Handbook of Applied Mathematics

WANG Zeyu[1]

May 17, 2025

[1]Email: zeyu.wang.0117@outlook.com

Contents

Part 1

Mathematical Foundation

Chapter 1

Analysis

1.1 Calculus

Definition 1.1. A number x is a **lower bound** of a nonempty set S if $\forall s \in S, x \leq s$.

Definition 1.2. A number x is a **upper bound** of a nonempty set S if $\forall s \in S, x \geq s$.

Definition 1.3. Let S be a nonempty set, denoted by $\inf S$ the **infimum** of S where

(1) $\forall s \in S, s \geq \inf S$;

(2) $\forall y > \inf S, \exists s \in S \ \text{ s.t. } \ s < y$.

Definition 1.4. Let S be a nonempty set, denoted by $\inf S$ the **supremum** of S where

(1) $\forall s \in S, s \leq \sup S$;

(2) $\forall y < \sup S, \exists s \in S \ \text{ s.t. } \ s > y$.

Theorem 1.5. Let $S_1 \subseteq S_2$, then $\inf S_1 \geq \inf S_2, \sup S_1 \leq \sup S_2$.

Corollary 1.6. $\inf \emptyset = +\infty, \sup \emptyset = -\infty$.

Theorem 1.7. A set Ω is **closed** if it contains all the limits of convergent sequences of points in Ω.

Definition 1.8. A set Ω is **bounded** if there exists $R \in \mathbb{R}^+$ such that $\Omega \subseteq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq R\}$.

Theorem 1.9. (Bolzano-Weierstrass) Let $\Omega \subset \mathbb{R}^n$ a bounded closed set. If $\left\{\mathbf{x}^{[k]}\right\}_{k=1}^{\infty} \subseteq \Omega$, then there exists $\mathbf{x}^* \in \Omega$ and a subsequence $\left\{\mathbf{x}^{[k_i]}\right\}_{i=1}^{\infty}$ such that

$$\lim_{i \to \infty} \mathbf{x}^{[k_i]} = \mathbf{x}^*.$$

Definition 1.10. A bounded closed set in $\mathbb{R}^n$ is called a **compact set**.

Theorem 1.11. Let Ω be a nonempty set and $f \in C(\Omega)$, then f achieves its infimum and supremum over Ω, i.e.

$$\exists x, y \in \Omega, f(x) = \inf_{\Omega} f, f(y) = \sup_{\Omega} f.$$

Theorem 1.12. (Rolle's theorem) Let $f \in C([a,b]) \cap C^1((a,b))$, if $f(a) = f(b)$, then there exists a point $\xi \in (a,b)$ such that $f'(\xi) = 0$.

Theorem 1.13. (Generalized Rolle's theorem) Given $n \geq 2$ and $f \in C^{n-1}([a,b])$ with $f^{(n)}(x)$ exists at each point of (a,b), if $f(x_0) = \cdots f(x_n) = 0$ for $a \leq x_0 < \cdots < x_n \leq b$, then there exists a point $\xi \in (a,b)$ such that $f^{(n)}(\xi) = 0$.

Theorem 1.14. (Taylor's theorem with remainder term) Let f be $n+1$ times differentiable on an open interval containing $[a,b]$, then there exists $\xi \in (a,b)$,

$$f(b) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(b-a)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(b-a)^{n+1}$$

Theorem 1.15. (High-dimensional Taylor's theorem with remainder term) Let $f \in C^1(\mathbb{R}^n), \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then there exists $\xi \in (0,1)$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + (\nabla f((1-\xi)\mathbf{x} + \xi\mathbf{y}))^T(\mathbf{y} - \mathbf{x}).$$

Let $f \in C^2(\mathbb{R}^n), \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then there exists $\xi \in (0,1)$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + (\nabla f(\mathbf{x}))^T(\mathbf{y} - \mathbf{x}) + (\mathbf{y} - \mathbf{x})^T \nabla^2 f((1-\xi)\mathbf{x} + \xi\mathbf{y})(\mathbf{y} - \mathbf{x}).$$

Theorem 1.16. Let $f \in C^2(\mathbb{R}^n)$ and there exists L such that $L \geq \left\|\nabla^2 f(\mathbf{x})\right\|_2$ for all $x \in \mathbb{R}^n$, then

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \|\nabla f(u) - \nabla f(v)\|_2 \leq L\|\mathbf{u} - \mathbf{v}\|_2.$$

Theorem 1.17. Let $h \in C^2(\mathbb{R}^m)$ and let $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$. Define $f(x) = h(Ax - b)$ then $f \in C^2(\mathbb{R}^n)$ and $\nabla f(x) = A^T \nabla h(Ax - b)$, $\nabla^2 f(x) = A^T \nabla^2 h(Ax - b)A$.

Theorem 1.18. (Subdifferential inequality) Let h be convex C^1, then

$$\forall x, y \in \mathbb{R}^n, h(y) - h(x) \geq (\nabla h(x))^T(y - x).$$

1.1.1 Generalized derivative

Definition 1.19. (Generalized derivative) For $f(x) \in L^1_{\text{loc}}(\Omega)$, then $g(x) \in L^1_{\text{loc}}(\Omega)$ is called the $|\alpha|$-th order generalized derivative of $f(x)$ if

$$\forall \varphi(x) \in C^\alpha(\Omega), \int_\Omega g(x)\varphi(x)\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f(x)\partial^\alpha \varphi(x)\mathrm{d}x,$$

with the notation

$$D^\alpha f(x) = g(x).$$

Theorem 1.20. If $f(x) \in C^\alpha(\Omega)$, then $D^\alpha f(x) = f^\alpha(x)$.

Theorem 1.21. Given $\Omega = \Omega_1 \cup \Omega_1$, $m(\Omega_1 \cap \Omega_1) = 0$, and $f \in C\left(\overline{\Omega}\right) \cap C^1(\Omega_1) \cap C^1(\Omega_2)$, then $D^\alpha f(x)$ exists for $|\alpha| = 1$, and for all $x \in \text{int } \Omega_1 \cup \text{int } \Omega_2$, $D^\alpha f(x) = f^\alpha(x)$.

1.1.2 Convex sets and functions

Definition 1.22. A set $\Omega \subseteq \mathbb{R}^n$ is said to be **convex** if for any $\mathbf{x}, \mathbf{y} \in \Omega$, and $\lambda \in (0,1)$, it holds that $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \Omega$.

Theorem 1.23. Let $\Omega \subseteq \mathbb{R}^n$ be a nonempty closed convex set and $\mathbf{y} \in \mathbb{R}^n$, then there exists a unique solution to the following optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2 \ \text{ s.t. } \ \mathbf{x}\in\Omega.$$

The unique solution is called the projection of $\mathbf{y}$ onto Ω, denoted by $P_\Omega(y)$.

Theorem 1.24. Let $\Omega \subseteq \mathbb{R}^n$ be a nonempty closed convex set, $\mathbf{y}\in\mathbb{R}^n$ and $\mathbf{u}\in\Omega$, then

$$(\mathbf{y}-P_\Omega(\mathbf{y}))^T(\mathbf{u}-P_\Omega(\mathbf{y})) \le 0.$$

Theorem 1.25. (Separation) Let $\Omega \subseteq \mathbb{R}^n$ be a nonempty closed convex set and $\mathbf{y}\in\mathbb{R}^n\setminus\Omega$, then there exists $\mathbf{v}\in\mathbb{R}^n\setminus\{0\}$ and $\alpha\in\mathbb{R}$ so that

$$\mathbf{v}^T\mathbf{y} > \alpha > \mathbf{v}^T\mathbf{u}$$

for all $\mathbf{u}\in\Omega$.

Theorem 1.26. Let $A\in\mathbb{R}^{m\times n}$, then the set $S=\{A\mathbf{y} : \forall i=1,...,n, \mathbf{y}_i \ge 0\}$ is closed and convex.

Definition 1.27. A function $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ is called

- Proper if $\operatorname{dom}(f)=\{\mathbf{x} : f(x)<\infty\}\neq\emptyset$;
- Convex if $\operatorname{epi}(f)=\{(\mathbf{x},r) : r\ge f(\mathbf{x})\}$ is convex;
- Closed if is lower semicontinuous $\underline{\lim}_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \ge f(\mathbf{x}_0)$ (same as $\operatorname{epi}(f)$ is closed).

Theorem 1.28. Let $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$, it is convex iff for any $\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ and $\lambda\in(0,1)$, it holds that

$$f(\lambda\mathbf{u}+(1-\lambda)\mathbf{v}) \le \lambda f(\mathbf{u})+(1-\lambda)f(\mathbf{v}).$$

Theorem 1.29. (First-order condition under convexity) Let $f\in C^1(\mathbb{R}^n)$, if f is convex and $\nabla f(\mathbf{x})=0$, then $\mathbf{x}$ isa global minimizer of f.

Theorem 1.30. Let $f\in C^2(\mathbb{R}^n)$, then f is convex iff $\nabla^2 f(\mathbf{x})\succeq 0$ for all $\mathbf{x}\in\mathbb{R}^n$.

Proposition 1.31. Let $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$, $g:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ both be convex, $A\in\mathbb{R}^{n\times p}$, $b\in\mathbb{R}^n$, $H(\mathbf{x})=A\mathbf{x}-b$ and $\alpha>0$, then the following functions are convex:

$$f+g,\ \alpha f,\ f\circ H = f(A\mathbf{x}-\mathbf{b}),\ \max\{f,g\},\ \|\cdot\|.$$

Proposition 1.32. Let $f:\mathbb{R}^n\to[0,+\infty)$ and $g:[0,+\infty)\to\mathbb{R}$ both be convex and non-decreasing, then $g\circ f = g(f(\mathbf{x}))$ is convex.

1.1.3 Mean value theorem

Theorem 1.33. (Rolle's theorem) Given $n\ge 2$ and $f\in C^{n-1}([a,b])$ with $f^{(n)}(x)$ exists at each point of (a,b), suppose that $f(x_0)=\cdots f(x_n)=0$ for $a\le x_0<\cdots<x_n\le b$, then there is a point $\xi\in(a,b)$ such that $f^{(n)}(\xi)=0$.

Theorem 1.34. (Lagrange's mean value theorem) Given $f\in C^1([a,b])$, then there exists $\xi\in(a,b)$ such that

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Theorem 1.35. (Cauchy's mean value theorem) Given $f, g \in C^1([a, b])$, then there exists $\xi \in (a, b)$ such that

$$(f(b) - f(a))g'(\xi) = (g(b) - g(a))f'(\xi).$$

If $g(a) \neq g(b)$ and $g(\xi) \neq 0$, this is equivalent to

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Theorem 1.36. (First mean value theorems for definite integrals) Given $f \in C([a, b])$ and g integrable and does not change sign on $[a, b]$, then there exists ξ in (a, b) such that

$$\int_a^b f(x)g(x)\mathrm{d}x = f(\xi) \int_a^b g(x)\mathrm{d}x.$$

Theorem 1.37. (Second mean value theorems for definite integrals) Given f a integrable function and g a positive monotonically decreasing function, then there exists ξ in (a, b) such that

$$\int_a^b f(x)g(x)\mathrm{d}x = g(a) \int_a^\xi f(x)\mathrm{d}x.$$

If g is a positive monotonically increasing function, then there exists ξ in (a, b) such that

$$\int_a^b f(x)g(x)\mathrm{d}x = g(b) \int_\xi^b f(x)\mathrm{d}x.$$

If g is a monotonically function, then there exists ξ in (a, b) such that

$$\int_a^b f(x)g(x)\mathrm{d}x = g(a) \int_a^\xi f(x)\mathrm{d}x + g(b) \int_\xi^b f(x)\mathrm{d}x.$$

1.1.4 Series

Definition 1.38. A series $\sum_{n=1}^\infty a_n$ is **absolute convergent** if the series of absolute values $\sum_{n=1}^\infty |a_n|$ converges.

Theorem 1.39. If a series is absolute convergent, then any reordering of it converges to the same limit.

Theorem 1.40. (n-th term test) If $\lim\limits_{n\to\infty} a_n \neq 0$, then the series divergent.

Theorem 1.41. (Direct comparison test) If $\sum_{n=1}^\infty b_n$ is convergent and exists $N > 0$, for all $n > N$, $0 \leq a_n \leq b_n$, then $\sum_{n=1}^\infty a_n$ is convergent; if $\sum_{n=1}^\infty b_n$ is divergent and exists $N > 0$, for all $n > N$, $0 \leq b_n \leq a_n$, then $\sum_{n=1}^\infty a_n$ is divergent.

Theorem 1.42. (Limit comparison test) Given two series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ with $a_n \geq 0, b_n > 0$. Then if $\lim_{n\to\infty} \frac{a_n}{b_n} = c \in (0, \infty)$, then either both series converge or both series diverge.

Theorem 1.43. (Ratio test) Given $\sum_{n=1}^{\infty} a_n$ and

$$R = \limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|, r = \liminf_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|,$$

if $R < 1$, then the series converges absolutely; if $r > 1$, then the series diverges.

Theorem 1.44. (Root test) Given $\sum_{n=1}^{\infty} a_n$ and

$$R = \limsup_{n\to\infty} (|a_n|)^{\frac{1}{n}},$$

if $R < 1$, then the series converges absolutely; if $R > 1$, then the series diverges.

Theorem 1.45. (Integral test) Given $\sum_{n=1}^{\infty} f(n)$ where f is monotone decreasing, then the series converges iff the improper integral

$$\int_1^{\infty} f(x)\mathrm{d}x$$

is finite. In particular,

$$\int_1^{\infty} f(x)\mathrm{d}x \leq \sum_{n=1}^{\infty} f(n) \leq f(1) + \int_1^{\infty} f(x)\mathrm{d}x$$

Theorem 1.46. (Alternating series test) Given $\sum_{n=1}^{\infty} (-1)^n a_n$ where a_n are all positive or negative, then the series converges if $|a_n|$ decreases monotonically and $\lim_{n\to\infty} a_n = 0$.

1.1.5 Multivariable calculus

Theorem 1.47. (Green's theorem) Let Ω be the region in a plane with $\partial\Omega$ a positively oriented, piecewise smooth, simple closed curve. If P and Q are functions of (x, y) defined on an open region containing Ω and have continuous partial derivatives there, then

$$\oint_{\partial\Omega} (P\mathrm{d}x + Q\mathrm{d}y) = \iint_{\Omega} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathrm{d}x\mathrm{d}y$$

where the path of integration along C is anticlockwise.

Theorem 1.48. (Stokes' theorem) Let Ω be a smooth oriented surface in $\mathbb{R}^3$ with $\partial\Omega$ a piecewise smooth, simple closed curve. If $\mathbf{F}(x, y, z) = \left(F_x(x, y, z), F_y(x, y, z), F_z(x, y, z)\right)$ is defined and has continuous first order partial derivatives in a region containing Ω, then

$$\iint_{\Omega} (\nabla \times \mathbf{F}) \cdot \mathrm{d}S(x) = \oint_{\partial\Omega} \mathbf{F} \cdot \mathrm{d}x$$

Theorem 1.49. (Gauss-Green theorem (Divergence theorem)) For a bounded open set $\Omega \in \mathbb{R}^n$ that $\partial\Omega \in C^1$ and a function $\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), ..., F_n(\mathbf{x})) : \overline{\Omega} \to \mathbb{R}^n$ satisfies $\mathbf{F}(\mathbf{x}) \in C^1(\Omega) \cap C\left(\overline{\Omega}\right)$,

$$\int_\Omega \text{div } \mathbf{F}(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\partial\Omega} \mathbf{F}(\mathbf{x}) \cdot \mathbf{n}\mathrm{d}S(x),$$

where $\mathbf{n}$ is outward pointing unit normal vector at $\partial\Omega$.

Definition 1.50. An **implicit function** is a function of the form

$$F(x_1, ..., x_n) = 0,$$

where $x_1, ..., x_n$ are variables.

Theorem 1.51. Let $F(\mathbf{x}, \mathbf{y}) : \mathbb{R}^{n+m} \to \mathbb{R}^m$ be a differentiable function of two variables, and $(\mathbf{x}_0, \mathbf{y}_0)$ the point that $F(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$. If the Jacobian matrix

$$J_{F,\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = \left(\frac{\partial F_i}{\partial y_j}(\mathbf{x}_0, \mathbf{y}_0)\right)$$

is invertible, then there exists an open set $\Omega \subseteq \mathbb{R}^n$ containing $\mathbf{x}_0$ such that there exists a unique function $f : \Omega \to \mathbb{R}^m$ such that $f(\mathbf{x}_0) = \mathbf{y}_0$ and $F(\mathbf{x}, f(\mathbf{y})) = \mathbf{0}$ for all $\mathbf{x} \in \Omega$.

Moreover, f is continuously differentiable and, denoting the left-hand panel of the Jacobian matrix shown in the previous section as

$$J_{F,\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0) = \left(\frac{\partial F_i}{\partial x_j}(\mathbf{x}_0, \mathbf{y}_0)\right),$$

the Jacobian matrix of partial derivatives of f in Ω is given by

$$\left(\frac{\partial f_i}{\partial x_j}(\mathbf{x})\right)_{m\times n} = -\big(J_{F,\mathbf{y}}(\mathbf{x}, f(\mathbf{x}))\big)^{-1}_{m\times m}\big(J_{F,\mathbf{x}}(\mathbf{x}, f(\mathbf{x}))\big)_{m\times n}.$$

1.2 Real Analysis

1.2.1 Lebesgue Measure

Definition 1.52. Given an bounded interval $I \in \mathbb{R}$, denoted by $\ell(I)$ the **length** of the interval defined as the distance of its endpoints,

$$\ell([a, b]) = \ell((a, b)) = b - a.$$

Definition 1.53. For any subset $E \subset \mathbb{R}$, the **Lebesgue outer measure** $m^*(E)$ is defined as

$$m^*(E) = \inf\left\{\sum_{i=1}^n \ell(I_i) : \{I_i\}_{i=1}^n \text{ is a sequence of open intervals that } E \subset \bigcup_{i=1}^n I_i\right\}.$$

Theorem 1.54. If $E_1 \subset E_2 \subset \mathbb{R}$, then $m^*(E_1) \leq m^*(E_2)$.

Theorem 1.55. Given an interval $I \subset \mathbb{R}$, $m^*(I) = \ell(I)$.

Theorem 1.56. Given $\{E_i \subset \mathbb{R}\}_{i=1}^n$, $m^*\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n m^*(E_i)$.

Definition 1.57. The sets E are said to be **Lebesgue-measurable** if

$$\forall A \subset \mathbb{R}, m^*(A) = m^*(A \cap X) + m^*(A \cap (\mathbb{R} \setminus A))$$

and its Lebesgue measure is defined as its Lebesgue outer measure: $m(E) = m^*(E)$.

Theorem 1.58. The set of all measurable sets $E \subset \mathbb{R}$ forms a σ-algebra $\mathcal{F}$ where

- $\mathcal{F}$ contains the sample space: $\mathbb{R} \in \mathcal{F}$;
- $\mathcal{F}$ is closed under complements: if $A \in \mathcal{F}$, then also $(\mathbb{R} \setminus A) \in \mathcal{F}$;
- $\mathcal{F}$ is closed under countable unions: if $A_i \in \mathcal{F}, i = 1, ...,$ then also $(\cup_{i=1}^{\infty} A_i) \in \mathcal{F}$.

Definition 1.59. A **measurable space** is a tuple $(X, \mathcal{F})$ consisting of an arbitrary non-empty set X and a σ-algebra $\mathcal{F} \subseteq 2^X$.

1.3 Complex Analysis

Definition 1.60. Given an open set Ω and a function $f(z) : \Omega \to \mathbb{C}$, the **derivative** of $f(z)$ at a point $z_0 \in \Omega$ is defined as the limits

$$f'(z) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0},$$

and the function is said to be **complex differentiable** at z_0.

