface.

Example. Let $\gamma : [a,b] \to \mathbb{R}^3$ be a smooth map with image in the xz plane, so

$$\gamma(t) = (f(t), 0, g(t))$$

such that γ is injective, $\gamma' \neq 0$, and $f > 0$. The *surface of revolution* of γ has allowable parametrisation

$$\sigma(u,v) = (f(u)\cos v, f(u)\sin v, g(u))$$

where $(u,v) \in (a,b) \times (\theta, \theta + 2\pi)$ for a fixed θ. Note that $\sigma_u = (f_u \cos v, f_u \sin v, g_u)$ and $\sigma_v = (-f \sin v, f \cos v, 0)$, and we can check $\|\sigma_u \times \sigma_v\| = f^2((f')^2 + (g')^2)$ which is nonzero on γ, so this really is an allowable parametrisation.

Example. The orthogonal group $O(3)$ acts on S^2 by diffeomorphisms. Indeed, any $A \in O(3)$ defines a linear (hence smooth) map $\mathbb{R}^3 \to \mathbb{R}^3$ preserving S^2. Hence, the induced map on S^2 is by a homeomorphism which is smooth according to the above definition. This is analogous to the action of the Möbius group on $S^2 = \mathbb{C} \cup \{\infty\}$.

3.4. Orientability

Definition. Let V, V' be open sets in $\mathbb{R}^2$. Let $f : V \to V'$ be a diffeomorphism. Then at every point $x \in V$, $D f|_x \in GL(2,\mathbb{R})$; it is invertible since f is a diffeomorphism. Let $GL^+(2,\mathbb{R})$ be the subgroup of matrices with positive determinant. We say that f is *orientation-preserving* if its derivative belongs to this subgroup for all points $x \in V$.

Definition. An abstract smooth surface Σ is *orientable* if it admits an atlas $\{(U_i, \varphi_i)\}$ where the transition maps are all orientation-preserving. A choice of such an atlas is an *orientation* of Σ; Σ can be called *oriented* when such an orientation is given.

Remark. An orientable atlas belongs to a maximal compatible oriented smooth atlas.

Lemma. If Σ_1 and Σ_2 are diffeomorphic abstract smooth surfaces, then Σ_1 is orientable if and only if Σ_2 is orientable.

Proof. Let $f : \Sigma_1 \to \Sigma_2$ be a diffeomorphism. Suppose Σ_2 is orientable and equipped with an oriented smooth atlas. Consider the atlas on Σ_1 of charts of the form $(f^{-1}(U), \phi \circ f|_{f^{-1}(U)})$, where (U, ψ) is a chart at $f(p)$ in the oriented atlas for Σ_2. Then, the transition map between two such charts is exactly a transition map between charts in the Σ_2 atlas.

In other words, in the maximal smooth atlas that exists a priori for Σ_1, we will allow charts of the form $(\widetilde{U}, \widetilde{\psi})$ when for all charts (U, ψ) at $f(p)$ in the Σ_2 atlas, the map $\psi \circ f \circ (\widetilde{\psi})^{-1}$ is orientation-preserving. Informally, if the atlas on Σ_2 was maximal as an oriented atlas, we can recover the previous set of charts. □

Remark. There is no sensible classification of the set of all smooth surfaces. For instance, $\mathbb{R}^2 \setminus Z$ for a closed set Z can be shown to yield uncountably many types of homeomorphisms. However, *compact* smooth surfaces may be classified by their Euler characteristic and their orientability, up to diffeomorphism. This theorem will not be proven in this course.

There is a definition of orientation-preserving *homeomorphism* that does not rely on the determinant, but that instead relies on some algebraic topology which is not covered in this course. The Möbius band is the surface

where the dashed lines represent the absence of edges. It is provable that an abstract smooth surface is orientable if and only if it contains no subsurface homeomorphic to the Möbius band. We can therefore say that a topological surface is orientable if and only if it contains no subsurface (an open set) homeomorphic to a Möbius band.

We can define other structures on an abstract smooth surface by considering smooth atlases such that if $\varphi_1\varphi_2^{-1}$ is a transition map, then $D(\varphi_1\varphi_2^{-1})$ at x belongs to a specific subgroup $G \leq GL(2, \mathbb{R})$. For example, defining $G = \{e\}$ leads to Euclidean surfaces. The group $GL(1, \mathbb{C})$ identified as a subgroup of $GL(2, \mathbb{R})$ yields the Riemann surfaces.

Example. For S^2 with the atlas of two stereographic projections, we can find the transition map to be

$$(u, v) \mapsto \left(\frac{u}{u^2 + v^2}, \frac{v}{u^2 + v^2}\right)$$

on $\mathbb{R}^2 \setminus \{0\}$. This has positive determinant, so S^2 is orientable.

For the torus T^2, we previously found an atlas such that the transition maps are translations of $\mathbb{R}^2$. Hence the torus is an oriented surface, and also a Euclidean surface.

3.5. Tangent planes

Recall that an *affine* subspace of a vector space is a translate of a linear subspace.

Definition. Let Σ be a smooth surface in $\mathbb{R}^3$, and $p \in \Sigma$. Let $\sigma\colon V \to U \subseteq \Sigma$ be an allowable parametrisation of Σ near p, so V is an open subset of $\mathbb{R}^2$ and U is open in Σ, such that $\sigma(0) = p$. The *tangent plane* $T_p\Sigma$ to p at Σ is the image of $\left(D\,\sigma|_0\right) \subseteq \mathbb{R}^3$, which is a two-dimensional vector subspace of $\mathbb{R}^3$. The *affine tangent plane* is $p + T_p\Sigma$, which is an affine subspace of $\mathbb{R}^3$.

Remark. The affine tangent plane is the 'best' linear approximation to a surface Σ at a given point.

Lemma. $T_p\Sigma$ is independent of the choice of allowable parametrisation.

Proof (i). Suppose $\sigma\colon V \to U$ and $\tilde{\sigma}\colon \widetilde{V} \to \widetilde{U}$ are allowable parametrisations with $\sigma(0) = \tilde{\sigma}(0) = p$. There exists a transition map $\sigma^{-1} \circ \tilde{\sigma}$, which is a diffeomorphism of open sets in $\mathbb{R}^2$. Therefore,

$$\tilde{\sigma} = \sigma \circ \underbrace{(\sigma^{-1} \circ \tilde{\sigma})}_{\text{diffeomorphism}}$$

Hence $D\,(\sigma^{-1} \circ \tilde{\sigma})\big|_0$ is an isomorphism. Thus, the images of $D\,\tilde{\sigma}|_0$ and $D\,\sigma|_0$ agree. □

Proof (ii). Let $\gamma\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^3$ be a smooth map such that γ has image inside Σ, and $\gamma(0) = p$. We will show that $\gamma'(0) \in T_p\Sigma$. If $\sigma\colon V \to U$ is an allowable parametrisation with $\sigma(0) = p$ as above, and ε is sufficiently small such that $\operatorname{Im}\gamma \subseteq U$, then $\gamma(t) = \sigma(u(t), v(t))$ for some smooth functions $u, v\colon (-\varepsilon, \varepsilon) \to V$. Then $\gamma'(t) = \sigma_u u'(t) + \sigma_v v'(t)$ is in the image of $D\,\sigma|_t$. Thus, $T_p\Sigma = \operatorname{span}\{\gamma'(0) : \gamma \text{ as above}\}$. □

Definition. If Σ is a smooth surface in $\mathbb{R}^3$ and $p \in \Sigma$, the *normal direction* to Σ at p is $(T_p\Sigma)^\perp$, the Euclidean orthogonal complement to the tangent plane at p.

Remark. For all $p \in \Sigma$, there exist exactly two normalised normal vectors.

Definition. A smooth surface in $\mathbb{R}^3$ is *two-sided* if it admits a continuous global choice of unit normal vector.

Lemma. A smooth surface in $\mathbb{R}^3$ is orientable (as an abstract smooth surface) if and only if it is two-sided (as a smooth surface in $\mathbb{R}^3$).

Proof. Let $\sigma\colon V \to U \subseteq \Sigma$ be an allowable parametrisation. Let $\sigma(0) = p$. We will define the positive unit normal with respect to σ at p to be the normal vector $n_\sigma(p)$ with the property that $\{\sigma_u, \sigma_v, n_\sigma(p)\}$ and $\{e_1, e_2, e_3\}$ are related by a positive determinant change of basis matrix, where $\{e_1, e_2, e_3\}$ are the standard basis vectors. In other words,

$$n_\sigma(p) = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}$$

Consider an alternative parametrisation $\tilde{\sigma}\colon \widetilde{V} \to \widetilde{U}$, where $\tilde{\sigma}(0) = p$, such that $\tilde{\sigma}$ belongs to the same oriented and smooth atlas as σ. Hence, $\sigma = \tilde{\sigma} \circ \varphi$ for some transition map φ. Let

$$D\,\varphi\Big|_0 = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

Hence,

$$\sigma_u = \alpha\tilde{\sigma}_u + \gamma\tilde{\sigma}_v; \quad \sigma_v = \beta\tilde{\sigma}_u + \delta\tilde{\sigma}_v$$

This gives

$$\sigma_u \times \sigma_v = \det\left(D\,\varphi\Big|_0\right)\tilde{\sigma}_u \times \tilde{\sigma}_v \tag{$\dagger$}$$

The determinant here is positive since the charts in question belong to an oriented atlas. Thus, the positive normal depends on the orientation of Σ, but does not depend on the choice of parametrisation. The expression for $n_\sigma(p)$ is continuous since the cross product is continuous, hence Σ is two-sided.

Conversely, if Σ is two-sided and there exists a global continuous choice of normal vector, we can consider the subatlas of the natural smooth atlas with the property that we allow (U, φ) if the associated parametrisation $\sigma = \varphi^{-1}$ satisfies $\{\sigma_u, \sigma_v, n\}$ is a positive basis for $\mathbb{R}^3$, where n is the given choice of normal. By ($\dagger$), the transition maps between such charts are orientation-preserving. Hence Σ is orientable. □

Lemma. If Σ is a smooth surface in $\mathbb{R}^3$ and $A\colon \mathbb{R}^3 \to \mathbb{R}^3$ is a smooth map which preserves Σ setwise, then $D\,A|_p \in L(\mathbb{R}^3, \mathbb{R}^3)$ maps $T_p\Sigma$ to $T_{A(p)}\Sigma$ for $p \in \Sigma$.

Proof. Let $\gamma\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^3$ be a smooth map such that its image lies on Σ, and $\gamma(0) = p$. Recall that $T_p\Sigma$ is spanned by $\gamma'(0)$ for such curves γ. Now, consider $A \circ \gamma\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^3$, which also has image Σ, and

$$D\,A\Big|_{\gamma(0)} \circ D\,\gamma\Big|_0 = D\,A\Big|_p(\gamma'(0)) = D\,(A \circ \gamma)\Big|_0 \in T_{A(p)}\Sigma$$

□

Example. Let S^2 be the unit sphere. The normal vector at p is the line through the origin and p; indeed, since SO_3 acts transitively on S^2, it suffices to check at one point, such as the north pole. We can choose the outward-facing normal vector to be the positive normal, denoted $n(p)$. S^2 is two-sided by the construction of this normal vector, hence S^2 is orientable.

Example. One embedding of the Möbius band in $\mathbb{R}^3$ is

$$\sigma(t, \theta) = \left(\left(1 - t\sin\frac{\theta}{2}\right)\cos\theta, \left(1 - t\sin\frac{\theta}{2}\right)\sin\theta, t\cos\frac{\theta}{2}\right)$$

where $(t, \theta) \in V_1 = \left\{t \in \left(-\frac{1}{2}, \frac{1}{2}\right), \theta \in (0, 2\pi)\right\}$ or $(t, \theta) \in V_2 = \left\{t \in \left(-\frac{1}{2}, \frac{1}{2}\right), \theta \in (-\pi, \pi)\right\}$. We begin with the unit circle $x^2 + y^2 = 1$, for $t = 0$. Then, at each point on the circle, we

consider an open interval of unit length, which will rotate as we move around the circle, such that at the point θ on the circle it has rotated by $\frac{\theta}{2}$. We can check that if σ_i is σ on V_i, then σ_i is allowable. Further,

$$\sigma_t \times \sigma_\theta = \left(-\cos\theta\cos\frac{\theta}{2}, -\sin\theta\cos\frac{\theta}{2}, -\sin\frac{\theta}{2}\right) \equiv n_\theta$$

which is already normalised. As $\theta \to 0$ from above, $n_\theta \to (-1, 0, 0)$. As $\theta \to 2\pi$ from below, $n_\theta \to (1, 0, 0)$. Hence, the surface is not two-sided.

4. Geometry of surfaces in $\mathbb{R}^3$

4.1. First fundamental form

Let $\gamma\colon (a,b) \to \mathbb{R}^3$ be smooth. The *length* of γ is

$$L(\gamma) = \int_a^b \|\gamma'(t)\| \,\mathrm{d}t$$

This result is independent of the choice of parametrisation. Let $s\colon (A,B) \to (a,b)$ be a monotonically increasing function, and let $\tau(t) = \gamma(s(t))$. Then

$$L(\tau) = \int_A^B \|\tau'(t)\| \,\mathrm{d}t = \int_A^B \|\gamma'(s(t))\||s'(t)| \,\mathrm{d}t = \int_a^b \|\gamma'(t')\| \,\mathrm{d}t' = L(\gamma)$$

Lemma. If $\gamma\colon (a,b) \to \mathbb{R}^3$ is continuously differentiable and $\gamma'(t) \neq 0$, then γ can be parametrised by arc length.

The proof is left as an exercise. Let Σ be a smooth surface in $\mathbb{R}^3$, and let $\sigma\colon V \to U \subseteq \Sigma$ be an allowable parametrisation. If $\gamma\colon (a,b) \to \mathbb{R}^3$ is smooth and its image is contained within U, then there exist functions $(u(t), v(t))\colon (a,b) \to V$ such that $\gamma(t) = \sigma(u(t), v(t))$. Hence $\gamma'(t) = \sigma_u u'(t) + \sigma_v v'(t)$, giving

$$\|\gamma'(t)\|^2 = Eu'(t)^2 + 2Fu'(t)v'(t) + Gv'(t)^2$$

for functions

$$E = \langle \sigma_u, \sigma_u \rangle; \quad F = \langle \sigma_u, \sigma_v \rangle = \langle \sigma_v, \sigma_u \rangle; \quad G = \langle \sigma_v, \sigma_v \rangle$$

where $\langle \cdot\,, \cdot \rangle$ represents the usual Euclidean inner product. Note that E, F, G depend only on σ and not on γ.

Definition. The *first fundamental form* of Σ in the parametrisation σ is the expression

$$E \,\mathrm{d}u^2 + 2F \,\mathrm{d}u \,\mathrm{d}v + G \,\mathrm{d}v^2$$

This notation is illustrative of the fact that if γ has image in the image of $\sigma(v)$, we find

$$L(\gamma) = \int_a^b \sqrt{E(u')^2 + 2Fu'v' + G(v')^2} \,\mathrm{d}t$$

where $\gamma(t) = \sigma(u(t), v(t))$.

Remark. The Euclidean inner product on $\mathbb{R}^3$ provides an inner product on the subspace $T_p\Sigma$. Choosing a parametrisation σ, we can say $T_p\Sigma = \operatorname{Im} D\,\sigma|_0 = \operatorname{span}\{\sigma_u, \sigma_v\}$ where $\sigma(0) = p$. The first fundamental form is a symmetric bilinear form on the tangent spaces $T_p\Sigma$, varying smoothly in p. However, we choose to express this in a basis coming from the parametrisation σ. In particular, we can think about the matrix expression

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

Example. The plane $\mathbb{R}^2_{xy} \subset \mathbb{R}^3$ has the parametrisation $(u, v) \mapsto (u, v, 0)$. Hence, $\sigma_u = e_1$ and $\sigma_v = e_2$, hence the first fundamental form is $\mathrm{d}u^2 + \mathrm{d}v^2$. We could also use polar coordinates, using $\sigma(r, \theta) = (r\cos\theta, r\sin\theta, 0)$. This parametrises the plane without the origin. This gives $\sigma_r = (\cos\theta, \sin\theta, 0)$ and $\sigma_\theta = (-r\sin\theta, r\cos\theta, 0)$. The first fundamental form is $\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2$.

Definition. Let Σ, Σ' be smooth surfaces in $\mathbb{R}^3$. We say that they are *isometric* if there exists a diffeomorphism $f : \Sigma \to \Sigma'$ that preserves the lengths of all curves. More formally, for every smooth curve $\gamma : (a, b) \to \Sigma$, the length of γ is the same as the length of $f \circ \gamma$.

Example. Let $\Sigma' = f(\Sigma)$ where $f : \mathbb{R}^3 \to \mathbb{R}^3$ is a global isometry, or rigid motion, of $\mathbb{R}^3$; that is, $v \mapsto Av + b$ for an orthogonal matrix A. These isometries preserve the Euclidean inner product on $\mathbb{R}^3$, hence f preserves length. However, in the definition, we need not map all of $\mathbb{R}^3$ to itself, just $\Sigma \to \Sigma'$.

Definition. We say that Σ and Σ' are *locally isometric* near points $p \in \Sigma$ and $q \in \Sigma'$ if there exist open neighbourhoods U of p and V of q such that U and V are isometric. We can also say that Σ and Σ' are locally isometric if they are locally isometric at all points; that is, each point of Σ is locally isometric to some point on Σ'.

Lemma. Smooth surfaces Σ, Σ' in $\mathbb{R}^3$ are locally isometric near $p \in \Sigma$ and $q \in \Sigma'$ if and only if there exist allowable parametrisations $\sigma : V \to U \subseteq \Sigma$ and $\sigma' : V \to U' \subseteq \Sigma'$ such that the first fundamental forms are equivalent.

Proof. By definition, the first fundamental form of Σ determines the lengths of all curves on Σ that lie in U. We will now show that lengths of curves determine the first fundamental form of a parametrisation. Given $\sigma : V \to U$, without loss of generality let $V = B(0, \delta)$ for some $\delta > 0$, where $\sigma(0) = p$. Consider, for all $\varepsilon < \delta$, the curve

$$\gamma_\varepsilon : [0, \varepsilon] \to U; \quad t \mapsto \sigma(t, 0)$$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} L(\gamma_\varepsilon) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \int_0^\varepsilon \sqrt{E(t, 0)}\,\mathrm{d}t = \sqrt{E(\varepsilon, 0)}$$

Hence,

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} L(\gamma_\varepsilon) = \sqrt{E(0, 0)}$$

So we can determine E at p by looking at lengths of curves. We can similarly consider

$$\chi_\varepsilon : [0, \varepsilon] \to U; \quad t \mapsto \sigma(0, t)$$

which determines G. Finally, consider

$$\lambda_\varepsilon : [0, \varepsilon] \to U; \quad t \mapsto \sigma(t, t)$$

which determines $\sqrt{(E + 2F + G)(0, 0)}$ which gives F implicitly. □

Example. The sphere of radius a, given by $\{x^2 + y^2 + z^2 = a^2\}$, has an open set with allowable parametrisation

$$\sigma(u, v) = (a \cos u \cos v, a \cos u \sin v, a \sin u)$$

where $u \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and $v \in (0, 2\pi)$. This parametrises the complement of a half great circle. Here,

$$\sigma_u = (-a \sin u \cos v, -a \sin u \sin v, a \cos u); \quad \sigma_v = (-a \cos u \sin v, a \cos u \cos v, 0)$$

Hence,

$$E = a^2; \quad F = 0; \quad G = a^2 \cos^2 u$$

which gives the first fundamental form as

$$a^2 \, \mathrm{d}u^2 + a^2 \cos^2 u \, \mathrm{d}v^2$$

Example. Consider the surface of revolution given by a curve

$$\eta(t) = (f(t), 0, g(t))$$

rotated about the z axis. The resulting surface has parametrisation

$$\sigma(u, v) = (f(u) \cos v, f(u) \sin v, g(u))$$

Hence,

$$\sigma_u = (f_u \cos v, f_u \sin v, g_u); \quad \sigma_v = (-f \sin v, f \cos v, 0)$$

which gives

$$(f_u^2 + g_u^2) \, \mathrm{d}u^2 + f^2 \, \mathrm{d}v^2$$

Example. Consider the cone with angle $\arctan a$ to the vertical. For $u > 0$ and $v \in (0, 2\pi)$, we define

$$\sigma(u, v) = (au \cos v, au \sin v, u)$$

The first fundamental form is

$$(1 + a^2) \, \mathrm{d}u^2 + a^2 u^2 \, \mathrm{d}v^2$$

Consider cutting the cone along the line $v = 0$ and flattening it into a plane sector. The circumference of the sector is $2\pi a$ and the radius is $\sqrt{1 + a^2}$, hence the angle traced out by the sector is $\theta_0 = \frac{2\pi a}{\sqrt{1+a^2}}$. We can parametrise this subset of the plane by

$$\sigma(r, \theta) = \left(\sqrt{1 + a^2} r \cos\left(\frac{a\theta}{\sqrt{1 + a^2}}\right), \sqrt{1 + a^2} r \sin\left(\frac{a\theta}{\sqrt{1 + a^2}}\right), 0\right)$$

for $r > 0$ and $\theta \in (0, 2\pi)$. We can then check that the first fundamental form here is

$$(1 + a^2) \, \mathrm{d}r^2 + r^2 a^2 \, \mathrm{d}\theta^2$$

which matches the first fundamental form for the cone itself. Hence the cone and the plane are locally isometric. However, the cone and plane are not globally isometric, since the two topological spaces are not homeomorphic, so no diffeomorphism that preserves lengths can be constructed.

Lemma. Let Σ be a smooth surface in $\mathbb{R}^3$, and let $p \in \Sigma$. Suppose we have two allowable parametrisations $\sigma : V \to U$ and $\sigma' : V' \to U$ for the same open neighbourhood of p. The two parametrisations differ by a transition map $F = {\sigma'}^{-1} \circ \sigma$ which is a diffeomorphism of open subsets of $\mathbb{R}^2$. There exist first fundamental forms for both parametrisations. Then,

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix} = (DF)^\intercal \begin{pmatrix} E' & F' \\ F' & G' \end{pmatrix} (DF)$$

Proof. By definition,

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix} = \begin{pmatrix} \sigma_u \cdot \sigma_u & \sigma_u \cdot \sigma_v \\ \sigma_v \cdot \sigma_u & \sigma_v \cdot \sigma_v \end{pmatrix} = (D\sigma)^\intercal (D\sigma)$$

Now, $\sigma = \sigma' \circ F$ hence the result follows. □

4.2. Conformality

If $v, w \in \mathbb{R}^3$, we have $v \cdot w = |v| \cdot |w| \cdot \cos\theta$. This allows us to deduce the angle θ between two vectors given their dot product and lengths. This can also be done when v, w are in the tangent plane $T_p\Sigma$, and then we can express the angle in terms of the first fundamental form. Let σ be an allowable parametrisation for Σ near p, such that $D\,\sigma|_0$ evaluates to v at v_0 and w at w_0.

$$\cos\theta = \frac{v \cdot w}{|v| \cdot |w|} = \frac{I(v_0, w_0)}{\sqrt{I(v_0, v_0)}\sqrt{I(w_0, w_0)}}$$

where I denotes the first fundamental form of σ at zero.

Lemma. Let Σ be a smooth surface in $\mathbb{R}^3$, and let $\sigma : V \to U$ be an allowable parametrisation of Σ near p. Then σ is *conformal* if $E = G$ and $F = 0$ in the first fundamental form.

Proof. Consider curves $\gamma : t \mapsto (u(t), v(t))$ and $\tilde{\gamma} : t \mapsto (\tilde{u}(t), \tilde{v}(t))$ in V, where $\gamma(0) = \tilde{\gamma}(0) = 0 \in V$. Let σ be a parametrisation $V \to U \subseteq \Sigma$ such that $\sigma(0) = p \in \Sigma$. Then the curves $\sigma \circ \gamma$ and $\sigma \circ \tilde{\gamma}$ meet at angle θ on Σ, where

$$\cos\theta = \frac{E\dot{u}\dot{\tilde{u}} + F\left(\dot{u}\dot{\tilde{v}} + \dot{v}\dot{\tilde{u}}\right) + G\dot{v}\dot{\tilde{v}}}{\sqrt{E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2}\sqrt{E\dot{\tilde{u}}^2 + 2F\dot{\tilde{u}}\dot{\tilde{v}} + G\dot{\tilde{v}}^2}}$$

In particular, if σ is conformal, suppose $\gamma(t) = (t, 0)$ and $\tilde{\gamma}(t) = (0, t)$. Then, we have that the curves meet at $\frac{\pi}{2}$ in V, so they meet at $\frac{\pi}{2}$ in Σ, so we find that $\cos\theta = 0 \implies F = 0$. Similarly, if $\gamma(t) = (t, t)$ and $\tilde{\gamma}(t) = (t, -t)$, we find $\cos\theta = 0 \implies E = G$.

Conversely, suppose there exists a parametrisation σ such that $E = G$ and $F = 0$. Then, in this parametrisation, the first fundamental form is of the form $\rho\left(\mathrm{d}u^2 + \mathrm{d}v^2\right)$ for $\rho = E : V \to \mathbb{R}$. Hence, the first fundamental form is a pointwise rescaling of the Euclidean fundamental form $\mathrm{d}u^2 + \mathrm{d}v^2$. Rescaling the plane does not change angles, so σ is conformal as required. □

Remark. Conformality in charts is historically important for cartography. The existence of conformal charts is closely connected to Riemann surfaces, which are topological surfaces locally modelled on $\mathbb{C}$ instead of $\mathbb{R}^2$.

4.3. Area

Recall that a parallelogram spanned by vectors v, w has area $|v \times w| = \langle v, v\rangle\langle w, w\rangle - \langle v, w\rangle^2$, where $\times$ denotes the cross product. Let $\sigma : V \to U \subseteq \Sigma$ be an allowable parametrisation with $\sigma(0) = p$, and consider $\sigma_u, \sigma_v \in T_p\Sigma$. The square of the area of the infinitesimal parallelogram spanned by σ_u, σ_v is given by

$$\langle \sigma_u, \sigma_u\rangle\langle \sigma_v, \sigma_v\rangle - \langle \sigma_u, \sigma_v\rangle^2 = EG - F^2$$

Definition. Let Σ be a smooth surface in $\mathbb{R}^3$, and $\sigma : V \to U \subseteq \Sigma$ an allowable parametrisation. Then,

$$\operatorname{area}(U) = \int_V \sqrt{EG - F^2}\,\mathrm{d}u\,\mathrm{d}v$$

Remark. This is independent of parametrisation. Indeed, suppose $\sigma : V \to U$ and $\tilde{\sigma} : \widetilde{V} \to U$ are allowable. Then $\tilde{\sigma} = \sigma \circ \varphi$ for some transition map $\varphi : \widetilde{V} \to V$. We know then that

$$\begin{pmatrix} \widetilde{E} & \widetilde{F} \\ \widetilde{F} & \widetilde{G} \end{pmatrix} = (D\tilde{\sigma})^\intercal (D\tilde{\sigma}) = (D\varphi)^\intercal \begin{pmatrix} E & F \\ F & G \end{pmatrix} (D\varphi)$$

Hence,

$$\sqrt{\widetilde{E}\widetilde{G} - \widetilde{F}^2} = |\det(D\varphi)|\sqrt{EG - F^2}$$

The usual change of variables formula for integration, combined with the fact that φ is a diffeomorphism, gives

$$\int_V \sqrt{EG - F^2}\,\mathrm{d}u\,\mathrm{d}v = \int_{\widetilde{V}} \sqrt{\widetilde{E}\widetilde{G} - \widetilde{F}^2}\,\mathrm{d}u\,\mathrm{d}v$$

Note, we can compute the area of an open set $U \subseteq \Sigma$, not necessarily lying in a single parametrisation, by covering the set by a finite amount of open subsets which lie in single charts. For instance, if Σ is compact, we can compute the area of Σ itself.

Example. Consider the graph $\Sigma = \{(u, v, f(u, v)) : (u, v) \in \mathbb{R}^2\}$, where $f : \mathbb{R}^2 \to \mathbb{R}$ is a smooth function. This has a global parametrisation $\sigma(u, v) = (u, v, f(u, v))$. Here, $\sigma_u = (1, 0, f_u)$ and $\sigma_v = (0, 1, f_v)$, hence

$$\sqrt{EG - F^2} = \sqrt{1 + f_u^2 + f_v^2}$$

Let $U_R \subseteq \Sigma$ be the part of the graph lying inside the disc $B(0, R) \subseteq \mathbb{R}^2$. Then

$$\text{area}(U_R) = \int_{B(0,R)} \sqrt{1 + f_u^2 + f_v^2}\, \mathrm{d}u\, \mathrm{d}v \geq \pi R^2$$

with equality exactly when $f_u = f_v = 0$, so f is constant and U_R is contained inside a plane perpendicular to the z axis. Hence, the projection from Σ to $\mathbb{R}^2_{xy}$ is not area-preserving, unless Σ is a plane perpendicular to the z axis.

Example. Consider the sphere enclosed exactly by a cylinder. The cylindrically radial projection from the sphere to the cylinder is area-preserving. This is explored further in the example sheets.

4.4. Second fundamental form

Let $\sigma : V \to U \subseteq \Sigma$ be allowable. By using Taylor's theorem, we can write

$$\begin{aligned}\sigma(u + h, v + \ell) = \sigma(u, v)&\\ &+ h\sigma_u(u, v) + \ell\sigma_v(u, v)\\ &+ \frac{1}{2}\big(h^2\sigma_{uu}(u, v) + 2h\ell\sigma_{uv}(u, v) + \ell^2\sigma_{vv}(u, v)\big)\\ &+ O(h^3, \ell^3)\end{aligned}$$

where h, ℓ are small, and $(u + h, v + \ell) \in V$. Recall that if $p = \sigma(u, v)$, we have $T_p\Sigma = \langle\{\sigma_u, \sigma_v\}\rangle$. Hence, the orthogonal distance from $\sigma(u + h, v + \ell)$ to the affine tangent plane $T_p\Sigma + p$ is given by projection to the normal direction.

$$\langle n, \sigma(u + h, v + \ell) - \sigma(u, v)\rangle = \frac{1}{2}\big(\langle n, \sigma_{uu}\rangle h^2 + 2\langle n, \sigma_{uv}\rangle h\ell + \langle n, \sigma_{vv}\rangle \ell^2\big) + O(h^3, \ell^3)$$

Definition. The *second fundamental form* of Σ in the allowable parametrisation σ is the quadratic form

$$L\,\mathrm{d}u^2 + 2M\,\mathrm{d}u\,\mathrm{d}v + N\,\mathrm{d}v^2$$

where

$$L = \langle n, \sigma_{uu}\rangle; \quad M = \langle n, \sigma_{uv}\rangle; \quad N = \langle n, \sigma_{vv}\rangle$$

and

$$n = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}$$

We can write this as the matrix

$$\begin{pmatrix} L & M \\ M & N \end{pmatrix}$$

which defined a quadratic form on $T_p\Sigma$ which varies smoothly in p.

Lemma. Let V be connected and $\sigma\colon V \to U \subseteq \Sigma$ be an allowable parametrisation such that the second fundamental form vanishes identically with respect to σ. Then U lies in an affine plane.

Remark. The first fundamental form is a non-degenerate symmetric bilinear form on $T_p\Sigma$, whereas the second fundamental form may be degenerate.

Proof. By definition,

$$\langle n, \sigma_u \rangle = 0 = \langle n, \sigma_v \rangle$$

Hence, by differentiating, we find

$$0 = \langle n_u, \sigma_u \rangle + \langle n, \sigma_{uu} \rangle = \langle n_v, \sigma_v \rangle + \langle n, \sigma_{vv} \rangle = \langle n_v, \sigma_u \rangle + \langle n, \sigma_{uv} \rangle$$

Some of these terms appear in the definition of the second fundamental form:

$$\begin{aligned} L &= \langle n, \sigma_{uu} \rangle = -\langle n_u, \sigma_u \rangle \\ M &= \langle n, \sigma_{uv} \rangle = -\langle n_v, \sigma_u \rangle = -\langle n_u, \sigma_v \rangle \\ N &= \langle n, \sigma_{vv} \rangle = -\langle n_v, \sigma_v \rangle \end{aligned}$$

If the second fundamental form vanishes, then n_u is orthogonal to σ_u, σ_v, and n itself. Since σ_u, σ_v, n form a basis for $\mathbb{R}^3$, we have $n_u = 0$. Similarly, $n_v = 0$, hence n is constant by the mean value theorem. □

Remark. The first fundamental form in parametrisation σ can be written $(D\sigma)^\intercal(D\sigma)$. We can similarly write the second fundamental form as

$$-(Dn)^\intercal(D\sigma) = \begin{pmatrix} L & M \\ M & N \end{pmatrix} = -\begin{pmatrix} n_u \cdot \sigma_u & n_u \cdot \sigma_v \\ n_v \cdot \sigma_u & n_v \cdot \sigma_v \end{pmatrix}$$

Hence, if $\sigma\colon V \to \Sigma$ and $\tilde{\sigma}\colon \widetilde{V} \to \Sigma$ are allowable parametrisations for an open set $U \subseteq \Sigma$ with transition map $\varphi\colon \widetilde{V} \to V$ given by $\varphi = \sigma^{-1} \circ \tilde{\sigma}$, we can use the above expression to find

$$\begin{pmatrix} \tilde{L} & \widetilde{M} \\ \widetilde{M} & \widetilde{N} \end{pmatrix} = \pm(D\varphi)^\intercal \begin{pmatrix} L & M \\ M & N \end{pmatrix} (D\varphi)$$

The change in sign depends on whether the transition map preserves or reverses orientation. If the normal vectors agree, there is no negative sign.

$$n_{\sigma\circ\varphi}\Big|_{(\tilde{u},\tilde{v})} = \pm\, n_\sigma\Big|_{\varphi(\tilde{u},\tilde{v})}$$

for $(\tilde{u}, \tilde{v}) \in \widetilde{V}$. In particular, if $\det(D\varphi) < 0$, we arrive at a negative sign. If we assume that $V, \widetilde{V}$ are connected, the determinant $\det(D\varphi)$ does not change sign.

Example. Consider the cylinder with allowable parametrisation

$$\sigma(u, v) = (a\cos u, a\sin u, v)$$

where $u \in (0, 2\pi), v \in \mathbb{R}$. Note that $\sigma_{uv} = \sigma_{vv} = 0$, hence $M = N = 0$. We can show that the second fundamental form is given by

$$\begin{pmatrix} -a & 0 \\ 0 & 0 \end{pmatrix}; \quad -a\,\mathrm{d}u^2$$

4.5. Gauss maps

Definition. Let Σ be a smooth oriented surface in $\mathbb{R}^3$. The *Gauss map* $n : \Sigma \to \mathbb{S}^2$ is the map $p \mapsto n(p)$, where the normal vector is normalised and hence lies in the unit sphere.

Lemma. The Gauss map is smooth.

Proof. Since smoothness is a local property, it suffices to check the smoothness of the map on an arbitrary parametrised part of Σ. Let $\sigma : V \to U \subseteq \Sigma$ be allowable and compatible with a chosen orientation. Then

$$n(p) = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}$$

Since σ is allowable, the denominator is non-vanishing. Hence, $n(p)$ is smooth as required. □

Remark. If $\Sigma = F^{-1}(0)$ for some function $F : \mathbb{R}^3 \to \mathbb{R}$ with nonzero derivative DF at all points $x \in \Sigma$ (which was required for Σ to be a smooth surface in $\mathbb{R}^3$), then we can explicitly calculate the Gauss map to be

$$n(p) = \frac{\nabla F}{\|\nabla F\|}$$

Note that,

$$T_p\Sigma = T_{n(p)}S^2 = (n(p))^\perp$$

since the two planes are orthogonal to the same vector. More concretely, if $v \in T_p\Sigma$ is $\gamma'(0)$ where $\gamma : (-\varepsilon, \varepsilon) \to \Sigma$, $\gamma(0) = p$ for a smooth curve γ, we can apply the Gauss map to γ and find

$$n \circ \gamma : (-\varepsilon, \varepsilon) \to S^2; \quad (n \circ \gamma)(0) = n(p)$$

Then, by the chain rule,

$$D\,n\Big|_p\,(v) = (n \circ \gamma)'(0) \in T_{n(p)}S^2 = T_p\Sigma$$

Thus, the derivative of the Gauss map is $D\,n|_p : T_p\Sigma \to T_p\Sigma$. This can be viewed as an endomorphism of a fixed (with respect to parametrisation choice) two-dimensional subspace of $\mathbb{R}^3$.

To summarise, let Σ be an oriented smooth surface in $\mathbb{R}^3$. Then,

(i) The first fundamental form is a symmetric bilinear form $\langle \cdot\,,\, \cdot \rangle = \mathrm{I}_p : T_p\Sigma \times T_p\Sigma \to \mathbb{R}$, which is the restriction of the Euclidean inner product to this space $T_p\Sigma$. We can write $\mathrm{I}_p(v, w)$, where $v, w \in T_p\Sigma$.

(ii) The second fundamental form is also a symmetric bilinear form $\mathrm{II}_p : T_p\Sigma \times T_p\Sigma \to \mathbb{R}$, given by

$$\mathrm{II}_p(v, w) = \mathrm{I}_p\left(-D\,n\Big|_p(v), w\right)$$

where n is the Gauss map.

If we choose an allowable parametrisation (which for the second fundamental form must be correctly oriented) $\sigma : V \to U \subseteq \Sigma$ near $p \in \Sigma$, and if

$$D\,\sigma\Big|_0(\hat{v}) = v; \quad D\,\sigma\Big|_0(\hat{w}) = w; \quad \sigma(0) = p$$

Then,

$$\mathrm{I}_p(v, w) = \hat{v}^\intercal \begin{pmatrix} E & F \\ F & G \end{pmatrix} \hat{w}; \quad \mathrm{II}_p(v, w) = \hat{v}^\intercal \begin{pmatrix} L & M \\ M & N \end{pmatrix} \hat{w}$$

where E, F, G, L, M, N depend on the choice of σ. Note that the functions I_p and II_p are independent of σ.

Lemma. The derivative of the Gauss map is self-adjoint. More precisely, viewing the map $D\,n|_p : T_p\Sigma \to T_p\Sigma$ as an endomorphism over the inner product space with the first fundamental form, this linear map satisfies

$$\mathrm{I}_p\left(D\,n\Big|_p(v), w\right) = \mathrm{I}_p\left(v, D\,n\Big|_p(w)\right)$$

for all $v, w \in T_p\Sigma$.

Proof. From expressions for local parametrisations, we can show that I_p and II_p are symmetric. Hence,

$$\mathrm{I}_p(D\,n\Big|_p(v), w) = -\,\mathrm{II}_p(v, w) = -\,\mathrm{II}_p(w, v) = \mathrm{I}_p(D\,n\Big|_p(w), v) = \mathrm{I}_p(v, D\,n\Big|_p(w))$$

□

Remark. The *fundamental theorem of surfaces in* $\mathbb{R}^3$ states that a smooth oriented connected surface in $\mathbb{R}^3$ is determined completely, up to rigid motion, by the two fundamental forms.

4.6. Gauss curvature

Definition. Let Σ be a smooth surface in $\mathbb{R}^3$. The *Gauss curvature* $\kappa\colon \Sigma \to \mathbb{R}$ of Σ is the function defined by

$$\kappa(p) = \det\left(D\, n\Big|_p\right)$$

Remark. This is always well-defined, even if Σ is not oriented. This is because Σ is always locally orientable, and the two normals differ by sign. In two dimensions, $\det(-A) = \det(A)$, so the determinant is invariant.

We can compute κ directly. Let Σ be a smooth surface in $\mathbb{R}^3$, and σ an allowable parametrisation for an open neighbourhood of a point p. Recall that

$$\mathrm{I}_p\colon T_p\Sigma \to T_p\Sigma;\quad (v,w) \mapsto \langle v, w\rangle;\quad \mathrm{II}_p\colon T_p\Sigma \to T_p\Sigma;\quad (v,w) \mapsto \mathrm{I}_p(-Dn\Big|_p(v), w)$$

and $Dn|_p : T_p\Sigma \to T_p\Sigma$. The choice of parametrisation σ for an open neighbourhood U of p provides a preferred basis $\{\sigma_u, \sigma_v\}$ for $T_p\Sigma$. We can therefore write the fundamental forms as matrices with respect to this basis. Let $A = \mathrm{I}_p, B = \mathrm{II}_p, \mathbb{S} = Dn|_p$ in this basis. In matrix form, we can write $\mathrm{II}_p = \mathrm{I}_p(-Dn|_p(v), w)$ as

$$B = -\mathbb{S}^\intercal A \implies \kappa(p) = \det(\mathbb{S}) = \det\left(-A^{-1}B\right) = \frac{LN - M^2}{EG - F^2}$$

If $\sigma, \tilde{\sigma}$ are allowable and $\varphi = \sigma^{-1} \circ \sigma$ is a transition map, then

$$\widetilde{A} = (D\varphi)^\intercal A (D\varphi);\quad \widetilde{B} = \pm(D\varphi)^\intercal B(D\varphi)$$

Since the sign vanishes under taking determinants, κ is intrinsic and does not depend on the choice of parametrisation.

Example. For a cylinder $\{x^2 + y^2 = 1\}$ the Gauss map $n\colon \Sigma \to S^2$ has image which lies in the equator. Its derivative $Dn|_p : T_p\Sigma \to T_p\Sigma$ has one-dimensional image, since any $\gamma\colon (-\varepsilon, \varepsilon) \to \Sigma$ has $n \circ \gamma \subseteq S^1$. Hence its Gauss curvature is zero.

Definition. A smooth surface in $\mathbb{R}^3$ with vanishing Gauss curvature everywhere is *flat*.

Remark. If $\sigma\colon V \to U$ is allowable, and n_σ is defined to be $n \circ \sigma\colon V \to S^2$, then

$$Dn_\sigma\Big|_0 : \sigma_u \mapsto (n_\sigma)_u;\quad \sigma_v \mapsto (n_\sigma)_v$$

In particular, $\kappa(p) = \kappa(\sigma(0))$ vanishes if and only if $(n_\sigma)_u \times (n_\sigma)_v = 0$. Usually, we will write n to denote n_σ. In this case, the condition for flatness is that $n_u \times n_v = 0$.

Example. If Σ is the graph of a smooth function f, then on the example sheets we show that

$$\kappa = \frac{f_{uu}f_{vv} - f_{uv}^2}{(1 + f_u^2 + f_v^2)^2}$$

Hence, the curvature depends on the derivative and the Hessian of f. For instance, let $f(u,v) = \sqrt{r^2 - u^2 - v^2}$. Here, the graph is a piece of a sphere of radius r. We can find

$$f_{uu}\Big|_0 = f_{vv}\Big|_0 = \frac{-1}{r}; \quad f_{uv}\Big|_0 = 0 \implies \kappa(0,0,r) = \frac{1}{r^2}$$

Since $O(3)$ acts transitively on S^2, and the fundamental forms are preserved by such global isometries, $\kappa = \frac{1}{r^2}$ everywhere on the sphere of radius r.

Example. Let Σ be the smooth surface given by $\{z = x^2 + y^2\}$. We claim that, for the inward facing choice of orientation, the image of the Gauss map is the open northern hemisphere. Note that Σ is invariant under rotations about the z axis. Also, we can show that if R is a rotation, $n \circ R = R \circ n$. Therefore, it suffices to consider an arbitrary point with $y = 0$.

Here, $\Sigma = F^{-1}(0)$ for the function $F(x,y,z) = z - x^2 - y^2$, which has nonvanishing derivative at the points $p \in \Sigma$. Hence, at $p = (x, 0, x^2)$, we have

$$n(p) = \frac{\nabla F}{\|\nabla F\|} = \frac{(-2x, 0, 1)}{\sqrt{1 + 4x^2}}$$

We can check explicitly that this map has image which an arc lying in the open northern hemisphere.

4.7. Elliptic, hyperbolic, and parabolic points

Definition. Let Σ be a smooth surface in $\mathbb{R}^3$ and $p \in \Sigma$. We say that p is

(i) *elliptic* if $\kappa(p) > 0$;

(ii) *hyperbolic* if $\kappa(p) < 0$;

(iii) *parabolic* if $\kappa(p) = 0$.

Lemma. In a sufficiently small neighbourhood of an elliptic point p, Σ lies entirely on one side of $p + T_p\Sigma$. If p is hyperbolic, Σ lies on both sides of $p + T_p\Sigma$.

Proof. Let σ be a local parametrisation near p. Here,

$$\kappa = \frac{LN - M^2}{EG - F^2}$$

The denominator is always positive, since it is the determinant of a positive definite symmetric bilinear form I_p. Hence, the sign of κ depends on the sign of $LN - M^2$. If $w = h\sigma_u + \ell\sigma_v \in T_p\Sigma$, then $\frac{1}{2}\,\mathrm{II}_p(w,w)$ measures the signed distance from $\sigma(h,l)$ to $p + T_p\Sigma$. If p is elliptic, then II_p has eigenvalues of the same sign, so it is either positive or negative definite at p. Since II_p varies smoothly in p, it remains positive or negative definite in a small neighbourhood of p. Hence, in such a neighbourhood, the signed distance has the same sign as required. Conversely, if p is hyperbolic, $\mathrm{II}_p(w,w)$ takes both signs in a neighbourhood of p. □

Remark. We cannot conclude anything about parabolic points *a priori*. For instance, the cylinder is flat (all points are parabolic), and the surface lies on one side of the tangent plane at every point. Consider also the *monkey saddle* defined by

$$\sigma(u, v) = (u, v, u^3 - 3v^2u)$$

which has a parabolic point at the origin, but Σ lies on both sides of the tangent plane in every open neighbourhood of the origin. At $p = \sigma(0, 0)$, the Gauss curvature vanishes, but the surface lies locally on both sides of the tangent plane.

Proposition. Let Σ be a compact smooth surface in $\mathbb{R}^3$. Then Σ has an elliptic point.

Proof. Since Σ is compact, it is closed and bounded as a subset of $\mathbb{R}^3$. Hence, for R' sufficiently large, Σ lies entirely within $\overline{B(0, R')}$. Let R be the minimal such R'. Up to a global isometry of $\mathbb{R}^3$, there exists a point $p = (0, 0, R) \in \Sigma$ on the sphere $S^2(R)$ of radius R. Here, $T_p\Sigma = T_pS^2$. Hence, locally near p, we can view Σ as the graph of a smooth function $f : V \to \mathbb{R}^3$ on the x, y coordinates with the property that $f - \sqrt{R^2 - u^2 - v^2} \leq 0$. This expresses the fact that Σ lies underneath the sphere of radius R.

We can now consider the Taylor series of f. Note that $(0, 0)$ is a maximum point of f, hence $f_u = f_v = 0$ at 0. Thus, for sufficiently small u, v,

$$\frac{1}{2}\left(f_{uu}u^2 + 2f_{uv}uv + f_{vv}v^2\right) + \frac{1}{2R}(u^2 + v^2) \leq 0$$

Hence, the second fundamental form is locally negative definite near $(0, 0)$. Hence, $\kappa(p) > 0$, so p is elliptic as required. In particular, the curvature at this point is greater than that of the sphere. □

Theorem. Let Σ be a smooth surface in $\mathbb{R}^3$, and let $p \in \Sigma$ such that $\kappa(p) \neq 0$. Let U be an open neighbourhood of p, and a decreasing sequence $A_i \subseteq U$ of neighbourhoods that 'shrink to p', in the sense that for all $\varepsilon > 0$, $A_i \subseteq B(p, \varepsilon)$ for sufficiently large i. Then,

$$|\kappa(p)| = \lim_{i \to \infty} \frac{\text{area}_{S^2}(n(A_i))}{\text{area}_\Sigma(A_i)}$$

In other words, the Gauss curvature is an infinitesimal measure of how much the Gauss map n distorts area.

Remark. Around hyperbolic points, the signed area of $n(A_i)$ is reversed, since curves γ reverse direction under n. We can alternatively define the *signed area* of $n(A_i)$ to be the area of $n(A_i)$ if $\kappa > 0$ and the negation of this area if $\kappa < 0$. The above theorem holds when $\kappa = 0$, but this will not be proven.

Proof. Let σ be an allowable parametrisation near $p \in \Sigma$. Using σ, we can define the open sets $\sigma^{-1}(A_i) = V_i \subset V$. Since the A_i shrink to p, we have that $\bigcap V_i = \{(0, 0)\}$. We have

$$\text{area}_\Sigma(A_i) = \int_{V_1} \sqrt{EG - F^2}\, \mathrm{d}u\, \mathrm{d}v = \int_{V_i} \|\sigma_u \times \sigma_v\|\, \mathrm{d}u\, \mathrm{d}v$$

Recall from the chain rule applied to $n \circ \gamma$ that

$$Dn\Big|_{(u,v)}(\sigma_u) = n_u; \quad Dn\Big|_{(u,v)}(\sigma_v) = n_v$$

Since $\kappa(p) = \kappa(\sigma(0,0)) \neq 0$, $n \circ \sigma \colon V \to S^2$ has derivative of rank 2. This defines an allowable parametrisation for an open neighbourhood of $n((0,0))$ by the inverse function theorem. Therefore,

$$\operatorname{area}_{S^2}(n(A_i)) = \int_{V_i} \|n_u \times n_v\| \,\mathrm{d}u \,\mathrm{d}v$$

for sufficiently large i such that $\sigma^{-1}A_i = V_i$ lies in the open neighbourhood of $(0,0)$ where $n \circ \sigma$ is a diffeomorphism.

$$\begin{aligned}\int_{V_i} \|n_u \times n_v\| \,\mathrm{d}u \,\mathrm{d}v &= \int_{V_i} \|Dn(\sigma_u) \times Dn(\sigma_v)\| \,\mathrm{d}u \,\mathrm{d}v \\ &= \int_{V_i} |\det(Dn)| \cdot \|\sigma_u \times \sigma_v\| \,\mathrm{d}u \,\mathrm{d}v \\ &= \int_{V_i} |\kappa(u,v)| \cdot \|\sigma_u \times \sigma_v\| \,\mathrm{d}u \,\mathrm{d}v\end{aligned}$$

As κ is continuous, given $\varepsilon > 0$ there exists $\delta > 0$ such that $|\kappa(u,v) - \kappa(0,0)| < \varepsilon$ for all $(u,v) \in B((0,0),\delta)$. In particular, for sufficiently large i, we have

$$|\kappa(u,v)| \in (|\kappa(p)| - \varepsilon, |\kappa(p)| + \varepsilon)$$

Hence,

$$\begin{aligned}(|\kappa(p)| - \varepsilon) \int_{V_i} \|\sigma_u \times \sigma_v\| \,\mathrm{d}u \,\mathrm{d}v &\leq \int_{V_i} |\kappa(u,v)| \cdot \|\sigma_u \times \sigma_v\| \,\mathrm{d}u \,\mathrm{d}v \\ &\leq (|\kappa(p)| + \varepsilon) \int_{V_i} \|\sigma_u \times \sigma_v\| \,\mathrm{d}u \,\mathrm{d}v\end{aligned}$$

In other words,

$$|\kappa(p)| - \varepsilon \leq \frac{\operatorname{area}_{S^2}(n(A_i))}{\operatorname{area}_{\Sigma}(A_i)} \leq |\kappa(p)| + \varepsilon$$

Letting $i \to \infty$ gives the result as required. □

Theorem (*theorema egregium*)**.** The Gauss curvature of a smooth surface in $\mathbb{R}^3$ is isometry invariant. In other words, if $f \colon \Sigma_1 \to \Sigma_2$ is a diffeomorphism of surfaces in $\mathbb{R}^3$ which is an isometry, then $\kappa(p) = \kappa(f(p))$ for all p.

Remark. Isometries rely on only the first fundamental form, but Gauss curvature is defined using both fundamental forms. We can do a direct proof by simply differentiating the formula and rearranging until the result follows. This proof is given in Part II.

Alternatively, we can consider a different question: are some allowable parametrisations of a smooth surface in $\mathbb{R}^3$ 'better' than others in some way? If we have a parametrisation $\sigma : V \to U \subseteq \Sigma$, this defines certain distinguished curves, which are the images of $\sigma(t, 0)$ and $\sigma(0, t)$. In this sense, looking for a 'best' parametrisation is equivalent to looking for 'best' distinguished curves near a point. This leads to the study of geodesics. We will later show that every smooth surface in $\mathbb{R}^3$ admits local parametrisations such that the first fundamental form has form $\mathrm{d}u^2 + G\,\mathrm{d}v^2$, so $E = 1$ and $F = 0$. We will also see (on an example sheet) that if such a local parametrisation exists, then κ can be expressed as a function just of G. This allows us to approach the proof of the *theorema egregium* from a more conceptual way, since we have expressed κ in terms of the first fundamental form alone.

Theorem (Gauss–Bonnet theorem)**.** If Σ is a compact smooth surface in $\mathbb{R}^3$, then

$$\int_\Sigma \kappa \,\mathrm{d}A_\Sigma = 2\pi\chi(\Sigma)$$

5. Geodesics

5.1. Definitions

Recall that we defined, for a smooth curve $\gamma : [a, b] \to \mathbb{R}^3$,

$$\text{length}(\gamma) = \int_a^b \|\gamma'(t)\| \, \mathrm{d}t$$

Definition. The *energy* of γ is given by

$$E(\gamma) = \int_a^b \|\gamma'(t)\|^2 \, \mathrm{d}t$$

Definition. Let $\gamma : [a, b] \to \Sigma$, where Σ is a smooth surface in $\mathbb{R}^3$. A *one-parameter variation* (with fixed endpoints) of γ is a smooth map $\Gamma : (-\varepsilon, \varepsilon) \times [a, b] \to \Sigma$, such that if $\gamma_s = \Gamma(s, \cdot)$, then $\gamma_0(t) = \gamma(t)$, and $\gamma_s(a)$ and $\gamma_s(b)$ are independent of s.

Definition. A smooth curve $\gamma : [a, b] \to \Sigma$ is a *geodesic* if, for every variation (γ_s) of γ with fixed endpoints as above, we have $\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} E(\gamma_s) = 0$. Alternatively, γ is a critical point of the energy functional on curves from $\gamma(a)$ to $\gamma(b)$.

5.2. The geodesic equations

Let γ have image contained within the image of an allowable parametrisation $\sigma : V \to U$. Then, for sufficiently small s, we can write $\gamma_s(t) = \sigma(u(s, t), v(s, t))$. Suppose that the first fundamental form, with respect to σ, is

$$E \, \mathrm{d}u^2 + 2F \, \mathrm{d}u \, \mathrm{d}v + G \, \mathrm{d}v^2$$

Let

$$R = E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2$$

By definition,

$$E(\gamma_s) = \int_a^b R \, \mathrm{d}t$$

where R depends on s. Hence,

$$\begin{aligned}\frac{\partial R}{\partial s} = &\left(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2\right)\frac{\partial u}{\partial s} + \left(E_v\dot{v}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2\right)\frac{\partial v}{\partial s} \\ &+ 2(E\dot{u} + F\dot{v})\frac{\partial \dot{u}}{\partial s} + 2(F\dot{u} + G\dot{v})\frac{\partial \dot{v}}{\partial s}\end{aligned}$$

This gives

$$\frac{\mathrm{d}}{\mathrm{d}s}E(\gamma_s) = \int_a^b \frac{\partial R}{\partial s} \, \mathrm{d}t$$

We can integrate by parts. Note that $\frac{\partial u}{\partial s}$ and $\frac{\partial v}{\partial s}$ vanish at a, b. Hence,

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} E(\gamma_s) = \int_a^b \left(A\frac{\partial u}{\partial s} + B\frac{\partial v}{\partial s}\right)\mathrm{d}t$$

where

$$A = E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2 - 2\frac{\partial}{\partial t}(E\dot{u} + F\dot{v})$$
$$B = E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2 - 2\frac{\partial}{\partial t}(F\dot{u} + G\dot{v})$$

Corollary. A smooth curve $\gamma\colon [a,b] \to \Sigma$ with image in $\operatorname{Im}\sigma$ is a geodesic if and only if it satisfies the *geodesic equations*:

$$\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u} + F\dot{v}) = \frac{1}{2}\left(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2\right)$$
$$\frac{\mathrm{d}}{\mathrm{d}t}(F\dot{u} + G\dot{v}) = \frac{1}{2}\left(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2\right)$$

Note that these equations are evaluated at $s = 0$, so no choice of variation is required.

Remark. Solving a differential equation is a local procedure. The original definition of the geodesic seems to be a global property. However, we can always consider a sub-curve of γ to also be a geodesic, since its variations are variations of γ. So the definition can be thought of as local.

Energy is sensitive to reparametrisation. If $f, g\colon [a,b] \to \mathbb{R}$ are smooth, the Cauchy–Schwarz inequality gives that

$$\left(\int_a^b fg\,\mathrm{d}t\right)^2 \le \int_a^b f^2\,\mathrm{d}t \cdot \int_a^b g^2\,\mathrm{d}t$$

Let us apply this to $f = \sqrt{R}$, $g = 1$ to find

$$\operatorname{length}(\gamma)^2 \le E(\gamma)(b-a)$$

Since equality holds only when the two functions are proportional, we must have that $\|\gamma'(t)\|$ is constant for the equality to hold. In other words, γ must be parametrised proportional to arc length.

Corollary. If γ has constant speed and locally minimises length, then it is a geodesic. Further, if γ globally minimises energy, then it must globally minimise length, and is parametrised with constant speed.

Remark. We would like geodesics to be a local property, but not necessarily global length minimisers. For example, all arcs of great circles will be shown to be geodesics, even if large arcs are not global length minimisers between fixed endpoints.

5.3. Geodesics on the plane

The plane $\mathbb{R}^2$ has parametrisation $\sigma(u, v) = (u, v, 0)$ and first fundamental form $\mathrm{d}u^2 + \mathrm{d}v^2$. The geodesic equations here are

$$\ddot{u} = 0; \quad \ddot{v} = 0$$

In particular, the geodesics on the plane are given by

$$u(t) = \alpha t + \beta; \quad v(t) = \gamma t + \delta$$

This is a straight line, parametrised at constant speed.

5.4. Geodesics on the sphere

Consider the unit sphere with parametrisation

$$\sigma(u, v) = (\cos u \cos v, \cos u \sin v, \sin u); \quad u \in \left(-\frac{\pi}{2}\right); \quad v \in (0, 2\pi)$$

This has first fundamental form

$$\mathrm{d}u^2 + \cos^2 u \, \mathrm{d}v^2 \implies E = 1;\ F = 0;\ G = \cos^2 u$$

The geodesic equations give

$$\frac{\mathrm{d}}{\mathrm{d}t}(\dot{u}) = \frac{1}{2} 2 \cos u \sin u \dot{v}^2; \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\cos^2 u \dot{v}\right) = 0$$

This gives

$$\ddot{u} + \sin u \cos u \dot{v}^2 = 0; \quad \ddot{v} - 2 \tan u \dot{u} \dot{v} = 0$$

Since geodesics are parametrised at constant speed, we can assume that it is parametrised at unit speed without loss of generality.

$$\|\gamma'(t)\| = 1 \implies \dot{u} + \cos^2 u \dot{v}^2 = 1$$

Hence,

$$\frac{\ddot{v}}{\dot{v}} = 2 \tan u \dot{u} \implies \ln \dot{v} = -2 \ln \cos u + \text{constant} \implies \dot{v} = \frac{C}{\cos^2 u}$$

Substituting into the unit speed equation,

$$\dot{u}^2 = 1 - \frac{C^2}{\cos^2 u} \implies \dot{u} = \sqrt{\frac{\cos^2 u - C^2}{\cos^2 u}}$$

Then,

$$\frac{\dot{v}}{\dot{u}} = \frac{\mathrm{d}v}{\mathrm{d}u} = \frac{C}{\cos u \sqrt{\cos^2 u - C^2}}$$

Hence,

$$v = \int \frac{\mathrm{d}v}{\mathrm{d}u} \, \mathrm{d}u = \int \frac{C \sec^2 u}{\sqrt{1 - C^2 sec^2 u}} \, \mathrm{d}u$$

Using the substitution $w = \frac{C \tan u}{\sqrt{1-C^2}}$, we find

$$v = \int \frac{w}{1-w^2}\,\mathrm{d}w = \arcsin w + \text{constant} = \arcsin(\lambda \tan u) + \delta$$

for some constants λ, δ. Hence,

$$\sin(v - \delta) = \lambda \tan u$$

Rewriting using the angle addition formula,

$$\underbrace{(\sin v \cos u)}_{x} \cos\delta - \underbrace{(\cos v \cos u)}_{y} \sin\delta - \lambda \underbrace{\sin u}_{z} = 0$$

Hence, the geodesic γ lies on a plane through the origin, since this is a linear equation in x, y, z. Such planes intersect the sphere in great circles.

5.5. Geodesics on the torus

Consider the surface of revolution of a circle in the xz-plane centred at $(a, 0, 0)$ about the z axis, giving a torus. An allowable parametrisation for this surface is

$$\sigma(u, v) = ((a + \cos u)\cos v, (a + \cos u)\sin v, \sin u)$$

The first fundamental form is

$$\mathrm{d}u^2 + (a + \cos u)^2\,\mathrm{d}v^2 \implies E = 1;\ F = 0;\ G = (a + \cos u)^2$$

Note that if we were to take $a = 0$, we would arrive at the unit sphere and its first fundamental form. We can follow the same procedure as above with the sphere, or formally replace $\cos u$ with $a + \cos u$ in the result.

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{C}{(a + \cos u)\sqrt{(a + \cos u)^2 - C^2}}$$

which cannot be integrated using classical functions. This leads to the study of elliptic functions.

5.6. Equivalent characterisation of geodesics

We have so far restricted our analysis to the first fundamental form, without considering its embedding in $\mathbb{R}^3$. Intuitively, we know that straight lines in $\mathbb{R}^2$ are not just locally shortest but also locally straightest. We would expect this to hold for other surfaces as well. We can characterise this notion via stating that the change in the tangent vector to a curve is as small as it could be, subject to the constraint that it lies on the surface.

Proposition. Let Σ be a smooth surface in $\mathbb{R}^3$. A smooth curve $\gamma\colon [a,b] \to \Sigma$ is a geodesic if and only if $\ddot{\gamma}(t)$ is everywhere normal to the surface Σ.