Definition 1.61. A function $f(z)$ is holomorphic on an open set Ω if it is complex differentiable at every point of Ω.

Theorem 1.62. If a complex function $f(x + \mathbf{i}y) = u(x, y) + \mathbf{i}v(x, y)$ is holomorphic, then u and v have first partial derivatives, and satisfy the Cauchy–Riemann equations,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = \frac{\partial v}{\partial x},$$

or equivalently,

$$\frac{\partial f}{\partial \bar{z}} = 0.$$

Theorem 1.63. (Cauchy's integral theorem) Given a simply connected domain Ω and a holomorphic function $f(z)$ on it, for any simply closed contour C in Ω,

$$\int_C f(z) \mathrm{d}x = 0.$$

Theorem 1.64. (Residue formula) Suppose that f is holomorphic in an open set containing a toy contour γ and its interior, except for some points $z_1, ..., z_n$ inside γ, then

$$\int_\gamma f(z) \mathrm{d}z = 2\pi \mathbf{i} \sum_{k=1}^{n} \mathrm{res}_{z_k} f,$$

where for a pole z_0 of order n,

$$\operatorname{res}_{z_0} f = \lim_{z \to z_0} \frac{1}{(n-1)!} \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^{n-1} (z - z_0)^n f(z).$$

1.4 Important Inequalities

1.4.1 Fundamental inequality

Theorem 1.65. (Fundamental inequality)

$$\forall x, y \in \mathbb{R}^+, \frac{2}{\frac{1}{a} + \frac{1}{b}} \leq \sqrt{ab} \leq \frac{a+b}{2} \leq \sqrt{\frac{a^2+b^2}{2}}, \text{ equality holds iff } \ a = b.$$

1.4.2 Triangle inequality

Theorem 1.66. (Triangle inequality)

$$\begin{aligned} &a, b \in \mathbb{C}, \quad ||a| - |b|| \leq |a \pm b| \leq |a| + |b|, \\ &\mathbf{a}, \mathbf{b} \in \mathbb{R}^n, |\|\mathbf{a}\| - \|\mathbf{b}\|| \leq \|\mathbf{a} \pm \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|. \end{aligned}$$

1.4.3 Bernoulli inequality

Theorem 1.67. (Bernoulli inequality)

$$\begin{aligned} &\forall x \in (-1, +\infty), \forall a \in [1, +\infty), &&(1+x)^a \geq 1 + ax, \\ &\forall x \in (-1, +\infty), \forall a \in (0, 1), &&(1+x)^a \leq 1 + ax, \\ &\forall x \in (-1, +\infty), \forall a \in (-1, 0), &&(1+x)^a \geq 1 + ax, \\ &\forall x_i \in \mathbb{R}, i \in \{1, ..., n\}, &&\prod_{i=1}^{n} (1 + x_i) \geq 1 + \sum_{i=1}^{n} x_i, \\ &\forall y \geq x > 0, &&(1+x)^y \geq (1+y)^x. \end{aligned}$$

1.4.4 Jensen's inequality

Theorem 1.68. (Jensen's inequality) For a real convex function $f(x) : [a, b] \to \mathbb{R}$, numbers $x_1 ..., x_n \in [a, b]$ and weights $a_1, ..., a_n$, the Jensen's inequality can be start as

$$\frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \geq f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i} \right).$$

And for concave function f,

$$\frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \leq f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i} \right).$$

Equality holds iff $x_1 = \cdots = x_n$ or f is linear on $[a, b]$.

1.4.5 Cauchy–Schwarz inequality

Theorem 1.69. (Cauchy–Schwarz inequality)
Discrete form. For real numbers $a_1, ...a_n, b_1, ...b_n \in \mathbb{R}, n \geq 2$

$$\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2 \geq \left(\sum_{i=1}^{n} a_i b_i\right).$$

Equality holds iff $\frac{a_1}{b_1} = \cdots = \frac{a_n}{b_n}$ or $a_i = 0$ or $b_i = 0$.

Inner product form. For a inner product space V with a norm induced by the inner product,

$$\forall \mathbf{a}, \mathbf{b} \in V \ \|\mathbf{a}\| \cdot \|\mathbf{b}\| \geq |\langle \mathbf{a}, \mathbf{b} \rangle|.$$

Equality holds iff $\exists k \in \mathbb{R}, \text{ s.t. } k\mathbf{a} = \mathbf{b} \text{ or } \mathbf{a} = k\mathbf{b}$.

Probability form. For random variables X and Y,

$$\sqrt{E(X^2)} \cdot \sqrt{E(Y^2)} \geq |E(XY)|.$$

Equality holds iff $\exists k \in \mathbb{R}, \text{ s.t. } kX = Y \text{ or } X = kY$.

Integral form. For integrable functions $f, g \in L^2(\Omega)$,

$$\left(\int_\Omega f^2(x) \mathrm{d}x\right)\left(\int_\Omega g^2(x) \mathrm{d}x\right) \geq \left(\int_\Omega f(x)g(x) \mathrm{d}x\right)^2.$$

Equality holds iff $\exists k \in \mathbb{R}, \text{ s.t. } kf(x) = g(x) \text{ or } f(x) = kg(x)$.

1.4.6 Hölder's inequality

Theorem 1.70. (Hölder's inequality)
Discrete form. For real numbers $a_1, ...a_n, b_1, ...b_n \in \mathbb{R}, n \geq 2$ and $p, q \in [1, +\infty)$ that $\left(\frac{1}{p}\right) + \left(\frac{1}{q}\right) = 1$,

$$\left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} b_i^q\right)^{\frac{1}{q}} \geq \left(\sum_{i=1}^{n} a_i b_i\right).$$

Equality holds iff $\exists c_1, c_2 \in \mathbb{R}, c_1^2 + c_2^2 \neq 0, \text{ s.t. } c_1 a_i^p = c_2 b_i^q$.

Integral form. For functions $f \in L^p(\Omega), g \in L^q(\Omega)$ and $p, q \in [1, +\infty)$ that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\left(\int_\Omega |f(x)|^p \ \mathrm{d}x\right)^{\frac{1}{p}} \left(\int_\Omega |g(x)|^q \ \mathrm{d}x\right)^{\frac{1}{q}} \geq \int_\Omega f(x)g(x) \mathrm{d}x.$$

1.4.7 Young's inequality

Theorem 1.71. (Young's inequality) For $p, q \in [1, +\infty)$ that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\forall a, b \in \mathbb{R}^*, \frac{a^p}{p} + \frac{b^q}{q} \geq ab.$$

Equality holds iff $a^p = b^q$.

1.4.8 Minkowski inequality

Theorem 1.72. (Minkowski inequality) For a metric space S,

$$\forall f, g \in L^p(S), p \in [1, +\infty], \|f\|_p + \|g\|_p \geq \|f+g\|_p.$$

For $p \in (1, +\infty)$, equality holds iff $\exists k \geq 0,$ s.t. $f = kg$ or $kf = g$.

1.4.9 Friedriches inequality

Theorem 1.73. (Friedriches inequality) Given a bounded simply connected region $\Omega \subset \mathbb{R}^n$, with the diameter d, then for $u \in H_0^1(\Omega)$,

$$\|u\|_{L^2(\Omega)} \leq d\|\nabla u\|_{L^2(\Omega)}.$$

1.5 Special Functions

1.5.1 Gaussian function

Definition 1.74. A **Gaussian function**, or a Gaussian, is a function of the form

$$f(x) = a \exp\left(-\frac{(x-b)^2}{2c^2}\right),$$

where $a \in \mathbb{R}^+$ is the height of the curve's peak, $b \in \mathbb{R}$ is the position of the center of the peak and $c \in \mathbb{R}^+$ is the standard deviation or the Gaussian root mean square width.

Theorem 1.75. The integral of a Gaussian is

$$\int_{-\infty}^{+\infty} a \exp\left(-\frac{(x-b)^2}{2c^2}\right) \mathrm{d}x = ac\sqrt{2\pi}.$$

Definition 1.76. A **normal distribution** or a **Gaussian distribution** is a continuous probability distribution of the form

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-((x-\mu)^2)(2\sigma^2)),$$

where μ is the mean and σ is the standard deviation.

1.5.2 Dirac delta function

Definition 1.77. The **Dirac delta function** centered at $\overline{x}$ is

$$\delta(x-\overline{x}) = \lim_{\varepsilon \to 0} f_{\overline{x},\varepsilon}(x-\overline{x}),$$

where $f_{\overline{x},\varepsilon}$ is a normal distribution with its mean at $\overline{x}$ and its standard deviation as ε.

Theorem 1.78. The Dirac delta function satisfies

$$\delta(x-\overline{x}) = \begin{cases} +\infty, & x = \overline{x} \\ 0, & x \neq \overline{x} \end{cases} \quad \int_{-\infty}^{x} \delta(x-\overline{x})\mathrm{d}x = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

where $H(x) = \int_{-\infty}^{x} \delta(x-\overline{x})\mathrm{d}x$ is called **Heaviside function** or **step function**.

Theorem 1.79. If $f : \mathbb{R} \to \mathbb{R}$ is continuous, then

$$\int_{-\infty}^{+\infty} \delta(x - \overline{x}) f(x) \mathrm{d}x = f(\overline{x}).$$

1.5.3 Gamma function

Definition 1.80. The **Gamma function** defined on $\mathbb{C}$ is

$$\Gamma(z) = \int_{0}^{+\infty} t^{z-1} e^{-t} \mathrm{d}t,$$

where Re $(z) > 0$.

Theorem 1.81. The Gamma function satisfies

$$\forall x \in \mathbb{C},\ \ \Gamma(x+1) = x\Gamma(x),$$
$$\forall n \in \mathbb{N}^*, \Gamma(n) = (n-1)!.$$

Theorem 1.82. The Gamma function satisfies

$$\forall x \in (0,1), \Gamma(1-x)\Gamma(x) = \frac{\pi}{\sin(\pi x)},$$

which implies

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

1.5.4 Beta Function

Definition 1.83. For $p, q \in \mathbb{R}^+$, the **Beta function** is defined as

$$B(p,q) = \int_{0}^{1} x^{p-1}(1-x)^{q-1} \mathrm{d}x.$$

Theorem 1.84. The Beta function satisfies

$$\forall p, q \in \mathbb{R}^+, B(p,q) = B(q,p) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Theorem 1.85. The Beta function satisfies

$$\forall p > 0, \forall q > 1, B(p,q) = \frac{q-1}{p+q-1} B(p, q-1),$$
$$\forall p > 1, \forall q > 0, B(p,q) = \frac{p-1}{p+q-1} B(p-1, q),$$
$$\forall p > 1, \forall q > 1, B(p,q) = \frac{(p-1)(q-1)}{(p+q-1)(p+q-2)} B(p-1, q-1).$$

Chapter 2
Algebra

2.1 Linear Space

Definition 2.1. (Linear Space) A **linear space** over a field $\mathbb{F}$ is a nonempty set V with a addition and a scalar multiplication that satisfies

(1) Associativity of addition: $\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$,
(2) Commutativity of addition: $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V, (\mathbf{x}+\mathbf{y})+\mathbf{z} = \mathbf{x}+(\mathbf{y}+\mathbf{z})$,
(3) Identity element of addition: $\exists \mathbf{0} \in V, \forall \mathbf{x}, \mathbf{x}+\mathbf{0} = \mathbf{x}$,
(4) Inverse elements of addition: $\forall \mathbf{x} \in V, \exists \mathbf{y} \in V,$ s.t. $\mathbf{x}+\mathbf{y} = 0$,
(5) Compatibility of multiplication: $\forall \mathbf{x} \in V, a, b \in \mathbb{F}, (ab)\mathbf{x} = a(b\mathbf{x})$,
(6) Identity element of multiplication: $\exists 1 \in \mathbb{F}, \forall \mathbf{x} \in V, 1\mathbf{x} = \mathbf{x}$,
(7) Distributivity: $\forall \mathbf{x} \in V, a, b \in \mathbb{F}, (a+b)\mathbf{x} = a\mathbf{x}+b\mathbf{x}$,
(8) Distributivity: $\forall \mathbf{x}, \mathbf{y} \in V, a \in \mathbb{F}, a(\mathbf{x}+\mathbf{y}) = a\mathbf{x}+b\mathbf{y}$.

Notation 2.2. The **dimension** of a linear space V is written as $\dim(V)$.

Definition 2.3. Denoted by $V_1, ..., V_n$ linear spaces over a field $\mathbb{F}$, the **product of linear spaces** is defined as

$$V_1 \times \cdots \times V_n = \{(\mathbf{v}_1, ..., \mathbf{v}_n) : \mathbf{v}_1 \in V_1, ..., \mathbf{v}_n \in V_n\},$$

which is also a linear space over $\mathbb{F}$.

Definition 2.4. Given a linear space V, a subspace $U \subset V$ and $\mathbf{v} \in V$, the **coset** (or **affine subset**) is defined as

$$\overline{\mathbf{v}} = \{\mathbf{w} \in V : \mathbf{w} = \mathbf{v}+\mathbf{u}, \mathbf{u} \in U\}.$$

Definition 2.5. Given a linear space V and a subspace $U \subset V$, the **quotient space** is defined as

$$V/U = \{\mathbf{v}+U : \mathbf{v} \in V\}.$$

2.1.1 Linear map

Definition 2.6. Denoted by V and W the linear spaces over a field $\mathbb{F}$, a function $f : V \to W$ is called a linear map between V and W if it satisfies

(1) Additivity: $\forall \mathbf{x}, \mathbf{y} \in V, f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x})+f(\mathbf{y})$;
(2) Homogeneity: $\forall \mathbf{x} \in V, \forall k \in \mathbb{F}, f(k\mathbf{x}) = kf(\mathbf{x})$.

Notation 2.7. Denoted by $\mathcal{L}(V, W)$ the set of all linear maps between V and W (it also be written as $\mathcal{L}(V)$ if $V = W$).

Theorem 2.8. For linear space V, W over a field $\mathbb{F}$ and linear maps $f, g \in \mathcal{L}(V, W)$, if we define

$$\forall \mathbf{x} \in V, \forall k \in \mathbb{F}, (f+g)(\mathbf{x}) = f(\mathbf{x})+g(\mathbf{x}) \quad \text{and} \quad (kf)(\mathbf{x}) = kf(\mathbf{x}),$$

then $\mathcal{L}(V, W)$ is a linear space.

Theorem 2.9. For a linear map $f \in \mathcal{L}(V, W)$, $f(\mathbf{0}) = f(0\mathbf{v}) = 0f(\mathbf{v}) = 0$.

Theorem 2.10. Given $\mathbf{v}_1, ...\mathbf{v}_n$ the basis of linear space V and $\mathbf{w}_1, ...\mathbf{w}_n$ the basis of linear space W, then there exists the only linear map $f \in \mathcal{L}(V, W)$ such that

$$\forall i \in \{1, ..., n\}, f(\mathbf{v}_i) = \mathbf{w}_i.$$

Definition 2.11. For a linear map $f \in \mathcal{L}(V, W)$, the **kernal** (or **null space**) of f is defined as

$$\ker(f) = \{\mathbf{v} \in V : f(\mathbf{v}) = \mathbf{0}\},$$

where $\ker(f)$ is a subspace of V and the number $\dim(\ker(f))$ is the **nullity** of f which also written as $\text{nullity}(f)$

Definition 2.12. For a linear map $f \in \mathcal{L}(V, W)$, the **image** of f is defined as

$$\text{im}(f) = \{\mathbf{w} \in W : \mathbf{w} = f(\mathbf{v}), \mathbf{v} \in V\},$$

where $\text{im}(f)$ is a subspace of W and the number $\dim(\text{im}(f))$ is the **dimension** (or **rank**) of f which also written as $\text{rank}(f)$

Theorem 2.13. (Rank–nullity theorem) For a linear map $f \in \mathcal{L}(V, W)$,

$$\dim(\ker(f)) + \dim(\text{im}(f)) = \dim(V).$$

Definition 2.14. A **isomorphism** is a invertible linear map.

Definition 2.15. Two linear spaces are called **isomorphic** if there exists a invertible linear map between them.

Theorem 2.16. Two linear spaces V, W over a field $\mathbb{F}$ are isomorphic iff $\dim(V) = \dim(W)$.

Theorem 2.17. For a linear space V that $\dim(V) < +\infty$ and a linear map $f \in \mathcal{L}(V)$, the following statements are equivalent:

(1) f is invertible;
(2) f is injective;
(3) f is surjective.

2.2 Metric Space

Definition 2.18. (Metric) For a nonempty set X, the **metric** is a function $d : X \times X \to \mathbb{R}$ that satisfies

(1) Positive definiteness: $\forall \mathbf{x}, \mathbf{y} \in X, d(\mathbf{x}, \mathbf{y}) \geq 0, d(\mathbf{x}, \mathbf{y}) \Leftrightarrow \mathbf{x} = \mathbf{y}$,
(2) Symmetry: $\forall \mathbf{x}, \mathbf{y} \in X, d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$,
(3) Triangle inequality: $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V, d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$,

Definition 2.19. (Metric space) A **metric space** is a set X provided with a metric.

Notation 2.20. (Neighbourhood) For a metric space X, the **neighbourhood** of $\mathbf{x} \in X$ with radius $\varepsilon > 0$ is defined as

$$U_X(\mathbf{x}, \varepsilon) = \{t : d(\mathbf{x}, t) < \varepsilon, t \in X\}.$$

Notation 2.21. (Punctured neighbourhood) For a metric space X, the **punctured neighbourhood** of $\mathbf{x} \in X$ with radius $\varepsilon > 0$ is defined as

$$U_X^\circ(\mathbf{x}, \varepsilon) = U_X(\mathbf{x}, \varepsilon) \setminus \{\mathbf{x}\} = \{t : d(\mathbf{x}, t) < \varepsilon, t \in X \setminus \{\mathbf{x}\}\}.$$

2.2.1 Completeness & Compactness

Theorem 2.22. (Cauchy's convergence test) A sequence $\{\mathbf{x}_n\}$ in a metric space X is convergent (or said a **cauchy sequence**) iff

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \text{ s.t. } \forall m, n > N, \|\mathbf{x}_n - \mathbf{x}_m\| < \varepsilon.$$

Definition 2.23. (Completeness) A metric space X is **complete** iff all cauchy sequence of X is convergent in X.

Theorem 2.24. (Supremum and infimum principle) For a nonempty set X, if the upper/ lower bound of X exists, then the supremum/infimum of X exists.

Theorem 2.25. (The monotone bounded convergence Theorem) For a bounded sequence $\{\mathbf{x}_n\}$, if it is increased, then

$$\lim_{n \to \infty} \mathbf{x}_n = \sup\{\mathbf{x}_n : n \in \mathbb{N}\}.$$

If it is decreased, then

$$\lim_{n \to \infty} \mathbf{x}_n = \inf\{\mathbf{x}_n : n \in \mathbb{N}\}.$$

2.2.2 Cover

Definition 2.26. (Cover) For a metric space $S \subseteq X$, A **cover** of S is a set of open sets $\{D_n\}$ satisfies

$$\forall \mathbf{x} \in X, \exists D_n, \text{ s.t. } \mathbf{x} \in D_n.$$

Definition 2.27. (Compactness) A metric space X is called **compact** if every open cover of X has a finite subcover.

2.2.3 Cantor's intersection Theorem

Theorem 2.28. (Cantor's intersection Theorem) For a decreasing sequence of nested non-empty compact, closed subsets $S_n \subseteq X, n \in \mathbb{N}$ of a metric space, if $\{S_n\}$ satisfies

$$S_0 \supset S_1, \cdots, \supset S_n \supset \cdots,$$

then

$$\bigcap_{k=0}^{\infty} S_k \neq \emptyset.$$

where there is only one point $\mathbf{x} \in \bigcap_{k=0}^{\infty} S_k$ for a complete metric space.