Remark. This proposition makes use of the tangent plane, a notion that exists only because we have an embedding in $\mathbb{R}^3$. Note that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \dot{\gamma}, \dot{\gamma}\rangle = 2\left\langle \underbrace{\dot{\gamma}}_{\text{tangent to }\Sigma}, \underbrace{\ddot{\gamma}}_{\text{normal to }\Sigma}\right\rangle = 0$$

Hence, $\langle \dot{\gamma}, \dot{\gamma}\rangle$ is constant, giving that geodesics are parametrised proportional to arc length.

Proof. The property of being a geodesic as we previously defined is a local property, and so is the condition in the proposition. Hence, we may work entirely within an allowable parametrisation $\sigma\colon V \to U$. Suppose $\gamma(t) = \sigma(u(t), v(t))$. Hence,

$$\dot{\gamma} = \sigma_u \dot{u} + \sigma_v \dot{v}$$

$\ddot{\gamma}$ is normal to Σ when it is orthogonal to the tangent plane, which is spanned by σ_u, σ_v. This is true if and only if

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}(\sigma_u \dot{u} + \sigma_v \dot{v}), \sigma_u \right\rangle = 0 = \left\langle \frac{\mathrm{d}}{\mathrm{d}t}(\sigma_u \dot{u} + \sigma_v \dot{v}), \sigma_v \right\rangle$$

We will prove the first equality. This can be rewritten

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \sigma_u \dot{u} + \sigma_v \dot{v}, \sigma_u\rangle - \left\langle \sigma_u \dot{u} + \sigma_v \dot{v}, \frac{\mathrm{d}}{\mathrm{d}t}\sigma_u \right\rangle = 0$$

Note that $\langle \sigma_u, \sigma_u\rangle = E$ and $\langle \sigma_u, \sigma_v\rangle = F$.

$$\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u} + F\dot{v}) - \langle \sigma_u \dot{u} + \sigma_v \dot{v}, \sigma_{uu}\dot{u} + \sigma_{uv}\dot{v}\rangle = 0$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u} + F\dot{v}) - \left[\dot{u}^2 \langle \sigma_u, \sigma_{uu}\rangle + \dot{u}\dot{v}(\langle \sigma_u, \sigma_{uv}\rangle + \langle \sigma_v, \sigma_{uu}\rangle) + \dot{v}^2 \langle \sigma_v \sigma_{uv}\rangle\right] = 0$$

Note that $E_u = 2\langle \sigma_u, \sigma_{uu}\rangle$, $F_u = \langle \sigma_u, \sigma_{uv}\rangle + \langle \sigma_v, \sigma_{uu}\rangle$, and $G_u = 2\langle \sigma_v, \sigma_{uv}\rangle$. This gives

$$\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u} + F\dot{v}) = \frac{1}{2}\left(E_u \dot{u}^2 + 2F_u \dot{u}\dot{v} + G_u \dot{v}^2\right)$$

which is the first of the geodesic equations. By symmetry, we find the second geodesic equation similarly. □

5.7. Planes of symmetry

Let Σ be a smooth surface in $\mathbb{R}^3$ such that there exists a plane $\Pi \subseteq \mathbb{R}^3$ such that $\Pi \cap \Sigma$ is a smooth embedded curve $C \subseteq \Sigma$, and Σ is setwise preserved by reflection in the plane Π. We will show that C is a geodesic when parametrised at constant speed. Consider a point p on C. We can think of $\mathbb{R}^3 = \Pi \oplus \Pi^\perp$, where we change coordinates such that p is the origin. We can also write $\mathbb{R}^3 = T_p\Sigma \oplus \mathbb{R}n_p$, where $\mathbb{R}n_p$ is the vector subspace of $\mathbb{R}^3$ generated by n_p. Clearly, reflection in Π acts on Π by the identity, and on $\Pi^\perp$ by -1. Since reflection in Π fixes Σ setwise and fixes p, it must also preserve the subspace $T_p\Sigma$. Hence it also preserves $\mathbb{R}n_p$, so $\mathbb{R}n_p \subseteq \Pi$, since Π is not the identity on $T_p\Sigma$. Now, let us parametrise C locally near p using $t \mapsto \gamma(t) \in C$ at constant speed. Since $\gamma(t) \subseteq \Pi$, we have $\dot\gamma(t), \ddot\gamma(t) \in \Pi$. γ has constant speed, so $\langle \dot\gamma, \ddot\gamma \rangle = 0$. Hence $\dot\gamma$ lies in $\Pi \cap T_p\Sigma$ and $\ddot\gamma$ is orthogonal to this and lies in Π, so lies in $\mathbb{R}n_p \subseteq \Pi$. Hence γ is indeed a geodesic.

In particular, arcs of great circles are geodesics, since they lie in planes of symmetry.

5.8. Surfaces of revolution

Consider the surface of revolution given by $\eta(u) = (f(u), 0, g(u))$ where η is smooth and injective, and $f(u) > 0$, rotated about the z axis.

Definition. A circle obtained by rotating a point of η is called a *parallel*. A curve obtained by rotating η itself by a fixed angle about the z axis is called a *meridian*.

A plane in $\mathbb{R}^3$ containing the z axis is a plane of symmetry, hence meridians are geodesics by the previous discussion. Not all parallels are geodesics.

Lemma. A parallel given by $u = u_0$ is a geodesic when parametrised at constant speed if and only if $f'(u_0) = 0$.

Proof. Consider the allowable parametrisation

$$\sigma(u, v) = (f(u)\cos v, f(u)\sin v, g(u))$$

where $u \in (a, b)$ and $v \in (0, 2\pi)$. The first fundamental form is

$$\left[(f')^2 + (g')^2\right] \mathrm{d}u^2 + f^2 \, \mathrm{d}v^2$$

If without loss of generality we choose to parametrise η by arc length, this becomes

$$\mathrm{d}u^2 + f^2 \, \mathrm{d}v^2 \implies E = 1;\ F = 0;\ G = f^2$$

The geodesic equations are

$$\frac{\mathrm{d}}{\mathrm{d}t}\dot u = \ddot u = f f_u \dot v^2; \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(f^2 \dot v\right) = 0$$

and $\dot{u}^2 + f^2\dot{v}^2$ is a nonzero constant. Given that we want to only consider parallels of the surface of revolution, we can impose the constraint that $u = u_0$ is constant. Hence, the constant speed condition gives that

$$\dot{v} = \frac{\text{constant}}{f(u_0)} = \text{constant}$$

The second equation holds automatically. The first equation is

$$0 = f(u_0)f_u(u_0) \cdot \text{constant}$$

So this holds exactly when $f_u(u_0) = 0$. □

Consider a curve $\gamma(t)$ on Σ, making angle θ with the parallel of radius $\rho = f$.

Proposition (Clairaut's relation)**.** If γ is a geodesic, then $\rho \cos\theta$ is constant along γ.

Proof. Let $\gamma(t) = \sigma(u(t), v(t))$, so $\dot{\gamma} = \sigma_u\dot{u} + \sigma_v\dot{v}$. The tangent vector to the parallel is σ_v. By the earlier discussion on angles in terms of the first fundamental form,

$$\cos\theta = \frac{\langle \sigma_v, \sigma_u\dot{u} + \sigma_v\dot{v}\rangle}{\|\sigma_v\| \cdot \|\sigma_u\dot{u} + \sigma_v\dot{v}\|}$$

If γ is parametrised by arc length, $\|\dot{\gamma}\| = 1$, so $\|\sigma_u\dot{u} + \sigma_v\dot{v}\| = 1$. So, using our values for F, G above,

$$\cos\theta = |f(u)\dot{v}| = \rho\dot{v}$$

The second geodesic equation is exactly

$$\rho\cos\theta = \text{constant}$$

□

Example. Usually, for a surface of revolution, we take the assumption that η never intersects the z-axis, or that f is positive. This ensures that all points on the surface are locally smooth. However, we can allow η to meet the z-axis orthogonally, as in the ellipsoid or sphere.

Consider an ellipsoid of revolution. $\rho\cos\theta$ is constant along a geodesic γ. Suppose that at some point γ intersects a parallel of radius ρ_0 at angle θ_0, and that γ is not a meridian (so $\cos\theta \neq 0$). Hence $\theta_0 \in \left[0, \frac{\pi}{2}\right)$. In particular, for $\rho\cos\theta$ to be constant, we must have that ρ is bounded below. A geodesic which is not a meridian is therefore 'trapped' between parallels corresponding to the bound on the size of ρ. In particular, any geodesic through a pole is a meridian.

5.9. Local existence of geodesics

It is difficult to solve the geodesic equations globally. We can often intead prove local results about any geodesics that may arise.

Recall Picard's theorem from Analysis and Topology. Let $I = [t_0 - a, t_0 + a] \subseteq \mathbb{R}$, $B = \{x : \|x - x_0\| \leq b\} \subseteq \mathbb{R}^n$, and $f : I \times B \to \mathbb{R}^n$ that is continuous, and Lipschitz in the second variable.

$$\|f(t, x_1) - f(t, x_2)\| \leq N\|x_1 - x_2\|$$

Then the differential equation $\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x)$ with $x(t_0) = x_0$ has a unique solution for some time interval $|t - t_0| < h$, where $h = \min\left\{a, \frac{b}{s}\right\}$ where $s = \sup \|f\|$. Further, if f is smooth in all parameters, then the solution to the differential equation is smooth and depends smoothly on the initial condition.

Recall the geodesic equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u} + F\dot{v}) = \frac{1}{2}\left(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2\right)$$
$$\frac{\mathrm{d}}{\mathrm{d}t}(F\dot{u} + G\dot{v}) = \frac{1}{2}\left(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2\right)$$

We can write this as

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix}\begin{pmatrix} \ddot{u} \\ \ddot{v} \end{pmatrix} = R$$

where R is composed of smooth functions of u, v. The matrix on the left hand side is invertible, and the inverse map $A \mapsto A^{-1}$ on matrices is smooth. Hence, we can write the geodesic equations in the form

$$\ddot{u} = A(u, v, \dot{u}, \dot{v}); \quad \ddot{v} = B(u, v, \dot{u}, \dot{v})$$

In the usual way we can turn second-order equations into first-order equations by introducing $p = \dot{u}, q = \dot{v}$, and we find

$$\dot{u} = p; \quad \dot{v} = q; \quad \dot{p} = A(u, v, p, q); \quad \dot{q} = B(u, v, p, q)$$

This is a system of first-order ordinary differential equations as governed by Picard's theorem. Since A, B are smooth, a local bound on $\|DA\|$ and $\|DB\|$ will give the required Lipschitz condition.

Corollary. Let Σ be a smooth surface in $\mathbb{R}^3$. For $p \in \Sigma$ and $v \in T_p\Sigma$ nonzero, then there exists $\varepsilon > 0$ and a geodesic $\gamma : [0, \varepsilon) \to \Sigma$ such that

$$\gamma(0) = p; \quad \dot{\gamma}(0) = v$$

Moreover, this geodesic depends smoothly on p, v.

The local existence of geodesics gives rise to allowable parametrisations of Σ with 'nice' properties in terms of the first fundamental form. Let $p \in \Sigma$, and consider a geodesic arc γ starting at p and parametrised by arc length. At each point $\gamma(t)$ for small $t > 0$, we can consider a

geodesic arc γ_t starting at $\gamma(t)$, and $\gamma_t'(0)$ is orthogonal to $\gamma'(t)$, and also parametrised by arc length. Now, we define $\sigma(u,v) = \gamma_v(u)$, which is defined for $u \in [0,\varepsilon)$ and $v \in [0,\delta)$.

Lemma. For ε, δ sufficiently small, $\sigma : (u,v) \mapsto \gamma_v(u)$ defines an allowable parametrisation of an open set in Σ, taking the interior of the domain.

Proof. Smoothness follows from the addendum to Picard's theorem above. At the origin $(0,0)$, by construction we have σ_u, σ_v orthogonal. Hence, they stay linearly independent for sufficiently small ε, δ. So $D\sigma$ has full rank, and (on a smaller set if necessary) σ is injective. So σ is allowable. □

Corollary. Any smooth surface Σ in $\mathbb{R}^3$ admits local parametrisations for which the first fundamental form has form $\mathrm{d}u^2 + G(u,v)\,\mathrm{d}v^2$, so $E = 1$ and $F = 0$.

Proof. Consider the parametrisation $\sigma(u,v) = \gamma_v(u)$. For v_0 fixed, the curve $u \mapsto \gamma_{v_0}(u)$ is a geodesic parametrised at unit speed, so $E = 1$. One of the geodesic equations is

$$\frac{\mathrm{d}}{\mathrm{d}t}(F\dot{u} + G\dot{v}) = \frac{1}{2}\left(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2\right)$$

and consider $v(t) = v_0$, $u(t) = t$. $E_v = \dot{v} = 0$ and $\dot{u} = 1$, so

$$\frac{\mathrm{d}}{\mathrm{d}t}F = 0 \implies F_u\dot{u} = 0 \implies F_u = 0$$

So F is independent of u. At $u = 0$, then by construction of γ_v as being orthogonal to γ at $\gamma(v)$, we see $F = 0$. □

These coordinates are called *geodesic normal coordinates*. Note that by fixing u and letting v vary, the curve obtained is typically not a geodesic, except for $u = 0$ which is γ itself. In these coordinates, we can also find

$$G(0,v) = 1; \quad G_u(0,v) = 0$$

The first result holds since σ_v has unit length at $u = 0$. The second result holds because $u = 0$ yields a geodesic with arc length parametrisation, and then we can use one of the geodesic equations to find

$$\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u} + F\dot{v}) = \frac{1}{2}\left(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2\right) \implies 0 = \frac{1}{2}G_u(0,v)$$

5.10. Surfaces of constant curvature

In the example sheets, we show that for a smooth surface Σ in $\mathbb{R}^3$ with allowable parametrisation for which $E = 1$ and $F = 0$, we have the following result for the Gauss curvature.

$$\kappa = \frac{-\left(\sqrt{G}\right)_{uu}}{\sqrt{G}}$$

If $a : \mathbb{R}^3 \to \mathbb{R}^3$ is a dilation $a(x, y, z) = (ax, ay, az)$, then

$$\kappa_{a(\Sigma)} = \frac{1}{a^2}\kappa_\Sigma$$

since E, F, G rescale by a^2, and L, N, M rescale by a. This matches the results previously found for spheres of varying radii. By dilating, to understand surfaces of constant curvature it suffices to consider surfaces with constant curvature ± 1 or 0.

Proposition. Let Σ be a smooth surface in $\mathbb{R}^3$. Then,

(i) if $\kappa \equiv 0$, then Σ is locally isometric to $(\mathbb{R}^2, \mathrm{d}u^2 + \mathrm{d}v^2)$;

(ii) if $\kappa \equiv 1$, then Σ is locally isometric to $(S^2, \mathrm{d}u^2 + \cos^2 u\, \mathrm{d}v^2)$.

Proof. Σ admits an allowable parametrisation with $E = 1$ and $F = 0$ by using geodesic normal coordinates, so

$$\kappa = \frac{-\sqrt{G}_{uu}}{\sqrt{G}}; \quad G(0, v) = 1; \quad G_u(0, v) = 0$$

If $\kappa \equiv 0$, we have $\sqrt{G}_{uu} = 0$, so $\sqrt{G} = A(v)u + B(v)$, and the boundary conditions give $A \equiv 0, B \equiv 1$. In particular, $G \equiv 1$. The fundamental form then is $\mathrm{d}u^2 + \mathrm{d}v^2$, which is that of $\mathbb{R}^2$.

If $\kappa \equiv 1$, we find $\left(\sqrt{G}\right)_{uu} + \sqrt{G} = 0$ so $\sqrt{G} = A(v)\sin u + B(v)\cos u$. The boundary conditions then imply that $A \equiv 0, B \equiv 1$ and hence the fundamental form is $\mathrm{d}u^2 + \cos^2 u\, \mathrm{d}v^2$. This matches the first fundamental form of a sphere with parametrisation

$$\sigma(u, v) = (\cos u \cos v, \cos u \sin v, \sin u)$$

□

Remark. If $\kappa \equiv -1$, we will find the first fundamental form $\mathrm{d}u^2 + \cosh^2 u\, \mathrm{d}v^2$. There exists an object known as the tractoid, which is a smooth surface in $\mathbb{R}^3$, and has this first fundamental form. We could alternatively choose not to embed this surface in $\mathbb{R}^3$.

In fact, the change of variables $v = e^v \tanh u, w = e^v \operatorname{sech} u$ turns the fundamental form $\mathrm{d}u^2 + \cosh^2 u\, \mathrm{d}v^2$ into $\frac{\mathrm{d}V^2 + \mathrm{d}W^2}{W^2}$, which is a 'standard' presentation of the first fundamental form, which we will see more of later.

6. Riemannian metrics

6.1. Definitions

Definition. Let $V \subseteq \mathbb{R}^2$ be an open set. An *(abstract) Riemannian metric* is a smooth map from V to the set of positive definite symmetric bilinear forms, given by

$$v \mapsto \begin{pmatrix} E(v) & F(v) \\ F(v) & G(v) \end{pmatrix}$$

such that $E > 0$, $G > 0$, $EG - F^2 > 0$. The image of this map can be viewed as an open subset of $\mathbb{R}^4$.

If v is a vector at $p \in V$, we can compute its infinitesimal length by

$$\|v\|^2 = v^\intercal \begin{pmatrix} E(v) & F(v) \\ F(v) & G(v) \end{pmatrix} v$$

Thus, if $\gamma : [a, b] \to V$ is smooth,

$$\text{length}(\gamma) = \int_a^b \left(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2\right)^{\frac{1}{2}} \mathrm{d}t$$

where $\gamma(t) = (u(t), v(t))$.

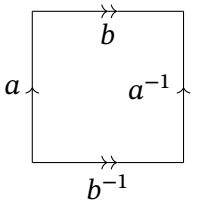
Definition. Let Σ be an abstract smooth surface, so $\Sigma = \bigcup_{i \in I} U_i$ for open sets U_i, with charts $\varphi_i : U_i \to V_i \subseteq \mathbb{R}^2$ which are homeomorphisms, and with smooth transition maps $\varphi_i \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$. A *Riemannian metric* on Σ, usually called g or $\mathrm{d}s^2$, is a choice of Riemannian metric in the above sense on each V_i, which are compatible in the following sense. Let $\sigma = \varphi_i^{-1}$ and $\tilde{\sigma} = \varphi_j^{-1}$ for some i, j, and define $f = \tilde{\sigma}^{-1} \circ \sigma$. Then we require

$$(Df)^\intercal \begin{pmatrix} \widetilde{E} & \widetilde{F} \\ \widetilde{F} & \widetilde{G} \end{pmatrix} (Df) = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

So Df defines an isometry from an open set in the chart $(U, \varphi(U) = V)$ to one in the chart $\left(\widetilde{U}, \widetilde{\varphi}(\widetilde{U}) = \widetilde{V}\right)$.

This compatibility condition is the transition law for first fundamental forms for smooth surfaces in $\mathbb{R}^3$.

Example. Recall the torus $T^2 = {}^{\mathbb{R}^2}\!/_{\mathbb{Z}^2}$.



We have an atlas of charts for which the transition maps are the restrictions of translations of open subsets of $\mathbb{R}^2$. For each $V_i \subseteq \mathbb{R}^2$, we associate the natural Euclidean metric $\mathrm{d}u^2 + \mathrm{d}v^2$. If f is a translation, Df is the identity, and so

$$(Df)^\intercal I(Df) = I$$

holds trivially. So this gives a global Riemannian metric on T^2. This metric is flat, since it is locally isometric to $\mathbb{R}^2$ at all points.

Conversely, consider the torus of revolution embedded in $\mathbb{R}^3$. As a compact smooth surface in $\mathbb{R}^3$, it must contain an elliptic point. Hence, the flat Riemannian metric described above is not the same (up to isometry) as the metric obtained by any possible embedding of the torus in $\mathbb{R}^3$.

The real projective plane $\mathbb{RP}^2$ admits a Riemannian metric with constant curvature $+1$. We have constructed a smooth atlas for $\mathbb{RP}^2$ where the charts were of the form (U, φ), with $U = q\hat{U}$ and $q\colon S^2 \to \mathbb{RP}^2$ the quotient map, $\hat{U} \subseteq S^2$ open and contained within an open hemisphere, and $\varphi\colon U\colon U \to V \subseteq \mathbb{R}^2$ is given by $\hat{\varphi} \circ q^{-1}|_U$ and $\hat{\varphi}\colon \hat{U} \to V$ a chart on S^2. The transition maps for this atlas were found to be locally the identity, or induced from the antipodal map. The antipodal map from S^2 to S^2 is an isometry, so both types of transition maps preserve the usual round metric on S^2.

In the first example sheet, we consider the Klein bottle. This has an atlas such that all transition maps are either translations or translations composed with a reflection. These preserve the flat metric in $\mathbb{R}^2$, so the Klein bottle inherits a flat Riemannian metric. The Klein bottle and $\mathbb{RP}^2$ are not embedded in $\mathbb{R}^3$, so we could not construct a 'non-abstract' Riemannian metric.

Definition. Let $(\Sigma_1, g_1), (\Sigma_2, g_2)$ be abstract smooth surfaces with abstract Riemannian metrics. A diffeomorphism $f\colon \Sigma_1 \to \Sigma_2$ is an *isometry* if it preserves the lengths of all curves, where lengths are taken with respect to these abstract Riemannian metrics.

Example. If (Σ_2, g_2) is given, and $f\colon \Sigma_1 \to \Sigma_2$ is a diffeomorphism, we can equip Σ_1 with a metric known as the *pullback* metric $g_1 = f^* g_2$ that gives that f is an isometry.

6.2. The length metric

Definition. Let (Σ, g) be a connected abstract smooth surface with an abstract Riemannian metric. The *length metric* is defined by

$$d_g(p, q) = \inf_\gamma L(\gamma)$$

where γ varies over piecewise smooth paths in Σ from p to q, and L is length computed using g.

Proposition. Let (Σ, g) be a connected abstract smooth surface with an abstract Riemannian metric. Then d_g is indeed a metric, and d_g induces a topology on Σ that agrees with the given topology.

Proof. Let $p, q \in \Sigma$. We will show that there exists some piecewise smooth path γ from p to q, so $d_g(p,q)$ is well-defined and finite. Connected surfaces are path-connected. There exists a continuous path γ and a finite set of charts (U_i, φ_i) with associated parametrisations $\sigma_i = \varphi_i^{-1} : V_i \to U_i \subset \Sigma$ such that $\operatorname{Im}\gamma \subseteq \bigcup_{i=1}^{N} U_i$. Consider points

$$p = x_0 \in U_1, x_1 \in U_1 \cap U_2, x_2 \in U_2 \in U_3, \dots, q = x_N \in U_N$$

Smooth paths in V_i from $\varphi_i(x_i)$ to $\varphi_{i+1}(x_{i+1})$ exist, since smooth paths between two points in $\mathbb{R}^2$ exist. Since the atlas is smooth, being a smooth path in some U_i is the same as being smooth in U_{i+1} whenever U_i and U_{i+1} intersect, since the transition maps are smooth. So $p, q \in \Sigma$ are joined by some piecewise smooth path.

For any piecewise smooth path from p to q there exists the inverse path parametrised in the opposite direction, which has the same length. We can also concatenate paths from p to q and from q to r, with length equal to the sum of the lengths. In both cases, the new paths are piecewise smooth. This then implies that d_g is symmetric, and satisfies the triangle inequality.

To show d_g is a metric, it now suffices to show that $d_g(p,q) = 0$ implies $p = q$, since the converse is trivial. Let $p \in \Sigma$ and fix a chart (U, φ) at p. Without loss of generality let $V = B(0,1)$, and $\varphi(p) = 0$. If $q \neq p \in \Sigma$, there exists $\varepsilon > 0$ such that $q \notin \varphi^{-1}\left(\overline{B(0,\varepsilon)}\right)$. Suppose $\gamma : [0,1] \to \Sigma$ is a piecewise smooth path from p to q. Certainly, γ must escape the disc $\varphi^{-1}\left(\overline{B(0,\varepsilon)}\right)$, since it must reach q. Length along paths is additive, so by the triangle inequality, it suffices to show that there exists $\delta > 0$ such that $d_g(p,r) > \delta$ for all $r \in \partial\varphi^{-1}\left(\overline{B(0,\varepsilon)}\right) = \varphi^{-1}\{\text{circle of radius } \varepsilon\}$. The data on the Riemannian metric g includes the non-degenerate symmetric bilinear form $\begin{pmatrix} E_z & F_z \\ F_z & G_z \end{pmatrix}$ for all $z \in \overline{B(0,\varepsilon)} \subseteq V$. We also have the usual Euclidean inner product on the disc, $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. For all $z \in \overline{B(0,\varepsilon)}$, these matrices are positive definite. Since $\overline{B(0,\varepsilon)}$ is compact, there exists $\delta > 0$ such that $\begin{pmatrix} E_z - \delta & F_z \\ F_z & G_z - \delta \end{pmatrix}$ is still positive definite for all $z \in \overline{B(0,\varepsilon)}$. In other words, the determinant $EG - F^2 > 0$ for all $z \in \overline{B(0,\varepsilon)}$, which is compact, so it is bounded below by some positive number. Hence, $\text{length}_g(\hat{\gamma}) \geq \text{length}_{\delta\cdot\text{Euclidean}}(\hat{\gamma})$ for any $\hat{\gamma}$ contained withing $\overline{B(o,\varepsilon)}$. Taking $\hat{\gamma} = \varphi\left[\gamma \cap \varphi^{-1}\left(\overline{B(o,\varepsilon)}\right)\right]$, which is the part of γ in $\overline{B(0,\varepsilon)}$ with respect to the chart, we have that $\text{length}_{\delta\cdot\text{Euclidean}}(\hat{\gamma}) \geq \delta\varepsilon$, so $d_g(p,q) \geq \delta\varepsilon$. □

Remark. The last step of the argument for the proof above, comparing the inner products $\begin{pmatrix} E_z & F_z \\ F_z & G_z \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ can be modified to show that d_g induces a topology on Σ that agrees with the given topology, which is given by local homeomorphisms to $\mathbb{R}^2$ everywhere.

6.3. The hyperbolic metric

Definition. Let

$$D = B(0,1) = \{z \in \mathbb{C} : |z| < 1\}$$

The abstract Riemannian metric g_{hyp} on D is given by

$$\frac{4(\mathrm{d}u^2 + \mathrm{d}v^2)}{(1-u^2-v^2)^2} = \frac{4|\mathrm{d}z|^2}{\left(1-|z|^2\right)^2}$$

Since there is only one chart, this holds for all of D. In particular, if $\gamma : [0,1] \to D$ is smooth, then

$$L_{g_{\text{hyp}}}(\gamma) = 2\int_0^1 \frac{|\dot\gamma(t)|}{1-|\gamma(t)|^2}\,\mathrm{d}t$$

If $\gamma(t) = (u(t), v(t))$, we can write

$$L(\gamma) = 2\int_0^1 \frac{\left(\dot u^2 + \dot v^2\right)^{\frac{1}{2}}}{1-u^2-v^2}\,\mathrm{d}t$$

This is very similar to a first fundamental form with $E = G = \frac{4}{(1-u^2-v^2)^2}$ and $F = 0$, but we do not claim that this fundamental form arises from an embedding in $\mathbb{R}^3$.

Note that the flat metric on $\mathbb{R}^2$ and the usual round metric on S^2 have large and transitive isometry groups. We will show that this metric also induces a large symmetry group, which is induced by the Möbius group. Recall that

$$\text{Möb} = \left\{z \mapsto \frac{az+b}{cz+d} : \begin{pmatrix} a & b \\ c & d\end{pmatrix} \in GL(2,\mathbb{C})\right\} \curvearrowright \mathbb{C} \cup \{\infty\}$$

Lemma. The subgroup of the Möbius group that preserves D,

$$\text{Möb}(D) = \{T \in \text{Möb} : T(D) = D\}$$

is also given by

$$\text{Möb}(D) = \left\{z \mapsto e^{i\theta}\frac{z-a}{1-\overline{a}z} : |a| < 1\right\} = \left\{\begin{pmatrix} a & b \\ \overline{b} & \overline{a}\end{pmatrix} \in \text{Möb} : |a|^2 - |b|^2 = 1\right\}$$

Proof. Note that

$$\begin{aligned}
\left|\frac{z-a}{1-\overline{a}z}\right| = 1 &\iff (z-a)(\overline{z}-\overline{a}) = (1-\overline{a}z)(1-a\overline{z}) \\
&\iff z\overline{z} - a\overline{z} - \overline{a}z + a\overline{a} = 1 - \overline{a}z - a\overline{z} + a\overline{a}z\overline{z} \\
&\iff |z|^2\left(1-|a|^2\right) = 1 - |a|^2 \\
&\iff |z| = 1
\end{aligned}$$

So these maps of the form

$$z \mapsto e^{i\theta}\frac{z-a}{1-\overline{a}z}$$

do indeed preserve the unit circle, and $a \in D$ is mapped to $0 \in D$. Hence, it preserves the entire disc. □

Lemma. The Riemannian metric g_{hyp} is invariant under Möb(D). In other words, the Möbius group Möb(D) acts by isometries on D.

Proof. Möb(D) is generated by $z \mapsto e^{i\theta}$ and $z \mapsto \frac{z-a}{1-\overline{a}z}$. The rotations preserve g_{hyp}, since it depends only on $|z|$ and not z itself. For the second type of transformation, let $w = \frac{z-a}{1-\overline{a}z}$. Here,

$$\mathrm{d}w = \frac{\mathrm{d}z}{1-\overline{a}z} + \frac{z-a}{(1-\overline{a}z)^2}\overline{a}\,\mathrm{d}z = \frac{\mathrm{d}z}{(1-\overline{a}z)^2}\left(1-|a|^2\right)$$

Then,

$$\frac{|\mathrm{d}w|}{1-|w|^2} = \frac{|\mathrm{d}z|}{|1-\overline{a}z|^2}\cdot\frac{1-|a|^2}{1-\left|\frac{z-a}{1-\overline{a}z}\right|^2} = \frac{|\mathrm{d}z|\left(1-|a|^2\right)}{|1-\overline{a}z|^2-|z-a|^2} = \frac{|\mathrm{d}z|}{1-|z|^2}$$

Hence the hyperbolic metric, which is a function of this $\frac{|\mathrm{d}z|}{1-|z|^2}$, is also invariant under this change of variables. □

Lemma. On (D, g_{hyp}),

(i) every pair of points in (D, g_{hyp}) is joined by a unique geodesic up to reparametrisation;

(ii) the geodesics are diameters of the disc and circular arcs orthogonal to the boundary ∂D.

The *whole* geodesics (ones that are defined on $\mathbb{R}$) are called *hyperbolic lines*.