Corollary 2.29. For decreasing sequence of nested non-empty compact, closed subsets $S_n \in X, n \in \mathbb{N}$ of a complete metric space and $\{\mathbf{x}\} = \bigcap_{k=0}^{\infty} S_k$, then

$$\forall \varepsilon > 0, \exists N > 0, \text{ s.t. } \forall n > N, X_n \subset U_X(x, \varepsilon).$$

2.2.4 Cluster point

Definition 2.30. (Cluster point) For a metric space $S \subseteq X$, the **cluster point** of S is the point $\mathbf{x} \in X$ satisfies

$$\forall \varepsilon > 0, U_X^{\circ}(\mathbf{x}, \varepsilon) \cup S \neq \emptyset.$$

Theorem 2.31. For a convergent sequence $\left\{\mathbf{x}_n : n \in \mathbb{N}, \forall i \neq j, \mathbf{x}_i \neq \mathbf{x}_j\right\} \subseteq X$, the point $x = \lim\limits_{n\to\infty} \mathbf{x}_n$ is a cluster point of X.

Theorem 2.32. (Bolzano–Weierstrass Theorem) For a metric sapce X and a bounded infinite subset $S \in X$, there exists at least one cluster point of X.

2.3 Normed Space

Definition 2.33. (Norm) For a linear space V over a field $\mathbb{F}$, the **norm** is a function $\|\cdot\| : V \to \mathbb{F}$ that satisfies

(1) Positive definiteness: $\forall \mathbf{x} \in V, \|\mathbf{x}\| \geq 0,\ \|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0$;
(2) Absolute homogeneity: $\forall \mathbf{x} \in V, k \in \mathbb{F}, \|k\mathbf{x}\| = |k|\|\mathbf{x}\|$;
(3) Triangle inequality: $\forall \mathbf{x}, \mathbf{y} \in V, \|\mathbf{x}\| + \|\mathbf{y}\| \geq \|\mathbf{x} + \mathbf{y}\|$.

Definition 2.34. (Equivalent norms) Two norms $p(\cdot), q(\cdot)$ on $\mathbb{R}^n$ are called **equivalent** if

$$\exists C_1, C_2 \in \mathbb{R}^+ \text{ s.t. } \forall \mathbf{x} \in V, C_1 q(\mathbf{x}) \leq p(\mathbf{x}) \leq C_2 q(\mathbf{x}).$$

Definition 2.35. (Normed space) A **normed space** is a linear space V over the the field $\mathbb{F}$ with a norm.

2.3.1 Vector norm and matrix norm

Example 2.36. The followings are some commonly used vector norms:

(1) l_1 norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |\mathbf{x}_i|$;
(2) l_2 norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |\mathbf{x_i}|^2}$;
(3) l_∞ norm: $\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |\mathbf{x}_i|$.

Theorem 2.37. Any two l_p norms $\|\cdot\|_p, \|\cdot\|_q$ on $\mathbb{R}^n$ are equivalent.

Example 2.38. For l_p norms on $\mathbb{R}^n$,

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2, \ \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty, \ \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty.$$

Definition 2.39. Let $\left\{\mathbf{x}^{[k]} \in \mathbb{R}^n\right\}_{i=1}^{\infty}$ be a sequences and $\mathbf{x}^* \in \mathbb{R}^n$, then

$$\lim_{i\to\infty} \mathbf{x}^{[i]} = \mathbf{x}^* \Leftrightarrow \forall 1 \leq k \leq n, \lim_{i\to\infty} \mathbf{x}_k^{[i]} = \mathbf{x}_k^*.$$

Corollary 2.40. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, $\left\{\mathbf{x}^{[i]}\right\}_{i=1}^{\infty} \subset \mathbb{R}^n$ be a sequences and $x^* \in \mathbb{R}^n$, then

$$\lim_{i\to\infty} \mathbf{x}^{[k]} = \mathbf{x}^* \Leftrightarrow \lim_{i\to\infty} \left\|\mathbf{x}^{[k]} - \mathbf{x}^*\right\| = 0.$$

Definition 2.41. A function $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ is called a **matrix norm** if

(1) Positive definiteness: $\forall A \in \mathbb{R}^{n\times n}, \|A\| \geq 0$, $\|A\| = 0$ iff $A = 0$;

(2) Absolute homogeneity: $\forall A \in \mathbb{R}^{n\times n}, k \in \mathbb{R}, \|kA\| = |k|\|A\|$;

(3) Triangle inequality: $\forall A, B \in \mathbb{R}^{n\times n}, \|A\| + \|B\| \geq \|A + B\|$;

(4) Sub-multiplicative: $\forall A, B \in \mathbb{R}^{n\times n}, \|A\|\|B\| \geq \|AB\|$.

Theorem 2.42. Let $\|\cdot\|$ be a vector norm, then the **matrix norm induced by the vector norm** can be written as

$$\|A\| = \max_{\mathbf{x}\in\mathbb{R}^n} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

Example 2.43. The followings are some commonly used matrix norms:

(1) $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$ (maximum of the l_1 norms of columns);

(2) $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$;

(3) $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$ (maximum of the l_1 norms of rows);

(4) $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}$ (**Frobenius norm**).

Definition 2.44. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then A is **positive semidefinite** if $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, A is **positive definite** if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

Notation 2.45. We write $A \succeq 0$ if A is positive semidefinite, $A \succ 0$ if A is positive definite. The set of $n \times n$ positive semidefinite matrices is denoted by S_+^n.

Theorem 2.46. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then the following statements are equivalent

(1) All eigenvalues of A are nonnegative;

(2) There exists $M \in \mathbb{R}^{n\times n}$ such that $A = M^T M$;

(3) A is positive semidefinite.

Theorem 2.47. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then the following statements are equivalent

(1) All eigenvalues of A are positive;

(2) There exists an invertible matrix $M \in \mathbb{R}^{n\times n}$ such that $A = M^T M$;

(3) A is positive definite.

Remark 2.48. Let $A \succ 0$, then

(1) $A^{-1} \succ 0$ and $\lambda_{\min}(A) = \inf\{\mathbf{x}^T A\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$;

(2) $\|A\|_2 = \lambda_{\max}(A) = \left(\lambda_{\min}(A^{-1})\right)^{-1}$.

2.4 Inner Product Space

Definition 2.49. (Inner product) For a linear space V over a field $\mathbb{F}$, the **inner product** on V is a function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{F}$ that satisfies

(1) Positive definiteness: $\forall \mathbf{x} \in V$, $\langle \mathbf{x}, \mathbf{x}\rangle \geq 0$, $\langle \mathbf{x}, \mathbf{x}\rangle = 0 \Leftrightarrow \mathbf{x} = 0$,
(2) Conjugate symmetry: $\langle \mathbf{x}, \mathbf{y}\rangle = \overline{\langle \mathbf{y}, \mathbf{x}\rangle}$,
(3) Linearity in the first argument: $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V, a, b \in \mathbb{F}$, $\langle a\mathbf{x} + b\mathbf{z}, \mathbf{y}\rangle = a\langle \mathbf{x}, \mathbf{y}\rangle + b\langle \mathbf{z}, \mathbf{y}\rangle$.

Definition 2.50. (Inner product space) An **inner product space** is a linear space V over the field $\mathbb{F}$ with an inner product.

Theorem 2.51. Given a inner product space V and the norm defined as $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$ satisfies

$$\forall \mathbf{x}, \mathbf{y} \in V, \|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\,\|\mathbf{x}\|^2 + 2\,\|\mathbf{y}\|^2.$$

2.4.1 Orthonormal system

Definition 2.52. A subset W of an inner product space V is called textsf{orthonormal} if

$$\forall \mathbf{u}, \mathbf{v} \in S, \langle \mathbf{u}, \mathbf{v}\rangle = \begin{cases} 0, & u \neq v \\ 1, & u = v. \end{cases}$$

Definition 2.53. The **Gram-Schmidt process** takes in a finite or infinite independent list $(\mathbf{u}_1, \mathbf{u}_2, ...)$ and output two other lists $(\mathbf{v}_1, \mathbf{v}_2, ...)$ and $(\mathbf{u}_1^*, \mathbf{u}_2^*, ...)$ by

$$\begin{aligned} \mathbf{v}_{n+1} &= \mathbf{u}_{n+1} - \sum_{i=1}^{n} \langle \mathbf{u}_{n+1}, \mathbf{u}_k^*\rangle \mathbf{u}_k^*, \\ \mathbf{u}_{n+1}^* &= \frac{\mathbf{v}_{n+1}}{\|\mathbf{v}_{n+1}\|}, \end{aligned}$$

with the recursion basis as $\mathbf{v}_1 = \mathbf{u}_1$.

Definition 2.54. Let $(\mathbf{u}_1^*, \mathbf{u}_2^*, ...)$ be a finite or infinite orthonormal list. The **orthogonal expansion** or **Fourier expansion** for an arbitrary $\mathbf{w}$ is the series

$$\sum_{i=1}^{n} \langle \mathbf{w}, \mathbf{u}_i^*\rangle \mathbf{u}_i^*,$$

where the constants $\langle \mathbf{w}, \mathbf{u}_i^*\rangle$ are known as the **Fourier coefficients** of $\mathbf{w}$ and the term $\langle \mathbf{w}, \mathbf{u}_i^*\rangle \mathbf{u}_i^*$ is the **projection** of $\mathbf{w}$ on $\mathbf{u}_i^*$.

Theorem 2.55. (Minimum properties of Fourier expansions) Let $\mathbf{u}_1^*, \mathbf{u}_2^*, ...$ be an orthonormal system and let $\mathbf{w}$ be arbitrary. Then

$$\forall a_1, ..., a_n, \|\mathbf{w} - \sum_{i=1}^{n} \langle \mathbf{w}, \mathbf{u}_i^* \rangle \mathbf{u}_i^*\| \le \|\mathbf{w} - \sum_{i=1}^{n} a_i \mathbf{u}_i^*\|,$$

where $\|\mathbf{w} - \sum_{i=1}^{n} a_i \mathbf{u}_i^*\|$ is minimized only when $a_i = \langle \mathbf{w}, \mathbf{u}_i^* \rangle$.

Theorem 2.56. (Bessel inequality) Let $\mathbf{u}_1^*, \mathbf{u}_2^*, ...$ be an orthonormal system and let $\mathbf{w}$ be arbitrary. Then

$$\sum_{i=1}^{n} |\langle \mathbf{w}, \mathbf{u}_i^* \rangle| \le \|\mathbf{w}\|^2.$$

2.5 Banach Space

Definition 2.57. (Banach space) A **Banach space** is a complete normed vector space.

2.6 Hilbert Space

Definition 2.58. (Hilbert space) A **Hilbert space** is a inner product space that is also ce with respect to the distance function induced by the inner product.a complete metric space.

2.7 Single Variable Polynomial

Definition 2.59. Denoted by $\mathbb{V}$ a linear space and x the variable, a **(single variable) polynomial** over $\mathbb{V}$ is defined as

$$p_{n(x)} = \sum_{i=0}^{n} c_i x^i,$$

where $c_0, ..., c_n \in \mathbb{V}$ are constants that called the **coefficients of the polynomial**.

Definition 2.60. Given a polynomial $p(x) = \sum_{i=0}^{n} c_i x^i$ where $c_n \neq 0$, the degree of $p(x)$ is marked as $\deg(p(x)) = n$. In particular, the degree of zero polynomial $p(x) = 0$ is $\deg(0) = -\infty$.

Theorem 2.61. Denoted by $\mathbb{P}_n = \{p : \deg(p) \le n\}$ the set of polynomials with degree no more than n ($n \ge 0$), and $\mathbb{P} = \bigcup_{n=0}^{\infty} \mathbb{P}_n$ the set contains all polynomials, then $\mathbb{P}_n$ is a linear space and satisfies

$$\{0\} = \mathbb{P}_0 \subset \mathbb{P}_1 \subset \cdots \subset \mathbb{P}_n \subset \cdots \mathbb{P}$$

Theorem 2.62. (Vieta's formulas) Given a polynomial $p \in \mathbb{P}_n$ with the coefficients being real or complex numbers, denoted by $x_1, ..., x_n$ the complex roots, then

$$\begin{cases} x_1 + \cdots + x_n & = -c_{n-1}, \\ \sum\limits_{i=1}^{n} \sum\limits_{j=i+1}^{n} x_i x_j & = c_{n-2}, \\ & \cdots \\ \prod\limits_{i=1}^{n} x_i & = (-1)^n c_0, \end{cases}$$

where $c_n = 1$ WLOG.

2.8 Orthogonal Polynomial

Definition 2.63. Given a weight function $\rho(x) : [a,b] \to \mathbb{R}^+$, satisfies

$$\int_a^b \rho(x)\mathrm{d}x > 0, \int_a^b x^k \rho(x)\mathrm{d}x > 0 \text{ exists.}$$

The set of **orthogonal polynomials** on $[a,b]$ with the weight function $\rho(x)$ is defined as

$$\{p_i, i \in \mathbb{N}\} \subset L_\rho([a,b]) = \left\{ f(x) : \int_a^b f^2(x)\rho(x)\mathrm{d}x < \infty \right\}.$$

where $\{p_i, i \in \mathbb{N}\}$ are calculate from $\{x^n, n \in \mathbb{N}\}$ using the Gram-Schmidt process with the inner product

$$\forall f, g \in L_\rho([a,b]), \langle f, g \rangle = \int_a^b \rho(x) f(x) g(x)\mathrm{d}x.$$

Theorem 2.64. Orthogonal polynomials $p_{n-1}(x), p_n(x), p_{n+1}(x)$ satisfies

$$p_{n+1}(x) = (a_n + b_n x)p_n(x) + c_n p_{n-1}(x).$$

where a_n, b_n, c_n are depends on $[a,b]$ and ρ.

Theorem 2.65. The orthogonal polynomial $p_n(x)$ on $[a,b]$ with the weight function $\rho(x)$ has n roots on (a,b).

2.8.1 Legendre polynomial

Definition 2.66. The **Legendre polynomial** is defined on $[-1,1]$ with the weight function $\rho(x) = 1$.

Theorem 2.67. The Legendre polynomials $\{p_i(x), i \in \mathbb{N}\}$ satisfies

$$\int_{-1}^1 p_i(x)p_j(x)\mathrm{d}x = \begin{cases} \frac{2}{2i+1}, & i = j \\ 0, & i \neq j. \end{cases}$$

Theorem 2.68. The Legendre polynomial p_{n-1}, p_n, p_{n+1} satisfies

$$p_{n+1}(x) = \frac{2n+1}{n+1} x p_n(x) - \frac{n}{n+1} p_{n-1}(x).$$

Example 2.69. The first three terms of Legendre polynomials is

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = \frac{3}{2}x^2 - \frac{1}{2}.$$

2.8.2 Chebyshev polynomial of the first kind

Definition 2.70. The **Chebyshev polynomial of the first kind** is defined on $[-1,1]$ with the weight function $\rho(x) = \frac{1}{\sqrt{1-x^2}}$.

Theorem 2.71. The Chebyshev polynomials of the first kind $\{p_i(x), i \in \mathbb{N}\}$ satisfies

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} p_i(x)p_j(x)\mathrm{d}x = \begin{cases} \pi & i=j=0 \\ \frac{\pi}{2} & i=j\neq 0 \\ 0 & i\neq j. \end{cases}$$

Theorem 2.72. The Chebyshev polynomial of the first kind p_{n-1}, p_n, p_{n+1} satisfies

$$p_{n+1}(x) = 2xp_n(x) - p_{n-1}(x).$$

Example 2.73. The first three terms of Chebyshev polynomials of the first kind is

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = 2x^2 - 1.$$

2.8.3 Chebyshev polynomial of the second kind

Definition 2.74. The **Chebyshev polynomial of the second kind** is defined on $[-1, 1]$ with the weight function $\rho(x) = \sqrt{1-x^2}$.

Theorem 2.75. The Chebyshev polynomials of the second kind $\{p_i(x), i \in \mathbb{N}\}$ satisfies

$$\int_{-1}^{1} \sqrt{1-x^2} p_i(x)p_j(x)\mathrm{d}x = \begin{cases} \frac{\pi}{2}, & i=j \\ 0, & i\neq j. \end{cases}$$

Theorem 2.76. The Chebyshev polynomial of the second kind p_{n-1}, p_n, p_{n+1} satisfies

$$p_{n+1}(x) = 2xp_n(x) - p_{n-1}(x).$$

Example 2.77. The first three terms of Chebyshev polynomials of the second kind is

$$p_0(x) = 1, \quad p_1(x) = 2x, \quad p_2(x) = 4x^2 - 1.$$

2.8.4 Laguerre polynomial

Definition 2.78. The **Laguerre polynomial** is defined on $[0, +\infty)$ with the weight function $\rho(x) = x^\alpha e^{-x}$.

Theorem 2.79. The Laguerre polynomial $\{p_i(x), i \in \mathbb{N}\}$ satisfies

$$\int_{0}^{+\infty} x^\alpha e^{-x} p_i(x)p_j(x)\mathrm{d}x = \begin{cases} \frac{\Gamma(n+\alpha+1)}{n!}, & i=j \\ 0, & i\neq j. \end{cases}$$

Theorem 2.80. For $\alpha = 0$, the Laguerre polynomial p_{n-1}, p_n, p_{n+1} satisfies

$$p_{n+1}(x) = (2n+1-x)p_n(x) - n^2 p_{n-1}(x).$$

Example 2.81. For $\alpha = 0$, the first three terms of Laguerre polynomial is

$$p_0(x) = 1, \quad p_1(x) = -x + 1, \quad p_2(x) = x^2 - 4x + 2.$$

2.8.5 Hermite polynomial (probability theory form)

Definition 2.82. The **Hermite polynomial** is defined on $(-\infty, +\infty)$ with the weight function $\rho(x) = \left(\frac{1}{\sqrt{2\pi}}\right)e^{-\frac{x^2}{2}}$.

Theorem 2.83. The Hermite polynomial $\{p_i(x), i \in \mathbb{N}\}$ satisfies

$$\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} p_i(x) p_j(x) \mathrm{d}x = \begin{cases} n!, & i = j \\ 0, & i \neq j. \end{cases}$$

Theorem 2.84. For $\alpha = 0$, the Hermite polynomial p_{n-1}, p_n, p_{n+1} satisfies

$$p_{n+1}(x) = xp_n(x) - np_{n-1}(x).$$

Example 2.85. For $\alpha = 0$, the first three terms of Hermite polynomial is

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - 1.$$

Chapter 3
Ordinary Differential Equation

Definition 3.1. Given a function F, an **explicit ordinary differential equation** of order n takes the form

$$\mathbf{F}\big(\mathbf{u}^{(n-1)}, ..., \mathbf{u}', \mathbf{u}, t\big) = \mathbf{u}^{(n)},$$

an **implicit ordinary differential equation** of order n takes the form

$$\mathbf{F}\big(\mathbf{u}^{(n)}, ..., \mathbf{u}', \mathbf{u}, t\big) = \mathbf{0},$$

Definition 3.2. An ODE is **autonomous** if it does not depend on the variable x.

Definition 3.3. A ODE is **linear** if can be written as

$$\sum_{i=0}^{n} A_i(t)\mathbf{u}^{(i)} + \mathbf{r}(t) = \mathbf{0},$$

where $A_i(t)$ and $r(t)$ are continuous functions of t.

Definition 3.4. A linear ODE is **homogeneous** if $\mathbf{r}(t) = 0$, and there is always the trivial solution $\mathbf{u} \equiv \mathbf{0}$.