Proof. Let $a \in \mathbb{R}_+ \cap D$ and γ a smooth path from the origin to a. Let $\gamma(t) = (u(t), v(t))$. Note that $\mathrm{Re}(\gamma)(t) = (u(t), 0)$ is also a smooth path from the origin to a. By definition of the hyperbolic metric,

$$\text{length}(\gamma) = \int_0^1 \frac{2|\dot\gamma|}{1-|\gamma|^2}\,\mathrm{d}t = \int_0^1 \frac{2\sqrt{\dot u^2+\dot v^2}}{1-u^2-v^2}\,\mathrm{d}t \geq \int_0^1 \frac{2|\dot u|}{1-u^2}\,\mathrm{d}t$$

where equality holds if and only if $\dot v \equiv 0$, and so $v \equiv 0$.

$$\text{length}(\gamma) \geq \int_0^1 \frac{2\dot u}{1-u^2}\,\mathrm{d}t$$

where equality holds in this expression if and only if u is monotonic. Hence, the arc of the diameter, parametrised monotonically, is a globally length-minimising path, and hence a geodesic. We can compute this integral to be

$$\text{length}(\gamma) = 2\,\text{artanh}\,a$$

Now, 0 and a in $\mathbb{R}_+ \cap D$ are joined by a unique geodesic, and Möb(D) acts transitively and by isometries, and can be used to send any two points $p, q \in D$ to $0, a \in \mathbb{R}_+ \cap D$. So every pair of points must be joined by a unique geodesic. Since Möbius maps send circles to circles, and they preserve angles and hence orthogonality to the boundary, we must have that all geodesics are diameters or circular arcs orthogonal to ∂D. □

Corollary. If $p, q \in D$, then the distance between them is

$$d_{\text{hyp}}(p, q) = 2 \operatorname{artanh} \left| \frac{p - q}{1 - \overline{p}q} \right|$$

6.4. The hyperbolic upper half-plane

Definition. The *hyperbolic upper half-plane* $(\mathfrak{h}, g_{\text{hyp}})$ is the set

$$\mathfrak{h} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$$

with the abstract Riemannian metric

$$\frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2} = \frac{|\mathrm{d}z|^2}{(\operatorname{Im} z)^2}$$

Lemma. The hyperbolic disc (D^2, g_{hyp}) and the hyperbolic upper half-plane $(\mathfrak{h}, g_{\text{hyp}})$ are isometric.

Proof. There exist maps $T : \mathfrak{h} \to D$ and $\widetilde{T} : D \to \mathfrak{h}$ given by

$$T(w) = \frac{w - i}{w + i}; \quad \widetilde{T}(z) = i\left(\frac{1 - z}{1 + z}\right)$$

which are inverse diffeomorphisms. Here,

$$T'(w) = \frac{1}{w + i} - \frac{w - i}{(w + i)^2} = \frac{2i}{(w + i)^2}$$

Considering $T(w) = z \in D$,

$$\frac{|\mathrm{d}z|}{1 - |z|^2} = \frac{|\mathrm{d}(Tw)|}{1 - |Tw|^2} = \frac{|T'(w)||\mathrm{d}w|}{1 - |Tw|^2} = \frac{2|\mathrm{d}w|}{|w + i|^2\left(1 - \left|\frac{w-i}{w+i}\right|^2\right)} = \frac{|\mathrm{d}w|}{2 \operatorname{Im} w}$$

Hence, under this coordinate change,

$$\frac{4|\mathrm{d}z|^2}{\left(1 - |z|^2\right)^2}$$

is the metric obtained under pullback by T from $\frac{\mathrm{d}w^2}{(\operatorname{Im} w)^2}$. □

Corollary. The hyperbolic upper half-plane is globally isometric to the hyperbolic disc, so every pair of points is joined by a unique geodesic, up to reparametrisation. The geodesics are arcs of circles orthogonal to the boundary, which are vertical straight lines and semicircles centred on a point in the real axis.

Proof. The isometry between $\mathfrak{h} \to D$ is given by a Möbius map. In particular, $\mathbb{R} \cup \{\infty\} \mapsto \partial D$, and Möbius maps preserve circles and orthogonality. □

Remark. When we discussed surfaces in $\mathbb{R}^3$ with constant Gauss curvature, we saw that if a surface had constant Gauss curvature, its first fundamental form in geodesic normal coordinates was of the form $\mathrm{d}u^2 + \cosh^2 \mathrm{d}v^2$, with a change of variables taking this form to $\frac{\mathrm{d}v^2+\mathrm{d}w^2}{w^2}$. This is exactly the form of the Riemannian metric on the hyperbolic upper half-plane. Gauss' *theorema egregium* implies that Gauss curvature makes sense for an abstract Riemannian metric, since it only depends on geodesics and hence the first fundamental form. We can therefore define the Gauss curvature for an abstract Riemannian metric to agree with this definition for surfaces in $\mathbb{R}^3$. Under this definition, we can show that the hyperbolic upper half-plane has constant curvature -1, and hence so does the disc.

Suppose we wanted to find a metric $d : D \times D \to \mathbb{R}_{\geq 0}$ on D^2 with the properties that it is invariant under the Möbius group $\mathrm{Möb}(D)$, and that the real diameter is length-minimising. By Möbius invariance, the distance between any two points is completely determined by knowing the distance from the origin to some point on the positive real axis a, which we will denote $p(a) = d(0, a)$. If $\mathbb{R}_+ \cap D$ is length-minimising, distance should be additive, so if $0 \leq a \leq b \leq 1$ we should have $d(0, a) + d(a, b) = d(0, b)$ so $d(a, b) = p\left(\frac{b-a}{1-ab}\right) = p(b) - p(a)$. If we furthermore constrain p to be differentiable, and we differentiate the above expression with respect to b and set $b = a$, we find the differential equation

$$p'(a) = \frac{p'(0)}{1 - a^2}$$

Hence, $p(a)$ is some constant multiple of $\operatorname{artanh} a$, since $p'(0)$ can be chosen freely. So, up to rescaling the length metric associated to g_{hyp} on D is the unique metric with these properties. The scale is chosen for g_{hyp} to enforce that the curvature is -1 precisely.

6.5. Isometries of hyperbolic space

We now would like to understand the full isometry group of the disc (D, g_{hyp}) or $(\mathfrak{h}, g_{\mathrm{hyp}})$. We will show that this group is precisely $\mathrm{Möb}(D)$ together with reflections in hyperbolic lines, which are called *inversions*.

Definition. Let $\Gamma \subseteq \widehat{\mathbb{C}}$ be a circle or line. We say that points $z, z' \in \widehat{\mathbb{C}}$ are *inverse for* Γ if every circle through z orthogonal to Γ also passes through z'.

Lemma. Such inverse points exist and are unique.

Proof. Recall that Möbius maps preserve circles in $\widehat{\mathbb{C}}$ and preserve angles. In particular, if z, z' are inverse for Γ and $T \in \text{Möb}$, then Tz and Tz' are inverse for the circle $T(\gamma)$. If $\Gamma = \mathbb{R} \cup \{\infty\}$, then $J(z) = \overline{z}$ gives inverse points; this map satisfies the definition above. Now, if $\Gamma \subseteq \widehat{\mathbb{C}}$ is any circle, there exists $T \in \text{Möb}$ such that $T(\mathbb{R} \cup \{\infty\}) = \Gamma$. We can therefore define inversion in Γ to be $J_\Gamma = T \circ (z \mapsto \overline{z}) \circ T^{-1}$. □

Definition. The map J_Γ in the proof above, sending z to the unique inverse point z' for z with respect to Γ, is called *inversion* in Γ.

This map fixes all points of Γ, and swaps points on the interior with points on the exterior.

Example. For Γ a straight line, this is simply reflection. For the unit circle, S^1, the map J_{S^1} maps $z \mapsto \frac{1}{\overline{z}}$ and $0 \mapsto \infty$.

Remark. The composition of two inversions is a Möbius map. Let C be the conjugation map $z \mapsto \overline{z}$, which is $J_{\mathbb{R}\cup\{\infty\}}$. If $\Gamma \subseteq \widehat{\mathbb{C}}$ is any circle, we have $J_\Gamma = T \circ C \circ T^{-1}$ where T is the Möbius transformation which maps $\mathbb{R} \cup \{\infty\}$ to Γ. If Γ_1, Γ_2 are circles, and T_1, T_2 are the transformations from $\mathbb{R} \cup \{\infty\}$ to Γ_1, Γ_2 respectively, then

$$\begin{aligned} J_{\Gamma_1} \circ J_{\Gamma_2} &= \left(J_{\Gamma_1} \circ C\right) \circ \left(C \circ J_{\Gamma_1}\right) \\ &= \left(C \circ J_{\Gamma_1}\right)^{-1} \circ \left(C \circ J_{\Gamma_1}\right) \end{aligned}$$

We have $C \circ J_\Gamma = C \circ T \circ C \circ T^{-1}$, so it suffices to show $C \circ T \circ C \in \text{Möb}$. If $T(z) = \frac{az+b}{cz+d}$, we have

$$(C \circ T \circ C)(z) = \frac{\overline{a}z + \overline{b}}{\overline{c}z + \overline{d}} \in \text{Möb}$$

Lemma. An orientation-preserving isometry of $(\mathbb{H}^2, g_{\text{hyp}})$ is an element of Möb($\mathbb{H}$), where $\mathbb{H}$ is D or $\mathfrak{h}$. The full isometry group is generated by inversions in hyperbolic lines.

Proof. It suffices to prove this in either model, so we will use the disc model. Inversion in the geodesic $\mathbb{R} \cap D$ is conjugation, which preserves g_{hyp}. Note that Möb($\mathbb{H}$) acts transitively by isometries on geodesics. Hence, if inversion in one geodesic preserves the metric, so does inversion in any geodesic.

Now, suppose α is some isometry of the hyperbolic disc D under the metric g_{hyp}. We have $\alpha(0) = a \in D$, and using $z \mapsto \frac{z-a}{1-\overline{a}z}$, so there exists $T \in \text{Möb}(D)$ such that $T \circ \alpha$ fixes the origin. There exists a rotation $R \in \text{Möb}(D)$ such that $R \circ T \circ \alpha$ maps $D \cap \mathbb{R}_+$ to itself. Composing with the conjugation map C if necessary, there exists an isometry A which is an inversion composed with a Möbius map such that $A \circ \alpha$ fixes $D \cap \mathbb{R}$ pointwise and fixes $D \cap i\mathbb{R}$ pointwise. The only such isometry is the identity, since every point in D is determined by its distance to these two lines. Hence, A is the inverse of α.

If α preserves orientation and fixes $\mathbb{R} \cap D$, then it necessarily fixes $i\mathbb{R} \cap D$ pointwise, so $\alpha = (R \circ T)^{-1} \in \text{Möb}$. In general, α was constructed from Möb($\mathbb{H}$) and inversions in hyperbolic lines.

So to show that the isometry group is generated by inversions, it suffices to show that all Möbius maps are compositions of inversions. This is presented on the example sheets. □

In the upper half-plane model of hyperbolic space,

$$\mathrm{Möb}(\mathfrak{h}) = \mathbb{P}SL(2,\mathbb{R}) = \left\{ z \mapsto \frac{az+b}{cz+d} : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{R}) \right\}; \quad d_{\mathrm{hyp}} = 2\,\mathrm{artanh}\left|\frac{b-a}{b-\overline{a}}\right|$$

6.6. Hyperbolic triangles

Definition. Let α be an orientation-preserving isometry of $\mathbb{H}$, which is equivalently an element of Möb($\mathbb{H}$). Suppose α is not the identity map. We say that α is

(i) *elliptic*, if α fixes some point $p \in \mathbb{H}$ (if $p = 0 \in D$, this behaves like a rotation);

(ii) *parabolic*, if α fixes a unique point $p \in \partial\mathbb{H}$ (if $p = \infty \in \mathfrak{h}$, this behaves like a translation);

(iii) *hyperbolic*, if α fixes two points on $\partial\mathbb{H}$, so it fixes the unique geodesic between these two points setwise, and so α must translate points across the geodesic; it is not an inversion in the geodesic because it is not the identity map.

All elements of Möb($\mathbb{H}$) are either elliptic, parabolic, or hyperbolic.

Definition. Let ℓ, ℓ' be hyperbolic lines. Then, we say

(i) *parallel*, if they meet at the boundary $\partial\mathbb{H}$ but never inside $\mathbb{H}$;

(ii) *ultra-parallel*, if they never meet in $\overline{\mathbb{H}}$;

(iii) *intersecting*, if they meet in $\mathbb{H}$.

All pairs of hyperbolic lines are either parallel, ultra-parallel, or intersecting. A *hyperbolic triangle* is a region bound by three geodesics, no two of which are ultra-parallel. Vertices that lie 'at infinity' (on $\partial\mathbb{H}$) are called *ideal vertices*.

Note that the points in $\partial\mathbb{H}$ are not contained within the hyperbolic plane, so in particular the ideal vertices are not points in $\mathbb{H}$. We typically denote side lengths by A, B, C, and denote the angles opposite these sides by α, β, γ. The vertices at α, β, γ are denoted a, b, c. The hyperbolic metric is conformal, since $E = G$ and $F = 0$. Hence, we can use Euclidean angles in place of hyperbolic angles.

Proposition (hyperbolic cosine formula)**.** For a hyperbolic triangle,

$$\cosh C = \cosh A \cosh B - \sinh A \sinh B \cos\gamma$$

Proof. To simplify, by an isometry we can let the vertex c at γ be placed at $0 \in D$, and the vertex b at β be placed at $\mathbb{R}_+ \cap D$. Hence, the sides A, B are straight Euclidean line segments

in D, and the angle between them is γ. We have

$$d_{\text{hyp}}(0,a) = 2\operatorname{artanh} a \implies a = \tanh\frac{A}{2};\quad b = e^{i\gamma}\tanh\frac{B}{2};\quad \left|\frac{b-a}{1-\overline{a}b}\right| = \tanh\frac{C}{2}$$

Recall that

$$t = \tanh\frac{\lambda}{2} \implies \cosh\lambda = \frac{1+t^2}{1-t^2};\quad \sinh\lambda = \frac{2t}{1-t^2}$$

Hence,

$$\cosh A = \frac{1+|a|^2}{1-|a|^2};\quad \cosh B = \frac{1+|b|^2}{1-|b|^2};$$

$$\cosh C = \frac{|1-\overline{a}b|^2 + |b-a|^2}{|1-\overline{a}b|^2 - |b-a|^2} = \frac{\left(1+|s|^2\right)\left(1+|b|^2\right) - 2\left(\overline{a}b + a\overline{b}\right)}{\left(1-|a|^2\right)\left(1-|b|^2\right)}$$

Note that $a \in \mathbb{R}$ and $b + \overline{b} = 2\operatorname{Re} b = 2b\cos\gamma$, so

$$\cosh C = \cosh A\cosh B - \sinh A\sinh B\cos\gamma$$

as required. □

Remark. If A, B, C are small, the standard approximations to the hyperbolic sine and cosine functions give

$$C^2 \approx A^2 + B^2 - 2AB\cos\gamma$$

which is the Euclidean cosine formula. Since a dilation of a surface in $\mathbb{R}^3$ rescales curvature, at small scales we can treat any abstract smooth surface with a Riemannian metric as flat.

Since $\cos\gamma \geq -1$, we have that

$$\cosh C \leq \cosh A\cosh B + \sinh A\sinh B = \cosh(A+B)$$

The hyperbolic cosine is increasing, so $C \leq A + B$. This is a more precise variant of the hyperbolic triangle inequality.

6.7. Area of triangles

Theorem. Let $T \subseteq \mathbb{H}^2$ be a hyperbolic triangle with internal angles α, β, γ defined as before. The area of T is

$$\text{area}_{\text{hyp}}(T) = \pi - \alpha - \beta - \gamma$$

Note that α, β, γ may be zero, so T may have ideal vertices, and the internal angle is zero for such vertices.

This is a version of the Gauss–Bonnet theorem for hyperbolic triangles.

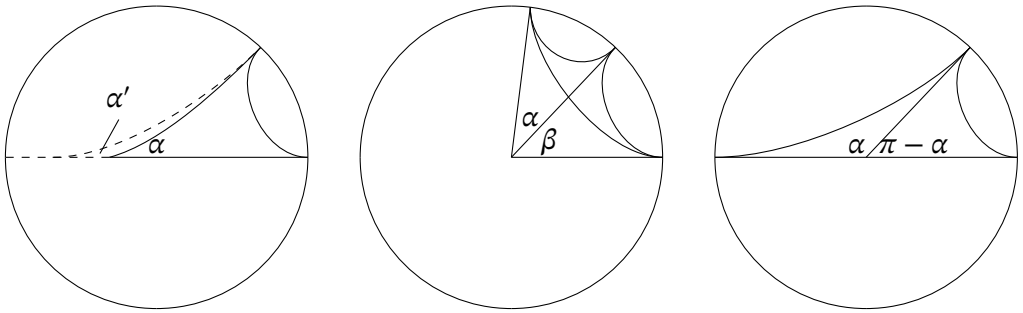
Proof. The Möbius group $\mathrm{Möb}(\mathbb{H}^2)$ acts transitively on triples of points in the boundary with the correct cycle order. In particular, there exists a single ideal triangle (with all vertices at infinity) up to isometry. Consider the ideal triangle in the hyperbolic upper half-plane with vertices $-1, +1, \infty$. Its area is

$$\mathrm{area}_{\mathrm{hyp}}(T) = \int_{-1}^{1} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2} \,\mathrm{d}y \,\mathrm{d}x$$

since $\sqrt{EG - F^2} = \frac{1}{y^2}$. We can compute this explicitly as

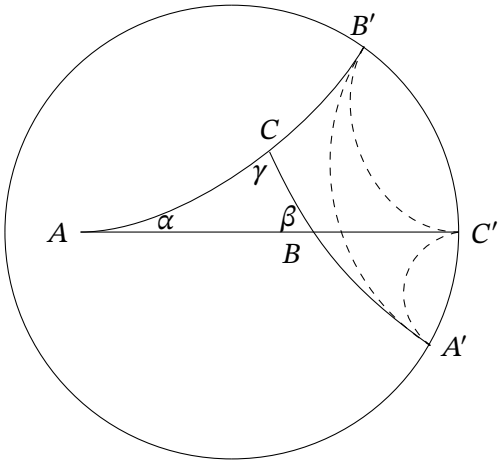
$$\mathrm{area}_{\mathrm{hyp}}(T) = \int_{-1}^{1} \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \pi$$

Now, let $A(\alpha)$ be the area of a triangle with angles $0, 0, \alpha$. We can see that $A(\alpha)$ is decreasing in α and continuous in α, by fixing two ideal vertices in the hyperbolic disc and translating the third vertex.



The first diagram shows that by moving the vertex α on the real line, the area must increase, since the triangle with angle $\alpha' < \alpha$ contains the triangle with angle α. From the second diagram, we see that $A(\alpha) + A(\beta) = A(\alpha + \beta) + \pi$ by comparing the different areas of triangles formed from hyperbolic lines in the diagram. By letting $F(\alpha) = \pi - A(\alpha)$, we have $F(\alpha) + F(\beta) = F(\alpha + \beta)$. Since F is continuous and increasing, we have that $F(\alpha) = \lambda\alpha$ for some fixed $\lambda > 0$. In particular, $A(\alpha) = \pi - \lambda\alpha$. Now, by considering the angles in the third diagram, we see that $A(\alpha) + A(\pi - \alpha) = \pi$. Hence, $\lambda = 1$, and so $A(\alpha) = \pi - \alpha$.

Finally, we consider the general case.



By writing ABC for $\text{area}_{\text{hyp}}(T)$ where T is the triangle with vertices A, B, C, we can see that

$$ABC + A'CB' + A'B'C' = \text{area of interior of diagram} = AB'C' + A'BC'$$

Equivalently,

$$ABC + \pi - (\pi - \gamma) + \pi = (\pi - \alpha) + (\pi - \beta) \implies ABC = \pi - \alpha - \gamma - \beta$$

as required. □

Note that if G is a hyperbolic n-gon, so it is a region bound by n hyperbolic geodesics, it may be decomposed into a union of hyperbolic triangles. Since any two points in $\mathbb{H}^2$ are joined by a unique geodesic, the area of G is given by

$$\text{area}_{\text{hyp}}(G) = (n-2)\pi - \sum_{i=1}^{n} \alpha_i$$

Lemma. If $g \geq 2$, then there exists a regular $4g$-gon in $\mathbb{H}^2$ with internal angle $\frac{2\pi}{4g} = \frac{\pi}{2g}$.

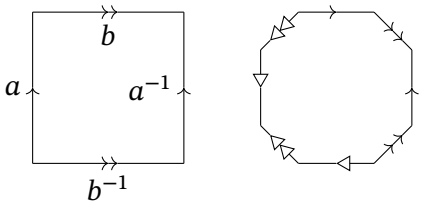
Proof. Consider an ideal $4g$-gon, whose vertices all lie at infinity, in the disc model of hyperbolic space. The ideal vertices can be placed at the $4g$-th roots of unity, such that this polygon is invariant under a rotational symmetry. By sliding each vertex radially inwards in $\mathbb{R}^2$, we obtain a continuous family of regular $4g$-gons, with areas which vary monotonically from $(4g-2)\pi$ to zero. The internal angle of the polygon therefore varies continuously from zero to $\beta_{\min}$ such that $(4g-2)\pi = 4g\beta_{\min}$. It therefore suffices to check that $\frac{\pi}{2g}$ lies in this interval $(0, \beta_{\min})$. □

6.8. Surfaces of constant negative curvature

Theorem. For each $g \geq 2$, there exists an abstract Riemannian metric on the compact surface of genus g with curvature $\kappa \equiv -1$ and locally isometric to $\mathbb{H}^2$.

Recall the the Euler characteristic of a surface of genus g is exactly $2-2g$. Note, if $g = 0$ we can construct a Riemannian metric with $\kappa \equiv +1$ since this is the sphere, and if $g = 1$ we can have $\kappa \equiv 0$ since this is the torus as a quotient $\mathbb{R}^2/\mathbb{Z}^2$. We will outline two proofs.

Proof. Recall that we can construct the torus and double torus by



Analogously, a $4g$-gon with side labels $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \ldots$ gives a surface of genus g.

We say that a *flag* comprises an oriented hyperbolic line, a point on that line, and a choice of side to that line. Given two such flags, there exists a hyperbolic isometry between them. So Möb($\mathbb{H}$) acts transitively on flags. In particular, we can swap the side of a flag using an inversion.

Consider a regular hyperbolic $4g$-gon with internal angle $\frac{\pi}{2g}$. We label this polygon with side labels as above to give a genus g surface. For each paired set of two edges, there exists a hyperbolic isometry taking one to the other, respecting orientations and, and taking the side corresponding to the inside of the polygon to the side corresponding to the outside of the polygon. This is possible since Möb($\mathbb{H}$) acts transitively on flags.

We can now define an atlas for Σ_g as follows.

- If p is in the interior of the polygon P, consider a small disc contained in the interior of the polygon. Then, include this disc into the hyperbolic disc D.
- If p is contained in an edge, let $\hat{p}$ be the corresponding point on the paired edge. We have an isometry γ from edge e_1 to edge e_2, exchanging sides, and mapping p to $\hat{p}$. We can use this to define the chart. Using γ, we can combine U, the intersection of P with an open neighbourhood of p, and $\widetilde{U}$, the intersection of P with an open neighbourhood of $\hat{p}$, such that the chart is an inclusion on U and is γ on $\widetilde{U}$. These agree on $U \cap \widetilde{U}$.
- All $4g$ vertices are identified to one point of Σ, and we need a chart at this point. Using a hyperbolic isometry, let one vertex v of P be at the origin in D, such that an edge e containing v is mapped to a subset of the real line. Since the polygon P has internal angle $\frac{\pi}{2g}$, the angle between $\mathbb{R}$ and the adjacent edge is $\frac{\pi}{2g}$. The fact that the internal angles sum to 2π means that we can construct hyperbolic isometries for each vertex

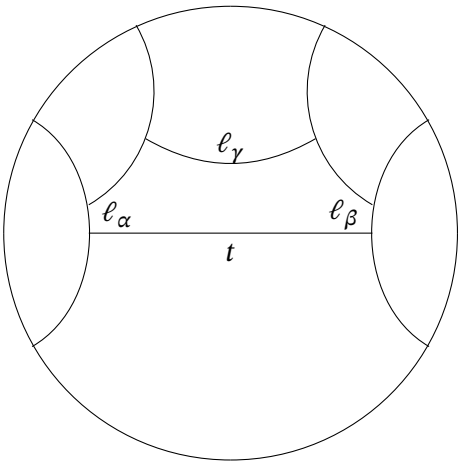
that join them exactly, giving an open neighbourhood of zero in D in the shape of a disc. The chart is defined at $[v] \in \Sigma_g$ by this identification.

All charts are obtained from inclusion or an inclusion composed with a hyperbolic isometry, therefore the transition maps are hyperbolic isometries. In particular, hyperbolic isometries are smooth, and preserve the locally defined hyperbolic metric. □

Remark. The torus can be given by $\mathbb{R}^2/\mathbb{Z}^2$. This characterisation was useful when describing the flat metric, precisely because its charts are easy to define. For Σ_g, we chose $2g$ hyperbolic isometries which paired sides. Hence, there is a group $\Gamma \leq \text{Möb}(\mathbb{H})$, generated by these isometries. In Part II Algebraic Topology, the surface Σ_g will be constructed by $\mathbb{H}/\Gamma$.

Lemma. For each $\ell_\alpha, \ell_\beta, \ell_\gamma > 0$, there exists a right-angled hyperbolic hexagon with side lengths $\ell_\alpha, a, \ell_\beta, b, \ell_\gamma, c$ for some a, b, c.

Proof. Given $t > 0$, there exists a pair of ultra-parallel hyperbolic lines a distance t apart. We show on the fourth example sheet that each pair of ultra-parallel hyperbolic lines has a unique common perpendicular geodesic. Given lengths ℓ_α, ℓ_β, construct new perpendicular geodesics orthogonal to the originals, having moved lengths ℓ_α, ℓ_β from the common perpendicular (in the same direction). If t is made large, the new geodesics $\sigma, \tilde{\sigma}$ can be made ultraparallel. Hence, by making t smaller, there exists a threshold t_0 by continuity such that the new geodesics are parallel. Now, for $t \in (t_0, \infty)$, the two new geodesics are ultra-parallel. So $\sigma, \tilde{\sigma}$ have a unique common perpendicular geodesic. As t increases above t_0, the length of this line increases monotonically from zero to infinity. So there exists a value of $t > t_0$ such that the new common perpendicular has length ℓ_γ.



This is exactly the right-angled hyperbolic hexagon as required. □

Definition. A *pair of pants* is a topological space homeomorphic to the complement of three open discs in S^2.

Note that this space has a boundary. Consider two right-angled hyperbolic hexagons with side lengths $\ell_\alpha, \ell_\beta, \ell_\gamma$ arranged as above. The original configuration of two ultra-parallel geodesics of a distance t apart is unique up to isometry. So the side lengths have a correspondence, and the hexagon with side lengths $\ell_\alpha, \ell_\beta, \ell_\gamma$ is unique up to isometry. Suppose that we glue together the corresponding unknown sides $t_{\alpha\beta}, t_{\beta\gamma}, t_{\gamma\alpha}$ with the same side identifications. Locally near ℓ_α, for instance, we arrive at a closed circle of length $2\ell_\alpha$, extended into a cylindrical shape with two seams $t_{\alpha\beta}, t_{\gamma\alpha}$. Since the hexagons were right-angled, we have constructed a hyperbolic pair of pants. The boundary circles are geodesics in the sense that, for any point on such a circle, the local neighbourhood is a point on a geodesic on a polygon in $\mathbb{H}$.

We will now construct Σ_g using a more flexible approach.

Proof. If P_1, P_2 are two hyperbolic 'surfaces' with geodesic boundaries, and if $\gamma_1 \subset P_1$ and $\gamma_2 \subset P_2$ are boundary circles of the same length (in the hyperbolic metric), we can glue P_1 and P_2 together along this common-length circle. P_1 and P_2 may be glued by any isometry of γ_1, γ_2. The result $P_1 \cup_{\gamma_1 \sim \gamma_2} P_2$ has a hyperbolic metric. For any point $p \in P_i$ not on the boundary γ_i, it already has a suitable open neighbourhood since P_i is hyperbolic. For any point $p \in \gamma_1 \sim \gamma_2$, we have a chart to a small disc in $\mathbb{H}$ using the fact that the boundary circles are geodesics. These charts are constructed analogously to the charts for points on edges of hyperbolic polygons under appropriate side identifications as seen above. Any compact surface of genus $g \geq 2$ can be built from glued pairs of pants, not necessarily uniquely.

Under this construction, we have many choices. For example, the lengths of circles in the original hyperbolic hexagons are now arbitrary. Also, the choice of 'pants decomposition' of a given surface is not unique, and the different possibilities are topologically different. □

6.9. Gauss–Bonnet theorem

Recall that in a spherical triangle with internal angles α, β, γ, we have seen in the example sheets that this has area $\alpha + \beta + \gamma - \pi$, and that a hyperbolic triangle with the same internal angles has area $\pi - \alpha - \beta - \gamma$. We have seen the convex Gauss–Bonnet theorem, which states

$$\int_\Sigma \kappa \, \mathrm{d}A = 4\pi$$

where Σ bounds a convex region in $\mathbb{R}^3$ and $\kappa_\Sigma > 0$. These are special cases of a pair of theorems as shown below.

Theorem (local Gauss–Bonnet theorem)**.** Let Σ be an abstract smooth surface with abstract Riemannian metric g. Let R be an n-sided *geodesic polygon* on Σ, which is a smooth disc with boundary decomposed into n geodesic arcs. Then

$$\int_{R \subseteq \Sigma} \kappa_\Sigma \, \mathrm{d}A = \sum_{i=1}^{n} \alpha_i - (n-2)\pi$$

where the α_i are the internal angles of the polygon.

It is important that γ_i be geodesics that cut out a disc; R must be homeomorphic to $\mathbb{R}^2$, and it cannot (for example) contain any holes.

Theorem (global Gauss–Bonnet theorem)**.** Let Σ be a compact smooth surface with abstract Riemannian metric g. Then

$$\int_\Sigma \kappa_\Sigma \, \mathrm{d}A = 2\pi\chi(\Sigma)$$

Remark. Gauss curvature can be defined using only the first fundamental form, or equivalently an abstract Riemannian metric.

For hyperbolic surfaces, we can construct Σ_g from a $4g$-gon with internal angles $\frac{\pi}{2g}$ in such a way that the total area of Σ is exactly the area of the polygon, so

$$\int_\Sigma 1 \, \mathrm{d}A = \text{area(polygon)} = (4g-2)\pi - \sum_1^{4g} \frac{\pi}{2g} = (4g-4)\pi$$

Since $\kappa \equiv -1$ and $\chi(\Sigma_g) = 2-2g$, this agrees with the Gauss–Bonnet theorem.

A right-angled hyperbolic hexagon has area

$$4\pi - \sum_1^6 \frac{\pi}{2} = \pi$$

Each pair of pants was constructed from two such polygons, and to construct a genus g surface we required $2g-2$ pairs of pants. So the total area is $4g-4\pi$, which agrees with the theorem.

The Gauss–Bonnet theorem also shows that the Euler characteristic does not depend on the choice of triangulation of Σ.

Suppose Σ is a flat surface and γ is a closed geodesic, so $\gamma : \mathbb{R} \to \Sigma$ and is periodic with some period T. Then γ cannot bound a smooth disc in Σ. Conversely, on S^2, the great circle is a closed geodesic, and bounds a hemisphere. For instance, for the flat torus $\mathbb{R}^2/\mathbb{Z}^2$, if γ is a closed curve on this torus bounding a closed disc R it is not a geodesic. Indeed, if we formally add two vertices to such a geodesic, we find a geodesic 2-gon with two internal angles π, but by the Gauss–Bonnet theorem we expect

$$0 = \int_R \kappa_\Sigma \, \mathrm{d}A = \sum_1^2 \alpha_i - (n-2)\pi = 2\pi$$

We can in fact deduce the global Gauss–Bonnet theorem from the local Gauss–Bonnet theorem, utilising the following lemma.