Definition 3.5. An ODE is **separable** if can be written as

$$P_1(x)Q_1(y) = P_2(x)Q_2(y)\frac{\mathrm{d}y}{\mathrm{d}x}.$$

Definition 3.6. For initial value $(\mathbf{u}_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$, $T \geq t_0$ and $\mathbf{f} : \mathbb{R}^n \times [t_0, T] \to \mathbb{R}^n$, the **initial value problem** (IVP) is to find $u(t) \in C^1([t_0, T])$ satisfies

$$\mathbf{u}' = \mathbf{f}(\mathbf{u}, t), \quad \mathbf{u}(t_0) = \mathbf{u}_0.$$

Theorem 3.7. Given an IVP, denoted by $u_0 = u$, $u_i, i = 1, ..., n$ the ith derivative of u, then the ODE

$$\mathbf{F}\big(\mathbf{u}^{(n-1)}, ..., \mathbf{u}', \mathbf{u}, t\big) = \mathbf{u}^{(n)}$$

can be written as an IVP,

$$\begin{pmatrix} \mathbf{u}_0' \\ \vdots \\ \mathbf{u}_{n-2}' \\ \mathbf{u}_{n-1}' \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{n-1} \\ \mathbf{F}(\mathbf{u}_{n-1}, ..., \mathbf{u}_1, \mathbf{u}_0, t) \end{pmatrix}.$$

3.1 General Theory

Theorem 3.8. (Peano existence theorem) Given an IVP with an open set $\Omega \subseteq \mathbb{R}^n \times \mathbb{R}$, if $\mathbf{f}(\mathbf{u}, t) \in C(\Omega)$ and $(\mathbf{u}_0, t_0) \in \Omega$, then there is a local solution $\tilde{\mathbf{u}} : U \to \mathbb{R}^n$ satisfies the IVP, where U is a neighbourhood of t_0 in $\mathbb{R}$.

Theorem 3.9. (Picard–Lindelöf theorem) Given an IVP with an open set $\Omega \subseteq \mathbb{R}^n \times \mathbb{R}$, if $\mathbf{f}(\mathbf{u}, t) : \Omega \to \mathbb{R}^n$ is continuous in t and Lipschitz continuous in $\mathbf{u}$ and $(\mathbf{u}_0, t_0) \in \Omega$, then there is a unique local solution $\tilde{\mathbf{u}} : U \to \mathbb{R}^n$ satisfies the IVP, where U is a neighbourhood of t_0 in $\mathbb{R}$.

Theorem 3.10. (Comparison theorem) Given two IVPs

$$\mathbf{u}_1' = \mathbf{f}_1(\mathbf{u}_1, t), \quad \mathbf{u}_1(t_0) = \mathbf{u}_0,$$
$$\mathbf{u}_2' = \mathbf{f}_2(\mathbf{u}_2, t), \quad \mathbf{u}_2(t_0) = \mathbf{u}_0,$$

and a open set $\Omega \subseteq \mathbb{R}^n \times \mathbb{R}$, if for all $(\mathbf{u}, t) \in \Omega$, $\mathbf{f}_1(\mathbf{u}, t) < \mathbf{f}_2(\mathbf{u}, t)$, then

$$\begin{cases} \mathbf{u}_1(t) > \mathbf{u}_2(t), & t > t_0, (\mathbf{u}_1(t), t), (\mathbf{u}_2(t), t) \in \Omega, \\ \mathbf{u}_1(t) < \mathbf{u}_2(t), & t < t_0, (\mathbf{u}_1(t), t), (\mathbf{u}_2(t), t) \in \Omega, \end{cases}$$

3.2 Exact solutions

Example 3.11. Given an initial point (y_0, x_0), and a separable equation

$$P_1(x)Q_1(y) = P_2(x)Q_2(y)\frac{\mathrm{d}y}{\mathrm{d}x},$$

the solution of the equation is

$$\int_{x_0}^{x} \frac{P_1(t)}{P_2(t)}\mathrm{d}t = \int_{y_0}^{y} \frac{Q_2(t)}{Q_1(t)}\mathrm{d}t.$$

Example 3.12. Given an initial point (y_0, x_0), and a first-order homogeneous equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = F\left(\frac{y}{x}\right),$$

the solution of the equation is

$$\int_{x_0}^{x} \frac{1}{x}\mathrm{d}x = \int_{\frac{y_0}{x_0}}^{\frac{y}{x}} \frac{1}{F(t) - t}\mathrm{d}t.$$

Example 3.13. Given an initial point (y_0, x_0), and a first-order separable equation

$$yM(xy) + xN(xy)\frac{\partial y}{\partial x} = 0,$$

the solution of the equation is

$$\int_{x_0}^{x} \frac{1}{x}\mathrm{d}x = \int_{y_0 x_0}^{yx} \frac{N(t)}{t(N(t) - M(t))}\mathrm{d}t,$$

where C is a constant.

Example 3.14. Given a nth-order, linear, inhomogeneous, constant coefficients equation

$$\sum_{i=0}^{n} a_i \frac{\partial^i y}{\partial x^i} = 0,$$

the solution of the equation is

$$\sum_{i=1}^{k} \left(\sum_{j=1}^{m_i} c_{ij} x^{j-1} \right) e^{\alpha_i x},$$

where $\{c_{ij}\}$ are constants and α_i is the root of

$$\sum_{i=0}^{n} a_i x^i = 0$$

that repeated m_i times.

3.3 Important ODEs

3.3.1 Bernoulli differential equation

Definition 3.15. The **Bernoulli differential equation** takes the form

$$y' + P(x)y = Q(x)y^n,$$

where $n \neq 0, 1$.

Theorem 3.16. The solution of the Bernoulli differential equation is

$$y = (z(x))^{\frac{1}{1-n}},$$

where $z(x)$ is the solution of

$$z' + (1-n)P(x)z + (1-n)Q(x) = 0.$$

3.3.2 Riccati equation

Definition 3.17. The **Riccati equation** takes the form

$$y' = q_0(x) + q_1(x)y + q_2(x)y^2,$$

where $q_0(x) \neq 0, q_2(x) \neq 0$.

Theorem 3.18. If u is one particular solution of the Riccati equation, the general solution is obtained as $y = u + \frac{1}{v}$, where v satisfies

$$v' + (q_1(x) + 2q_2(x)u)v + q_2(x).$$

Chapter 4
Partial Differential Equation

Definition 4.1. A **2th order partial differential equation** in $\mathbb{R}^n$ takes the form

$$\sum_{i=0}^{n}\sum_{j=0}^{n} a_{ij}(\mathbf{x})u_{x_i x_j} + \sum_{i=0}^{n} b_i(\mathbf{x})u_{x_i} + c(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x}),$$

where $a_{ij}(\mathbf{x}) = a_{ji}(\mathbf{x})$.

Definition 4.2. Let $A(\mathbf{x}) = \left(a_{ij}(\mathbf{x})\right)_{n\times n}$ be a symmetric matrix, and $\lambda_1 \geq \cdots \geq \lambda_n$ the eigenvalues of A at $\mathbf{x}_0$, then

- The equation is **elliptic** at $\mathbf{x}_0$ if for $i = 1, ..., n$, $\lambda_i < 0$
- The equation is **parabolic** at $\mathbf{x}_0$ if $\lambda_1 = 0$ and for $i = 2, ..., n$, $\lambda_i < 0$;
- The equation is **hyperbolic** at $\mathbf{x}_0$ if $\lambda_1 > 0$ and for $i = 2, ..., n$, $\lambda_i < 0$;

Definition 4.3. The boundary conditions for the unknown function y, constants c_0, c_1 specified by the boundary conditions, and known scalar functions g, h specified by the boundary conditions, where

- **Dirichlet boundary condition**: $y = g$;
- **Neumann boundary condition**: $\frac{\partial y}{\partial n} = g$;
- **Robin boundary condition**: $c_0 y + c_1 \frac{\partial y}{\partial n} = g$ where $c_0, c_1 \neq 0$;
- **Mixed boundary condition**: $y = g$ and $c_0 y + c_1 \frac{\partial y}{\partial n} = h$ where $c_0, c_1 \neq 0$;
- **Cauchy boundary condition**: $y = g$ and $\frac{\partial y}{\partial n} = h$.

4.1 Poisson's Equation

Definition 4.4. A **Poisson's equation** in $\mathbb{R}^n$ takes the form

$$-\Delta u = f(\mathbf{x}),$$

where Δ is the Laplace operator, $u, f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$.

4.2 Heat Equation

Definition 4.5. A **Heat equation** in $\mathbb{R}^n \times \mathbb{R}$ takes the form

$$\frac{\partial u}{\partial t} - a^2 \Delta u = f(\mathbf{x}, t),$$

where Δ is the Laplace operator on $\mathbb{R}^n$, $u, f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$.

4.3 Wave Equation

Definition 4.6. A **Wave equation** in $\mathbb{R}^n \times \mathbb{R}$ takes the form

$$\frac{\partial^2 u}{\partial t^2} - a^2 \Delta u = f(\mathbf{x}, t),$$

where Δ is the Laplace operator on $\mathbb{R}^n$, $u, f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$.

Chapter 5
Probability Theory

5.1 Probability

Definition 5.1. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ consisting of

- the sample space Ω: an arbitrary non-empty set;
- the σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$: a set of subsets of Ω, called events, such that
  - $\mathcal{F}$ contains the sample space: $\Omega \in \mathcal{F}$;
  - $\mathcal{F}$ is closed under complements: if $A \in \mathcal{F}$, then also $(\Omega \setminus A) \in \mathcal{F}$;
  - $\mathcal{F}$ is closed under countable unions: if $A_i \in \mathcal{F}, i = 1, \ldots$, then also $(\cup_{i=1}^{\infty} A_i) \in \mathcal{F}$;
- the probability measure $P : \mathcal{F} \to [0, 1]$: a function such that
  - P is countably additive (also called σ-additive): if $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$;
  - the measure of the entire sample space is equal to one: $P(\Omega) = 1$.

Definition 5.2. Given a probability space $(\Omega, \mathcal{F}, P)$, a **random variable** is a measurable function $X : \Omega \to \mathbb{R}$ that for all $t \in \mathbb{R}$,

$$\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}.$$

Definition 5.3. The **cumulative distribution function (cdf)** of a random variable X on probability space $(\Omega, \mathcal{F}, P)$ is

$$F_X(x) = P(X \leq x).$$

5.1.1 Continuous random variables

Definition 5.4. A **continuous random variables** is a random variables with the range of X is uncountable.

Definition 5.5. The **probability density function (pdf)** of a continuous random variables is

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

Theorem 5.6. Let X be a discrete random variables, its probability mass function satisfies

(1) $f(x) \geq 0$;
(2) $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$;
(3) $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$.

Theorem 5.7. Let X be a continuous random variables and $Y = g(X)$ is a differentiable bijection, denoted by $f_X(x), f_Y(y)$ the pdf's of X and Y, then

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|.$$

5.1.2 Discrete random variables

Definition 5.8. A **discrete random variables** is a random variables with the range of X is countable.

Definition 5.9. The **probability mass function (pmf)** of a discrete random variables is

$$p_{X(x)} = P(X = x).$$

Theorem 5.10. Let X be a discrete random variables, its probability mass function satisfies

$$0 \leq p_X(x) \leq 1 \text{ and } \sum_{x \in \text{ Range}(X)} p_X(x) = 1.$$

Theorem 5.11. Let X be a discrete random variables and $Y = g(X)$, denoted by $p_X(x), p_Y(y)$ the pmf's of X and Y, then

$$p_Y(y) = \sum_{x; g(x) = y} p_X(x).$$

In particular, if g is a bijection, then

$$p_Y(y) = p_X(g^{-1}(y)).$$

Remark 5.12. The discrete random variable X can be written in continuous form via Dirac delta function, i.e.

$$f_X(x) = \sum_{\overline{x} \in \text{ Range}(X)} p_X(x)\delta(x - \overline{x}).$$

5.1.3 Multivariate distributions

Definition 5.13. A **random vector** is a vector $(X_1, ..., X_n)$ where all X_k are random variables.

Definition 5.14. The **joint cdf** of a random vector $(X_1, ..., X_n)$ is defined as

$$F_{X_1,...,X_n}(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n).$$

Definition 5.15. The **joint pmf** of a random vector $(X_1, ..., X_n)$ is defined as

$$p_{X_1,...,X_n}(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n).$$

Definition 5.16. The **joint pdf** of a random vector $(X_1, ..., X_n)$ is defined as

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \frac{\partial F_{X_1,...,X_n}(x_1, ..., x_n)}{\partial x_1 \cdots \partial x_n}.$$

Theorem 5.17. A random vector $(X_1, ..., X_n)$ satisfies

(1) $F_{X_1,...,X_{n-1}}(x_1, ..., x_{n-1}) = F_{X_1,...,X_n}(x_1, ..., x_{n-1}, +\infty)$;

(2) $p_{X_1,...,X_{n-1}}(x_1, ..., x_{n-1}) = \sum_{x \in \text{ Range}(X_n)} p_{X_1,...,X_n}(x_1, ..., x_{n-1}, x)$ (discrete case);

(3) $f_{X_1,\dots,X_{n-1}}(x_1,\dots,x_{n-1}) = \int_{-\infty}^{+\infty} f_{X_1,\dots,X_n}(x_1,\dots,x_{n-1},x)\mathrm{d}x$ (continuous case);

(4) $p_{X_1,\dots,X_n \mid X_1}(x_1,\dots,x_n \mid x_1) = \frac{p_{X_1,\dots,X_n}(x_1,\dots,x_n)}{p_{X_1}(x_1)}$ (discrete case);

(5) $f_{X_1,\dots,X_n \mid X_1}(x_1,\dots,x_n \mid x_1) = \frac{f_{X_1,\dots,X_n}(x_1,\dots,x_n)}{f_{X_1}(x_1)}$ (continuous case).

Theorem 5.18. Given two random vectors $X = (X_1,\dots,X_n)$ and $Y = (Y_1,\dots,Y_n)$ and a series of bijection $\{g_i\}$ that $X_i = g_{i(Y_i)}$, then

$$f_{Y_1,\dots,Y_n}(y_1,\dots,y_n) = f_{X_1,\dots,X_n}(g_1(y_1,\dots,y_n),\dots,g_n(y_1,\dots,y_n))\left|\frac{\partial(x_1,\dots,x_n)}{\partial(y_1,\dots,y_n)}\right|.$$

Theorem 5.19. Two random vectors $X = (X_1,\dots,X_n)$ and $Y = (Y_1,\dots,Y_n)$ are mutually independent iff

$$\begin{cases} p_{X_1,X_2}(x_1,x_2) = p_{X_1}(x_1)p_{X_2}(x_2), & \text{(discrete case)}, \\ f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2), & \text{(continuous case)}. \end{cases}$$

5.1.4 Distributional quantities

Definition 5.20. Given a random variable X, the **expectation** of X is

$$E(X) = \sum_{x\in\ \mathrm{Range}(X)} xp(x), \quad \text{if} \sum_{x\in\ \mathrm{Range}(X)} |x|p(x) < \infty \qquad \text{(discrete case)},$$

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x, \quad \text{if} \int_{-\infty}^{+\infty} |x|f(x)\mathrm{d}x < \infty \qquad \text{(continuous case)}.$$

Definition 5.21. Given a random variable X, the ***k*-th moment** of X is $E(X^k)$, and the ***k*-th central moment** is $E((X-E(X))^k)$.

Example 5.22. The **variance** of random variable X is the **2-nd central moment** of X,

$$\sigma^2 = \mathrm{Var}(X) = E((X-E(X))^2) = E(X^2) - E(X)^2.$$

Definition 5.23. Given a random variable X, if $E(e^{tX})$ exists for $t \in \mathbb{R}$, then the **moment generating function (mgf)** of X is

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} \frac{t^k E(X^k)}{k!}.$$

Theorem 5.24. The moment generating function (mgf) of random variables X and Y satisfies

(1) For all $k \in \mathbb{N}^*$, $M^{(k)}(0) = E(X^k)$;
(2) If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

5.2 Characteristic functions

5.3 Probability limit theorems

5.4 Common distributions

5.4.1 Common discrete distributions

Definition 5.25. (Bernoulli distribution) If X is a random variable with Bernoulli$(p), p \in (0,1)$, then:

$$P(X=1)=p, P(X=0)=1-p.$$

Theorem 5.26. For Bernoulli(p), the expectation is $\mu = p$, the variance is $\sigma^2 = p(1-p)$, the moment generating function is $M(t) = (1-p) + pe^t$.

Definition 5.27. (Binomial distribution) If X is a random variable with Binomial$(n,p), p \in (0,1), n \in \mathbb{N}^*$, then:

$$P(X=x) = C_n^x p^x (1-p)^{n-x}, x = 0, 1, ..., n.$$

Theorem 5.28. For Binomial(n,p), the expectation is $\mu = np$, the variance is $\sigma^2 = np(1-p)$, the moment generating function is $M(t) = \left((1-p)+pe^t\right)^n$.

Definition 5.29. (Geometric distribution) If X is a random variable with Geometric$(p), p \in (0,1)$, then:

$$P(X=x) = p(1-p)^x, x \in \mathbb{N}.$$

Theorem 5.30. For Geometric(p), the expectation is $\mu = \frac{p}{1-p}$, the variance is $\sigma^2 = \frac{1-p}{p^2}$, the moment generating function is $M(t) = p\left(1-(1-p)e^t\right)^{-1}$.

Definition 5.31. (Hypergeometric distribution) If X is a random variable with Hypergeometric$(N,D,n), n = 1, 2, ..., \min(N,D)$, then:

$$P(X=x) = \frac{C_{N-D}^{n-x} C_D^x}{C_N^n}, x = 0, 1, ..., n.$$

Theorem 5.32. For Hypergeometric(N,D,n), the expectation is $\mu = \frac{nD}{N}$, the variance is $\sigma^2 = \frac{nD(N-D)(N-n)}{N^2(N-1)}$.

Definition 5.33. (Negative binomial distribution) If X is a random variable with NB$(r,p), r \in \mathbb{N}^*, p \in (0,1)$, then:

$$P(X=x) = C_{x+r-1}^{r-1} p^r (1-p)^x, x \in \mathbb{N}.$$

Theorem 5.34. For NB(p), the expectation is $\mu = \frac{rp}{1-p}$, the variance is $\sigma^2 = \frac{r(1-p)}{p^2}$, the moment generating function is $M(t) = p^r\left(1-(1-p)e^t\right)^{-r}$.

Definition 5.35. (Poisson distribution) If X is a random variable with Poisson$(\lambda), \lambda > 0$, then:

$$P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}, x \in \mathbb{N}.$$

Theorem 5.36. For Poisson(p), the expectation is $\mu = \lambda$, the variance is $\sigma^2 = \lambda$, the moment generating function is $M(t) = \exp\big(\lambda(e^t - 1)\big)$.

Chapter 6

Stochastic Process

6.1 Poisson process

6.2 Markov chain

Chapter 7

Statistics

Chapter 8

Graph

8.1 Shortest Path

8.2 Matching

8.3 Network Flow

8.4 Tree

Chapter 9

Combinatorics

9.1 Generating function

9.2 Inclusion–exclusion principle

9.3 Special Numbers

9.3.1 Catalan number

9.3.2 Stirling number

Part 2

Scientific Computing

Chapter 10

Interpolation

10.1 Polynomial Interpolation

10.1.1 Lagrange formula

Definition 10.1. To interpolate given points $(x_0, f(x_0)), ..., (x_n, f(x_n))$, the Lagrange formula is

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x),$$

where the **elementary Lagrange interpolation polynomial** (or **fundamental polynomial**) for pointwise interpolation $l_k(x)$ is

$$l_k(x) = \prod_{i=0}^{n} \frac{x - x_i}{x_k - x_i}.$$

In particular, for $n = 0, l_0(x) = 1$.

10.1.2 Newton formula

Definition 10.2. The kth divided difference ($k \in \mathbb{N}^+$) on the **table of divided differences**

$$\begin{array}{c|ccccc}
x_0 & f[x_0] & & & \\
x_1 & f[x_1] & f[x_0, x_1] & & \\
x_2 & f[x_2] & f[x_1, x_2] & f[x_0, x_1, x_2] & \\
x_3 & f[x_3] & f[x_2, x_3] & f[x_1, x_2, x_3] & f[x_0, x_1, x_2, x_3] \\
... & ... & ... & ... & ...
\end{array}$$

where the **divided differences** satisfy

$$\begin{aligned}
f[x_0] &= f(x_0), \\
f[x_0, ..., x_k] &= \frac{f[x_1, ..., x_k] - f\left[x_0, ..., x_{\{k-1\}}\right]}{x_k - x_0}.
\end{aligned}$$

Corollary 10.3. Suppose $(i_0, ..., i_k)$ is a permutation of $(0, ..., k)$. Then

$$f[x_0, ..., x_k] = f\left[x_{i_0}, ..., x_{i_k}\right].$$

Theorem 10.4. For distinct points $x_0, ..., x_n$ and x, we have

$$f(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, ..., x_n] \prod_{i=0}^{n-1} (x - x_i) + f[x_0, ..., x_n, x] \prod_{i=0}^{n} (x - x_i).$$

Definition 10.5. The **Newton formula** for interpolating the points $(x_0, f(x_0)), ..., (x_n, f(x_n))$ is

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, ..., x_n] \prod_{i=0}^{n-1} (x - x_i).$$

10.1.3 Neville-Aitken algorithm

Definition 10.6. Denote $p_0^{[i]}(x) = f(x_i)$ for $i = 0, ..., n$. For all $k = 0, ..., n-1$ and $i = 0, ..., n-k-1$, define

$$p_{k+1}^{[i]}(x) = \frac{(x - x_i)p_k^{[i+1]}(x) - (x - x_{x+k+1})p_k^{[i]}(x)}{x_{i+k+1} - x_i}.$$

Then each $p_k^{[i]}(x)$ is the interpolating polynomial for the function f at the points $x_i, ..., x_{\{i+k\}}$. In particular, $p_n^{[0]}(x)$ is the interpolating polynomial of degree n for the function f at the points $x_0, ..., x_n$.

10.1.4 Hermite interpolation

Definition 10.7. Given distinct points $x_0, ..., x_k$ in $[a, b]$, non-negative integers $m_0, ..., m_k$, and a function $f \in C^M[a, b]$ where $M = \max_{i=0,...,k}(m_i)$, the **Hermite interpolation problem** seeks a polynomial $p(x)$ of the lowest degree satisfies

$$\forall i \in \{0, ..., k\}, \forall \mu \in \{0, ..., m_i\}, p^{(\mu)}(x_i) = f^{(\mu)}(x_i).$$

Definition 10.8. (Generalized divided difference) Let $x_0, ..., x_k$ be $k+1$ pairwise distinct points with each x_i repeated $m_i + 1$ times; write $N = k + \sum_{i=0}^{k} m_i$. The Nth divided difference associated with these points is the cofficient of x^N in the polynomial p that uniquely solves the Hermite interpolation problem.

Corollary 10.9. The nth divided difference at $n+1$ “confluent” (i.e. identical) points is

$$f[x_0, ..., x_0] = \frac{1}{n!} f^{(n)}(x_0),$$

where x_0 is repeated $n+1$ times on the left-hand side.

10.1.5 Approximation

Definition 10.10. Given condition functions $c_0, ..., c_k : \mathbb{P}_n \to \mathbb{R}^+$, the **Approximation problem** seeks a polynomial $p_n(x)$ of the given degree n satisfies a unconstrained optimization

$$\min_{p_n \in \mathbb{P}_n} \sum_{i=0}^{k} c_i\left(p_n^{(m_i)}\right).$$

where condition function $c(p)$ includes but is not limited to

$$|p^{(m)}(x)|, \left(p_n^{(m)}(x)\right)^2, \int_a^b |p^{(m)}| \, \mathrm{d}x, \int_a^b \left(p^{(m)}\right)^2 \mathrm{d}x.$$

Example 10.11. For non-negative integers $m_0, ..., m_k$ and condition functions $c_i(p_n) = \left(p_n^{(m_i)}(x)\right)^2$, denote by

$$p_n(x) = \sum_{i=0}^{n} c_i x^i$$

the polynomial of the given degree n, then the mth derivative of p_n is

$$p_n^{(m)}(x) = \sum_{i=m}^{n} \frac{i!}{(i-m)!} c_i x^{i-m}.$$

All above implies the least squares system

$$\begin{cases} p_n^{(m_0)}(x) = \sum_{i=m_0}^{n} \frac{i!}{(i-m_0)!} c_i x^{i-m_0} = 0, \\ \dots \\ p_n^{(m_k)}(x) = \sum_{i=m_k}^{n} \frac{i!}{(i-m_k)!} c_i x^{i-m_k} = 0, \end{cases}$$

which can be solved by algorithms such as Householder transformation.

10.1.6 Error analysis

Theorem 10.12. Let $f \in C^n[a, b]$ and suppose that $f^{(n+1)}(x)$ exists at each point of (a, b). Let $p_n(x) \in \mathbb{P}_n$ denote the unique polynomial that coincides with f at $x_0, \dots, x_n$. Define

$$R_n(f; x) = f(x) - p_n(x),$$

as the **Cauchy remainder** of the polynomial interpolation.

If $a \le x_0 < \cdots < x_n \le b$, then there exists some $\xi \in (a, b)$ satisfies

$$R_n(f; x) = \frac{f^{\{(n+1)\}}(xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$$

where the value of ξ depends on $x, x_0, \dots, x_n$ and f.

Theorem 10.13. For the Hermite interpolation problem, denote $N = k + \sum_{i=0}^{k} m_i$. Denote by $p_N(x) \in \mathbb{P}_N$ the unique solution of the problem. Suppose $f^{(N+1)}(x)$ exists in (a, b). Then there exists some $\xi \in (a, b)$ satisfies

$$R_N(f; x) = \frac{f^{(N+1)}(\xi)}{(N+1)!} \prod_{i=0}^{k} (x - x_i)^{m_i+1}.$$

10.2 Spline

Definition 10.14. Given nonnegative integers n, k, and a strictly increasing sequence $a = x_1 < \cdots < x_N = b$, the set of **spline** functions of degree n and smoothness class k relative to the partition $\{x_i\}$ is

$$\mathbb{S}_n^k = \left\{ s : s \in C^k[a, b]; \forall i \in \{1, \dots, N-1\}, s \mid_{[x_i, x_{i+1}]} \in \mathbb{P}_n \right\},$$

where x_i is the **knot** of the spline.

10.2.1 Cubic spline

Definition 10.15. (Boundary conditions of splines) The followings are common boundary conditions of cubic splines.

- The **complete cubic spline** s satisfies $s'(a) = f'(a), s'(b) = f'(b)$;
- The **cubic spline with specified second derivatives** s satisfies $s''(a) = f''(a), s''(b) = f''(b)$;
- The **natural cubic spline** s satisfies $s''(a) = s''(b) = 0$;
- The **not-a-knot cubic spline** s satisfies $s'''(x)$ exists at $x = x_2$ and $x = x_{N-1}$.
- The **periodic cubic spline** s satisfies $s(a) = s(b), s'(a) = s'(b), s''(a) = s''(b)$.

Theorem 10.16. Denote $m_i = s'(x_i), M_i = s''(x_i)$ for $s \in \mathbb{S}_3^2$, then

$$\forall i = 2, 3, ..., N-1, \quad \lambda_i m_{i-1} + 2m_i + \mu_i m_i + 1 = 3\mu_i f[x_i, x_{i+1}] + 3\lambda_i f[x_{i-1}, x_i],$$
$$\forall i = 2, 3, ..., N-1, \quad \mu_i M_{i-1} + 2M_i + \lambda_i m_{i+1} = 6f[x_{i-1}, x_i, x_{i+1}],$$

where

$$\mu_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}, \quad \lambda_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}.$$

In particular, m_i and M_i should be replaced to the derivatives given at the boundary.

Theorem 10.17. Cubic spline $s \in \mathbb{S}_3^2$ from the linear system of $\lambda_i, \mu_i, m_i, M_i$ and the boundary conditions.

10.2.2 B-spline

Definition 10.18. B-splines are defined recursively by

$$B_i^{n+1}(x) = (x - x_{i-1})(x_{i+n} - x_{i-1})B_i^n(x) + \frac{x_{i+n+1} - x}{x_{i+n+1} - x_{i-1}} B_{i+1}^n(x),$$

where recursion base is the B-spline of degree zero

$$B_i^0(x) = \begin{cases} 1, & x \in (x_{i-1}, x_i], \\ 0, & \text{otherwise}. \end{cases}$$

Theorem 10.19. The $\{B_i^n(x)\}$ forms a basis of $\mathbb{S}_n^{n-1}$.

Definition 10.20. For $N \in \mathbb{N}^*$, the **support** of a $B_i^n(x)$ is

$$\text{supp}\ \{B_i^n(x)\} = \overline{\{x \in \mathbb{R} : B_i^n(x) \neq 0\}} = [x_{i-1}, x_{i+n}].$$

Theorem 10.21. (Integrals of B-splines) The average of a B-spline over its support only depends on its degree,

$$\frac{1}{t_{i+n} - t_{i-1}} \int_{t_{i-1}}^{t_{i+n}} B_i^n(x) \mathrm{d}x = \frac{1}{n+1}.$$

Theorem 10.22. (Derivatives of B-splines) For $n \geq 2$, we have

$$\forall x \in \mathbb{R}, \quad \frac{\mathrm{d}}{\mathrm{d}x} B_i^n(x) = \frac{nB_i^{n-1}(x)}{x_{i+n-1} - x_{i-1}} - \frac{nB_{i+1}^{n-1}(x)}{x_{i+n} - x_i}.$$

For $n = 1$, it holds for all x except x_{i-1}, t_i, t_{i+1}, where the derivative of $B_i^1(x)$ is not defined.

10.2.3 Error analysis

Theorem 10.23. Suppose a function $f \in C^4[a,b]$, is interpolated by a complete cubic spline or a cubic spline with specified second derivatives at its end points. Then

$$\forall m = 0, 1, 2, |f^{(m)}(x) - s^{(m)}(x)| \leq c_m h^{4-m} \max_{x \in [a,b]} |f^{(4)}(x)|,$$

where $c_0 = \frac{1}{16}, c_1 = c_2 = \frac{1}{2}$ and $h = \max_{i=1,\ldots,N-1} |x_{i+1} - x_i|$.

Chapter 11
Integration

Definition 11.1. A **weighted quadrature formula** $I_n(f)$ is a linear function

$$I_n(f) = \sum_{i=1}^{n} w_i f(x_i),$$

which approximates the integral of a function $f \in C[a, b]$,

$$I(f) = \int_a^b \rho(x) f(x) \mathrm{d}x,$$

where the weight function $\rho \in [a, b]$ satisfies $\forall x \in (a, b),\ \rho(x) > 0$. The points $\{x_i\}$ at which the integrand f is evaluated are called nodes or abscissas, and the multipliers $\{w_i\}$ are called weights or coefficients.

Definition 11.2. A weighted quadrature formula has (polynomial) **degree of exactness** d_E iff

$$\begin{aligned} &\forall f \in \mathbb{P}_{d_E}, \quad E_n(f) = 0, \\ &\exists g \in \mathbb{P}_{d_E+1}, \ \text{s.t.} \ \ E_n(g) \neq 0 \end{aligned}$$

where $\mathbb{P}_d$ denotes the set of polynomials with degree no more than d.

Theorem 11.3. A weighted quadrature formula $I_n(f)$ satisfies $d_E \leq 2n - 1$.

Definition 11.4. The **error** or **remainder** of $I_n(f)$ is

$$E_n(f) = I(f) - I_n(f),$$

where $I_n(f)$ is said to be convergent for $C[a, b]$ iff

$$\forall f \in C[a, b], \lim_{n \to +\infty} E_n(f) = 0.$$

Theorem 11.5. Let $x_1, ..., x_n$ be given as distinct nodes of $I_n(f)$. If $d_E \geq n - 1$, then its weights can be deduced as

$$\forall k \in \{1, ..., n\}, w_k = \int_a^b \rho(x) l_k(x) \mathrm{d}x,$$

where $l_k(x)$ is the elementary Lagrange interpolation polynomial for pointwise interpolation applied to the given nodes.

11.1 Newton-Cotes Formulas

Definition 11.6. A **Newton-Cotes formula** is a formula based on approximating $f(x)$ by interpolating it on uniformly spaced nodes $x_1, ..., x_n \in [a, b]$.

11.1.1 Midpoint rule

Definition 11.7. The **midpoint rule** is a formula based on approximating $f(x)$ by the constant $f\left(\frac{a+b}{2}\right)$.

For $\rho(x) \equiv 1$, it is simply

$$I_M(f) = (b-a) f\left(\frac{a+b}{2}\right).$$

Theorem 11.8. For $f \in C^2[a,b]$, with weight functino $\rho \equiv 1$, the error (remainder) of midpoint rule satisfies

$$\exists \xi \in [a,b], \text{ s.t. } E_M(f) = \frac{(b-a)^3}{24} f''(\xi).$$

Corollary 11.9. The midpoint rule has $d_E = 1$.

11.1.2 Trapezoidal rule

Definition 11.10. The **trapezoidal rule** is a formula based on approximating $f(x)$ by the straight line that connects $(a, f(a))$ and $(b, f(b))$.

For $\rho(x) \equiv 1$, it is simply

$$I_T(f) = \frac{b-a}{2}(f(a) + f(b)).$$

Theorem 11.11. For $f \in C^2[a,b]$, with weight functino $\rho \equiv 1$, the error (remainder) of trapezoidal rule satisfies

$$\exists \xi \in [a,b], \text{ s.t. } E_T(f) = -\frac{(b-a)^3}{12} f''(\xi).$$

Corollary 11.12. The trapezoidal rule has $d_E = 1$.

11.1.3 Simpson's rule

Definition 11.13. The **Simpson's rule** is a formula based on approximating $f(x)$ by the quadratic polynomial that goes through the points $(a, f(a))$, $\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right)$ and $(b, f(b))$.

For $\rho(x) \equiv 1$, it is simply

$$I_S(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Theorem 11.14. For $f \in C^4[a,b]$, with weight functino $\rho \equiv 1$, the error (remainder) of Simpson's rule satisfies

$$\exists \xi \in [a,b], \text{ s.t. } E_T(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi).$$

Corollary 11.15. The Simpson's rule has $d_E = 3$.

11.2 Gauss Formulas

Theorem 11.16. For an interval $[a, b]$ and a weight function $\rho : [a, b] \to \mathbb{R}$, the nodes for gauss formula $I_n(f)$ is the root of the nth order orthogonal polynomial on $[a, b]$ with the weight function $\rho(x)$.

Theorem 11.17. A Gauss formula $I_n(f)$ has $d_E = 2n - 1$.

Chapter 12
Optimization

12.1 Optimality Conditions

Definition 12.1. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^*$ is a **global minimizer** of f if $\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^*$ is a **local minimizer** of f if $\exists \delta > 0, \forall \mathbf{x} \in U(\mathbf{x}, \delta), f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

Theorem 12.2. (1st-order necessary conditions) Let $f \in C^1(\mathbb{R}^n)$ and $\mathbf{x}^*$ be a local minimizer of f, then $\nabla f(\mathbf{x}^*) = 0$.

Definition 12.3. Let $f \in C^1(\mathbb{R}^n)$ then $\mathbf{x}^*$ is called a **stationary point** of f if $\nabla f(\mathbf{x}^*) = 0$.

Theorem 12.4. (2nd-order necessary conditions) Let $f \in C^2(\mathbb{R}^n)$.

- If $\mathbf{x}^*$ is a local minimizer of f, then $\nabla^2 f(\mathbf{x}^*) \succeq 0$;
- If $\mathbf{x}^*$ is a stationary point of f and $\nabla^2 f(\mathbf{x}^*) \succ 0$, then $\mathbf{x}^*$ is a local minimizer.

Definition 12.5. Let $f \in C^1(\mathbb{R}^n)$ and $\mathbf{x} \in \mathbb{R}^n$. A $\mathbf{d} \in \mathbb{R}^n$ is called a **descent direction** at x if

$$(\nabla f(\mathbf{x}))^T \mathbf{d} < 0.$$

Specifically, $-\nabla f(x)$ is called the **steepest descent direction**.

Remark 12.6. Let $D \in \mathbb{R}^{n\times n}$ and $D \succ 0$, then $d = -D\nabla f(x)$ is a descent direction.

12.1.1 KKT Conditions

Definition 12.7. We say that x^* is a local minimizer of

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{s.t.}\ \ h_j(x) = 0, j = 1, ..., p, \\ &\qquad g_i(x) \leq 0, i = 1, ..., m. \end{aligned}$$

if x^* is feasible and exists $\varepsilon > 0$ such that $f(x) \geq f(x^*)$ whenever x is feasible and $\|x - x^*\|_2 \leq \varepsilon$.

Theorem 12.8. (Karush-Kuhn-Tucker conditions for the LP, KKT condition) Consider the linear program

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} c^T x \\ &\text{s.t.}\ \ Bx = d, \\ &\qquad Ax \leq b. \end{aligned}$$

where $c \in \mathbb{R}^n$, $B \in \mathbb{R}^{p\times n}$ and $A \in \mathbb{R}^{q\times n}$. Then $x^* \in \mathbb{R}^n$ is an optimal solution iff there exists $\lambda^* \in \mathbb{R}^q$ and $\mu^* \in \mathbb{R}^p$ such that the following conditions holds:

- (Primal feasibility) $Bx^* = d$ and $Ax \leq b$;
- (Dual feasibility) $B^T\mu^* + A^T\lambda^* + c = 0$ and $\lambda^* \geq 0$;

- (Complementary slackness) ${\lambda^*}^T(Ax^* - b) = 0$.

Theorem 12.9. (Mangasarian-Fromovitz constraint qualification) Consider the feasible set of

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} f(x)\\
&\text{s.t.}\ \ h_j(x) = 0, j = 1, ..., p,\\
&\qquad g_i(x) \le 0, i = 1, ..., m.
\end{aligned}$$

and let x^* be feasible. We say that the **Mangasarian-Fromovitz constraint qualification (MFCQ)** holds at x^* if the following conditions holds:

- If $\sum_{j\in J} \mu_j \nabla h_j(x^*) + \sum_{i\in I(x^*)} \lambda_i \nabla g_i(x^*) = 0$ and $\forall i \in I(x^*), \lambda_i \ge 0$ then $\lambda_i = 0$ for all $i \in I(x^*)$ and $\mu_j = 0$ for all $j \in J$.

Theorem 12.10. (KKT conditions for NLP) Consider

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} f(x)\\
&\text{s.t.}\ \ h_j(x) = 0, j = 1, ..., p,\\
&\qquad g_i(x) \le 0, i = 1, ..., m.
\end{aligned}$$

and let x^* be a local minimizer. Suppose that MFCQ holds at x^*. Then there exists $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that

- $\nabla f(x^*) + \sum_{j\in J} \mu_j^* \nabla h_j(x^*) + \sum_{i\in I(x^*)} \lambda_i^* \nabla g_i(x^*) = 0$ and $\forall i \in I, \lambda_i^* \ge 0, \lambda_i^* g_i(x^*) = 0$.