Lemma. A compact smooth surface admits subdivisions into geodesic polygons.

The proof of this lemma considers the exponential map, discussed in Part II. Given such a subdivision on Σ, we can find

$$\sum_{\text{polygons } P} \int_P \kappa_\Sigma \, \mathrm{d}A = \int_\Sigma \kappa_\Sigma \, \mathrm{d}A$$

By the local Gauss–Bonnet theorem, the left hand side is equal to

$$\sum_n \sum_{n\text{-gons } P} \left(\sum_{i=1}^n \alpha_i(P) - (n-2)\pi \right)$$

Since the angles at each point add to 2π, and each n-gon contains two edges which each separate two polygons, this is equal to $2\pi V + 2\pi F - 2\pi E = 2\pi\chi(\Sigma)$ as required.

6.10. Green's theorem (non-examinable)

The local Gauss–Bonnet theorem is very closely related to Green's theorem in $\mathbb{R}^2$. This discussion is non-examinable.

Theorem. Let $R \subseteq \mathbb{R}^2$ be a region bound by a piecewise smooth curve γ, and P, Q be smooth real-valued functions defined on an open set $V \supset R$. Then

$$\int_\gamma P \, \mathrm{d}u + Q \, \mathrm{d}v = \int_R (Q_u - P_v) \, \mathrm{d}u \, \mathrm{d}v$$

We will consider a geodesic polygon on Σ which lies in the domain of some local parametrisation defined on $V \subseteq \mathbb{R}^2$. Consider an orthonormal basis for $\mathbb{R}^2$ which varies from point to point, defined by $e = \sigma_u, f = \sigma_v/\sqrt{G}$ where we use geodesic normal coordinates u, v to give $E = 1, F = 0$. Then $T_p\Sigma = \operatorname{span}(e, f)$ if $p \in \operatorname{Im}\sigma$. We parametrise γ by arc length and consider

$$I = \int_\gamma \langle e, \dot{f} \rangle \, \mathrm{d}t$$

We will compute this in two ways. Note that

$$\dot{f} = f_u \dot{u} + f_v \dot{v}$$

Let $P = \langle e, f_u \rangle$ and $Q = \langle e, f_v \rangle$. Then

$$Q_u - P_v = \langle e_u, f_v \rangle - \langle f_v, e_u \rangle + \langle e, f_{uv} \rangle - \langle e, f_{uv} \rangle = \langle e_u, f_v \rangle - \langle f_u, e_v \rangle$$

which we can show to be equal to $-\left(\sqrt{G}\right)_{uu} = \kappa\sqrt{G}$. But $\sqrt{G}$ is the area element $\sqrt{EG - F^2}$, so

$$\int_R (Q_u - P_v) \, \mathrm{d}u \, \mathrm{d}v = \int_R \kappa_\Sigma \, \mathrm{d}A$$

Let $\theta(t)$ be the angle between $\dot\gamma(t)$ and $e(t)$, which is a function of t in the domain of γ. More precisely,

$$\dot\gamma = e\cos\theta(t) + f\sin\theta(t)$$

Thus

$$\ddot\gamma = \dot e\cos\theta + \dot f\sin\theta + \eta\dot\theta; \quad \eta = -e\sin\theta + f\cos\theta$$

γ is a piecewise geodesic, so if $\Sigma \subseteq \mathbb{R}^3$ was smooth, $\ddot\gamma$ is orthogonal to $T_p\Sigma = \operatorname{span} e, f$. But $\eta \in \langle e, f\rangle$, so $\ddot\gamma$ is orthogonal to η. By expanding,

$$\left\langle \dot e\cos\theta + \dot f\sin\theta + \eta\dot\theta, -e\sin\theta + f\cos\theta\right\rangle = 0$$

Since e, f are orthogonal unit vectors, we have $\langle e, \dot e\rangle = 0 = \left\langle f, \dot f\right\rangle$ and $\left\langle e, \dot f\right\rangle = 0 = \langle \dot e, f\rangle$, so we can expand to find

$$\langle \ddot\gamma, \eta\rangle = 0 \implies \dot\theta = \left\langle e, \dot f\right\rangle$$

Thus,

$$I = \int_\gamma \left\langle e, \dot f\right\rangle \mathrm{d}t = \int_\gamma \dot\theta(t)\,\mathrm{d}t = 2\pi - \sum(\text{external angles of } R)$$

since γ is composed of straight lines. Since external angles and internal angles sum to π, this is exactly the local Gauss–Bonnet theorem. Green's theorem suggests the study of non-geodesic polygons.

6.11. Alternate flat toruses

We have constructed a flat metric on the torus, viewed as ${\mathbb{R}^2}/{\mathbb{Z}^2}$, or as ${[0,1]^2}/{\sim}$ for a suitably defined equivalence relation. Importantly, opposite sides of the square $[0,1]^2$ were identified by translation, which allowed us to find a smooth atlas where transition maps preserve the usual Euclidean metric on $\mathbb{R}^2$. This construction is valid for any parallelogram; any such shape $Q \subseteq \mathbb{R}^2$ defines a flat metric g_Q on T^2. If one vertex is set to zero in $\mathbb{R}^2$ and the edges of this vertex are labelled by their endpoints v_1, v_2, then $(T^2, g_Q) = {\mathbb{R}^2}/{\mathbb{Z}v_1 \oplus \mathbb{Z}v_2}$ where $\mathbb{Z}v_1 \oplus \mathbb{Z}v_2$ is a viewed as a subgroup of the group $\mathbb{R}^2$ of translations.

The area with respect to g_Q of T^2 is the Euclidean area of the parallelogram Q. In particular, if two parallelograms have different areas, the two metrics cannot be globally isometric. However, this is not the only restriction for global isometries.

Lemma. Consider the torus based on $Q = [0,1]^2$ and the torus based on $\hat Q = [0,10]\times\left[0, \frac{1}{10}\right]$. The metrics $g_Q, g_{\hat Q}$ are not isometric, but both have unit total area.

Proof. Recall that geodesics in a flat torus correspond to straight lines in $\mathbb{R}^2$. By Picard's theorem, there exists a unique geodesic from a given point p for each direction in $T_p\Sigma$. We can therefore see that all geodesics through p are the images of straight lines in $\mathbb{R}^2$.

Recall that a *closed geodesic* is defined on $\mathbb{R}$ and is periodic. We can see that geodesics in $\mathbb{R}^2$ through $\hat p \in q^{-1}(p)$ define a closed geodesic if and only if they pass through another lift

$\hat{p}' \in q^{-1}(p)$ of p; that is, the line has rational gradient. The shortest closed geodesic on the surface in metric Q is of unit length, but the shortest closed geodesic with metric $\hat{Q}$ is $\frac{1}{10}$. So the surfaces are not globally isometric. □

We would like to understand all possible flat metrics on the torus T^2, up to global dilation and Euclidean isometries of Q, which lead to essentially the same geometry on the quotient torus. Given any parallelogram, we can set one vertex at zero and another at $(1,0) = 1 \in \mathbb{R}^2$ by performing dilation and a Euclidean isometry, and then the third lies at τ and the fourth at $1+\tau$, where τ has positive y-coordinate. This provides a metric on the torus, and now the only degree of freedom is τ. Hence, this defines a map from the upper half-plane to the set of flat metrics on T^2 up to dilation.

We can pull back metrics by diffeomorphisms. Metrics allow us to measure lengths of curves by integrating lengths of tangent vectors, so a metric can be viewed as an inner product on the tangent space at each point. If $f : \Sigma \to \Sigma'$ and $p \in \Sigma$, then for two small curves γ_1, γ_2 through p, the pullback metric f^*g was defined such that

$$\langle \dot{\gamma}_1, \dot{\gamma}_2 \rangle_{p, f^*g} = \langle f \circ \dot{\gamma}_1, f \circ \dot{\gamma}_2 \rangle_{f(p), g}$$

$SL(2,\mathbb{Z})$ acts on $\mathbb{R}^2$ preserving $\mathbb{Z}^2$, so it acts on $\mathbb{R}^2/\mathbb{Z}^2 = T^2$.

Lemma. $SL(2,\mathbb{Z})$ acts by diffeomorphisms on T^2.

Proof. Clearly $A \in SL(2,\mathbb{Z})$ acts smoothly (indeed, linearly) on $\mathbb{R}^2$, and the charts for the smooth atlas are such that A then acts smoothly with respect to these. □

Also, $SL(2,\mathbb{Z}) \subseteq SL(2,\mathbb{R})$ acts on the upper half-plane by Möbius maps.

Theorem. The map from the upper half-plane $\mathfrak{h}$ to the set of flat metrics on T^2 modulo dilation induces a map from $\mathfrak{h}/SL(2,\mathbb{Z})$ to the set of flat metrics on T^2 modulo dilation and diffeomorphism. This resulting map is a bijection. We say that $\mathfrak{h}/SL(2,\mathbb{Z})$ is the *moduli space* of flat metrics on T^2.

In the above theorem, 'diffeomorphism' is taken to mean 'orientation-preserving diffeomorphism'.

Remark. The left-hand side $\mathfrak{h}/SL(2,\mathbb{Z})$ is an object of hyperbolic geometry, yet the right-hand side is entirely concerned with flat metrics.

Similar results can be shown for surfaces of higher genus. The moduli space of hyperbolic metrics on Σ_g where $g \geq 2$ is perhaps the most studied space in all of geometry.

6.12. Further courses

There are four Part II courses that extend this course.

(i) Algebraic Topology. Spaces are studied through algebraic invariants, such as the Euler characteristic, and covering maps of surfaces like $S^2 \to \mathbb{RP}^2$ or $\mathbb{R}^2 \to T^2$.

(ii) Differential Geometry. While in IB Geometry the Gauss curvature $\kappa = \det(DN)$ is discussed, the trace $\operatorname{tr}(DN)$ is the *mean curvature*, discussed heavily in this course.

(iii) Riemann Surfaces. This course studies the fact that if $f : \mathbb{C} \to \mathbb{C}$ is holomorphic (or, indeed, entire) and $w \in \mathbb{C}$, then $f(z + w)$ is holomorphic, and if $f : D \to D$ is holomorphic and $A \in \text{Möb}(D)$, then $f \circ A$ is holomorphic.

(iv) General Relativity. This is the theory of light as geodesics.

XI. Statistics

Lectured in Lent 2022 by Dr. S. Bacallado

An estimator is a random variable that approximates a parameter. For instance, the parameter could be the mean of a normal distribution, and the estimator could be a sample mean. In this course, we study how estimators behave, what properties they have, and how we can use them to make conclusions about the real parameters. This is called parametric inference: the study of inferring parameters from statistics of sample data.

Towards the end of the course, we study the normal linear model, which is a useful way to model data that is believed to depend linearly on a vector of inputs, together with some normally distributed noise. Even nonlinear patterns can be analysed using this model, by letting the inputs to the model be polynomials in the real-world data.

Contents

1. Introdution and review of IA Probability 612
1.1. Introduction 612
1.2. Review of IA Probability 612
1.3. Standardised statistics 614
1.4. Moment generating functions 614
1.5. Limit theorems 615
1.6. Conditional probability 615
1.7. Change of variables in two dimensions 616
1.8. Common distributions 616
2. Estimation 617
2.1. Estimators 617
2.2. Bias-variance decomposition 617
2.3. Sufficiency 619
2.4. Factorisation criterion 619
2.5. Minimal sufficiency 620
2.6. Rao–Blackwell theorem 621
2.7. Maximum likelihood estimation 624
3. Inference 626
3.1. Confidence intervals 626
3.2. Interpreting the confidence interval 628
4. Bayesian analysis 629
4.1. Introduction 629
4.2. Inference from the posterior 630
4.3. Point estimation 630
4.4. Credible intervals 631
5. Hypothesis testing 632
5.1. Hypotheses 632
5.2. Testing hypotheses 632
5.3. Neyman–Pearson lemma 633
5.4. p-values 635
5.5. Composite hypotheses 635
5.6. Generalised likelihood ratio test 636
5.7. Wilks' theorem 637
5.8. Goodness of fit 638
5.9. Pearson statistic 638
5.10. Goodness of fit for composite null 639
5.11. Testing independence in contingency tables 640

5.12. Testing homogeneity in contingency tables 641
5.13. Tests and confidence sets 642
6. The normal linear model 644
6.1. Multivariate normal distribution 644
6.2. Orthogonal projections 645
6.3. Linear model 647
6.4. Matrix formulation 648
6.5. Assumptions 648
6.6. Least squares estimation 648
6.7. Fitted values and residuals 650
6.8. Normal linear model 650
6.9. Inference 651
6.10. F-tests 654
6.11. Analysis of variance 657
6.12. Simple linear regression 658

1. Introdution and review of IA Probability

1.1. Introduction

Statistics can be defined as the science of making informed decisions. The field comprises, for example:

- the design of experiments and studies;
- visualisation of data;
- formal statistical inference (which is the focus of this course);
- communication of uncertainty and risk; and
- formal decision theory.

This course concerns itself with *parametric inference*. Let $X_1, \dots, X_n$ be i.i.d. (independent and identically distributed) random variables, where we assume that the distribution of X_1 belongs to some family with parameter $\theta \in \Theta$. For instance, let $X_1 \sim \text{Poi}(\mu)$, where $\theta = \mu$ and $\Theta = (0, \infty)$. Another example is $X_1 \sim N(\mu, \sigma^2)$, and $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (0, \infty)$. We use the observed $X = (X_1, \dots, X_n)$ to make inferences about the parameter θ:

(i) we can estimate the value of θ using a *point estimate* written $\hat{\theta}(X)$;

(ii) we can make an *interval estimate* of θ, written $(\hat{\theta}_1(X), \hat{\theta}_2(X))$;

(iii) hypotheses about θ can be tested, for instance the hypothesis $H_0 : \theta = 1$, by checking whether there is evidence in the data X against the hypothesis H_0.

Remark. In general, we will assume that the family of distributions of the observations X_i is known *a priori*, and the parameter θ is the only unknown. There will, however, be some remarks later in the course where we can make weaker assumptions about the family.

1.2. Review of IA Probability

This subsection reviews material covered in the IA Probability course. Some keywords are measure-theoretic, and are not defined.

Let Ω be the *sample space* of outcomes in an experiment. A *measurable* subset of Ω is called an *event*, and we denote the set of events by $\mathcal{F}$. A *probability measure* $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies the following properties.

(i) $\mathbb{P}(\varnothing) = 0$;

(ii) $\mathbb{P}(\mathcal{F}) = 1$;

(iii) $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ if (A_i) is a sequence of disjoint events.

A *random variable* is a *measurable function* $X : \Omega \to \mathbb{R}$. The *distribution function* of a random variable X is the function $F_X(x) = \mathbb{P}(X \leq x)$. We say that a random variable is *discrete*

when it takes values in a countable set $\mathcal{X} \subset \mathbb{R}$. The *probability mass function* of a discrete random variable is the function $p_X(x) = \mathbb{P}(X = x)$. We say that X has a *continuous distribution* if it has a *probability density function* $f_X(x)$ such that $\mathbb{P}(x \in A) = \int_A f_X(x)\,\mathrm{d}x$ for 'nice' sets A.

The *expectation* of a random variable X is defined as

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in X} x p_X(x) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x & \text{if } X \text{ continuous} \end{cases}$$

If $g : \mathbb{R} \to \mathbb{R}$, we define $\mathbb{E}[g(X)]$ by considering the fact that $g(X)$ is also a random variable. For instance, in the continuous case,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$$

The *variance* of a random variable X is defined as $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

We say that a set of random variables $X_1, \dots, X_n$ are *independent* if, for all $x_1, \dots, x_n$, we have

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n)$$

If and only if $X_1, \dots, X_n$ have probability density (or mass) functions $f_1, \dots, f_n$, then the *joint probability density (respectively mass) function* is

$$f_X(x) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

If $Y = \max\{X_1, \dots, X_n\}$ where the X_i are independent, then the distribution function of Y is given by

$$\mathbb{P}(Y \le y) = \mathbb{P}(X_1 \le y) \cdots \mathbb{P}(X_n \le y)$$

The probability density function of Y (if it exists) is obtained by the differentiating the above.

Under a linear transformation, the expectation and variance have certain properties. Let $a = (a_1, \dots, a_n)^\intercal \in \mathbb{R}^n$ be a constant in $\mathbb{R}^n$.

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n] = \mathbb{E}[a^\intercal X] = a^\intercal \mathbb{E}[X]$$

where $\mathbb{E}[X]$ is defined componentwise. Note that independence of X_i is not required for linearity of the expectation to hold. Similarly,

$$\operatorname{Var}(a^\intercal X) = \sum_{i,j} a_i a_j \operatorname{Cov}(X_i, X_j) = a^\intercal \operatorname{Var}(X)\, a$$

where we define $\operatorname{Cov}(X, Y) \equiv \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$, and $\operatorname{Var}(X)$ is the *variance-covariance matrix* with entries $(\operatorname{Var}(X))_{ij} = \operatorname{Cov}(X_i, X_j)$. We can say that the variance is bilinear.

1.3. Standardised statistics

Suppose that $X_1, \dots, X_n$ are i.i.d. and $\mathbb{E}[X_1] = \mu$, $\operatorname{Var}(X_1) = \sigma^2$. We define

$$S_n = \sum_i X_i; \quad \overline{X_n} = \frac{S_n}{n}$$

where $\overline{X_n}$ is called the *sample mean*. By linearity of expectation and bilinearity of variance,

$$\mathbb{E}\left[\overline{X_n}\right] = \mu; \quad \operatorname{Var}\left(\overline{X_n}\right) = \frac{\sigma^2}{n}$$

We further define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

which has the properties that

$$\mathbb{E}\left[\overline{Z}_n\right] = 0; \quad \operatorname{Var}(Z_n) = 1$$

1.4. Moment generating functions

The *moment generating function* of a random variable X is the function $M_X(t) = \mathbb{E}\left[e^{tX}\right]$, provided that this function exists for t in some neighbourhood of zero, This can be thought of as the Laplace transform of the probability density function. Note that

$$\mathbb{E}[X^n] = \frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(t)\Big|_{t=0}$$

Under broad conditions, moment generating functions uniquely define a distribution function of a random variable. In other words, the Laplace transform is invertible. They are also useful for finding the distribution of sums of independent random variables. For instance, let $X_1, \dots, X_n$ be i.i.d. Poisson random variables with parameter μ. Then, the moment generating function of X_i is

$$M_{X_1}(t) = \mathbb{E}\left[e^{tX_i}\right] = \sum_{x=0}^{\infty} e^{tx} e^{-\mu}\frac{\mu^x}{x!} = e^{-\mu}\sum_{x=0}^{\infty}\frac{(e^t\mu)^x}{x!} = e^{-\mu}e^{\mu e^t} = e^{-\mu(1-e^t)}$$

Now,

$$M_{S_n}(t) = \mathbb{E}\left[e^{tS_n}\right] = \prod_{i=1}^{n}\mathbb{E}\left[e^{tX_i}\right] = e^{-n\mu(1-e^t)}$$

This defines a Poisson distribution with parameter $n\mu$ by inspection.

1.5. Limit theorems

The *weak law of large numbers* states that for all $\varepsilon > 0$, $\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \to 0$ as $x \to \infty$. Note that the event $\left|\overline{X}_n - \mu\right| > \varepsilon$ depends only on $X_1, \dots, X_n$.

The *strong law of large numbers* states that $\mathbb{P}\left(\overline{X}_n \to \mu\right) = 1$. In this formulation, the event depends on the whole sequence of random variables X_i, since the limit is inside the probability calculation.

The *central limit theorem* states that $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ is approximately a $\mathrm{N}(0,1)$ random variable when n is large. More precisely, $\mathbb{P}(Z_n \leq z) \to \Phi(z)$ for all $z \in \mathbb{R}$.

1.6. Conditional probability

If X, Y are discrete random variables, we can define the conditional probability mass function to be

$$p_{X|Y}(x \mid y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

when $\mathbb{P}(Y = y) \neq 0$. If X, Y are continuous, we define the joint probability density function to be $f_{X,Y}(x, y)$ such that

$$\mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y') \,\mathrm{d}y' \,\mathrm{d}x'$$

The conditional probability density function is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, y) \,\mathrm{d}x}$$

The denominator is sometimes referred to as the *marginal probability density function* of Y, written $f_Y(y)$. Now, we can define the conditional expectation by

$$\mathbb{E}[X \mid Y] = \begin{cases} \sum_x x p_{X|Y}(x \mid Y) & \text{if } X \text{ discrete} \\ \int_x x f_{X|Y}(x \mid Y) \,\mathrm{d}x & \text{if } X \text{ continuous} \end{cases}$$

The conditional expectation is itself a random variable, as it is a function of the random variable Y. The conditional variance is defined similarly, and is a random variable. The *tower property* is that

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

The *law of total variance* is that

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])$$

1.7. Change of variables in two dimensions

Suppose that $(x, y) \mapsto (u, v)$ is a differentiable bijection from $\mathbb{R}^2$ to itself. Then, the joint probability density function of U, V can be written as

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v))|\det J|$$

where J is the Jacobian matrix,

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}$$

1.8. Common distributions

X has the binomial distribution with parameters n, p if X represents the number of successes in n independent Bernoulli trials with parameter p.

X has the multinomial distribution with parameters $n; p_1, \dots, p_k$ if there are n independent trials with k types, where p_j is the probability of type j in a single trial. Here, X takes values in $\mathbb{N}^k$, and X_j is the amount of trials with type j. Each X_j is marginally binomially distributed.

X has the negative binomial distribution with parameters k, p if, in i.i.d. Bernoulli trials with parameter p, the variable X is the time at which the kth success occurs. The negative binomial with parameter $k = 1$ is the geometric distribution.

The Poisson distribution with parameter λ is the limit of the distribution $\mathrm{Bin}(n, \lambda/n)$ as $n \to \infty$.

If $X_i \sim \Gamma(\alpha_i, \lambda)$ for $i = 1, \dots, n$ with $X_1, \dots, X_n$ independent, then the distribution of S_n is given by the product of the moment generating functions. By inspection,

$$M_{S_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\sum_i \alpha_i}$$

or ∞ if $t \geq \lambda$. Hence the sum of these random variables is $S_n \sim \Gamma\left(\sum_i \alpha_i, \lambda\right)$, where the shape parameter α is constructed from the sum of the shape parameters of the original functions. We call λ the rate parameter, and λ^{-1} is called the scale parameter. If $X \sim \Gamma(\alpha, \lambda)$, then for all $b > 0$ we have $bX \sim \Gamma(x, \lambda/b)$. Special cases of the Γ distribution include:

- $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$;
- $\Gamma(k/2, 1/2) = \chi^2_k$ with k degrees of freedom, which is the distribution of a sum of k i.i.d. squared standard normal random variables.

2. Estimation

2.1. Estimators

Suppose $X_1, \dots, X_n$ are i.i.d. observations with a p.d.f. (or p.m.f.) $f_X(x \mid \theta)$, where θ is an unknown parameter in some parameter space Θ. Let $X = (X_1, \dots, X_n)$.

Definition. An *estimator* is a statistic, or a function of the data, written $T(X) = \hat{\theta}$, which is used to approximate the true value of θ. This does not depend (explicitly) on θ. The distribution of $T(X)$ is called its *sampling distribution*.

Example. Let $X_1, \dots, X_n \sim N(0,1)$ be i.i.d. Let $\hat{\mu} = T(X) = \overline{X}_n$. The sampling distribution is $T(X) \sim N\left(\mu, \frac{1}{n}\right)$. Note that this sampling distribution in general depends on the true parameter μ.

Definition. The *bias* of $\hat{\theta}$ is

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta\left[\hat{\theta}\right] - \theta$$

Note that $\hat{\theta}$ is a function only of $X_1, \dots, X_n$, and the expectation operator $\mathbb{E}_\theta$ assumes that the true value of the parameter is θ.

Remark. In general, the bias is a function of the true parameter θ, even though it is not explicit in the notation.

Definition. An estimator with zero bias for all θ is called an *unbiased estimator*.

Example. The estimator $\hat{\mu}$ in the above example is unbiased, since

$$\mathbb{E}_\mu\left[\hat{\mu}\right] = \mathbb{E}_\mu\left[\overline{X}_n\right] = \mu$$

for all $\mu \in \mathbb{R}$.

Definition. The *mean squared error* of θ is defined as

$$\text{mse}(\hat{\theta}) = \mathbb{E}_\theta\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Remark. Like the bias, the mean squared error is, in general, a function of the true parameter θ.

2.2. Bias-variance decomposition

The mean squared error can be written as

$$\text{mse}(\hat{\theta}) = \mathbb{E}_\theta\left[\left(\hat{\theta} - \mathbb{E}_\theta\left[\hat{\theta}\right] + \mathbb{E}_\theta\left[\hat{\theta}\right] - \theta\right)^2\right] = \text{Var}_\theta\left(\hat{\theta}\right) + \text{bias}^2(\hat{\theta})$$

Note that both the variance and bias squared terms are positive. This implies a tradeoff between bias and variance when minimising error.

Example. Let $X \sim \text{Bin}(n, \theta)$ where n is known and θ is an unknown probability. Let $T_U = X/n$. This is the proportion of successes observed. This is an unbiased estimator, since $\mathbb{E}_\theta[T_U] = \mathbb{E}_\theta[X]/n = \theta$. The mean squared error for the estimator is then

$$\text{Var}_\theta(T_n) = \text{Var}_\theta\left(\frac{X}{n}\right) = \frac{\text{Var}_\theta(X)}{n^2} = \frac{\theta(1-\theta)}{n}$$

Now, consider an alternative estimator which has some bias:

$$T_B = \frac{X+1}{n+2} = w\underbrace{\frac{X}{n}}_{T_U} + (1-w)\frac{1}{2}; \quad w = \frac{n}{n+2}$$

This interpolates between the estimator T_U and the fixed estimator $\frac{1}{2}$. Here,

$$\text{bias}(T_B) = \mathbb{E}_\theta[T_B] - \theta = \frac{n}{n+2}\theta - \frac{1}{n+2}\theta$$

The bias is nonzero for all but one value of θ. Further,

$$\text{Var}_\theta(T_B) = \frac{\text{Var}_\theta(X+1)}{(n+2)^2} = \frac{n\theta(1-\theta)}{(n+2)^2}$$

We can calculate

$$\text{mse}(T_B) = (1-w)^2\left(\frac{1}{2} - \theta\right)^2 + w^2\underbrace{\frac{\theta(1-\theta)}{n}}_{\text{mse}(T_U)}$$

There exists a range of θ such that T_B has a lower mean squared error, and similarly there exists a range such that T_U has a lower error. This indicates that prior judgement of the true value of θ can be used to determine which estimator is better.

It is not necessarily desirable that an estimator is unbiased.

Example. Suppose $X \sim \text{Poi}(\lambda)$ and we wish to estimate $\theta = \mathbb{P}(X=0)^2 = e^{-2\lambda}$. For some estimator $T(X)$ of θ to be unbiased, we need that

$$\mathbb{E}_\lambda[T(X)] = \sum_{x=0}^{\infty} T(x)\frac{\lambda^x e^{-\lambda}}{x!} = e^{-2\lambda}$$

Hence,

$$\sum_{x=0}^{\infty} T(x)\frac{\lambda^x}{x!} = e^{-\lambda}$$

But $e^{-\lambda}$ has a known power series expansion, giving $T(X) \equiv (-1)^X$ for all X. This is not a good estimator, for example because it often predicts negative numbers for a positive quantity.

2.3. Sufficiency

Definition. A statistic $T(X)$ is *sufficient* for θ if the conditional distribution of X given $T(X)$ does not depend on θ. Note that θ and $T(X)$ may be vector-valued, and need not have the same dimension.

Example. Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter θ where $\theta \in [0,1]$. The mass function is

$$f_X(x \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Note that this dependent only on x via the statistic $T(X) = \sum_{n=1}^{n} x_i$. Here,

$$f_{X|T=t}(x \mid \theta) = \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(x) = t)}$$

If $\sum x_i = t$, we have

$$f_{X|T=t}(x \mid \theta) = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-\sum x_i}} = \frac{1}{\binom{n}{t}}$$

Hence $T(X)$ is sufficient for θ.

2.4. Factorisation criterion

Theorem. T is sufficient for θ if and only if

$$f_X(x \mid \theta) = g(T(x), \theta)h(x)$$

for suitable functions g, h.

Proof. This will be proven in the discrete case; the continuous case can be handled analogously. Suppose that the factorisation criterion holds. Then, if $T(x) = t$,

$$\begin{aligned} f_{X|T=t}(x \mid T = t) &= \frac{\mathbb{P}_\theta(X = x, T(x) = t)}{\mathbb{P}_\theta(T(x) = t)} \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{x' \,:\, T(x')=t} g(T(x'), \theta)h(x')} \\ &= \frac{h(x)}{\sum_{x' \,:\, T(x')=t} h(x')} \end{aligned}$$

which does not depend on θ. By definition, $T(X)$ is sufficient.

Conversely, suppose that $T(X)$ is sufficient.

$$\begin{aligned} f_X(x \mid \theta) &= \mathbb{P}_\theta(X = x) \\ &= \mathbb{P}_\theta(X = x, T(X) = T(x)) \\ &= \underbrace{\mathbb{P}_\theta(X = x \mid T(X) = T(x))}_{h(x)} \underbrace{\mathbb{P}_\theta(T(X) = T(x))}_{g(T(X),\theta)} \end{aligned}$$

□

Example. Consider the above example with n Bernoulli random variables with mass function

$$f_X(x \mid \theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Let $T(X) = \sum x_i$, and then the above mass function is in the form of $g(T(X), \theta)$ and we can set $h(x) \equiv 1$. Hence $T(X)$ is sufficient.

Example. Let $X_1, \dots, X_n$ be i.i.d. from a uniform distribution on the interval $[0, \theta]$ for some $\theta > 0$. The mass function is

$$f_X(x \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}\{x_i \in [0,\theta]\} = \left(\frac{1}{\theta}\right)^n \mathbb{1}\left\{\min_i x_i \geq 0\right\} \mathbb{1}\left\{\max_i x_i \leq \theta\right\}$$

Let $T(X) = \max_i X_i$. Then

$$g(T(X), \theta) = \left(\frac{1}{\theta}\right)^n \mathbb{1}\left\{\max_i x_i \leq \theta\right\}; \quad h(x) \equiv \mathbb{1}\left\{\min_i x_i \geq 0\right\}$$

We can then conclude that $T(X)$ is sufficient for θ.

2.5. Minimal sufficiency

Sufficient statistics are not unique. For instance, any bijection applied to a sufficient statistic is also sufficient. Further, $T(X) = X$ is always sufficient. We instead seek statistics that maximally compress and summarise the relevant data in X and that discard extraneous data.

Definition. A sufficient statistic $T(X)$ for θ is *minimal* if it is a function of every other sufficient statistic for θ. More precisely, if $T'(X)$ is sufficient, $T'(x) = T'(y) \implies T(x) = T(y)$.

Remark. Any two minimal statistics S, T for the same θ are bijections of each other. That is, $T(x) = T(y)$ if and only if $S(x) = S(y)$.

Theorem. Suppose that $f_X(x \mid \theta)/f_X(y \mid \theta)$ is constant in θ if and only if $T(x) = T(y)$. Then T is minimal sufficient.

Remark. This theorem essentially states the following. Let $x \overset{1}{\sim} y$ if the above ratio of probability density or mass functions is constant in θ. This is an equivalence relation. Similarly, we can define $x \overset{2}{\sim} y$ if $T(x) = T(y)$. This is also an equivalence relation. The hypothesis in the theorem is that the equivalence classes of $\overset{1}{\sim}$ and $\overset{2}{\sim}$ are equal. Further, we may always construct a minimal sufficient statistic for any parameter since we can use the construction $\overset{1}{\sim}$ to create equivalence classes, and set T to be constant for all such equivalence classes.

Proof. Let $t \in \operatorname{Im} T$. Then let z_t be a representative of the equivalence class $\{x : T(x) = t\}$. Then

$$f_X(x \mid \theta) = f_X(z_{T(x)} \mid \theta) \frac{f_X(x \mid \theta)}{f_X(z_{T(x)} \mid \theta)}$$

By the hypothesis, the ratio on the right hand side does not depend on θ, so let this ratio be $h(x)$. Further, the other term depends only on $T(x)$, so it may be $g(T(x), \theta)$. Hence T is sufficient by the factorisation criterion.

To prove minimality, let S be any other sufficient statistic, and then by the factorisation criterion there exist g_S and h_S such that $f_X(x \mid \theta) = g_S(S(x), \theta) h_S(x)$. Now, suppose $S(x) = S(y)$ for some x, y. Then,

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)} = \frac{g_S(S(x), \theta) h_S(x)}{g_S(S(y), \theta) h_S(y)} = \frac{h_S(x)}{h_S(y)}$$

which is constant in θ. Hence, $x \overset{1}{\sim} y$. By the hypothesis, we have $x \overset{2}{\sim} y$, so $T(x) = T(y)$, which is the requirement for minimality. □

Example. Let $X_1, \dots, X_n$ be normal with unknown μ, σ^2.

$$\begin{aligned}\frac{f_X(x \mid \mu, \sigma^2)}{f_X(y \mid \mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\right\}}{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2 \sum_i (y_i - \mu)^2}\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_i x_i - \sum_i y_i\right)\right\}\end{aligned}$$

Hence, for minimality, this is constant in the parameters μ, σ^2 if and only if $\sum_i x_i^2 = \sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$. Thus, a minimal sufficient statistic is $\left(\sum_i x_i^2, \sum_i x_i\right)$ is a minimal sufficient statistic. A more common way of expressing the minimal sufficient statistic is

$$S(x) = \left(\overline{X}_n, S_{xx}\right); \quad \overline{X}_n = \frac{1}{n} \sum_i x_i; \quad S_{xx} = \sum_i \left(X_i - \overline{X}_n\right)^2$$

which is a bijection of the above.

Example. θ and a minimal statistic T need not have the same dimension. Consider $X_1, \dots, X_n \sim N(\mu, \mu^2)$. Here, there is a single parameter μ but the minimal sufficient statistic is still $S(x)$ as defined above.

2.6. Rao–Blackwell theorem

Previously, the notation $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ have been used to denote expectations and probabilities under the model where the observations are i.i.d. with p.d.f. or p.m.f. f_X. From now, we omit this subscript, as it will be implied for much of the remainder of the course.

Theorem. Let T be a sufficient statistic for θ, and define an estimator $\tilde{\theta}$ with $\mathbb{E}\left[\tilde{\theta}^2\right] < \infty$ for all θ. Now we define another estimator

$$\hat{\theta} = \mathbb{E}\left[\tilde{\theta} \mid T(x)\right]$$

Then, for all values of θ, we have

$$\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \leq \mathbb{E}\left[\left(\tilde{\theta} - \theta\right)^2\right]$$

In other words, the mean squared error of $\hat{\theta}$ is not greater than the mean squared error of $\tilde{\theta}$. Further, the inequality is strict unless $\tilde{\theta}$ is a function of T.

Remark. Starting from any estimator $\tilde{\theta}$, if we condition on the sufficient statistic T we obtain a 'better' statistic $\hat{\theta}$. Note that T must be sufficient, otherwise $\hat{\theta}$ may be a function of θ and thus not an estimator:

$$\hat{\theta}(X) = \hat{\theta}(T) = \int \hat{\theta}(x) \underbrace{f_{X|T}(x \mid T)}_{\text{does not depend on } \theta \text{ as } T \text{ is sufficient}} \mathrm{d}x$$

Proof. By the tower property of the expectation, we can find

$$\mathbb{E}\left[\hat{\theta}\right] = \mathbb{E}\left[\mathbb{E}\left[\tilde{\theta} \mid T(x)\right]\right] = \mathbb{E}\left[\tilde{\theta}\right]$$

Hence, subtracting $\tilde{\theta}$ from both sides, we find $\operatorname{bias}\left(\hat{\theta}\right) = \operatorname{bias}\left(\tilde{\theta}\right)$. By the conditional variance formula,

$$\operatorname{Var}\left(\tilde{\theta}\right) = \mathbb{E}\left[\underbrace{\operatorname{Var}\left(\tilde{\theta} \mid T\right)}_{\geq 0}\right] + \underbrace{\operatorname{Var}\left(\mathbb{E}\left[\tilde{\theta} \mid T\right]\right)}_{\operatorname{Var}\left(\hat{\theta}\right)} \geq \operatorname{Var}\left(\hat{\theta}\right)$$

By the bias-variance decomposition, we know that $\operatorname{mse}\left(\tilde{\theta}\right) \geq \operatorname{mse}\left(\hat{\theta}\right)$. The inequality is strict unless $\operatorname{Var}\left(\tilde{\theta} \mid T\right) = 0$ almost surely. This requires that $\tilde{\theta}$ is a function of T. □

Example. Let $X_1, \ldots, X_n$ be i.i.d. Poisson random variables with parameter λ. Then let $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$. Here,

$$f_X(x \mid \lambda) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!} \implies f_X(x \mid \theta) = \frac{\theta^n (-\log \theta)^{\sum x_i}}{\prod x_i!}$$

Using the factorisation criterion, we find

$$g(T(x), \theta) = g\left(\sum x_i, \theta\right) = \theta^n (-\log \theta)^{\sum x_i}; \quad h(x) = \frac{1}{\prod x_i!}$$

so $T(x) = \sum x_i$ is sufficient. Note that $\sum X_i$ has a Poisson distribution with parameter $n\lambda$. Consider the estimator $\tilde{\theta} = \mathbb{1}\{X_1 = 0\}$. This depends only on X_1, hence it is a weak estimator.

However, it is unbiased, so when we apply the Rao–Blackwell theorem we will construct an unbiased $\hat{\theta}$, which is precisely

$$\begin{aligned}\hat{\theta} = \mathbb{E}\left[\tilde{\theta} \mid \sum X_i = t\right] &= \mathbb{P}\left(X_1 = 0 \mid \sum X_i = t\right) \\ &= \frac{\mathbb{P}\left(X_1 = 0, \sum X_i = t\right)}{\mathbb{P}\left(\sum X_i = t\right)} \\ &= \frac{\mathbb{P}\left(X_1 = 0\right)\mathbb{P}\left(\sum_{i=2}^{n} X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = t\right)} \\ &= \left(\frac{n-1}{n}\right)^t\end{aligned}$$

This may also be written

$$\hat{\theta} = \left(1 - \frac{1}{n}\right)^{\sum x_i}$$

which is an estimator with lower mean squared error than $\tilde{1}$ for all θ. Note that $\hat{\theta} = \left(1 = \frac{1}{n}\right)^{n\overline{X}_n}$ converges in the limit to $e^{-\overline{X}_n}$. By the strong law of large numbers, $\overline{X}_n \to \mathbb{E}\left[X_1\right] = \lambda$, so we arrive at $\hat{\theta} \to e^{-\lambda} = \theta$ almost surely.

Example. Let $X_1, \ldots, X_n$ be i.i.d. uniform random variables in an interval $[0, \theta]$. We wish to estimate $\theta > 0$. We observed that $T = \max X_i$ is sufficient for θ. Let $\tilde{\theta} = 2X_1$. This is an unbiased estimator of θ. Then the Rao–Blackwellised estimator $\hat{\theta}$ is

$$\begin{aligned}\hat{\theta} &= \mathbb{E}\left[\tilde{\theta} \mid T = t\right] \\ &= 2\mathbb{E}\left[X_1 \mid \max X_i = t\right] \\ &= 2\mathbb{E}\left[X_1 \mid \max X_i = t, X_1 = \max X_i\right]\mathbb{P}\left(X_1 = \max X_i \mid \max X_i = t\right) \\ &+ 2\mathbb{E}\left[X_1 \mid \max X_i = t, X_1 \neq \max X_i\right]\mathbb{P}\left(X_1 \neq \max X_i \mid \max X_i = t\right)\end{aligned}$$

Since $X_1, \ldots, X_n$ are i.i.d., the conditional probability $\mathbb{P}\left(X_1 = \max X_i \mid \max X_i = t\right)$ can be reduced to $\mathbb{P}\left(X_1 = \max X_i\right) = \frac{1}{n}$. The complementary event may be reduced in an analogous way. The expectation $\mathbb{E}\left[X_1 \mid \max X_i = t, X_1 = \max X_i\right]$ can be reduced to t.

$$\begin{aligned}\hat{\theta} &= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}\left[X_1 \mid X_1 < t, \max_{i=2}^{n} X_i = t\right] \\ &= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}\left[X_1 \mid X_1 < t\right] \\ &= \frac{2t}{n} + \frac{2(n-1)}{n}\frac{t}{2} \\ &= \frac{2t}{n} + \frac{t(n-1)}{n} = \frac{n+1}{n}\max_i X_i\end{aligned}$$

By the Rao–Blackwell theorem, the mean squared error of $\hat{\theta}$ is not greater than the mean squared error of $\tilde{\theta}$. This is also an unbiased estimator.

2.7. Maximum likelihood estimation

Let $X_1, \dots, X_n$ be i.i.d. random variables with mass or density function $f_X(x \mid \theta)$.

Definition. For fixed observations x, the *likelihood function* $L : \Theta \to \mathbb{R}$ is given by

$$L(\theta) = f_X(x \mid \theta) = \prod_{i=1}^{n} f_{X_i}(x_i \mid \theta)$$

We will denote the *log-likelihood* by

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i \mid \theta)$$

Definition. A *maximum likelihood estimator* is an estimator that maximises the likelihood function L over Θ. Equivalently, the estimator maximises ℓ.

Example. Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter p. The log-likelihood function is

$$\ell(p) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1 - p)] = \log p + \sum X_i + \log(1 - p)\Big(n - \sum X_i\Big)$$

The derivative is

$$\ell'(p) = \frac{\sum X_i}{p} + \frac{n - \sum X_i}{1 - p}$$

which has a single stationary point at $p = \frac{1}{n} \sum X_i = \overline{X}_n$. We have $\mathbb{E}[\hat{p}] = p$, so the maximum likelihood estimator in this case is unbiased.

Example. Let $X_1, \dots, X_n$ be i.i.d. normal random variables with unknown mean μ and variance σ^2.

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (X_i - \mu)^2$$

This function is concave in μ and σ^2, so there exists a unique maximiser. In particular, ℓ is maximised when $\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial \sigma^2} = 0$.

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2} \sum (X_i - \mu)$$

This is zero if $\mu = \overline{X}_n$. Further,

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (X_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (X_i - \overline{X}_n)^2$$

This is zero if and only if

$$\sigma^2 = \frac{1}{n} \sum (X_i - \overline{X}_n)^2 = \frac{S_{xx}}{n}$$

Hence, the maximum likelihood estimator is $(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{X}_n, \frac{1}{n} S_{xx}\right)$. We can show that $\hat{\mu}$ is unbiased. We will later prove that

$$\frac{S_{xx}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hence

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{n} \mathbb{E}\left[\chi^2_{n-1}\right] = \sigma^2 \frac{n-1}{n}$$

This is therefore a biased estimator, but the bias converges to zero as $n \to \infty$: $\hat{\sigma}^2$ is *asymptotically unbiased*.

Example. Let $X_1, \dots, X_n$ be i.i.d. uniform random variables on $[0, \theta]$. Here, we derived the unbiased estimator $\hat{\theta} = \frac{n+1}{n} \max X_i$. The likelihood is given by

$$L(\theta) = \frac{1}{\theta^n} \mathbb{1}\{\max X_i \leq \theta\}$$

This function is maximised at $\hat{\theta}_{\text{mle}} = \max X_i$. By comparison to the $\hat{\theta}$ derived from the Rao–Blackwell process, $\hat{\theta}_{\text{mle}}$ is biased. In particular,

$$\mathbb{E}\left[\hat{\theta}_{\text{mle}}\right] = \frac{n}{n+1} \mathbb{E}\left[\hat{\theta}\right] = \frac{n}{n+1} \theta$$

Remark. If T is a sufficient statistic for θ, then the maximum likelihood estimator is a function of T. Indeed, since X and T are fixed, the maximiser of $L(\theta) = g(T, \theta) h(X)$ depends on X only through T. If $\varphi = H(\theta)$ for a bijection H, then if $\hat{\theta}$ is the maximum likelihood estimator for θ, we have that $H(\hat{\theta})$ is the maximum likelihood estimator for φ.

Under some regularity conditions, as $n \to \infty$ the statistic $\sqrt{n}(\hat{\theta} - \theta)$ is approximately normal with mean zero and covariance matrix Σ. More precisely, for 'nice' sets A, we have

$$\mathbb{P}\left(\sqrt{n}(\hat{\theta} - \theta) \in A\right) \to \mathbb{P}(Z \in A); \quad Z \sim N(0, \Sigma)$$

We say that the maximum likelihood estimator is *asymptotically normal*. The limiting covariance matrix Σ is a known function of ℓ, which will not be defined in this course. In some sense, Σ is the smallest variance that any estimator can achieve asymptotically.

For practical purposes, this estimator can often be found numerically by maximising ℓ or L.

3. Inference

3.1. Confidence intervals

Definition. A $100\gamma\%$ *confidence interval* for a parameter θ is a random interval $(A(X), B(X))$ such that $\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma$ for all $\theta \in \Theta$. Note that the parameter θ is assumed to be fixed for the event $\{A(X) \leq \theta \leq B(X)\}$, and the confidence interval holds uniformly over θ.

Remark. Suppose that an experiment is repeated many times. On average, $100\gamma\%$ of the time, the random interval $(A(X), B(X))$ will contain the true parameter θ. This is the *frequentist* interpretation of the confidence interval.

A misleading interpretation is as follows. Given that a single value of X is observed, there is a probability γ that $\theta \in (A(x), B(x))$. This is wrong, as will be demonstrated later.

Example. Let $X_1, \dots, X_n$ be i.i.d. normal random variables with unit variance. We will find the 95% confidence interval for $\mu = \theta$. We have

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\Big(\theta, \frac{1}{n}\Big); \quad Z = \sqrt{n}\big(\overline{X} - \theta\big) \sim N(0,1)$$

Let a, b be numbers such that $\Phi(b) - \Phi(a) = 0.95$. Then

$$\mathbb{P}\left(a \leq \sqrt{n}\big(\overline{X} - \theta\big) \leq b\right) = 0.95 \implies \mathbb{P}\left(\overline{X} - \frac{b}{\sqrt{n}} \leq \theta \leq \overline{X} - \frac{a}{\sqrt{n}}\right) = 0.95$$

Hence, $\left(\overline{X} - \frac{b}{\sqrt{n}}, \overline{X} - \frac{a}{\sqrt{n}}\right)$ is a 95% confidence interval for θ. Typically, we wish to centre the interval around some estimator $\hat{\theta}$ such that its range is minimised for a given γ. In this case, we want to set $-a = b = z_{0.025} \approx 1.96$, where $z_\alpha = \Phi^{-1}(1-\alpha)$. Hence, the confidence interval is $\left(\overline{X} \pm \frac{1.96}{\sqrt{n}}\right)$.

Remark. In general, to find a confidence interval:

(i) Find a quantity $R(X, \theta)$ where the distribution $\mathbb{P}_\theta$ does not depend on θ. This is known as a *pivot*. In the example above, $R(X, \theta) = \sqrt{n}\big(\overline{X} - \theta\big)$.

(ii) Consider $\mathbb{P}(c_1 \leq R(X, \theta) \leq c_2) = \gamma$. Given some desired level of confidence γ, find c_1 and c_2 using the distribution function of the pivot.

(iii) Rearrange such that $\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma$, then $(A(X), B(X))$ is the confidence interval as required.

Proposition. Let T be a monotonically increasing function, and let $(A(X), B(X))$ be a $100\gamma\%$ confidence interval for θ. Then $(T(A(X)), T(B(X)))$ is a $100\gamma\%$ confidence interval for $T(\theta)$.

Remark. If θ is a vector, we can consider confidence sets instead of confidence intervals. A confidence set is a set $A(X)$ such that $\mathbb{P}(\theta \in A(X)) = \gamma$.

Example. Let $X_1, \dots, X_n$ be i.i.d. normal random variables with zero mean and unknown variance σ^2. We will find a 95% confidence interval for σ^2. Note that $\frac{X_1}{\sigma} \sim N(0,1)$ is a valid pivot, but it considers only one data point. We will instead consider

$$R(X, \sigma^2) = \sum_i \frac{X_i^2}{\sigma^2} \sim \chi_n^2$$

Now, we can define $c_1 = F_{\chi_n^2}^{-1}(0.025)$ and $c_2 = F_{\chi_n^2}^{-1}(0.975)$, giving

$$\mathbb{P}\left(c_1 \leq \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \leq c^2\right) = 0.95$$

Rearranging, we have

$$\mathbb{P}\left(\frac{\sum X_i^2}{c_2} \leq \sigma^2 \leq \frac{\sum X_i^2}{c_1}\right) = 0.95$$

Hence, the interval $\sum_{i=1}^{n} X_i^2\left(\frac{1}{c_2}, \frac{1}{c_1}\right)$ is a 95% confidence interval for σ^2.

Example. Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter p. Suppose n is large. We will find an approximate 95% confidence interval for p. The maximum likelihood estimator is

$$\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

By the central limit theorem, $\hat{p}$ is asymptotically distributed according to $N\left(p, \frac{p(1-p)}{n}\right)$. Hence,

$$\sqrt{n}\frac{\hat{p} - p}{\sqrt{p(1-p)}}$$

has approximately a standard normal distribution. We have

$$\mathbb{P}\left(-z_{0.025} \leq \sqrt{n}\frac{\hat{p} - p}{\sqrt{p(1-p)}} \leq z_{0.025}\right) \approx 0.95$$

Instead of directly rearranging the inequalities, we will make an approximation for the denominator of the central term, letting $\sqrt{p(1-p)} \mapsto \sqrt{\hat{p}(1-\hat{p})}$. When n is large, this approximation becomes more accurate.

$$\mathbb{P}\left(-z_{0.025} \leq \sqrt{n}\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}} \leq z_{0.025}\right) \approx 0.95$$

This is much easier to rearrange, leading to

$$\mathbb{P}\left(\hat{p} - z_{0.025}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + z_{0.025}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) \approx 0.95$$

This gives the approximate 95% confidence interval as required.

Remark. Note that the size of the confidence interval is maximised at $p = \frac{1}{2}$, with a length of $2z_{0.025}\frac{1}{2\sqrt{n}} \approx \frac{1}{\sqrt{n}}$. This is a *conservative* 95% confidence interval; it may be wider than necessary but holds for all values of θ.

3.2. Interpreting the confidence interval

Example. Let X_1, X_2 be i.i.d. uniform random variables in $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. We wish to estimate the value of θ with a 50% confidence interval. Observe that

$$\mathbb{P}(\theta \in (\min X_i, \max X_i)) = \mathbb{P}(X_1 \leq \theta \leq X_2) + \mathbb{P}(X_2 \leq \theta \leq X_1) = \frac{1}{2}$$

Hence, $(\min X_1, \max X_i)$ is a 50% confidence interval for θ. The frequentist interpretation is exactly correct; 50% of the time, θ will lie between X_1 and X_2. However, suppose that $|X_1 - X_2| > \frac{1}{2}$. Then we know that $\theta \in (\min X_i, \max X_i)$. Suppose $X_1 = 0.1, X_2 = 0.9$, then it is not sensible to say that there is a 50% chance that $\theta \in [0.1, 0.9]$.

4. Bayesian analysis

4.1. Introduction

Frequentist analysis considers the value θ to be fixed, and then we can make inferential statements about θ in the context of repeated experiments on a random variable X. Bayesian analysis is an alternative to frequentist analysis, where θ is itself treated as a random variable taking values in the parameter space Θ. We say that the *prior* distribution $\pi(\theta)$ is a distribution representing the beliefs of the investigator about θ before observing data. The data X has a p.d.f. or p.m.f. conditional on θ given by $f_X(\,\cdot \mid \theta)$. Having observed X, we can combine this information with the prior distribution to form the *posterior* distribution $\pi(\theta \mid X)$, which is the conditional distribution of θ given X. This contains updated information about the value of θ. By Bayes' rule,

$$\pi(\theta \mid x) = \frac{\pi(\theta) f_X(x \mid \theta)}{f_X(x)}$$

where $f_X(x)$ is the marginal distribution of X, defined by

$$f_X(x) = \begin{cases} \int_\Theta f_X(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta & \theta \text{ continuous} \\ \sum_\Theta f_X(x \mid \theta)\pi(\theta) & \theta \text{ discrete} \end{cases}$$

More simply,

$$\pi(\theta \mid X) \propto \pi(\theta) \cdot f_X(X \mid \theta)$$

The proportionality here is with respect to θ. So the posterior is proportional to the prior multiplied by the likelihood. It is often easy to recognise that the right hand side of this expression is in some family of distributions, such as N or Γ, up to some normalising constant.

Remark. By the factorisation criterion, if T is a sufficient statistic for θ, the posterior $\pi(\theta \mid x)$ depends on X only through T. More precisely,

$$\pi(\theta \mid X) \propto \pi(\theta) g(T(X), \theta) h(X) \propto \pi(\theta) g(T(C), \theta)$$

Example. Consider a patient who we will test for the presence of a disease, where we have no information about the health or lifestyle of the patient. Let θ take the value 1 if the patient is infected and 0 otherwise. We have a random variable X which takes the value 1 if a given test returns a positive result and 0 if the test is negative. We know the *sensitivity* of the test $f_X(X = 1 \mid \theta = 1)$, and the *specificity* of the test $f_X(X = 0 \mid \theta = 0)$. This fully specifies the likelihood function.

We now must choose a prior distribution. For example, let $\pi(\theta = 1)$ be the estimated proportion of the general population that have the given disease. The posterior is the probability of an infection given the test result.

$$\pi(\theta = 1 \mid X = 1) = \frac{\pi(\theta = 1) f_X(X = 1 \mid \theta = 1)}{\pi(\theta = 1) f_X(X = 1 \mid \theta = 1) + \pi(\theta = 0) f_X(X = 1 \mid \theta = 0)}$$

Even with a positive test result, the posterior distribution may still yield a low probability for θ, which may happen if $\pi(\theta = 1) \ll \pi(\theta = 0)$.

Example. Let θ be the mortality rate of a particular surgery, which will take values in $[0, 1]$. In the first ten operations, we observed that none of the patients died. We will model $X \sim B(10, \theta)$ and observe $X = 0$.

We must choose a prior. Suppose that we have data from other hospitals that suggests that the mortality for the surgery ranges from 3% to 20%, with an average of 10%. We can choose the prior to be the beta distribution, $\pi(\theta) \sim \text{Beta}(a, b)$, since the value of θ should range between zero and one. Let $a = 3$ and $b = 27$, which will give $\mathbb{E}[\theta] = 0.1$ and $\mathbb{P}(0.03 < \theta < 0.2) \approx 0.9$. In this case, the posterior is

$$\pi(\theta \mid X) \propto \pi(\theta) f_X(x = 0 \mid \theta) \propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} = \theta^{x+a-1}(1-\theta)^{b-n-x-1}$$

This is again a beta distribution with parameters $x + a$ and $n - x + b$. The normalising constant does not need to be explicitly calculated since the form of the distribution can be recognised.

With the above data, we obtain $\pi(\theta \mid x = 0) \sim \text{Beta}(3, 37)$. This posterior has a smaller variance than the prior, and a smaller expectation due to observing no deaths. In this case, the prior and posterior have the same distribution. This is known as *conjugacy*.

4.2. Inference from the posterior

The posterior distribution $\pi(\theta \mid x)$ represents information about θ after having observed some data X. This can be used to make decisions under uncertainty.

(i) We first choose some decision $\delta \in \Delta$. For instance, in the first example, a decision could be to ask the patient to isolate from others to reduce transmission.

(ii) We define a *loss function* $L(\theta, \delta)$, which defines what loss is incurred by making decision δ given the true value of θ. In the above example, $L(\theta = 1, \delta = 1)$ is the loss incurred by asking the patient to isolate given that they have the disease.

(iii) We can now choose the decision δ that minimises

$$\int_\Theta L(\theta, \delta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

which is the posterior expectation of the loss.

4.3. Point estimation

We can use Bayesian analysis to represent an estimate for the value of θ as a decision.

Definition. The *Bayes estimator* $\hat{\theta}^{(B)}$ minimises

$$h(\delta) = \int_\Theta L(\theta, \delta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

Example. Suppose the loss function is quadratic, given by $L(\theta, \delta) = (\theta - \delta)^2$. Here,

$$h(\delta) = \int_\Theta (\theta - \delta)^2\pi(\theta \mid x)\,\mathrm{d}\theta$$

Thus, $h(\delta) = 0$ if

$$\int_\Theta (\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta = 0 \iff \delta = \int_\Theta \theta\pi(\theta \mid x)\,\mathrm{d}x$$

Under the quadratic loss function, $\hat{\theta}^{(B)}$ can be described as the expectation of θ under the posterior distribution.

Example. Consider the absolute error loss, given by $L(\theta, \delta) = |\theta - \delta|$. In this case we have

$$h(\delta) = \int_\Theta |\theta - \delta|\pi(\theta \mid x)\,\mathrm{d}\theta = \int_{-\infty}^{\delta} -(\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta + \int_\delta^\infty (\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

We can differentiate, using the fundamental theorem of calculus, to find

$$h'(\delta) = \int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta - \int_\delta^\infty \pi(\theta \mid x)\,\mathrm{d}\theta$$

This is zero if and only if

$$\int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta = \int_\delta^\infty \pi(\theta \mid x)\,\mathrm{d}\theta$$

This yields the median of the posterior distribution.

4.4. Credible intervals

Definition. A $100\gamma\%$ *credible interval* $(A(x), B(x))$ satisfies

$$\pi(A(x) \le \theta \le B(x) \mid x) = \gamma$$

Remark. Unlike confidence intervals, credible intervals can be interpreted conditionally on the data. For example, we could say that given a specific observation x, we are $100\gamma\%$ certain that θ lies within $(A(x), B(x))$. This credible interval is also dependent on the choice of prior distribution.

5. Hypothesis testing

5.1. Hypotheses

Definition. A *hypothesis* is an assumption about the distribution of the data X. Scientific questions are often phrased as a decision between two hypotheses. The *null hypothesis* H_0 is usually a basic hypothesis, often representing the simplest possible distribution of the data. The *alternative hypothesis* H_1 is the alternative, if H_0 were found to be false.

Example. Let $X = (X_1, \dots, X_n)$ be i.i.d. Bernoulli random variables with parameter θ. We could take, for example, $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta = \frac{3}{4}$. Alternatively, we could take $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta \neq \frac{1}{2}$.

Example. Suppose X_i takes values $0, 1, \dots$. We can take $H_0 : X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$ for some λ, and $H_1 : X_i \overset{\text{iid}}{\sim} f_1$ for some other distribution f_1. This is known as a *goodness of fit* test, which checks how well the model used for the data fits.

Definition. A *simple hypothesis* is a hypothesis which fully specifies the p.d.f. or p.m.f. of the data. A hypothesis that is not simple is called *composite*.

Example. In the first example above, $H_0 : \theta = \frac{1}{2}$ is simple, and $H_1 : \theta \neq \frac{1}{2}$ is composite. In the second example, $H_0 : X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$ is composite since λ was not fixed.

5.2. Testing hypotheses

Definition. A *test* of the null hypothesis H_0 is defined by a *critical region* $C \subseteq \mathcal{X}$. When $X \in C$, we *reject* the null hypothesis. This is a positive result. When $X \notin C$ we *fail to reject* the null hypothesis, or find *no sufficient evidence against* the null hypothesis. This is the negative result.

A *type I* error, or a *false positive*, is the error made by rejecting the null hypothesis when it is true. A *type II* error, or a *false negative*, is the error made by failing to reject the null hypothesis when it is not true. When H_0, H_1 are simple, we define

$$\alpha = \mathbb{P}_{H_0}(H_0 \text{ is rejected}) = \mathbb{P}_{H_0}(X \in C); \quad \beta = \mathbb{P}_{H_1}(H_0 \text{ is not rejected}) = \mathbb{P}_{H_1}(X \notin C)$$

The *size* of a test is α, which is the probability of a type I error. The *power* of a test is $1 - \beta$, which is the probability of not finding a type II error.

There is typically a tradeoff between α and β. Often, statisticians will choose an 'acceptable' value for the probability of type I errors α, and then maximise the power with respect to this fixed α. Computing the size of a test is typically simpler since it does not depend on H_1.

5.3. Neyman–Pearson lemma

Let H_0 and H_1 be simple, and let X have a p.d.f. or p.m.f. f_i under H_i. The *likelihood ratio statistic* is defined by

$$\Lambda_x(H_0; H_1) = \frac{f_1(x)}{f_0(x)}$$

The *likelihood ratio test* is a test that rejects H_0 when Λ_x exceeds a set value k, or more formally, $C = \{x : \Lambda_x(H_0; H_1) > k\}$.

Lemma. Suppose that f_0, f_1 are nonzero on the same set, and suppose that there exists $k > 0$ such that the likelihood ratio test with critical region $C = \{x : \Lambda_x(H_0; H_1) > k\}$ has size α. Then out of all tests of size upper bounded by α, this test has the largest power.

Remark. A likelihood ratio test with size α does not always exist for any given α. However, in general we can find a *randomised test* with arbitrary size α. This is a test where, for some values of X, we reject the null hypothesis; for some values, we fail to reject the null hypothesis; and for some values we reject the null hypothesis with a random chance of rejecting the null hypothesis.