Definition 12.11. Consider

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} f(x)\\
&\text{s.t.}\ \ h_j(x) = 0, j = 1, ..., p,\\
&\qquad g_i(x) \le 0, i = 1, ..., m.
\end{aligned}$$

An $\overline{x}$ is called a stationary point if it is feasible and there exist $\overline{\lambda} \in \mathbb{R}^m$ and $\overline{\mu} \in \mathbb{R}^p$ such that

- $\nabla f(\overline{x}) + \sum_{j\in J} \overline{\mu_j} \nabla h_j(\overline{x}) + \sum_{i\in I(\overline{x})} \overline{\lambda_i} \nabla g_i(\overline{x}) = 0$ and $\forall i \in I, \overline{\lambda_i} \ge 0, \overline{\lambda_i} g_i(\overline{x}) = 0$.

Theorem 12.12. (MFCQ from Slater) Consider the set defined by

$$S = \{x \in \mathbb{R}^n : \forall i \in I, g_i(x) \le 0\},$$

where g_i are convex C^1. Suppose that there exist $\overline{x}$ satisfying

$$\forall i \in I, g_i(\overline{x}) < 0.$$

Then MFCQ holds at every point in S.

Theorem 12.13. (MFCQ from generalized Slater) Consider the set defined by

$$S = \{x \in \mathbb{R}^n : \forall i \in I, g_i(x) \le 0, Ax = b\},$$

where g_i are convex C^1 and $A \in \mathbb{R}^{p\times n}$. Suppose that there exist $\overline{x}$ satisfying

$$\forall i \in I, g_i(\overline{x}) < 0, A\overline{x} = b,$$

and A has full row rank. Then MFCQ holds at every point in S.

Theorem 12.14. (Sufficiency under convexity) Consider

$$\begin{aligned}\min_{x\in\mathbb{R}^n} \quad & f(x)\\ \text{s.t.} \quad & Ax=b,\\ & g_i(x)\le 0, i\in I,\\ \text{where } f \text{ and } g_i & \text{ are convex } C^1, A\in\mathbb{R}^{p\times n}.\end{aligned}$$

Suppose that there exist $x^*\in\mathbb{R}^n$, $\lambda^*\in\mathbb{R}^m$ and $\mu^*\in\mathbb{R}^p$ such that

- $\forall i\in I, g_i(x^*)\le 0, Ax^*=b$;
- $\nabla f(x^*)+\sum_{i\in I}\lambda_i^*\nabla g_i(x^*)+A^T\mu^*=0$;
- $\forall i\in I, \lambda_i^*\ge 0, \lambda_i^* g_i(x^*)=0$.

Then x^* is a global minimizer.

12.2 One-dimensional Line Search

12.2.1 Inexact line search

Definition 12.15. (Armijo rule) Let $\sigma\in(0,1)$, $\mathbf{x}\in\mathbb{R}^n$ and $\mathbf{d}\in\mathbb{R}^n$. Find $\alpha>0$ such that

$$f(\mathbf{x}+\alpha\mathbf{d})\le f(\mathbf{x})+\alpha\sigma(\nabla f(\mathbf{x}))^T\mathbf{d}.$$

Theorem 12.16. Let $f\in C^1(\mathbb{R}^n)$, $\mathbf{x}\in\mathbb{R}^n$, $\sigma\in(0,1)$ and $\mathbf{d}\in\mathbb{R}^n$ be a descent direction at $\mathbf{x}$. Then there exists $\alpha_1>0$ such that for all $\alpha\in[0,\alpha_1]$,

$$f(\mathbf{x}+\alpha\mathbf{d})\le f(\mathbf{x})+\alpha\sigma(\nabla f(\mathbf{x}))^T\mathbf{d}.$$

Method 12.17. (Armijo line search by backtracking) Fix $\sigma\in(0,1)$ and $\beta\in(0,1)$. Given $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{d}\in\mathbb{R}^n$ and $\overline{\alpha}>0$. Find the smallest nonnegative integer j such that

$$f(\mathbf{x}+\overline{\alpha}\beta^j\mathbf{d})\le f(\mathbf{x})+\overline{\alpha}\beta^j\sigma(\nabla f(\mathbf{x}))^T\mathbf{d},$$

then the stepsize generated is $\overline{\alpha}\beta^j$.

Theorem 12.18. (Convergence under Armijo rule) Let $f\in C^1(\mathbb{R}^n)$ with $\inf f>-\infty$. Let $\{\overline{\alpha}^{[k]}\}\subset\mathbb{R}$ satisfies $0<\inf_k\alpha^{[k]}\le\sup_k\alpha^{[k]}<\infty$, and fix $\sigma\in(0,1)$ and $\beta\in(0,1)$. Suppose $\{\mathbf{x}^{[k]}\}$ is generated as $\mathbf{x}^{[k+1]}=\mathbf{x}^{[k]}+\alpha^{[k]}\mathbf{d}^{[k]}$, where

- $\mathbf{d}^{[k]}=-D^{[k]}\nabla f(\mathbf{x}^{[k]})$, where $\{D^{[k]}\}$ is a bounded sequence of positive definite matrices with $D^{[k]}-\delta I\succeq 0$ for some δ;
- $\alpha^{[k]}$ is generated via the Armijo line search by backtracking.

Then any accumulation point of $\{x^{[k]}\}$ is a stationary point of f.

Definition 12.19. (Wolfe's condition) Let $0<c_1<c_2<1$, $\mathbf{x}\in\mathbb{R}^n$ and $\mathbf{d}\in\mathbb{R}^n$. Find α such that

$$\text{(Armijo rule)} \qquad f(\mathbf{x}+\alpha\mathbf{d})\le f(\mathbf{x})+\alpha c_1(\nabla f(\mathbf{x}))^T\mathbf{d},$$

$$\text{(curvature condition)} \quad -(\nabla f(\mathbf{x}+\alpha\mathbf{d}))^T\mathbf{d} \leq -c_2(\nabla f(\mathbf{x}))^T\mathbf{d}.$$

Theorem 12.20. (Wolfe's conditions are not void) Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$ and $\mathbf{d} \in \mathbb{R}^n$ be a descent direction at $\mathbf{x}$. Let $0 < c_1 < c_2 < 1$. Then there exists $\alpha > 0$ with

$$\begin{aligned} &\text{(Armijo rule)} & f(\mathbf{x}+\alpha\mathbf{d}) &\leq f(\mathbf{x}) + \alpha c_1(\nabla f(\mathbf{x}))^T\mathbf{d}, \\ &\text{(curvature condition)} & -(\nabla f(\mathbf{x}+\alpha\mathbf{d}))^T\mathbf{d} &\leq -c_2(\nabla f(\mathbf{x}))^T\mathbf{d}. \end{aligned}$$

Theorem 12.21. (Strong Wolfe conditions) Let $0 < c_1 < c_2 < \frac{1}{2}$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{d} \in \mathbb{R}^n$. Find $\alpha > 0$ such that

$$\begin{aligned} f(\mathbf{x}+\alpha\mathbf{d}) &\leq f(\mathbf{x}) + \alpha c_1 \nabla f(\mathbf{x})^T\mathbf{d}, \\ \left|\nabla f(\mathbf{x}+\alpha\mathbf{d})^T\mathbf{d}\right| &\leq c_2\left|\nabla f(\mathbf{x})^T\mathbf{d}\right|. \end{aligned}$$

12.2.2 Exact line search

Definition 12.22. Given a function $f : \mathbb{R}^n \to \mathbb{R}$, a initial point $\mathbf{x}$ and a direction $\mathbf{d}$, denoted by $\varphi(\alpha) = f(\mathbf{x}+\alpha\mathbf{d})$, a **exact line search** method solves the problem

$$\varphi(\alpha) = \min_{t\in\mathbb{R}^+} \varphi(t).$$

Method 12.23. (Success-failure method) For a one-dimensional line search problem, the **success-failure method** is an inexact one-dimensional line search method to solve the interval $[a, b] \in [0, +\infty)$ that exact solution $\alpha^* \in [a, b]$, where we

(1) Choose initial value $\alpha_0 \in [0, +\infty)$, $h_0 > 0$, $t > 0$(commonly choose $t = 2$), calculate $\varphi(\alpha_0)$ and let $k = 0$;
(2) Let $\alpha_{k+1} = \alpha_k + h_k$ and calculate $\varphi(\alpha_{k+1})$, if $\varphi(\alpha_{k+1}) < \varphi(\alpha_k)$, then go to (3), otherwise go to (4);
(3) Let $h_{k+1} = th_k$, $\alpha = \alpha_k$, $k = k + 1$, and go to (2);
(4) If $k = 0$, then let $h_k = -h_k$ and go to (2), otherwise stop and the solution $[a, b]$ satisfies

$$a = \min\{\alpha, \alpha_k\}, \quad b = \max\{\alpha, \alpha_k\}.$$

Definition 12.24. A general form of one-dimensional line search method is the following three steps:

(1) **Initialization**: given initial point $\mathbf{x}$ and acceptable error $\varepsilon > 0$, $\delta > 0$;
(2) **Iteration**: calculate the direction $\mathbf{d}$ and step size α that $f(\mathbf{x}+\alpha\mathbf{d}) = \min_{t\in\mathbb{R}^+} f(\mathbf{x}+t\mathbf{d})$ and let $\mathbf{x} = \mathbf{x} + \alpha\mathbf{d}$;
(3) **Stop condition**: if $\|\nabla f(\mathbf{x})\| \leq \varepsilon$ or $U_{\mathbb{R}^n}(x, \delta)$ includes the exact solution, then the current $\mathbf{x}$ is the solution.

where the iteration step are repeated until $\mathbf{x}$ satisfies the stop condition.

Definition 12.25. Given a method, denoted by $\{\mathbf{x}_k\}$ the sequence of the iteration and $\mathbf{x}^*$ the exact solution, the method is **(Q-)linear convergence** if

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} \in (0, 1),$$

the method is **(Q-)sublinear convergence** if

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 1,$$

the method is **(Q-)superlinear convergence** if

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0.$$

For a superlinear convergence method, the method is r-order linear convergence if

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|^r} \in [0, +\infty),$$

where when $r = 2$ is called **(Q-)quadratic convergence**.

Remark 12.26. There is another **R-convergence** for judging a sequence which use another Q-convergence sequence as the boundary of $\{\|\mathbf{x}_k - x^*\|\}$, but is not needed here.

Method 12.27. (Golden section method) Given the initial point $\mathbf{x}$, an interval $[a, b]$ and $\delta > 0$,

- The iteration step is:
  (1) Calculate the two testing points $\lambda = a + (1-k)(b-a)$ and $\mu = a + k(b-a)$ where $k = \frac{\sqrt{5}-1}{2}$ is the golden ratio;
  (2) If $\varphi(\lambda) > \varphi(\mu)$, let $a = \lambda$, otherwise let $b = \mu$.
- The stop condition is $b - a \le \delta$;
- The solution is $\mathbf{x} + \frac{a+b}{2}\mathbf{d}$.

Theorem 12.28. The golden section method is a **linear convergent** method.

Method 12.29. (Fibonacci method) Given the initial point $\mathbf{x}$, an interval $[a, b]$ and $\delta > 0$,

- The k-th iteration step is:
  (1) Calculate the two testing points $\lambda = a + \frac{F_k}{F_{k+2}}(b-a)$ and $\mu = a + \frac{F_{k+1}}{F_{k+2}}(b-a)$ where F_k is the k-th fibonacci number and k;
  (2) If $\varphi(\lambda) > \varphi(\mu)$, let $a = \lambda$, otherwise let $b = \mu$.
- The stop condition is $b - a \le \delta$;
- The solution is $\mathbf{x} + \frac{a+b}{2}\mathbf{d}$.

Theorem 12.30. The Fibonacci method is a **linear convergent** method.

Method 12.31. (Bisection method) Given the initial point $\mathbf{x}$, an interval $[a, b]$ and $\delta > 0$,

- The iteration step is:
  (1) Calculate the midpoint $m = \frac{a+b}{2}$ and $\varphi(m)$;
  (2) If $\nabla f(m) \cdot d < 0$, let $a = m$, otherwise let $b = m$.
- The stop condition is $b - a \le \delta$;
- The solution is $\mathbf{x} + \frac{a+b}{2}\mathbf{d}$.

Theorem 12.32. The bisection method is a **linear convergent** method.

Method 12.33. (Newton's method) Given the initial point $\mathbf{x}$ and $\varepsilon > 0$,

- The iteration step is:
  (1) Calculate $(\nabla^2 f(\mathbf{x}))^T \cdot \mathbf{d}$ and $(\nabla f(\mathbf{x}))^T \cdot \mathbf{d}$;
  (2) Let $\mathbf{x} = \mathbf{x} - \frac{(\nabla f(\mathbf{x}))^T \cdot \mathbf{d}}{(\nabla^2 f(\mathbf{x}))^T \cdot \mathbf{d}}$;
- The stop condition is $(\nabla f(\mathbf{x}))^T \cdot \mathbf{d} \leq \varepsilon$;
- The solution is $\mathbf{x}$.

Theorem 12.34. The Newton's method is a **quadratic convergent** method.

12.3 Unconstrained Optimization

Definition 12.35. Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a **unconstrained optimization** method solves the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

by

(1) **Initialization**: given initial point $\mathbf{x}$ and acceptable error $\varepsilon > 0$, $\delta > 0$;
(2) **Iteration**: calculate the direction $\mathbf{d}$ and step size α, then let $\mathbf{x} = \mathbf{x} + \alpha \mathbf{d}$;
(3) **Stop condition**: if $\|\nabla f(\mathbf{x})\| \leq \varepsilon$ or $U_{\mathbb{R}^n}(\mathbf{x}, \delta)$ includes the exact solution, then the current $\mathbf{x}$ is the solution.

12.3.1 Gradient descent method

Method 12.36. (Gradient descent with exact line search)
Given $f \in C^1(\mathbb{R}^n)$,

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$,

For $k \in \mathbb{N}$,

(1) Set $\mathbf{d}^{[k]} = -\nabla f(\mathbf{x}^{[k]})$;

(2) Pick $\alpha^{[k]} \in \arg\min_{\alpha \in \mathbb{R}^+} \{ f(\mathbf{x}^{[k]} + \alpha \mathbf{d}^{[k]}) \}$;

(3) Set $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} + \alpha^{[k]} \mathbf{d}^{[k]}$.

Corollary 12.37. (Gradient descent with constant stepsize) Let $f \in C^2(\mathbb{R}^n)$ with $\inf f > -\infty$. Suppose that there exists $L > 0$ such that

$$\forall \mathbf{x} \in \mathbb{R}^n, L \geq \left\| \nabla^2 f(\mathbf{x}) \right\|.$$

For any fixed $\gamma \in (0, 2)$, and the sequence generated as

$$x^{[k+1]} = x^{[k]} - \frac{\gamma}{L} \nabla f(x^{[k]}),$$

then any accumulation point of $\{x^{[k]}\}$ is a stationary point of f.

Theorem 12.38. The gradient descent method is a **linear convergent** method.

12.3.2 Newton's method

Method 12.39. (Newton's method) Given $f \in C^2(\mathbb{R}^n)$,

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$,

For $k \in \mathbb{N}$,

(1) $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} - \left(\nabla^2 f\left(\mathbf{x}^{[k]}\right)\right)^{-1} \nabla f\left(\mathbf{x}^{[k]}\right)$.

Theorem 12.40. The Newton's method is a **quadratic convergent** method.

Example 12.41. (Failure of Newton's method) Given function $g = x - x^3$ and starting point $x^{[0]} = \frac{1}{\sqrt{5}}$, then the sequence of iteration is

$$x^{[1]} = -\frac{1}{\sqrt{5}}, x^{[2]} = \frac{1}{\sqrt{5}}, \ldots.$$

12.3.3 Quasi-Newton methods

Method 12.42. (Secant method) To solve $g(x) = 0$ where $g(x) \in C^1(\mathbb{R})$. Let $x^{[0]}, x^{[1]} \in \mathbb{R}$ and $g\left(x^{[0]}\right) \neq g\left(x^{[1]}\right)$, for $k = 1, \ldots,$, use finite difference to approximate g' in Newton's method, i.e.

$$x^{[k+1]} = x^{[k]} - g\left(x^{[k]}\right) \frac{x^{[k]} - x^{[k-1]}}{g\left(x^{[k]}\right) - g\left(x^{[k-1]}\right)}.$$

Definition 12.43. (Secant equations) Let $f \in C^2(\mathbb{R}^n)$ and given $\mathbf{x}^{[k+1]}$ and $\mathbf{x}^{[k]}$, we expect

$$\nabla^2 f\left(\mathbf{x}^{[k+1]}\right)\left(\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right) \approx \nabla f\left(\mathbf{x}^{[k+1]}\right) - \nabla f\left(\mathbf{x}^{[k]}\right).$$

Let $\mathbf{s}^{[k]} = \mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}$, $\mathbf{y}^{[k]} = \nabla f\left(\mathbf{x}^{[k+1]}\right) - \nabla f\left(\mathbf{x}^{[k]}\right)$ and $B^{[k+1]} = \left(H^{[k+1]}\right)^{-1}$ be the matrix constructed to approximate $\nabla^2 f\left(\mathbf{x}^{[k+1]}\right)$,

$$B^{[k+1]}\mathbf{s}^{[k]} = \mathbf{y}^{[k]}, \ H^{[k+1]}\mathbf{y}^{[k]} = \mathbf{s}^{[k]}.$$

Example 12.44. (Popular update formula) Initialize $B^{[0]}$ or $H^{[0]}$ at a positive definite matrix, then update by

DFP:

$$B^{[k+1]} = \left(I - \frac{\mathbf{y}^{[k]}\mathbf{s}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}}\right) B^{[k]} \left(I - \frac{\mathbf{s}^{[k]}\mathbf{y}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}}\right) + \frac{\mathbf{y}^{[k]}\mathbf{y}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}},$$

$$H^{[k+1]} = H^{[k]} + \frac{\mathbf{s}^{[k]}\mathbf{s}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}} - \frac{H^{[k]}\mathbf{y}^{[k]}\mathbf{y}^{[k]T}H^{[k]}}{\mathbf{y}^{[k]T}H^{[k]}\mathbf{y}^{[k]}};$$

BFGS:

$$B^{[k+1]} = B^{[k]} + \frac{\mathbf{y}^{[k]}\mathbf{y}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}} - \frac{B^{[k]}\mathbf{s}^{[k]}\mathbf{s}^{[k]T}B^{[k]}}{\mathbf{s}^{[k]T}B^{[k]}\mathbf{s}^{[k]}},$$

$$H^{[k+1]} = \left(I - \frac{\mathbf{s}^{[k]}\mathbf{y}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}}\right) H^{[k]} \left(I - \frac{\mathbf{y}^{[k]}\mathbf{s}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}}\right) + \frac{\mathbf{s}^{[k]}\mathbf{s}^{[k]T}}{\mathbf{y}^{[k]T}\mathbf{s}^{[k]}};$$

SR1:

$$B^{[k+1]} = B^{[k]} + \frac{\left(\mathbf{y}^{[k]} - B^{[k]}\mathbf{s}^{[k]}\right)\left(\mathbf{y}^{[k]} - B^{[k]}\mathbf{s}^{[k]}\right)^T}{\left(\mathbf{y}^{[k]} - B^{[k]}\mathbf{s}^{[k]}\right)^T \mathbf{s}^{[k]}},$$

$$H^{[k+1]} = H^{[k]} + \frac{\left(\mathbf{s}^{[k]} - H^{[k]}\mathbf{y}^{[k]}\right)\left(\mathbf{s}^{[k]} - H^{[k]}\mathbf{y}^{[k]}\right)^T}{\left(\mathbf{s}^{[k]} - H^{[k]}\mathbf{y}^{[k]}\right)^T \mathbf{y}^{[k]}}.$$

Remark 12.45.

- DFP and BFGS are rank-2 updates, while SR1 is rank-1 update.
- Since $B^{[0]}$ and $H^{[0]}$ are symmetric, all $B^{[k]}$ and $H^{[k]}$ are symmetric by induction.
- In practice, BFGS usually performs better.

Method 12.46. (Basic Quasi-Newton method) Given $f \in C^1(\mathbb{R}^n)$,

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$ and $B^{[0]} \succ 0$ (or $H^{[0]} \succ 0$),

For $k \in \mathbb{N}$,

(1) Find $\mathbf{d}^{[k]}$ via $B^{[k]}\mathbf{d}^{[k]} = -\nabla f\left(\mathbf{x}^{[k]}\right)$ (or $\mathbf{d}^{[k]} = -H^{[k]}\nabla f\left(\mathbf{x}^{[k]}\right)$);

(2) Update $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} + \alpha^{[k]}\mathbf{d}^{[k]}$ where $\alpha^{[k]} > 0$;

(3) Set $y^{[k]} = \nabla f\left(\mathbf{x}^{[k+1]}\right) - \nabla f\left(\mathbf{x}^{[k]}\right)$, $s^{[k]} = \mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}$ and compute $B^{[k+1]}$ (or $H^{[k+1]}$).

Proposition 12.47. Let $H^{[k]} \succ 0$ and $\mathbf{y}^{[k]^T}\mathbf{s}^{[k]} > 0$ and $H^{[k+1]}$ be given by BFGS update, then $H^{[k+1]} \succ 0$.

The same conclusion holds if $H^{[k]}$ and $H^{[k+1]}$ are replaced by $B^{[k]}$ and $B^{[k+1]}$, respectively.

Method 12.48. (Quasi-Newton method with Wolfe line search) Given $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$,

Initialize: $0 < c_1 < c_2 < 1$, $x^{[0]} \in \mathbb{R}^n$, and $H^{[0]} = \eta I$ for some $\eta > 0$,

For $k \in \mathbb{N}$,

(1) Find $d^{[k]}$ via $d^{[k]} = -H^{[k]}\nabla f\left(x^{[k]}\right)$;

(2) Compute $\alpha^{[k]}$ that satisfies the Wolfe's condition;

(3) Update $x^{[k+1]} = x^{[k]} + \alpha^{[k]}d^{[k]}$;

(4) Set $y^{[k]} = \nabla f\left(x^{[k+1]}\right) - \nabla f\left(x^{[k]}\right)$, $s^{[k]} = x^{[k+1]} - x^{[k]}$ and compute $H^{[k+1]}$ as in BFGS.

Theorem 12.49. (Zoutendijk's theorem) For $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$, $x^{[0]} \in \mathbb{R}^n$ and exists $l > 0$ such that for all x, y with $\max\{f(x), f(y)\} \le f\left(x^{[0]}\right)$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \le l\|x - y\|_2.$$

Then for a sequence $\left\{x^{[k]}\right\}$ with non-stationary points generated as

$$x^{[k+1]} = x^{[k]} + \alpha^{[k]}d^{[k]},$$

with $d^{[k]}$ a descent direction and $\alpha^{[k]}$ satisfying the Wolfe's condition, then it holds that

$$\sum_{k=0}^{\infty} \cos^2\left(\theta^{[k]}\right)\left\|\nabla f\left(x^{[k]}\right)\right\|_2^2 < \infty,$$

where

$$\cos\left(\theta^{[k]}\right) = \frac{-\left(\nabla f\left(x^{[k]}\right)\right)^T d^{[k]}}{\left\|\nabla f(x^{[k]})\right\|_2 \left\|d^{[k]}\right\|_2}.$$

Corollary 12.50. If there exists $k > 0$ such that $\cos\left(\theta^{[k]}\right) \geq \delta$ for all k, then $\lim_{k\to\infty} \left\|\nabla f\left(x^{[k]}\right)\right\|_2 = 0$. Hence, any accumulation point of $\left\{x^{[k]}\right\}$ is stationary.

For BFGS, if there exists $M > 0$ such that for all $k \in \mathbb{N}$ $\left\|H^{[k]}\right\|_2 \left\|\left(H^{[k]}\right)^{-1}\right\|_2 < M$, then $\lim_{k\to\infty} \left\|\nabla f\left(x^{[k]}\right)\right\|_2 = 0$.

Theorem 12.51. The Quasi-Newton method is a **superlinear convergent** method.

12.4 Linear Programming

Theorem 12.52. (Strong duality for LP, version I) Let $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^n$. Consider

$$v_p = \sup_{\mathbf{x}\in\mathbb{R}^n} \left\{\mathbf{c}^T\mathbf{x} : A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0\right\}, v_d = \inf_{\mathbf{y}\in\mathbb{R}^m} \left\{\mathbf{b}^T\mathbf{y} : \mathbf{c} \leq A^T\mathbf{y}\right\}.$$

Suppose that there exists $\hat{\mathbf{x}} \geq 0$ with $A\hat{\mathbf{x}} = \mathbf{b}$. Then $v_p = v_d$.

Theorem 12.53. (Strong duality for LP, version I) Let $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^n$. Consider

$$v_p = \sup_{\mathbf{x}\in\mathbb{R}^n} \left\{\mathbf{c}^T\mathbf{x} : A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0\right\}, v_d = \inf_{\mathbf{y}\in\mathbb{R}^m} \left\{\mathbf{b}^T\mathbf{y} : \mathbf{c} \leq A^T\mathbf{y}\right\}.$$

Suppose that either

- there exists $\hat{\mathbf{x}} \geq 0$ with $A\hat{\mathbf{x}} = b$; or
- there exists $\hat{\mathbf{y}}$ with $\mathbf{c} \leq A^T\hat{\mathbf{y}}$.

Then $v_p = v_d$ and both optimal values are attained when finite.

Remark 12.54. Recipe for writing dual problems:

	$\max \mathbf{c}^T\mathbf{x}$ s.t. $A\mathbf{x}\clubsuit\mathbf{b}$ $\mathbf{x}\diamondsuit$	$\min \mathbf{b}^T\mathbf{y}$ s.t. $A^T\mathbf{y}\diamondsuit c$ $\mathbf{y}\clubsuit$
$\clubsuit$	i-th constraint $\leq$ i-th constraint $\geq$ i-th constraint $=$	i-th variable ≥ 0 i-th variable ≤ 0 i-th variable unrestricted
$\diamondsuit$	j-th variable ≥ 0 j-th variable ≤ 0 j-th variable unrestricted	j-th constraint $\geq$ j-th constraint $\leq$ j-th constraint $=$

12.5 Semidefinite Programming

Definition 12.55. The **primal-dual SDP pairs** is defined as:

Primal	$\min_{X \in S^n} \operatorname{tr}(CX)$, s.t. $\operatorname{tr}(A_i X) = b_i, i = 1, ..., m$, $X \succeq 0$
Dual	$\max_{\mathbf{y} \in \mathbb{R}^n} \mathbf{b}^T\mathbf{y}$, s.t. $C - \sum_{i=1}^{m} \mathbf{y}_i A_i \succeq 0$

where $A_i, C \in S^n$ for all i. Let v_p and v_d denote their optimal values.

Theorem 12.56. (Strong duality for SDPs) Consider the primal-dual SDP pairs, then the following statements holds:

- If there exists $\overline{X} \succ 0$ such that $\operatorname{tr}\left(A_i \overline{X}\right) = \mathbf{b}_i$ for all i, then $v_p = v_d$ and v_d is attained while finite.
- If there exists $\overline{\mathbf{y}} \in \mathbb{R}^m$ such that $C - \sum_{i=1}^{m} \overline{\mathbf{y}}_i A_i \succeq 0$, then $v_p = v_d$ and v_p is attained while finite.

Remark 12.57. It always holds that $v_p \geq v_d$, indeed, for any primal feasible X and dual feasible $\mathbf{y}$, we have

$$\mathbf{b}^T y = \sum_{i=1}^{m} \mathbf{b}_i \mathbf{y}_i = \sum_{i=1}^{m} \operatorname{tr}(A_i X)\mathbf{y}_i = \operatorname{tr}\left(\sum_{i=1}^{m} \mathbf{y}_i A_i X\right) = \operatorname{tr}\left(\left(\sum_{i=1}^{m} \mathbf{y}_i A_i - C\right) X\right) + \operatorname{tr}(CX).$$

Theorem 12.58. Let $A, C \in S_+^n$, then $\operatorname{tr}(AC) \geq 0$.

Proposition 12.59. Consider the primal-dual SDP pairs and the set

$$\hat{\Upsilon} = \left\{[\operatorname{tr}(CX), \operatorname{tr}(A_1 X), ..., \operatorname{tr}(A_m X)]^T \in \mathbb{R}^{m+1} : X \succ 0\right\}$$

on the previous slide. Suppose that there exists $\overline{y} \in \mathbb{R}^m$ such that $C - \sum_{i=1}^{m} \overline{\mathbf{y}}_i A_i \succ 0$. Then $\hat{\gamma}$ is closed.

Theorem 12.60. (Schur complement) Let $A \in S^m$, $C \in S^n$, $B \in \mathbb{R}^{m \times n}$ and $A \succ 0$, then

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0 \Leftrightarrow C - B^T A^{-1} B \succeq 0.$$

We call $C - B^T A^{-1} B$ the Schur complement of A in $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$.

12.6 Penalty/Barrier Methods

Definition 12.61. (Penalty functions) A function $P : \mathbb{R}^n \to \mathbb{R}$ is a penalty function for the constraint set $\{x : \forall i \in I, g_i(\mathbf{x}) \leq 0\}$ if

- $\forall \mathbf{x} \in \mathbb{R}^n, P(\mathbf{x}) \geq 0$;
- $P(\mathbf{x}) = 0$ iff $\forall i \in I, g_i(\mathbf{x}) \leq 0$.

Method 12.62. (Penalty method: basic version) Let $c > 0$ and $\eta > 1$.

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$, $c_1 = c$,

For $k \in \mathbb{N}$,

(1) Find a minimizer $\mathbf{x}^{[k]}$ of $q_{c_k}(\mathbf{x}) = f(\mathbf{x}) + \frac{c_k}{2} \sum_{i=1}^{m} (\max(g_i(\mathbf{x}), 0))^2$, using $\mathbf{x}^{[k-1]}$ as the initial point for the iterative method;
(2) Update $c_{k+1} = \eta c_k$.

Theorem 12.63. Consider

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} && f(\mathbf{x}) \\ &\text{s.t.} && g_i(x) \leq 0, i \in I = \{1, \cdots, m\}, \\ &\text{where } && f, g_i \in C^1, \{x : \forall i \in I, g_i(\mathbf{x}) \leq 0\} \neq \emptyset. \end{aligned}$$

and suppose that $\inf f > 1$. Let $\{\mathbf{x}^{[k]}\}$ be generated by the basic version penalty method. Then any accumulation point x^* of $\{x^{[k]}\}$ is a globally optimal solution.

Method 12.64. (Penalty method: practical version) Let $c > 0$ and $\eta > 1$.

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$, $c_1 = c$,

For $k \in \mathbb{N}$,

(1) Find an $\mathbf{x}^{[k]}$ such that $\nabla q_{c_k}(\mathbf{x}^{[k]}) \approx 0$, using $\mathbf{x}^{[k-1]}$ as the initial point for the iterative method;

(2) Update $c_{k+1} = \eta c_k$.

Method 12.65. (Barrier method: basic version) Let $\mu > 0$ and $\eta > 1$.

Initialize: $\mathbf{x}^{[0]} \in \mathbb{S}^0$, $\mu_1 = \mu$,

For $k \in \mathbb{N}$,

(1) Find a minimizer $\mathbf{x}^{[x]}$ of $f_{\mu_k}(x) = f(x) - \mu_k \sum_{i=1}^{m} \ln(-g_i(\mathbf{x}))$, using $\mathbf{x}^{[k-1]}$ as the initial point for the iterative method;

(2) Update $\mu_{k+1} = \frac{\mu_k}{\eta}$.

12.7 Conjugate Gradient Method

Method 12.66. (Conjugate gradient method: Conceptual version)

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$, $\mathbf{d}^{[0]} = -\nabla f(\mathbf{x}^{[0]}) = \mathbf{b} - A\mathbf{x}^{[0]}$,

For $k \in \mathbb{N}$,

(1) If $d^{[k]} = 0$, terminate;

(2) Pick α_k so that $\alpha_k \in \arg\max_{\alpha \geq 0}\{f(\mathbf{x}^{[k]} + \alpha\mathbf{d}^{[k]})\}$;

(3) Set $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} + \alpha_k \mathbf{d}^{[k]}$ and $d^{[k+1]} = -\nabla f(\mathbf{x}^{[k+1]}) - \sum_{i=0}^{k} \frac{-\nabla f(\mathbf{x}^{[k+1]})^T A\mathbf{d}^{[j]}}{\mathbf{d}^{[j]T} A\mathbf{d}^{[j]}} \mathbf{d}^{[j]}$.

Theorem 12.67. Let $A \succ 0$ and $\mathbf{x}^{[0]} \in \mathbb{R}^n$. Set $d^{[0]} = -\nabla f(\mathbf{x}^{[0]})$. For $k \in \mathbb{N}$, suppose that $\mathbf{d}^{[0]}, ..., \mathbf{d}^{[k]} \neq 0$, where for each $i = 0, ..., k-1$,

$$\mathbf{d}^{[i+1]} = -\nabla f\left(\mathbf{x}^{[i+1]}\right) - \sum_{j=0}^{i} \frac{-\nabla f\left(\mathbf{x}^{[i+1]}\right)^T A\mathbf{d}^{[j]}}{\mathbf{d}^{[j]T} A\mathbf{d}^{[j]}}\mathbf{d}^{[j]},$$

with $\mathbf{x}^{[i+1]} = \mathbf{x}^{[i]} + \alpha_i \mathbf{d}^{[i]}$ and α_i coming from exact line search. Then for $j < k+1$, $\nabla f\left(\mathbf{x}^{[j]}\right)^T \nabla f\left(\mathbf{x}^{[k+1]}\right) = 0$ and $\mathbf{d}^{[j]T} \nabla f\left(\mathbf{x}^{[k+1]}\right) = 0$.

Theorem 12.68. For $k \in \mathbb{N}$, $\mathbf{x}^{[k]}, \mathbf{d}^{[k]}$ are generated by conjugate gradient method, then

$$\mathbf{d}^{[k+1]} = -\nabla f\left(\mathbf{x}^{[k+1]}\right) + \frac{\left\|\nabla f\left(\mathbf{x}^{[k+1]}\right)\right\|_2^2}{\left\|\nabla f\left(\mathbf{x}^{[k]}\right)\right\|_2^2}\mathbf{d}^{[k]}.$$

Method 12.69. (Conjugate gradient method: Formal version)

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$, $\mathbf{d}^{[0]} = -\nabla f\left(\mathbf{x}^{[0]}\right) = \mathbf{b} - A\mathbf{x}^{[0]}$,

For $k \in \mathbb{N}$,

(1) If $d^{[k]} = 0$, terminate;

(2) Pick α_k so that $\alpha_k \in \arg\max_{\alpha \geq 0}\left\{f\left(\mathbf{x}^{[k]} + \alpha\mathbf{d}^{[k]}\right)\right\}$;

(3) Set $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} + \alpha_k \mathbf{d}^{[k]}$ and $\mathbf{d}^{[k+1]} = -\nabla f\left(\mathbf{x}^{[k+1]}\right) + \frac{\left\|\nabla f\left(\mathbf{x}^{[k+1]}\right)\right\|_2^2}{\left\|\nabla f\left(\mathbf{x}^{[k]}\right)\right\|_2^2}\mathbf{d}^{[k]}$.

Method 12.70. (Conjugate gradient method: Actual version)

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$, $\mathbf{r}^{[0]} = \mathbf{d}^{[0]} = -\nabla f\left(\mathbf{x}^{[0]}\right) = \mathbf{b} - A\mathbf{x}^{[0]}$,

For $k \in \mathbb{N}$,

(1) If $\left\|r^{[k]}\right\|$ (or less commonly, $\left\|d^{[k]}\right\|$) is below a tolerance, terminate;

(2) Compute $\alpha_k = \frac{\mathbf{r}^{[k]T}\mathbf{r}^{[k]}}{\mathbf{d}^{[k]T}A\mathbf{d}^{[k]}}$, $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} + \alpha_k \mathbf{d}^{[k]}$, $\mathbf{r}^{[k+1]} = \mathbf{r}^{[k]} - \alpha_k A\mathbf{d}^{[k]}$;

(3) Compute $\beta_k = \frac{\mathbf{r}^{[k+1]T}\mathbf{r}^{[k+1]}}{\mathbf{r}^{[k]T}\mathbf{r}^{[k]}}$, $\mathbf{d}^{[k+1]} = \mathbf{r}^{[k+1]} + \beta_k \mathbf{d}^{[k]}$.

Theorem 12.71. (Luenberger) Consider the conjugate gradient method for minimizing $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}$ for some $\mathbf{b} \in \mathbb{R}^n$ and $A \succ 0$. Let $\left\{\mathbf{x}^{[k]}\right\}$ be the sequence generated and let $\mathbf{x}^*$ be the minimizer of f. If A has eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, then

$$\lambda_1\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^*\right\|_2^2 \leq \left(\mathbf{x}^{[k+1]} - \mathbf{x}^*\right)^T A\left(\mathbf{x}^{[k+1]} - \mathbf{x}^*\right) \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 \left(\mathbf{x}^{[0]} - \mathbf{x}^*\right)^T A\left(\mathbf{x}^{[0]} - \mathbf{x}^*\right)$$

Method 12.72. (Nonlinear conjugate gradient method: Conceptual version)

Initialize: $\mathbf{x}^{[0]} \in \mathbb{R}^n$, $\mathbf{d}^{[0]} = -\nabla f\left(\mathbf{x}^{[0]}\right)$,

For $k \in \mathbb{N}$,

(1) If $d^{[k]}$ is small, terminate;

(2) Pick α_k judiciously (e.g. exact line search, strong Wolfe conditions);

(3) Set $\mathbf{x}^{[k+1]} = \mathbf{x}^{[k]} + \alpha_k \mathbf{d}^{[k]}$ and $\mathbf{d}^{[k+1]} = -\nabla f\left(\mathbf{x}^{[k+1]}\right) + \frac{\left\|\nabla f\left(\mathbf{x}^{[k+1]}\right)\right\|_2^2}{\left\|\nabla f\left(\mathbf{x}^{[k]}\right)\right\|_2^2}\mathbf{d}^{[k]}$.

Chapter 13
Initial Value Problem

Notation 13.1. To numerically solve the IVP, we are given initial condition $\mathbf{u}_0 = \mathbf{u}(t_0)$, and want to compute approximations $\{\mathbf{u}_k, k = 1, 2, ...\}$ such that

$$\mathbf{u}_k \approx \mathbf{u}(t_k),$$

where k is the uniform time step size and $t_n = nk$.

13.1 Linear Multistep Method

Definition 13.2. For solving the IVP, an s-step **linear multistep method** (LMM) has the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{u}_{n+j} = k \sum_{j=0}^{s} \beta \mathbf{f}\left(\mathbf{u}_{n+j}, t_{n+j}\right),$$

where $\alpha_s = 1$ is assumed WLOG.

Definition 13.3. An LMM is **explicit** if $\beta_s = 0$, otherwise it is **implicit**.