Proof. Let $\overline{C}$ be the complement of C in $\mathcal{X}$. Then, the likelihood ratio test has

$$\alpha = \int_C f_0(x)\,\mathrm{d}x\,; \quad \beta = \int_{\overline{C}} f_1(x)\,\mathrm{d}x$$

Let $C^\star$ be a critical region for a different test, with type I and II error probabilities $\alpha^\star, \beta^\star$. Here,

$$\alpha^\star = \int_{C^\star} f_0(x)\,\mathrm{d}x\,; \quad \beta^\star = \int_{\overline{C^\star}} f_1(x)\,\mathrm{d}x$$

Suppose $\alpha^\star \le \alpha$. Then, we will show $\beta \le \beta^\star$.

$$\beta - \beta^\star = \int_{\overline{C}} f_1(x)\,\mathrm{d}x - \int_{\overline{C^\star}} f_1(x)\,\mathrm{d}x$$

By cancelling the integrals on the intersection, and using the definition of C,

$$\begin{aligned}
\beta - \beta^\star &= \int_{\overline{C} \cap C^\star} f_1(x)\,\mathrm{d}x - \int_{\overline{C^\star} \cap C} f_1(x)\,\mathrm{d}x \\
&= \int_{\overline{C} \cap C^\star} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\leq k} f_0(x)\,\mathrm{d}x - \int_{\overline{C^\star} \cap C} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\geq k} f_0(x)\,\mathrm{d}x \\
&\leq k\left[\int_{\overline{C} \cap C^\star} f_0(x)\,\mathrm{d}x - \int_{\overline{C}^\star \cap C} f_0(x)\,\mathrm{d}x\right] \\
&= k\left[\int_{\overline{C} \cap C^\star} f_0(x)\,\mathrm{d}x + \int_{C \cap C^\star} f_0(x)\,\mathrm{d}x - \int_{C \cap C^\star} f_0(x)\,\mathrm{d}x - \int_{\overline{C}^\star \cap C} f_0(x)\,\mathrm{d}x\right] \\
&= k\left[\int_{\cap C^\star} f_0(x)\,\mathrm{d}x - \int_C f_0(x)\,\mathrm{d}x\right] \\
&= k[\alpha^\star - \alpha] \\
&\leq 0
\end{aligned}$$

□

Example. Let $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d., where σ_0^2 is known and μ is an unknown. We wish to find the most powerful test of fixed size α for the hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1 > \mu_0$. The likelihood ratio is

$$\begin{aligned}
\Lambda_x(H_0; H_1) &= \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma_0^2} \sum (x_i - \mu_0)^2\right\}}{(2\pi\sigma_0^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma_0^2} \sum (x_i - \mu_1)^2\right\}} \\
&= \exp\left\{\underbrace{\frac{\mu_1 - \mu_0}{\sigma_0^2}}_{\geq 0} n\overline{X} + \frac{n(\mu_0 - \mu_1)^2}{2\sigma_0^2}\right\}
\end{aligned}$$

which depends only on $\overline{X}$, and is monotonically increasing with respect to the sample mean $\overline{X}$. Therefore, this is also monotonically increasing with respect to the statistic

$$Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma_0}$$

Thus, $\Lambda_x > k$ if and only if $Z > k'$ for some k'. Hence, the likelihood ratio test has critical region $\{x : Z(x) > k'\}$ for some k'. It thus suffices to find a critical region of Z with size α in order to construct the most powerful test of this size. Under H_0, $Z \sim N(0, 1)$. Hence, the critical region is given by $k' = \Phi^{-1}(1 - \alpha)$. This is known as a *Z-test*, since we are using the Z statistic.

5.4. p-values

Definition. Let C be a critical region of the form $\{x : T(x) > k\}$ for some test statistic T. Let $x^\star$ denote the observed data. Then, the *p-value* is

$$\mathbb{P}_{H_0}(T(X) > T(x^\star))$$

Typically, when reporting the results of a test, we describe the conclusion of the test as well as the p-value. In the example above, suppose $\mu_0 = 5$, $\mu_1 = 6$, $\alpha = 0.05$, and $x^\star = (5.1, 5.5, 4.9, 5.3)$. Here, $\overline{x^\star} = 5.2$ and $z^\star = 0.4$. The likelihood ratio test has critical region

$$\left\{x : Z(x) > \Phi^{-1}(0.95) \approx 1.645\right\}$$

The conclusion of the test here is to not reject H_0. The p-value is $1 - \Phi(z^\star) \approx 0.35$.

Proposition. Under the null hypothesis H_0, the p-value is a uniform random variable in $[0, 1]$.

Proof. Let F be the distribution of the test statistic T, which we will assume for this proof is continuous. Then,

$$\begin{aligned}
\mathbb{P}_{H_0}(p < u) &= \mathbb{P}_{H_0}(1 - F(T) < u) \\
&= \mathbb{P}_{H_0}(F(T) > 1 - u) \\
&= \mathbb{P}_{H_0}\left(T > F^{-1}(1 - u)\right) \\
&= 1 - F(F^{-1}(1 - u)) = u
\end{aligned}$$

□

5.5. Composite hypotheses

Let $X \sim f_X(\cdot \mid \theta)$ where $\theta \in \Theta$. Let $H_0 = \theta \in \Theta_0 \subset \Theta$ and $H_1 = \theta \in \Theta_1 \subseteq \Theta$. The probabilities of type I and type II error are now dependent on the precise value of θ, rather than simply on which hypothesis is taken.

Definition. The *power function* for a test C is

$$W(\theta) = \mathbb{P}_\theta(X \in C)$$

The *size* of a test C is

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta)$$

A test is *uniformly most powerful* of size α if, for any test $C^\star$ with power function $W^\star$ and size upper bounded by α, for all $\theta \in \Theta_1$ we have $W(\theta) \geq W^\star(\theta)$. Such tests need not exist. In simple models, many likelihood ratio tests are uniformly most powerful.

Example (one-sided test for normal location)**.** Let $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. where σ_0^2 is known and μ is unknown. Let $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ for some fixed μ_0. We claim that the simple hypothesis test given by $H_0' : \mu = \mu_0$ and $H_1' : \mu = \mu_1 > \mu_0$ is uniformly most powerful for H_0 and H_1. The power function is

$$\begin{aligned} W(\mu) &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma_0} = Z < z_\alpha = \Phi^{-1}(1-\alpha)\right) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma_0} > z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}\right) \\ &= 1 - \Phi\left(x_\alpha + \sqrt{n}\frac{\mu_0 - \mu}{\sigma_0}\right) \end{aligned}$$

The test has size α since $\sup_{w \in \Theta_0} W(\mu) = \alpha$. It remains to show that this power function dominates all other power functions $W^\star$ of size α in the alternative space Θ_1. First, observe that the critical region depends only on μ_0, and not on μ_1. In particular, for any $\mu_1 > \mu_0$, we have that the critical region C is the likelihood ratio test for the simple hypothesis test $H_0' : \mu = \mu_0$ and $H_1' : \mu = \mu_1$. We can also see $C^\star$ as a test of H_0' versus H_1', and for these simple hypotheses, $C^\star$ has size

$$W^\star(\mu_0) \leq \sup_{\mu < \mu_0} W^\star(\mu) \leq \alpha$$

By the Neyman–Pearson lemma, C has power no smaller than $C^\star$ for H_0' against H_1':

$$W(\mu_1) \geq W^\star(\mu_1)$$

Since this is true for all $\mu_1 > \mu_0$, the result holds, and the test C satisfies the property for being uniformly most powerful.

5.6. Generalised likelihood ratio test

Definition. Suppose we have *nested hypotheses*, i.e. $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where $\Theta_0 \subset \Theta_1$. The *generalised likelihood ratio* is given by

$$\Lambda_x(H_0; H_1) = \frac{\sup_{\theta \in \Theta_1} f_X(x \mid \theta)}{\sup_{\theta \in \Theta_0} f_X(x \mid \theta)}$$

Large values indicate a better fit under the alternative hypothesis. The *generalised likelihood ratio test* rejects the null hypothesis when Λ_x is sufficiently large.

Example (two-sided test for normal location)**.** Let $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. where σ_0^2 is known and μ is unknown. Let $H_0 : \mu = \mu_0$ and $H_1 : \mu \in \mathbb{R}$ for some fixed μ_0. In this

model, the generalised likelihood ratio is

$$\Lambda_x(H_0; H_1) = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma_0^2}\Sigma_{i=1}^n (x_i - \overline{X})^2\right\}}{(2\pi\sigma_0^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma_0^2}\Sigma_{i=1}^n (x_i - \mu_0)^2\right\}}$$
$$2\log\Lambda_x = \frac{n}{\sigma_0^2}(\overline{X} - \mu_0)^2$$

Under H_0, $\sqrt{n}\frac{\overline{X}-\mu_0}{\sigma_0} \sim N(0,1)$. Hence, $2\log\Lambda_x \sim \chi_1^2$. Therefore, the critical region of this generalised likelihood ratio test is

$$C = \left\{x : \frac{n}{\sigma_0^2}(\overline{X} - \mu_0)^2 > \chi_1^2(\alpha)\right\}$$

where $\chi_1^2(\alpha)$ is the upper α point of χ_1^2. This is called a *two-sided test* since there are two tails on the critical region, plotting with respect to $\sqrt{n}\frac{\overline{X}-\mu_0}{\sigma_0}$.

5.7. Wilks' theorem

Definition. The *dimension* of a hypothesis $H_0 : \theta \in \Theta_0$ is the number of 'free parameters' in this space.

Example. If $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_1 = \cdots = \theta_p = 0\}$, then the dimension of H_0 is $k - p$.

Let $A \in \mathbb{R}^{p\times k}$ be a $p \times k$ matrix with linearly independent rows. Let $b \in \mathbb{R}^p$ for $p < k$, then we define $\Theta_0 = \{\theta \in \mathbb{R}^k : A\theta = b\}$. Then the dimension of θ is $k - p$.

Let Θ_0 be a Riemannian manifold. We use differential geometry to deduce the dimensionality of such a manifold.

Theorem. Suppose $\Theta_0 \subset \Theta_1$, and $\dim\Theta_1 - \dim\Theta_0 = p$. Let $X = (X_1, \dots, X_n)$ be i.i.d. random variables under $f_x(\cdot \mid \theta)$ where $\theta \in \Theta_0^\circ$. Then, under some regularity conditions, as $n \to \infty$ we have

$$2\log\Lambda_x \sim \chi_p^2$$

More precisely, for all $\ell \in \mathbb{R}_+$,

$$\lim_{n\to\infty} \mathbb{P}_\theta(2\log\Lambda_x \le \ell) = \mathbb{P}(\Xi \le \ell); \quad \Xi \sim \chi_p^2$$

Remark. If n is large, this theorem allows us to implement a generalised likelihood ratio test even if we cannot find the exact distribution of $2\log\Lambda_x$. Frequentist guarantees obtained from such a test will be approximate.

Example. In the two-sided test for normal location, $\dim\Theta_1 = 1$ and $\dim\Theta_0 = 0$ hence the difference in dimensions is 1. Then, Wilks' theorem implies that $2\log\Lambda_x$ is approximately distributed according to χ_1^2, although the result is exact in this particular case.

5.8. Goodness of fit

Let $X_1, \dots, X_n$ be i.i.d. samples taking values in $\{1, \dots, k\}$. Let $p_i = \mathbb{P}(X_1 = i)$, and let N_i be the number of samples equal to i, so $\sum_i p_i = 1$ and $\sum_i N_i = n$. The parameters here are $p = (p_1, \dots, p_k)$, which has $k - 1$ dimensions. A *goodness of fit test* has a null hypothesis of the form $H_0 : p_i = \widetilde{p}_i$ for all i, for a fixed $\widetilde{p} = (\widetilde{p}_1, \dots, \widetilde{p}_k)$. The alternative hypothesis H_1 does not constrain p.

The model is $(N_1, \dots, N_k) \sim \text{Multi}(n; p_1, \dots, p_k)$. The likelihood function is

$$L(p) \propto p_1^{N_1} \cdots p_k^{N_k} \implies \ell(p) = \text{constant} + \sum_i N_i \log p_i$$

The generalised likelihood ratio is

$$2 \log \Lambda_x = 2\left(\sup_{p \in \Theta_1} \ell(p) - \sup_{p \in \Theta_0} \ell(p) \right) = 2(\ell(\hat{p}) - \ell(\widetilde{p}))$$

where $\hat{p}$ is the maximum likelihood estimator under H_1. To find $\hat{p}$, we typically use the method of Lagrange multipliers.

$$\mathcal{L}(p, \lambda) = \sum_i N_i \log p_i - \lambda\Big(\sum p_i - 1\Big)$$

We can compute that

$$\hat{p}_i = \frac{N_i}{n}$$

This is simply the fraction of observed samples of type i.

5.9. Pearson statistic

Let $o_i = N_i$ be the observed number of samples of type i, and $e_i = n\widetilde{p}_i$ be the expected value under the null hypothesis of the number of samples of type i. Here, we can write

$$2 \log \Lambda = 2 \sum_i N_i \log\left(\frac{N_i}{n\widetilde{p}_i} \right) = 2 \sum_i o_i \log \frac{o_i}{e_i}$$

Let $\delta_i = o_i - e_i$. Then

$$2 \log \Lambda = 2 \sum_i (e_i + \delta_i) \log \left(1 + \underbrace{\frac{\delta_i}{e_i}}_{\text{small when } n \text{ large}} \right)$$

By taking the Taylor expansion, we arrive at

$$2 \sum_i \left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i} \right)$$

Note that $\sum_i \delta_i = \sum_i (o_i - e_i) = n - n = 0$, so we can simplify and find

$$\sum_i \frac{\delta_i^2}{e_i} = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

This is *Pearson's χ^2 statistic*. This is also referred to a χ^2_{k-1} when performing a hypothesis test.

Example. Mendel performed an experiment in which 556 different pea plants were created from a small set of ancestors. Each descendent was either yellow or green, and either wrinkled or smooth, giving four possibilities in total. The observed result was

$$N = \left(\underbrace{315}_{SG}, \underbrace{108}_{SY}, \underbrace{102}_{WG}, \underbrace{31}_{WY}\right)$$

Mendel's theory gives a null hypothesis $H_0 : \ p = \widetilde{p} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$. Here,

$$2 \log \Lambda = 0.618; \quad \sum_i \frac{(o_i - e_i)^2}{e_i} = 0.604$$

These are referred to a χ^2_3 distribution. We observe that $\chi^2_3(0.05) = 7.815$, so we fail to reject the null hypothesis with a test of size 5%. We can compute that the p-value is $\mathbb{P}\left(\chi^2_3 > 0.6\right) \approx 0.96$, so there is a very high probability of observing a more extreme value than observed.

5.10. Goodness of fit for composite null

Suppose $H_0 : \ p_i = p_i(\theta)$ for some $\theta \in \Theta_0$, and $H_1 : \ p$ has any distribution on $\{1, \dots, k\}$. We can compute

$$2 \log \Lambda = 2\left(\sup_p \ell(p) - \sup_{\theta \in \Theta} \ell(p(\theta))\right)$$

We can sometimes compute these quantities explicitly, and hence find a test which refers this test statistic to a χ^2_p distribution where $p = \dim \Theta_1 - \dim \Theta_0 = (k-1) - \dim \Theta_0$.

Example. Consider a population of individuals who may have one of three genotypes, which occur with probabilities $(p_1, p_2, p_3) = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)$. In this case, we can find the maximum likelihood estimator under the null hypothesis to be

$$\hat{\theta} = \frac{2N_1 + N_2}{2n}$$

Hence,

$$2 \log \Lambda = 2(\ell(\hat{p}) - \ell(\hat{\theta}))$$

where $\hat{p}_i = \frac{N_1}{n}$ as found previously. This can be computed explicitly and referred to a χ^2_1 distribution. We can check that, in this model,

$$2 \log \Lambda = \sum_i o_i \log \frac{o_i}{e_i}$$

where $o_i = N_i$ and $e_i = np_i(\hat{\theta})$. We can approximate this using the Pearson statistic, $\sum_i \frac{(o_i - e_i)^2}{e_i}$.

5.11. Testing independence in contingency tables

Suppose we have observations $(X_1, Y_1), \dots, (X_n, Y_n)$ which are i.i.d., where the X_i take values in $1, \dots, r$ and the Y_i take values in $1, \dots, c$. We wish to test whether the X_i and Y_i are independent. We will summarise this data into a sufficient statistic known as a *contingency table* N, given by

$$N_{ij} = |\{\ell : 1 \le \ell \le n, (X_\ell, Y_\ell) = (i, j)\}|$$

So N_{ij} is the number of samples of type (i, j).

Example. Suppose we observe n samples, and each sample has probability p_{ij} of being of type (i, j). Flattening (N_{ij}) into a vector, this has a multinomial distribution with parameters (p_{ij}) (also flattened into a vector). The null hypothesis is $H_0 : p_{ij} = p_{i+}p_{+j}$ where $p_{i+} = \sum_j p_{ij}$ and $p_{+j} = \sum_i p_{ij}$. The alternative hypothesis places no restrictions on the p_{ij} apart from that it sums to 1 and has nonnegative entries. We find

$$2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}$$

where $\hat{p}_{ij}$ is the maximum likelihood estimator under H_1, and where $\hat{p}_{i+}$ and $\hat{p}_{+j}$ are the maximum likelihood estimators under H_0. These can be found using the method of Lagrange multipliers. In particular,

$$\hat{p}_{ij} = \frac{N_{ij}}{n}; \quad \hat{p}_{i+} = \frac{N_{i+}}{n} = \frac{1}{n} \sum_{j=1}^{c} N_{ij}; \quad \hat{p}_{+j} = \frac{N_{+j}}{n} = \frac{1}{n} \sum_{i=1}^{r} N_{ij}$$

Writing $o_{ij} = N_{ij}$ and $e_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$,

$$2 \log \Lambda = \sum_{i,j} o_{ij} \log \frac{o_{ij}}{e_{ij}} \approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

By Wilks' theorem, these test statistics have an approximate χ^2_p distribution, where $p = \dim \Theta_1 - \dim \Theta_0 = (rc - 1) - (r - 1 + c - 1) = (r - 1)(c - 1)$.

The χ^2 test for independence has a number of weaknesses.

(i) The χ^2 approximation requires n to be large. A reasonable heuristic is to require $N_{ij} \ge 5$ for *all* i, j. If this is not possible, we can perform an *exact test* (which is non-examinable).

(ii) The χ^2 test often has a low power. Heuristically, this is because the alternative space Θ_1 is too large, and there are many possible models that lie in this space.

Note that this test also applies when n is a random variable with a Poisson distribution. This is often the case when we do not fix the number of samples. The proof is not provided in this course.

5.12. Testing homogeneity in contingency tables

Example. Suppose we perform a clinical trial on 150 patients, who are randomly assigned to one of three groups of equal size. The first two sets take a drug with different doses, and the third set takes a placebo.

	improved	no difference	worse	
placebo	18	17	15	50
half dose	20	10	20	50
full dose	25	13	12	50

In the previous section, we fixed the total number of samples. Here, we fix the total number of samples, and the total number of samples in each row. We suppose

$$N_{i1}, \dots, N_{ic} \sim \text{Multinomial}(n_{i+}; p_{i1}, \dots, p_{ic})$$

which are independent for each row i of the table. The null hypothesis for homogeneity is that $p_{1j} = p_{2j} = \cdots = p_{rj}$ for all j. The alternative hypothesis assumes that $p_{i1}, \dots, p_{ic}$ is any arbitrary probability vector for each row i. Under the alternative hypothesis,

$$L(p) = \prod_{i=1}^{r} \frac{n_{i+}!}{N_{i1}! \cdots N_{ic}!} p_{i1}^{N_{i1}} \cdots p_{ic}^{N_{ic}}$$

Hence,

$$\ell(p) = \text{constant} + \sum_{i,j} N_{ij} \log p_{ij}$$

This is the same likelihood as the independence test above. To define the maximum likelihood estimator we can again use the method of Lagrange multipliers with constraints $\sum_j p_{ij} = 1$ for each i. We find

$$\hat{p}_{ij} = \frac{N_{ij}}{n_{i+}}$$

Under the null hypothesis, we let $p_j = p_{ij}$ for any i.

$$\ell(p) = \text{constant} + \sum_{i,j} N_{ij} \log p_j = \sum_j N_{+j} \log p_j$$

We have the constraint $\sum_j p_j = 1$. Using the method of Lagrange multipliers,

$$\hat{p}_j = \frac{N_{+j}}{n_{++}}$$

Hence,

$$2\log\Lambda = 2\sum_{i,j} N_{ij}\log\frac{\hat{p}_{ij}}{\hat{p}_j} = 2\sum_{i,j} N_{ij}\log\frac{N_{ij}}{n_{i+}N_{+j}/n_{++}}$$

This is precisely the same test statistic as the test for independence above. The only difference is that n_{i+} is fixed in this model. Further, if $o_{ij} = N_{ij}$ and $e_{ij} = n_{i+}\hat{p}_j = \frac{n_{i+}N_{+j}}{n_{++}}$, we have

$$2\log\Lambda = 2\sum_{i,j} o_{ij}\log\frac{o_{ij}}{e_{ij}} \approx \sum_{i,j}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

By Wilks' theorem, this is asymptotically a χ^2_p distribution. Here,

$$p = \dim\Theta_1 - \dim\Theta_0 = r(c-1) - (c-1) = (r-1)(c-1)$$

This is again exactly the same as in the χ^2 test for independence. Operationally, the tests for homogeneity and independence are therefore completely identical; we reject the null hypothesis for one test if and only if we reject the null for the other. In the example above,

$$2\log\Lambda = 5.129; \quad \sum_{i,j}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 5.173$$

Referring this to a χ^2_4 distribution, the upper 0.05-point is 9.488. Hence, we do not reject the null hypothesis at the 5% significance level.

5.13. Tests and confidence sets

Definition. The *acceptance region* A of a test is the complement of the critical region.

Theorem. Let $X \sim f_X(\cdot \mid \theta)$ for some $\theta \in \Theta$. Suppose that for each $\theta_0 \in \Theta$, there exists a test of size α with acceptance region $A(\theta_0)$ for the null hypothesis $\theta = \theta_0$. Then

$$I(X) = \{\theta : X \in A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence set.

Now suppose there exists a set $I(X)$ which is a $100(1-\alpha)\%$ confidence set for θ. Then

$$A(\theta_0) = \{x : \theta_0 \in I(x)\}$$

is the acceptance region of a test of size α for the hypothesis $\theta = \theta_0$.

Proof. Observe that for both parts of the theorem,

$$\theta_0 \in I(X) \iff X \in A(\theta_0) \iff \text{fail to reject } H_0 \text{ with data } X$$

For the first part, we assume that $\mathbb{P}_\theta(\text{fail to reject } H_0 \text{ with data } X) = 1-\alpha$, and we want to show $\mathbb{P}_\theta(\theta_0 \in I(X)) = 1-\alpha$. The second part is the converse. □

Example. Let $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. with σ_0^2 known and μ unknown. We found that a $100(1-\alpha)\%$ confidence interval for μ is

$$I(X) = \left(\overline{X} \pm \frac{Z_{\alpha/2}\sigma_0}{\sqrt{n}} \right)$$

Hence, by the second part of the theorem above, we can find a test for $H_0 : \mu = \mu_0$ with size α by

$$A(\mu_0) = \{x : \mu_0 \in I(x)\} = \left\{ x : \mu_0 \in \left[\overline{x} \pm \frac{Z_{\alpha/2}\sigma_0}{\sqrt{n}} \right] \right\}$$

This is equivalent to rejecting H_0 when

$$\left| \sqrt{n} \frac{\mu_0 - \overline{X}}{\sigma_0} \right| > Z_{\alpha/2}$$

This is a two-sided test for normal location.

6. The normal linear model

6.1. Multivariate normal distribution

Let $X = (X_1, \dots, X_n)$ be a vector of random variables. Then we define

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}; \quad \operatorname{Var}(X) = \left(\mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}\left[X_j\right])\right]\right)_{i,j}$$

The familiar linearity results are

$$\mathbb{E}[AX + b] = A\mathbb{E}[X] + b; \quad A \operatorname{Var}(X) A^\intercal$$

where $A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k$ are constant.

Definition. We say that X has a *multivariate normal distribution* if, for any fixed $t \in \mathbb{R}^n$, we have $t^\intercal X \sim N(\mu, \sigma^2)$ for some parameters μ, σ^2.

Proposition. Let X be multivariate normal. Then $AX + b$ is multivariate normal, where $A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k$ are constant.

Proof. Let $t \in \mathbb{R}^k$. Then,

$$t^\intercal(Ax + b) = \underbrace{(A^\intercal t)^\intercal X}_{\sim N(\mu, \sigma^2)} + t^\intercal b$$

which is the sum of a normal random variable and a constant. So this is $N(\mu + t^\intercal b, \sigma^2)$. □

Proposition. A multivariate normal distribution is fully specified by its mean and covariance matrix.

Proof. Let X_1, X_2 be multivariate normal vectors with the same mean μ and the same covariance matrix Σ. We will show that these two random variables have the same moment generating function, and hence the same distribution.

$$M_{X_1}(t) = \mathbb{E}\left[e^{1 \cdot t^\intercal X_1}\right]$$

Note that $t^\intercal X_1$ is univariate normal. Hence, this is equal to

$$M_{X_1}(t) = \exp\left(1 \cdot \mathbb{E}[t^\intercal X_1] + \frac{1}{2} \operatorname{Var}(t^\intercal X_1) \cdot 1^2\right) = \exp\left(t^\intercal \mu + \frac{1}{2} t^\intercal \Sigma t\right)$$

This depends only on μ and Σ, and we obtain the same moment generating function for X_2. □

6.2. Orthogonal projections

Definition. A matrix $P \in \mathbb{R}^{n\times n}$ is an *orthogonal projection* onto its column space $\operatorname{col}(P)$ if, for all $v \in \operatorname{col}(P)$, we have $Pv = v$, and for all $w \in \operatorname{col}(P)^\perp$, we have $Pw = 0$.

Proposition. P is an orthogonal projection if and only if it is idempotent and symmetric.

Proof. If P is idempotent and symmetric, let $v \in \operatorname{col}(P)$, so $v = Pa$ for some $a \in \mathbb{R}^n$. Then, $Pv = PPa = Pa = v$. Now, let $w \in \operatorname{col}(P)^\perp$. By definition, $P^\intercal w = 0$. By symmetry, $Pw = 0$.

Now, suppose P is an orthogonal projection. Any vector $a \in \mathbb{R}^n$ can be uniquely written as $a = v + w$ where $v \in \operatorname{col}(P)$ and $w \in \operatorname{col}(P)^\perp$. Then $PPa = PPv + PPw = Pv = P(v + w) = Pa$. As this holds for all a, we have that P is idempotent. Let $u_1, u_2 \in \mathbb{R}^n$, and note $(Pu_1) \cdot ((I - P)u_2) = 0$, as $Pu_1 \in \operatorname{col}(P)$ and $(I - P)u_2 \in \operatorname{col}(P)^\perp$. We have $u_1^\intercal P^\intercal (I - P)u_2 = 0$. Since this holds for all u_1, u_2, $P^\intercal(I - P) = 0$ so $P^\intercal = P^\intercal P$. Note that $P^\intercal P$ is symmetric, so $P^\intercal$ is symmetric, and hence P is symmetric. □

Corollary. Let P be an orthogonal projection matrix. Then $I - P$ is also an orthogonal projection matrix.

Proof. Clearly, if P is symmetric, so is $I - P$, so it suffices to prove idempotence. We have $(I - P)(I - P) = I - 2P + P^2 = I - 2P + P = I - P$ as required. □

Proposition. If P is an orthogonal projection, then $P = UU^\intercal$ where the columns of U are an orthonormal basis for the column space of P.

Proof. First, we show that $UU^\intercal$ is an orthogonal projection. This is clearly symmetric. It is idempotent: $UU^\intercal UU^\intercal = UU^\intercal$ since $U^\intercal U = I$, as the columns of U form an orthonormal basis for the column space of P. Further, the column space of P is exactly the column space of $UU^\intercal$. □

Proposition. The rank of an orthogonal projection matrix is equal to its trace.

Proof. The rank is the dimension of the column space, which is $\operatorname{rank} P = \operatorname{rank}(U^\intercal U) = \operatorname{tr}(U^\intercal U) = \operatorname{tr}(UU^\intercal) = \operatorname{tr} P$. □

Theorem. Let X be multivariate normal, where $X \sim N(0, \sigma^2 I)$, and let P be an orthogonal projection. Then

(i) $PX \sim N(0, \sigma^2 P)$, and $(I - P)X \sim N(0, \sigma^2(I - P))$, and these two random variables are independent;

(ii) $\frac{\|PX\|^2}{\sigma^2} \sim \chi^2_{\operatorname{rank} P}$.

Proof. The vector $(P, I-P)^\intercal X$ is multivariate normal, since it is a linear function of X. This distribution is fully specified by its mean and variance.

$$\mathbb{E}\left[\begin{pmatrix} PX \\ (I-P)X \end{pmatrix}\right] = \begin{pmatrix} P \\ I-P \end{pmatrix} \mathbb{E}[X] = 0$$

Further,

$$\operatorname{Var}\left(\begin{pmatrix} PX \\ (I-P)X \end{pmatrix}\right) = \begin{pmatrix} P \\ I-P \end{pmatrix} \sigma^2 I \begin{pmatrix} P \\ I-P \end{pmatrix}^\intercal = \sigma^2 \begin{pmatrix} P^2 & P(I-P) \\ P(I-P) & (I-P)^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I-P \end{pmatrix}$$

Now we must show that the variables $PX, (I-P)X$ are independent. Let $Z \sim N(0, \sigma^2 P), Z' \sim N(0, \sigma^2(I-P))$ be independent. Then we can see that $(Z, Z')^\intercal$ is multivariate normal with

$$\mu = 0; \quad \Sigma = \begin{pmatrix} P & 0 \\ 0 & I-P \end{pmatrix}$$

Hence $(PX, (1-P)X)^\intercal$ is equal in distribution to $(Z, Z')^\intercal$. So PX is independent of $(I-P)X$.

We must show that $\frac{\|PX\|^2}{\sigma^2} \sim \chi^2_{\operatorname{rank} P}$. Note that

$$\frac{\|PX\|^2}{\sigma^2} = \frac{X^\intercal P^\intercal PX}{\sigma^2} = \frac{X^\intercal (UU^\intercal)^\intercal UU^\intercal X}{\sigma^2} = \frac{\|U^\intercal X\|^2}{\sigma^2}$$

Note, $U^\intercal X \sim N(0, \sigma^2 U^\intercal U) = N(0, \sigma^2 I_{\operatorname{rank} P})$. So

$$\frac{(U^\intercal X)_i}{\sigma} \overset{\text{iid}}{\sim} N(0,1)$$

for $i = 1, \dots, \operatorname{rank} P$. Hence

$$\frac{\|PX\|^2}{\sigma^2} = \sum_{i=1}^{\operatorname{rank} P} \left(\frac{(U^\intercal X)_i}{\sigma}\right)^2 \sim \chi^2_{\operatorname{rank} P}$$

□

Theorem. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The maximum likelihood estimators for μ and σ are

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_i X_i; \quad \hat{\sigma}^2 = \frac{S_{xx}}{n} = \frac{\sum_i \left(X_i - \overline{X}\right)^2}{n}$$

Further,

(i) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;

(ii) $\frac{S_{xx}}{\sigma^2} \sim \chi^2_{n-1}$;

(iii) $\overline{X}, S_{xx}$ are independent.