13.2 Runge-Kutta Method

Definition 13.4. An s-stage **Runge-Kutta method** (RK) is a one-step method of the form

$$\begin{aligned} \mathbf{y}_i &= \mathbf{f}\left(\mathbf{u}_n + k\sum_{j=1}^{s} a_{ij}\mathbf{y}_j, t_n + c_i k\right), \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + k\sum_{j=1}^{s} b_j \mathbf{y}_j, \end{aligned}$$

where $i = 1, ..., s$ and $a_{ij}, b_j, c_i \in \mathbb{R}$.

Definition 13.5. The textsf{Butcher tableau} is one way to organize the coefficients of an RK method as follows

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

The matrix $A = \left(a_{ij}\right)_{s\times s}$ is called the RK matrix and $\mathbf{b} = (b_1, ..., b_s)^T, \mathbf{c} = (c_1, ..., c_s)^T$ are called the RK weights and the RK nodes.

Definition 13.6. An s-stage **collocation method** is a numerical method for solving the IVP, where we

(1) choose s distinct collocation parameters $c_1, ..., c_s$,

(2) seek s-degree polynomial p satisfying

$$\forall i = 1, 2, ..., s, \quad \mathbf{p}(t_n) = \mathbf{u}_n \;\text{ and }\; \mathbf{p}'(t_n + c_i k) = \mathbf{f}(\mathbf{p}(t_n + c_i k), t_n + c_i k),$$

(3) set $\mathbf{u}_{n+1} = \mathbf{p}(t_{n+1})$.

Theorem 13.7. The s-stage collocation method is an s-stage IRK method with

$$a_{ij} = \int_0^{c_i} l_j(\tau)\mathrm{d}\tau, \quad b_j = \int_0^1 l_j(\tau)\mathrm{d}\tau,$$

where $i, j = 1, ..., s$ and $l_k(\tau)$ is the elementary Lagrange interpolation polynomial.

13.3 Theoretical analysis

Definition 13.8. A function $\mathbf{f} : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$ is **Lipschitz continuous** in its first variable over some domain

$$\Omega = \{(\mathbf{u}, t) : \|\mathbf{u} - \mathbf{u}_0\| \le a, t \in [0, T]\}$$

iff

$$\exists L \ge 0, \text{ s.t. } \forall (\mathbf{u}, t) \in \Omega, \quad \|\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t) \le \|\mathbf{u} - \mathbf{v}\|.$$

13.3.1 Error analysis

Definition 13.9. The **local truncation error** τ is the error caused by replacing continuous derivatives with numerical formulas.

Definition 13.10. A numerical formulas is **consistent** if $\lim_{k \to 0} \tau = 0$.

13.3.2 Stability

Definition 13.11. The **region of absolute stability** (RAS) of a numerical method, applied to

$$\mathbf{u}' = \lambda \mathbf{u}, \quad \mathbf{u}_0 = \mathbf{u}(t_0),$$

is the region Ω that

$$\forall \mathbf{u}_0, \quad \forall \lambda k \in \Omega, \quad \lim_{n \to +\infty} \mathbf{u}_n = 0.$$

Definition 13.12. The **stability function** of a one-step method is a function $R : \mathbb{C} \to \mathbb{C}$ that satisfies

$$\mathbf{u}_{n+1} = R(z)\mathbf{u}_n$$

for the $\mathbf{u}' = \lambda \mathbf{u}$ where Re $(E(\lambda)) \le 0$ and $z = k\lambda$.

Definition 13.13. A numerical method is **stable** or **zero stable** iff its application to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\forall T > 0, \quad \lim_{k \to 0, Nk = t} \|\mathbf{u}_n\| < \infty.$$

Definition 13.14. A numerical method is **A($\boldsymbol{\alpha}$)-statble** if the region of absolute stability Ω satisfies

$$\{z \in \mathbb{C} : \pi - \alpha \leq \arg(z) \leq \pi + \alpha\} \subseteq \Omega.$$

Definition 13.15. A numerical method is **A-statble** if the region of absolute stability Ω satisfies

$$\{z \in \mathbb{C} : \text{Re } (z) \leq 0\} \subseteq \Omega.$$

Definition 13.16. A one-step method is **L-stable** if it is A-stable, and its stability function satisfies

$$\lim_{z \to \infty} |R(z)| = 0.$$

Definition 13.17. An one-step method is **I-stable** iff its stability function satisfies

$$\forall y \in \mathbb{R}, |R(y\mathbf{i})| \leq 1.$$

Definition 13.18. An one-step method is **B-stable** (or **contractive**) if for any contractive ODE system, every pair of its numerical solutions $\mathbf{u}_n$ and $\mathbf{v}_n$ satisfy

$$\forall n \in \mathbb{N}, \|u_{n+1} - v_{n+1}\| \leq \|u_n - v_n\|.$$

Definition 13.19. An RK method is **algebraically stable** iff the RK weights $b_1, ..., b_s$ are nonnegative, the **algebraic stability matrix** $M = \left(b_i a_{ij} + b_i a_{ji} - b_i b_j\right)_{s \times s}$ is positive semidefinite.

Theorem 13.20. The order of accuracy of an implicit A-stable LMM satisfies $p \leq 2$. An explicit LMM cannot be A-stable.

Theorem 13.21. No ERK method is A-stable.

Theorem 13.22. An RK method is A-stable if and only if it is I-stable and all poles of its stability function $R(z)$ have positive real parts.

Theorem 13.23. If an A-stable RK method with a nonsingular RK matrix A is stiffly accurate, then it is L-stable.

Theorem 13.24. If an A-stable RK method with a nonsingular RK matrix A satisfies

$$\forall i \in \{1, ..., s\}, \quad a_{i1} = b_i,$$

then it is L-stable.

Theorem 13.25. B-stable one-step methods are A-stable.

Theorem 13.26. An algebraically stable RK method is B-stable and A-stable.

13.3.3 Convergence

Definition 13.27. A numerical method is convergent iff its application to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\forall T > 0, \quad \lim_{k \to 0, nk = T} \mathbf{u}_n = \mathbf{u}(T).$$

Theorem 13.28. A numerical method is convergent iff it is consistent and stable.

13.4 Important Methods

13.4.1 Forward Euler's method

Definition 13.29. The **forward Euler's method** solves the IVP by

$$\mathbf{u}_{n+1} = \mathbf{u}_n + k\mathbf{f}(\mathbf{u}_n, t_n).$$

Theorem 13.30. The region of absolute stability for forward Euler's method is

$$\{z \in \mathbb{C} : |1 + z| \leq 1\}.$$

13.4.2 Backward Euler's method

Definition 13.31. The **backward Euler's method** solves the IVP by

$$\mathbf{u}_{n+1} = \mathbf{u}_n + k\mathbf{f}(\mathbf{u}_{n+1}, t_{n+1}).$$

Theorem 13.32. The region of absolute stability for backward Euler's method is

$$\{z \in \mathbb{C} : |1 - z| \geq 1\}.$$

13.4.3 Trapezoidal method

Definition 13.33. The **trapezoidal method** solves the IVP by

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \frac{k}{2}(\mathbf{f}(\mathbf{u}_n, t_n) + \mathbf{f}(\mathbf{u}_{n+1}, t_{n+1})).$$

Theorem 13.34. The region of absolute stability for trapezoidal method is

$$\left\{ z \in \mathbb{C} : \left| \frac{2+z}{2-z} \right| \geq 1 \right\}.$$

13.4.4 Midpoint method (Leapfrog method)

Definition 13.35. The **midpoint method (Leapfrog method)** solves the IVP by

$$\mathbf{u}_{n+1} = \mathbf{u}_{n-1} + 2k\mathbf{f}(\mathbf{u}_n, t_n).$$

Theorem 13.36. The region of absolute stability for midpoint method is

$$\left\{ z \in \mathbb{C} : \left| z \pm \sqrt{1 + z^2} \right| \leq 1 \right\} \stackrel{?}{=} \{0\}.$$

13.4.5 Heun's third-order RK method

Definition 13.37. The **Heun's third-order formula** is an ERK method of the form

$$\begin{cases} \mathbf{y}_1 &= \mathbf{f}(\mathbf{u}_n, t_n), \\ \mathbf{y}_2 &= \mathbf{f}\big(\mathbf{u}_n + \frac{k}{3}\mathbf{y}_1, t_n + \frac{k}{3}\big), \\ \mathbf{y}_3 &= \mathbf{f}\big(\mathbf{u}_n + \frac{2k}{3}\mathbf{y}_2, t_n + \frac{2k}{3}\big), \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + \frac{k}{4}(\mathbf{y}_1 + 3\mathbf{y}_3). \end{cases} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

13.4.6 Classical fourth-order RK method

Definition 13.38. The **classical fourth-order RK method** is an ERK method of the form

$$\begin{cases} \mathbf{y}_1 &= \mathbf{f}(\mathbf{u}_n, t_n), \\ \mathbf{y}_2 &= \mathbf{f}\big(\mathbf{u}_n + \frac{k}{2}\mathbf{y}_1, t_n + \frac{k}{2}\big), \\ \mathbf{y}_3 &= \mathbf{f}\big(\mathbf{u}_n + \frac{k}{2}\mathbf{y}_2, t_n + \frac{k}{2}\big), \\ \mathbf{y}_4 &= \mathbf{f}(\mathbf{u}_n + k\mathbf{y}_3, t_n + k), \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + \frac{k}{6}(\mathbf{y}_1 + 2\mathbf{y}_2 + 2\mathbf{y}_3 + \mathbf{y}_4). \end{cases} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

13.4.7 Third-order strong-stability preserving RK method

Definition 13.39. The **third-order strong-stability preserving RK method** is an ERK method of the form

$$\begin{cases} \mathbf{y}_1 &= \mathbf{u}_n + k\mathbf{f}(\mathbf{u}_n, t_n), \\ \mathbf{y}_2 &= \frac{3}{4}\mathbf{u}_n + \frac{1}{4}\mathbf{y}_1 + \frac{1}{4}k\mathbf{f}(\mathbf{y}_1, t_n + k), \\ \mathbf{u}_{n+1} &= \frac{1}{3}\mathbf{u}_n + \frac{2}{3}\mathbf{y}_2 + \frac{2}{3}k\mathbf{f}\big(\mathbf{y}_2, t_n + \frac{k}{2}\big). \end{cases}$$

which can also be written as

$$\begin{cases} \mathbf{y}_1 &= \mathbf{f}(\mathbf{u}_n, t_n), \\ \mathbf{y}_2 &= \mathbf{f}(\mathbf{u}_n + k\mathbf{y}_1, t_n + k), \\ \mathbf{y}_3 &= \mathbf{f}\big(\mathbf{u}_n + \frac{1}{4}k\mathbf{y}_1 + \frac{1}{4}k\mathbf{y}_2, t_n + \frac{1}{2}\big), \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + \frac{k}{6}(\mathbf{y}_1 + \mathbf{y}_2 + 4\mathbf{y}_3). \end{cases} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}$$

13.4.8 TR-BDF2 method

Definition 13.40. The **TR-BDF2 method** is an one-step method of the form

$$\begin{cases} \mathbf{u}_* &= \mathbf{u}_n + \frac{k}{4}\big(\mathbf{f}(\mathbf{u}_n, t_n) + \mathbf{f}\big(\mathbf{u}_*, t_n + \frac{k}{2}\big)\big), \\ \mathbf{u}_{n+1} &= \frac{1}{3}\big(4\mathbf{u}_* - \mathbf{u}_n + k\mathbf{f}(\mathbf{u}_{n+1}, t_{n+1})\big). \end{cases}$$

Chapter 14

Finite Element Method

Definition 14.1. A function $a(\cdot,\cdot)$ is a **bilinear function** if for all $u, v \in V$, $k_1, k_2 \in F$,

- $a(k_1 u + k_2 v, w) = k_1 a(u, w) + k_2 a(v, w)$,
- $a(u, k_1 v + k_2 w) = k_1 a(u, v) + k_2 a(u, w)$.

A bilinear function is **bounded** or **continuous** if $\|\cdot\|$ is the norm on V, and for all $u, v \in V$, exists $M > 0$, such that

$$|a(u, v)| \leq M\|u\|\|v\|.$$

A bilinear function is **symmetric** if for all $u, v \in V$, $a(u, v) = a(v, u)$.

A bilinear function is ***V*-elliptic** if exists $\alpha > 0$, for all $v \in V$,

$$\alpha\|v\|^2 \leq a(v, v).$$

Definition 14.2. Given a normed linear space V with a bounded bilinear function $a(\cdot,\cdot)$ on it and $f \in V^*$, then for $U \subset V$,

$$J(u) = \inf_{v \in U} J(v), J(v) = \frac{1}{2}a(v, v) - f(v).$$

Theorem 14.3. The solution to problem 14.2. exists and unique if

- V is complete,
- U is a closed convex subset of V,
- $a(\cdot,\cdot)$ is symmetric and V-elliptic.

Theorem 14.5. If u is the solution to problem 14.2., if and only if

$$\forall v \in U, a(u, v - u) \geq f(v - u).$$

where $a(u, u) = f(u)$ if U is a convex cone with the apex $\mathbf{0}$, $\forall v \in U, a(u, v) = f(v)$ if U is a closed subset of V.

Theorem 14.7. (Lax-Milgram lemma) Given a Hilbert space V, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a continuous V-elliptic bilinear function, $f : V \to \mathbb{R}$ a continuous linear functional, then there exists only $u \in V$, such that

$$\forall v \in V, a(u, v) = f(v).$$

Lemma 14.8. Given $u \in H_0^2(\Omega)$, $|\cdot|$ is the Sobolev seminorm, $\|\cdot\|$ is the Sobolev norm, then

$$\|\Delta u\|_{0,\Omega}^2 = |u|_{2,\Omega}^2.$$

Theorem 14.9. (Poincare-Friedrichs) Given a bounded set Ω, $v \in H_0^m(\Omega)$, then there exists a constant $C(\Omega)$, such that

$$\|v\|_{0,\Omega} \leq C(\Omega)|v|_{m,\Omega}.$$

Chapter 15

Number Theory

15.1 Prime Number

Definition 15.1. A **prime number** (or a **prime**) is a natural number greater than 1 that is not a product of two smaller natural numbers.

Definition 15.2. A **composite number** (or a **composite**) is a natural number greater than 1 that is a product of two smaller natural numbers.

15.1.1 Primality testing

Theorem 15.3. For a integer $n \in \mathbb{N}$, if it is a product of two natural number a and b thar $a \leq b$, then

$$1 \leq a \leq \sqrt{n} \leq b \leq n.$$

Method 15.4. (Trial division) Given a integer n, the **trial division method** divides n by each integer from 2 up to $\sqrt{n}$. Any such integer dividing n evenly establishes n as composite, otherwise it is prime.

Theorem 15.5. (Fermat's little theorem) For a prime number p and a number a that $\gcd(a, p) = 1$, then $a^{p-1} \equiv 1 (\text{mod } p)$

Method 15.6. The **Miller-Rabin** algorithm is a method of primality testing, where given a number n, where we

(1) determine directly for small numbers such as $p = 2$.
(2) factorize the number $p = u \times 2^t$;
(3) choose a number a that $\gcd(a, p) = 1$, and calculate $a^u, a^{u\times 2}, a^{u\times 2^2}, ..., a^{u\times 2^{t-1}}$;
(4) if $a^u \equiv 1(\text{mod } p)$, or $\exists a^{u\times k}, k < t$ that $a^{u\times k} \equiv p - 1(\text{mod } p)$ then p passes the test, otherwise, p is a composite number;
(5) repeat above steps to eliminate error.

For numbers less than 2^{32}, choose $a \in \{2, 7, 61\}$ is enough, for numbers less than $2^{\{64\}}$, choose $a \in \{2, 325, 9375, 28178, 450775, 9780504, 1795265022\}$ is enough.

15.1.2 Sieves

Method 15.7. (Sieve of Eratosthenes) Given a upper limit n, the **sieve of Eratosthenes** solves all the prime numbers up to n by marking composite numbers, where we

(1) create a list of consecutive integers from 2 to n: $\{2, 3, 4, ..., n\}$;
(2) initially, let $p = 2$, the smallest prime number;
(3) enumerate the multiples of p by counting in increments of p from $2p$ to n, and mark them in the list;
(4) find the smallest number in the list greater than p that is not marked;

(5) if there was no such number, the method is terminated and the numbers remaining not marked in the list are all the primes below n, otherwise let p now equal the new number which is the next prime, and repeat from step (3).

Part 3

Machine Learning

Chapter 16

Regression

16.1 Linear Regression

Definition 16.1. Given a data set $\{(\mathbf{x}_i, y_i), i \in \{1, ..., m\}\}$ where $\mathbf{x}_i \in \mathbb{R}^n$, the linear regression seeks $\tilde{\mathbf{w}} \in \mathbb{R}^n$ and $\tilde{b} \in \mathbb{R}$ such that

$$f(\mathbf{x}_i) = \tilde{\mathbf{w}}^T \mathbf{x}_i + \tilde{b} \approx y_i.$$

In general, we choose mean square error to estimate the error between $f(\mathbf{x}_i)$ and y_i, which implies

$$(\tilde{\mathbf{w}}, \tilde{b}) = \underset{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}}{\arg\min} \sum_{i=1}^{m} \left(f(\mathbf{x}_i) - y_i\right)^2 = \underset{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}}{\arg\min} \sum_{i=1}^{m} \left(\mathbf{w}^T x + b - y_i\right)^2.$$

Theorem 16.2. Given a data set $\{(\mathbf{x}_i, y_i), i \in \{1, ..., m\}\}$ where $\mathbf{x}_i \in \mathbb{R}^n$, let

$$X = \begin{pmatrix} \mathbf{x}_1^T & 1 \\ \vdots & 1 \\ \mathbf{x}_m^T & 1 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix},$$

if $X^T X$ is invertible, the solution of linear regression can be written as

$$\begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} = \left(X^T X\right)^{-1} X^T \mathbf{y}.$$

Chapter 17

Decision Tree

Chapter 18
Support Vector Machine

Chapter 19

Cluster

19.1 K-means

Definition 19.1. Given points $\mathbf{x}_1, ..., \mathbf{x}_m \in \mathbb{R}^n$, **k-means clustering** aims to partition the points into $k \leq n$ sets $S = \{S_1, ..., S_k\}$ satisfies

$$S = \arg\min_S \left\{ \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \right\},$$

where $\boldsymbol{\mu}_i$ is the mean (centroid) of points in S_i, i.e. denoted by $|S_i|$ the size of S_i,

$$\boldsymbol{\mu}_i = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x}.$$

Theorem 19.2. Denoted by $\mathbf{x}_1, ..., \mathbf{x}_m \in \mathbb{R}^n$ the points and $S = \{S_1, ..., S_k\}$ sets given by K-means,

$$S = \arg\min_S \left\{ \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{\mathbf{x}, \mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2 \right\}.$$

Method 19.3. (K-means clustering) Denoted by $S^{(t)} = \left\{ S_1^{(t)}, ..., S_k^{(t)} \right\}$ the sets given by k-means at t-th step and $\boldsymbol{\mu}_i^{(t)}$ the mean of $S_i^{(t)}$, the algorithm proceeds by

(1) **Assignment**: Assign each point to the cluster with the nearest mean,

$$S_i^{(t)} = \left\{ \mathbf{x}_p : \forall j \in \{1, ..., k\}, \|\mathbf{x}_p - \boldsymbol{\mu}_i^{(t)}\|^2 \leq \|\mathbf{x}_p - \boldsymbol{\mu}_j^{(t)}\|^2 \right\};$$

(2) **Update**: Recalculate means (centroids) of each cluster,

$$\boldsymbol{\mu}_i^{(t)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x} \in S_i^{(t)}} \mathbf{x}.$$

Chapter 20

Neural Networks