Proof. Let P be the square $n \times n$ matrix with all entries $\frac{1}{n}$. This is an orthogonal projection matrix, as it is symmetric and idempotent. Note that

$$PX = \begin{pmatrix} \overline{X} \\ \vdots \\ \overline{X} \end{pmatrix}$$

We will write the observations X as

$$X = \underbrace{\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}}_{M} + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2 I)$$

Note that $\overline{X}$ is a function of $P\varepsilon$, since $\overline{X} = (PX)_1 = (PM + P\varepsilon)_1$. Further,

$$S_{xx} = \sum_i \left(X_i - \overline{X}\right)^2 = \|X - PX\|^2 = \|(I - P)X\|^2 = \|(I - P)\varepsilon\|^2$$

Hence S_{xx} is a function of $(I - P)\varepsilon$. Since $P\varepsilon$ and $(I - P)\varepsilon$ are independent, $\overline{X}$ and S_{xx} are independent. Since $I - P$ is a projection with rank equal to its trace $n - 1$, we apply the previous theorem to obtain

$$S_{xx} = \|(I - P)\varepsilon\|^2 \chi^2_{n-1}$$

□

6.3. Linear model

Suppose we have data in pairs $(x_1, Y_1), \dots, (x_n, Y_n)$, where $Y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$. The Y_i are known as the *response* variables, or the *dependent* variables. The x_{i1}, x_{ip} are the *predictors*, or *independent* variables. We will model the expectation of the response Y_i as a linear function of the predictors $(x_{i1}, \dots, x_{ip})$.

Example. Let Y_i be the number of insurance claims that driver i makes in a given year, and $x_{i1}, \dots, x_{ip}$ is a set of variables about the specific driver. Predictors include age, the number of years they have held their license, and the number of points on their license, for instance.

We assume that

$$Y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where $\alpha \in \mathbb{R}$ is an *intercept*, β_i are the *coefficients*, and ε is a *noise vector*, which is a random variable. The intercept and coefficients are the parameters of interest. We will often eliminate the intercept by making one of the predictors $x_{i1} = 1$ for all i, so β_1 plays the role of the intercept.

Note that we can use a linear model to model nonlinear relationships. For example, suppose $Y_i = a + bz_i + cz_i^2 + \varepsilon_i$. We can rephrase this as a linear model with $x_i = (1, z_i, z_i^2)$.

The coefficient β_j can be interpreted as the effect on Y_i of increasing x_{ij} by one, while keeping all other predictors fixed. This cannot be interpreted as a causal relationship, unless this is a randomised control experiment.

6.4. Matrix formulation

Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}; \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}; \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

We call X the *design matrix*. The linear model is that

$$Y = X\beta + \varepsilon$$

$X\beta$ is considered fixed. Since ε is random, this makes Y into a random variable.

6.5. Assumptions

We make a number of *moment assumptions* on the noise vector ε. This allows us to deduce more results about the linear model.

(i) $\mathbb{E}[\varepsilon] = 0 \implies \mathbb{E}[Y_i] = x_i^\intercal \beta$;

(ii) $\operatorname{Var}(\varepsilon) = \sigma^2 I$, which is equivalent to both $\operatorname{Var}(\varepsilon_i) = \sigma^2$ and $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$. This property is known as *homoscedasticity*.

We will always assume that the design matrix X has full rank p, or equivalently, that it has linearly independent columns. Since $X \in \mathbb{R}^{n \times p}$, this requires that $n \geq p$, so we need at least as many samples as we have predictors.

6.6. Least squares estimation

Definition. The *least squares estimator* $\hat{\beta}$ minimises the *residual sum of squares*, which is

$$S(\beta) = \|Y - X\beta\|^2 = \sum_i \left(Y_i - x_i^\intercal \beta\right)^2$$

The term $Y_i - x_i^\intercal \beta$ is called the ith residual.

Since $S(\beta)$ is a positive definite quadratic in β, it is minimised at the stationary point.

$$\left.\frac{\partial S(\beta)}{\partial \beta_k}\right|_{\beta=\hat{\beta}} = 0 \iff \forall k,\ -2\sum_{i=1}^{n} x_{ik}\left(Y_i - \sum_k x_{ij}\hat{\beta}_j\right) = 0 \iff X^\intercal X\hat{\beta} = X^\intercal Y$$

As X has full column rank, $X^\intercal X$ is invertible.

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$$

This is notably a linear function of Y, given fixed X. Note that

$$\mathbb{E}\left[\hat{\beta}\right] = (X^\intercal X)^{-1} X^\intercal \mathbb{E}\left[Y\right] = (X^\intercal X)^{-1} X^\intercal X \beta = \beta$$

So $\hat{\beta}$ is an unbiased estimator. Further,

$$\begin{aligned}
\operatorname{Var}\left(\hat{\beta}\right) &= (X^\intercal X)^{-1} X^\intercal \operatorname{Var}\left(Y\right) \left[(X^\intercal X)^{-1} X^\intercal\right]^\intercal \\
&= (X^\intercal X)^{-1} X^\intercal \sigma^2 I \left[(X^\intercal X)^{-1} X^\intercal\right]^\intercal \\
&= \sigma^2 (X^\intercal X)^{-1}
\end{aligned}$$

Theorem (Gauss–Markov theorem)**.** Let an estimator $\beta^\star$ of β be unbiased and a linear function of Y, so $\beta^\star = CY$. Then, for any fixed $t \in \mathbb{R}^p$, we have

$$\operatorname{Var}\left(t^\intercal \hat{\beta}\right) \leq \operatorname{Var}\left(t^\intercal \beta^\star\right)$$

where $\hat{\beta}$ is the least squares estimator. We say that $\hat{\beta}$ is the best linear unbiased estimator (BLUE).

Remark. We can think of $t \in \mathbb{R}^p$ as a vector of predictors for a new sample. Then $t^\intercal \hat{\beta}$ is the prediction for $\mathbb{E}\left[Y_i\right]$ for this new sample, using the least squares estimator. $t^\intercal \beta^\star$ is the prediction with $\beta^\star$. In both cases, the prediction is unbiased.

Proof. Note that

$$\operatorname{Var}\left(t^\intercal \beta^\star\right) - \operatorname{Var}\left(t^\intercal \hat{\beta}\right) = t^\intercal \left[\operatorname{Var}\left(\beta^\star\right) - \operatorname{Var}\left(\hat{\beta}\right)\right] t$$

To prove that this quantity is always non-negative, we must show that $\operatorname{Var}\left(\beta^\star\right) - \operatorname{Var}\left(\hat{\beta}\right)$ is positive semidefinite. Let $A = C - (X^\intercal X)^{-1} X^\intercal$. Note that $\mathbb{E}\left[AY\right] = \mathbb{E}\left[\beta^\star\right] - \mathbb{E}\left[\hat{\beta}\right] = 0$. Also, $\mathbb{E}\left[AY\right] = A\mathbb{E}\left[Y\right] = AX\beta$. This holds for all β, so $AX = 0$. Now, since $X^\intercal X$ is symmetric,

$$\begin{aligned}
\operatorname{Var}\left(\beta^\star\right) &= \operatorname{Var}\left(CY\right) \\
&= \operatorname{Var}\left((A + (X^\intercal X)^{-1} X^\intercal) Y\right) \\
&= \left[A + (X^\intercal X)^{-1} X^\intercal\right] \operatorname{Var}\left(Y\right) \left[A + (X^\intercal X)^{-1} X^\intercal\right]^\intercal \\
&= \left[A + (X^\intercal X)^{-1} X^\intercal\right] \sigma^2 I \left[A + (X^\intercal X)^{-1} X^\intercal\right]^\intercal \\
&= \sigma^2 \left(AA^\intercal + (X^\intercal X)^{-1} + AX(X^\intercal X)^{-1} + (X^\intercal X)^{-1} X^\intercal A^\intercal\right) \\
&= \sigma^2 AA^\intercal + \operatorname{Var}\left(\hat{\beta}\right) \\
\operatorname{Var}\left(\beta^\star\right) - \operatorname{Var}\left(\hat{\beta}\right) &= \sigma^2 AA^\intercal
\end{aligned}$$

Note that the outer product $AA^\intercal$ is always positive semidefinite. □

6.7. Fitted values and residuals

Definition. The *fitted values* are $\hat{Y} = X\hat{\beta} = X(X^\intercal X)^{-1}X^\intercal Y$, where $P = X(X^\intercal X)^{-1}X^\intercal$ is the *hat matrix*. The *residuals* are $Y - \hat{Y} = (I - P)Y$.

Proposition. P is the orthogonal projection onto the column space of the design matrix.

Proof. If v is in the column space of X, then $v = Xb$ for some b. Hence

$$Pv = X(X^\intercal X)^{-1}X^\intercal Xb = Xb = v$$

If w is in the orthogonal complement, then

$$Pw = X(X^\intercal X)^{-1}\underbrace{X^\intercal w}_{0} = 0$$

□

Corollary. The fitted values are an orthogonal projection of the response variables to the column space of the design matrix. The residuals are orthogonal to the column space.

6.8. Normal linear model

The normal linear model is a linear model under the assumption that $\varepsilon \sim N(0, \sigma^2 I)$, where σ^2 is unknown. The parameters in the model are now (β, σ^2). The likelihood function in the normal linear model is

$$L(\beta, \sigma^2) = f_Y(y \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - x_i^\intercal \beta)^2\right\}$$

The log-likelihood is

$$\ell(\beta, \sigma^2) = \text{constant} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

To maximise this as a function of β for any fixed σ^2, we must minimise the residual sum of squares $S(\beta) = \|Y - X\beta\|^2$. So $\hat{\beta} = (X^\intercal X)^{-1}X^\intercal Y$ is the maximum likelihood estimator of β. Further, $\hat{\sigma}^2 = n^{-1}\left\|Y - X\hat{\beta}\right\|^2 = n^{-1}\left\|\hat{Y} - Y\right\|^2 = n^{-1}\|(I - P)Y\|^2$.

Theorem. In the normal linear model,

(i) $\hat{\beta} \sim N(\beta, \sigma^2(X^\intercal X)^{-1})$;

(ii) $n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$;

(iii) $\hat{\beta}, \hat{\sigma}^2$ are independent.

Proof. We prove each part separately.

(i) We already know that $\mathbb{E}\left[\hat{\beta}\right] = \beta$, and $\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2(X^\intercal X)^{-1}$. So it suffices to show that $\hat{\beta}$ is a normal vector. Since $\hat{\beta} = (X^\intercal X)^{-1}X^\intercal Y$, it is a linear function of a normal vector, so is a normal vector.

(ii) Observe that

$$n\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|(I-P)Y\|^2}{\sigma^2} = \frac{\|(I-P)(X\beta+\varepsilon)\|^2}{\sigma^2}$$

Since $(I-P)X = 0$ as P is the orthogonal projection onto the column space of X,

$$n\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|(I-P)\varepsilon\|^2}{\sigma^2} \sim \chi^2_{\mathrm{tr}(I-P)}$$

where $\mathrm{tr}(I-P) = \mathrm{tr}\, I - \mathrm{tr}\, P = n - p$ since $X \in \mathbb{R}^{n\times p}$ is assumed to have full rank.

(iii) Note that $\hat{\sigma}^2$ is a function of $(I-P)\varepsilon$, and

$$\begin{aligned}\hat{\beta} &= (X^\intercal X)^{-1}X^\intercal Y\\ &= (X^\intercal X)^{-1}X^\intercal(X\beta+\varepsilon)\\ &= \beta + (X^\intercal X)^{-1}X^\intercal\varepsilon\\ &= \beta + (X^\intercal X)^{-1}X^\intercal P\varepsilon\end{aligned}$$

is a function of $P\varepsilon$. Since $(I-P)\varepsilon$ and $P\varepsilon$ are independent, so are $\hat{\beta}, \hat{\sigma}^2$.

□

Note,

$$\mathbb{E}\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = \mathbb{E}\left[\chi^2_{n-p}\right] = n-p \implies \mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2\cdot\frac{n-p}{n} < \sigma^2$$

Hence this $\hat{\sigma}^2$ is a biased estimator, but asymptotically unbiased.

6.9. Inference

Definition. Let $U \sim N(0,1)$ and $V \sim \chi^2_n$ be independent random variables. Then

$$T = \frac{U}{\sqrt{\frac{V}{n}}}$$

has a *t_n-distribution*.

As $n \to \infty$, this approaches the standard normal distribution.

Definition. Let $V \sim \chi^2_n$ and $W \sim \chi^2_m$ be independent random variables. Then

$$F = \frac{V/n}{W/m}$$

has an *$F_{n,m}$-distribution*.

Example. We consider a $100(1-\alpha)\%$ confidence interval for one of the coefficients β in the normal linear model $Y = X\beta + \varepsilon$. Without loss of generality, we will consider β_1.

We begin by finding a *pivot*, which is a distribution that does not depend on the parameters of the model. By standardising the above form of $\hat{\beta}$,

$$\frac{\beta_1 - \hat{\beta}_1}{\sqrt{\sigma^2 (X^\intercal X)^{-1}_{11}}} \sim N(0,1)$$

where M^{-1}_{11} is the top left entry in the matrix M^{-1}. This random variable is independent from $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Now, to construct a pivot, we find

$$\frac{\frac{\beta_1 - \hat{\beta}_1}{\sqrt{\sigma^2 (X^\intercal X)^{-1}_{11}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2} \cdot \frac{n}{n-p}}} \sim \frac{U}{\sqrt{\frac{V}{n}}} \sim t_{n-p}$$

The σ^2 terms cancel, so the statistic is a function only of β_1 and functions of the data. Then,

$$\mathbb{P}_{\beta,\sigma^2}\left(-t_{n-p}\left(\frac{\alpha}{2}\right) \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{(X^\intercal X)^{-1}_{11}}} \sqrt{\frac{n-p}{n\hat{\sigma}^2}} \le t_{n-p}\left(\frac{\alpha}{2}\right) \right) = 1 - \alpha$$

since the t distribution is symmetric about zero. Rearranging to find an interval for β_1,

$$\mathbb{P}_{\beta,\sigma^2}\left(\hat{\beta}_1 - t_{n-p}\left(\frac{\alpha}{2}\right) \frac{\sqrt{(X^\intercal X)^{-1}_{11}\hat{\sigma}^2}}{\sqrt{(n-p)/n}} \le \beta_1 \le \hat{\beta}_1 + t_{n-p}\left(\frac{\alpha}{2}\right) \frac{\sqrt{(X^\intercal X)^{-1}_{11}\hat{\sigma}^2}}{\sqrt{(n-p)/n}} \right) = 1 - \alpha$$

Hence,

$$I = \left[\hat{\beta}_1 \pm t_{n-p}\left(\frac{\alpha}{2}\right) \frac{\sqrt{(X^\intercal X)^{-1}_{11}\hat{\sigma}^2}}{\sqrt{(n-p)/n}} \right]$$

is a $100(1-\alpha)\%$ confidence interval for β_1.

Consider a test for $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. By connecting tests and confidence intervals, we can test H_0 with size α by rejecting this null hypothesis when zero is not contained within the confidence interval I.

Consider a special case where $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where μ, σ^2 are unknown, and we want to infer results about μ. Note that this is a special case of the normal linear model where

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}; \quad \beta = (\mu)$$

So we can infer a confidence interval for μ using the above statistic.

Example. Consider a $100(1-\alpha)\%$ confidence set for β as a whole. Note that

$$\hat{\beta} - \beta \sim N(0, \sigma^2 (X^\intercal X)^{-1})$$

Then,

$$(X^\intercal X)^{1/2}(\hat{\beta} - \beta) \sim N(0, \sigma^2 (X^\intercal X)^{1/2}(X^\intercal X)^{-1}(X^\intercal X)^{1/2}) \sim N(0, \sigma^2 I)$$

where $(X^\intercal X)^{1/2}$ is obtained using the eigendecomposition of the positive definite matrix $X^\intercal X$. Hence,

$$\frac{\left\|(X^\intercal X)^{1/2}(\hat{\beta} - \beta)\right\|^2}{\sigma^2} \sim \chi^2_p$$

We can also write this as

$$\frac{\left\|(X^\intercal X)^{1/2}(\hat{\beta} - \beta)\right\|^2}{\sigma^2} = \frac{\left\|X(\hat{\beta} - \beta)\right\|^2}{\sigma^2}$$

Since this is a function of $\hat{\beta}$, this is independent of any function of $\hat{\sigma}^2$. In particular, it is independent of $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Thus, we can form a pivot by

$$\frac{\left\|X(\hat{\beta} - \beta)\right\|^2/(\sigma^2 p)}{\hat{\sigma}^2 n/(\sigma^2(n-p))} \sim \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \sim F_{p,n-p}$$

This does not depend on σ^2. For all β, σ^2,

$$\mathbb{P}_{\beta,\sigma^2}\left(\frac{\left\|X(\hat{\beta} - \beta)\right\|^2/p}{\hat{\sigma}^2 n/(n-p)} \leq F_{p,n-p}(\alpha)\right) = 1 - \alpha$$

because the F distribution has support only on the positive real line. It is nontrivial to express this as a region for β since it is vector-valued. We can say, however, that

$$\left\{\beta' \in \mathbb{R}^p : \frac{\left\|X(\hat{\beta} - \beta)\right\|^2/p}{\hat{\sigma}^2 n/(n-p)} \leq F_{p,n-p}(\alpha)\right\}$$

is a $100(1-\alpha)\%$ confidence set for β.

This set is an ellipsoid centred at $\hat{\beta}$. The shape of the ellipsoid depends on the design matrix X; the principal axes are given by eigenvectors of $X^\intercal X$.

The above two results are exact; no approximations were made.

6.10. F-tests

We wish to test whether a collection of predictors β_i are equal to zero. Without loss of generality, we will take the first $p_0 \leq p$ predictors. We have $H_0 : \beta_1 = \cdots = \beta_{p_0} = 0$, and $H_1 = \beta \in \mathbb{R}^p$. We denote $X = (X_0, X_1)$ as a block matrix with $X_0 \in \mathbb{R}^{n\times p_0}$ and $X_1 \in \mathbb{R}^{n\times(p-p_0)}$, and we denote $\beta = (\beta^0, \beta^1)^\intercal$ similarly. The null model has $\beta^0 = 0$. This is a linear model $Y = X\beta + \varepsilon = X_1\beta^1 + \varepsilon$. We will write $P = X(X^\intercal X)^{-1}X^\intercal$ and $P_1 = X_1(X_1^\intercal X_1)^{-1}X_1^\intercal$. Note that as X and P have full rank, so must X_1, P_1.

Lemma. $(I - P)(P - P_1) = 0$, and $P - P_1$ is an orthogonal projection with rank p_0.

Proof. $P - P_1$ is symmetric since P and P_1 are symmetric. It is also idempotent, since

$$(P - P_1)(P - P_1) = P^2 - P_1P - PP_1 + P_1^2 = P - P_1 - P_1 + P_1 = P - P_1$$

since P_1 projects onto the column space of X_1. Hence $P - P_1$ is indeed an orthogonal projection matrix. The rank is $\operatorname{rank}(P - P_1) = \operatorname{tr}(P - P_1) = \operatorname{tr} P - \operatorname{tr} P_1 = p - (p - p_0) = p_0$. Also,

$$(I - P)(P - P_1) = P - P_1 - P + PP_1 = P - P_1 - P + P_1 = 0$$

□

Recall that the maximum log-likelihood in the normal linear model is given by

$$\ell(\hat\beta, \hat\sigma^2) = \frac{-n}{2}\log\hat\sigma^2 - \frac{n}{2}\cdot\text{constant} = \frac{-n}{2}\log\frac{\|(I-P)Y\|^2}{n} + \text{constant}$$

The generalised likelihood ratio statistic is

$$\begin{aligned} 2\log\Lambda &= 2\sup_{\beta\in\mathbb{R}^p,\sigma^2>0}\ell(\beta,\sigma^2) - 2\sup_{\beta_0=0,\beta_1\in\mathbb{R}^{p-p_0},\sigma^2>0}\ell(\beta,\sigma^2) \\ &= n\left[-\log\frac{\|(I-P)Y\|^2}{n} + \log\frac{\|(I-P_1)Y\|^2}{n}\right] \end{aligned}$$

Wilks' theorem applies here, showing that $2\log\Lambda \sim \chi^2_{p_0}$ asymptotically as $n \to \infty$ with p, p_0 fixed. However, we can find an exact test, so using Wilks' theorem will not be necessary. $2\log\Lambda$ is monotone in

$$\begin{aligned} \frac{\|(I-P_1)Y\|^2}{\|(I-P)Y\|^2} &= \frac{\|(I-P+P-P_1)Y\|^2}{\|(I-P)Y\|^2} \\ &= \frac{\|(I-P)Y\|^2 + \|(P-P_1)Y\|^2 + 2Y^\intercal(I-P)(P-P_1)}{\|(I-P)Y\|^2} \\ &= \frac{\|(I-P)Y\|^2 + \|(P-P_1)Y\|^2}{\|(I-P)Y\|^2} \\ &= 1 + \frac{\|(P-P_1)Y\|^2}{\|(I-P)Y\|^2} \end{aligned}$$

The generalised likelihood ratio test rejects when the F-statistic

$$F = \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \cdot \frac{1/p_0}{1/(n-p)}$$

is large.

Theorem. Under $H_0 : \beta_1 = \cdots = \beta_{p_0} = 0$, in the normal linear model,

$$F = \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \cdot \frac{1/p_0}{1/(n-p)} \sim F_{p_0, n-p}$$

Proof. Recall that

$$\|(I - P)Y\|^2 = \|(I - P)\varepsilon\|^2 \sim \chi^2_{n-p} \cdot \sigma^2$$

Therefore it suffices to show that $\|(P - P_1)Y\|^2$ is an independent $\chi^2_{p_0} \cdot \sigma^2$ random variable. Under H_0, we have that

$$(P - P_1)Y = (P - P_1)(X\beta + \varepsilon) = (P - P_1)(X_1\beta^1 + \varepsilon) = (P - P_1)\varepsilon$$

since P, P_1 preserve X_1. Hence, $\|(P - P_1)Y\|^2 = \|(P - P_1)\varepsilon\|^2 \sim \chi^2_{\operatorname{rank}(P-P_1)} \cdot \sigma^2 = \chi^2_{p_0} \cdot \sigma^2$. We must now show independence between $(I-P)Y$ and $(P-P_1)Y$. The vectors $(I-P)\varepsilon, (P-P_1)\varepsilon$ are independent; indeed,

$$E = \begin{pmatrix} (I - P)\varepsilon \\ (P - P_1)\varepsilon \end{pmatrix}$$

is a multivariate normal vector, and

$$\mathbb{E}[E] = 0; \quad \operatorname{Var}(E) = \begin{pmatrix} I - P & (I - P)(P - P_1) \\ (I - P)(P - P_1) & P - P_1 \end{pmatrix} = \begin{pmatrix} I - P & 0 \\ 0 & P - P_1 \end{pmatrix}$$

and since $(I - P)\varepsilon$ and $(P - P_1)\varepsilon$ are elements of a multivariate normal vector and are uncorrelated, they are independent as required. □

The generalised likelihood ratio test of size α rejects H_0 when $F > F^{-1}_{p_0,n-p}(\alpha)$. This is an exact test for all n, p, p_0. Previously, we found a test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. This is a special case of the F-test derived above, where $p_0 = 1$. The previous test of size α rejects H_0 when

$$|\hat{\beta}| > t_{n-p}\Big(\frac{\alpha}{2}\Big)\sqrt{\frac{\hat{\sigma}^2 n (X^\intercal X)^{-1}_{11}}{n-p}}$$

We will show that these two tests are equivalent; they reject H_0 in the same critical region. The t-test rejects if and only if

$$\hat{\beta}_1^2 > t_{n-p}\Big(\frac{\alpha}{2}\Big)^2 \frac{\hat{\sigma}^2 n (X^\intercal X)^{-1}_{11}}{n-p}$$

Note that $t_{n-p}\left(\frac{\alpha}{2}\right)^2 = F_{1,n-p}(\alpha)$, since

$$U \sim N(0,1);\ W \sin \chi^2_n \implies T = \frac{U}{\sqrt{W/n}} \implies T^2 = \frac{U^2}{W/n} = \frac{V/1}{W/n} \sim F_{1,n}$$

where $V \sim \chi^2_1$. Hence,

$$\frac{\hat{\beta}_1/(X^\intercal X)^{-1}_{11}}{\hat{\sigma}^2 n/(n-p)} > F_{1,n-p}(\alpha)$$

It suffices to show that

$$\frac{\hat{\beta}_1}{(X^\intercal X)^{-1}_{11}} = \frac{\|(P-P_1)Y\|^2}{\underbrace{p_0}_{=1}};\quad \frac{\hat{\sigma}^2 n}{n-p} = \frac{\|(I-P)Y\|^2}{n-p}$$

We have already shown the latter part. For $\hat{\beta}_1$, note that in this case, $P - P_1$ is a projection of rank 1 onto the one-dimensional subspace spanned by the vector $v = (I-P)X^0$ where X^0 is the first column in the matrix X. First, note the following identity.

$$X_0^\intercal (I-P_1) = v^\intercal = v^\intercal (P-P_1) = X_0^\intercal (I-P_1)(P-P_1) = X_0^\intercal (I-P_1)P$$

Then,

$$\begin{aligned}
\|(P-P_1)Y\|^2 &= \left\| \frac{v}{\|v\|}\left(\frac{v}{\|v\|}\right)^\intercal Y \right\|^2 \\
&= \frac{(v^\intercal Y)^2}{\|v\|^2} = \frac{(X_0^\intercal (I-P_1)Y)^2}{\|(I-P_1)X_0\|^2} \\
&= \frac{(X_0^\intercal (I-P_1)PY)^2}{\|(I-P_1)X_0\|^2} \\
&= \frac{(X_0^\intercal (I-P_1)X\hat{\beta})^2}{\|(I-P_1)X_0\|^2}
\end{aligned}$$

Note that $(I-P_1)X = [(I-P_1)X_0, 0, \dots, 0]$. Hence,

$$\begin{aligned}
\|(P-P_1)Y\|^2 &= \frac{\|(I-P_1)X_0\|^4 \hat{\beta}_1}{\|(I-P_1)X_0\|^2} \\
&= \|(I-P_1)X_0\|^2 \hat{\beta}_1
\end{aligned}$$

Finally, we show that

$$(X^\intercal X)^{-1}_{11} = \frac{1}{\|(I-P_1)X_0\|^2}$$

using the Woodbury identity for blockwise matrix inversion. Hence,

$$\frac{\hat{\beta}_1^2}{(X^\intercal X)^{-1}_{11}} = \|(P-P_1)Y\|^2$$

as required.

6.11. Analysis of variance

Suppose we investigate responses of patients after receiving one of three treatments, including a control, which will be given index 1. We will consider only one predictor, denoting which treatment a given patient received. Consider the linear model

$$Y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$$

where $j = 1, 2, 3$ is the treatment index, and $i = 1, \dots, N$ is the index of a patient in a given group. Let $(\varepsilon_{ij}) \sim N(0, \sigma^2)$ be independent. Without loss of generality, we can set $\mu_1 = 0$, since we have an additional parameter α; this is known as a *corner point* constraint. Then, μ_j should be interpreted as the effect of treatment j relative to treatment 1, which in this case is the control.

Definition. The *analysis of variance (ANOVA)* test on the linear model

$$Y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$$

where $\mu_1 = 0$ is given by

$$H_0 : \mu_2 = \mu_3 = \cdots = 0; \quad H_1 : \mu_2, \mu_3, \cdots \in \mathbb{R}$$

In particular, H_0 gives $\mathbb{E}\left[Y_{ij}\right] = \alpha$.

In our example, $H_0 : \mu_2 = \mu_3 = 0$ and $H_1 : \mu_2, \mu_3 \in \mathbb{R}$. This is a special case of the F-test, since we are testing whether the coefficients μ_i are equal to zero.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} X_1 & X_0 \end{pmatrix}$$

The first column of X, denoted X_1, represents α, and the other columns, denoted X_0, represent μ_2, μ_3. X_0 is eliminated under the null hypothesis. The predictor can be called *categorical*; it is discrete, and entirely dependent on which treatment category a given patient is placed in. Note that X has $3N$ rows, where each block of N consecutive rows is identical. Recall that the F-test uses the test statistic

$$F = \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \cdot \frac{1/p_0}{1/(n-p)} \sim F_{p_0, n-p}$$

For this test, P projects onto the space of vectors in $\mathbb{R}^{3N}$ which are constant over treatment groups. In other words, let

$$\overline{Y}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}$$

Then,

$$PY = \left(\underbrace{\overline{Y}_1, \dots, \overline{Y}_1}_{N \text{ entries}}, \underbrace{\overline{Y}_2, \dots, \overline{Y}_2}_{N \text{ entries}}, \underbrace{\overline{Y}_3, \dots, \overline{Y}_3}_{N \text{ entries}} \right)^\intercal$$

P_1 projects onto the subspace of constant vectors in $\mathbb{R}^{3N}$, so

$$\overline{Y} = \frac{1}{3N} \sum_{i=1}^{N} \sum_{j=1}^{3} Y_{ij} \implies P_1 Y = \left(\underbrace{\overline{Y}, \dots, \overline{Y}}_{3N \text{ entries}} \right)^\intercal$$

Hence, we can write the F statistic as

$$F = \frac{\sum_{j=1}^{3} N\left(\overline{Y}_j - \overline{Y}\right)^2 / 2}{\sum_{i=1}^{N} \sum_{j=1}^{3} \left(Y_{ij} - \overline{Y}_j\right)^2 / (3N - 3)}$$

We can generatlise this to the case where there are $J > 3$ treatment groups:

$$F = \frac{\sum_{j=1}^{J} N\left(\overline{Y}_j - \overline{Y}\right)^2 / (J-1)}{\sum_{i=1}^{N} \sum_{j=1}^{J} \left(Y_{ij} - \overline{Y}_j\right)^2 / (JN - J)} = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

Remark. This test is sometimes called *one-way* analysis of variance. *Two-way* analysis of variance is a similar analysis in an experiment where groups are defined according to two variables. For instance, the response could be a student's performance in an exam, where the treatments are

(i) completion of supervisions (zero representing not complete, one representing complete); and

(ii) whether a monetary incentive was given (zero representing no incentive, one representing an incentive).

Here, we would have the result Y_{ijk} as the number of marks of student i in group (j, k). The model would be

$$Y_{ijk} = \alpha + \mu_j + \lambda_k + \varepsilon_{ijk}$$

with a constraint without loss of generality that $\mu_0 = \lambda_0 = 0$. The two-way analysis of variance test is then

$$H_0 : \mu_1 = \lambda_1 = 0; \quad H_1 : \mu_1, \lambda_1 \in \mathbb{R}$$

6.12. Simple linear regression

In a linear regression model, we often centre predictors to simplify certain expressions.

$$Y_i = \alpha + \beta(x - \overline{x}) + \varepsilon_i$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the ε_i independently have the usual $N(0, \sigma^2)$ distribution. In this case, the maximum likelihood estimator $(\hat{\alpha}, \hat{\beta})$ takes a simple form. Recall that $(\hat{\alpha}, \hat{\beta})$ minimises

$$S(\alpha, \beta) = \sum_{i=1}^{n} \big(Y_i - \alpha - \beta(x_i - \overline{x})\big)^2$$

Hence,

$$\frac{\partial S(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{n} -2\big(Y_i - \alpha - \beta(x_i - \overline{x})\big) = \sum_{i=1}^{n} -2(Y_i - \alpha)$$

This gives the simple expression

$$\alpha = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$$

Now,

$$\left.\frac{\partial S(\alpha, \beta)}{\partial \beta}\right|_{\alpha=\hat{\alpha}} = \sum_{i=1}^{n} -2\Big(Y_i - \overline{Y} - \beta(x_i - \overline{x})\Big)(x_i - \overline{x})$$

This vanishes when

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \big(Y_i - \overline{Y}\big)(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Note that $\frac{S_{xy}}{n}$ is the sample covariance of X and Y, and $\frac{S_{xx}}{n}$ is the sample variance of X.