Notes on Numerical Analysis
数值分析

Qinghai Zhang
张庆海

Spring 2023
2023年春季学期

Contents

Preface

This book comes out of my teaching of the course "Numerical Analysis" (formerly "Numerical Approximation") in the spring semesters of 2018, 2019, and 2020 and in the fall semesters of 2016 and 2021 at the school of mathematical sciences in Zhejiang University.

In writing this book, I have made special efforts to

- collect the prerequisites in the appendices so that students can quickly brush up on the preliminaries,
- emphasize the connection between numerical analysis and other branches of mathematics such as elementary analysis and linear algebra,
- arrange the contents carefully with the hope that the total entropy of this book is minimized,
- encourage the students to understand the motivations of definitions, to formally verify all major theorems on her/his own, to think actively about the contents, to relate mathematical theory to realworld physics, and to form a habit to tell logical and coherent stories.

In the whole progress of my teaching, many students asked for clarifications, pointed out typos, reported errors, raised questions, and suggested improvements. Each and every comment contributed to a better writing and/or teaching, be it small or big, negative or positive, subjective or objective.

关于数学学习的几点建议

A. 深入理解每一个知识点：证明或推导的每一步从哪里来的？争取做到“无一处无出处”，这有助于培养缜密的逻辑思维能力。

B. 寻找新内容和已知内容或其他数学分支之间的联系。我们学习数值分析已经用到的其它分支包括分析基础和线性代数等。学习的本质是把新内容和已经牢固掌握的知识联系起来！

C. 深入思考每一个知识点：一个定义捕捉到了什么？一个定理是否可以弱化条件？如果不能的话这些条件在证明中是在哪里出现的？作用是什么？一个定理的结论是否可以再加强？如果不能原因是什么？一个数学方法的适用范围是什么？局限性在哪里？

D. 精准识记核心的定义定理，再以一定的逻辑关系把相关知识点串成一个故事，这些关系可以包括继承、组合、蕴含、特例等；构建这样一个脉络的目的是使自己知识体系的熵（混乱度）最小。

E. 在完成知识体系构建的基础上尽可能地多做习题，但是构建知识体系永远比做题本身重要。

F. 将新知识以一种和已有知识相容的方式纳入自己的知识体系。学数学的过程是盖一座大楼不是在一个平面上搭很多帐篷；一座大楼的高度取决于基础以及每一层的坚固度。

G. 任何一门数学都包括内容和形式，两者相互依赖，互为补充。

H. “骐骥一跃，不能十步；驽马十驾，功在不舍。锲而舍之，朽木不折；锲而不舍，金石可镂。”
One baby step at a time!
Do the simpliest thing that could possibly work, then keep asking more and refining your answers.

I. “一阴一阳之谓道，继之者善也，成之者性也。仁者见之谓之仁，知者见之谓之知，百姓日用不知，故君子之道鲜矣！” —— 《易经系辞上》

J. “Think globally, act locally.”

K. “重剑无锋，大巧不工”—— 《神雕侠侣》

Chapter 1

Solving Nonlinear Equations

1.1 The bisection method

Algorithm 1.1. For the root-finding problem of a continuous function $f : \mathbb{R} \to \mathbb{R}$, the *bisection method* repeatedly reduce by half the given interval where the root must lie and returns the midpoint of the last interval as the solution.

Input: $f : [a, b] \to \mathbb{R},\ a \in \mathbb{R},\ b \in \mathbb{R},$
$M \in \mathbb{N},\ \delta \in \mathbb{R}^+,\ \epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}[a, b],$
$\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$
Output: c, h, k
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

```
u ← f(a)
v ← f(b)
for k = 0 : M do
    h ← b − a
    c ← a + h/2
    if |h| < δ or k = M then break
    w ← f(c)
    if |w| < ϵ then
        break
    else if sgn(w) ≠ sgn(u) then
        b ← c
        v ← w
    else
        a ← c
        u ← w
    end
end
```

1.2 The signature of an algorithm

Definition 1.2. An *algorithm* is a step-by-step procedure that takes some set of values as its *input* and produces some set of values as its *output*.

Definition 1.3. A *precondition* is a condition that holds for the input prior to the execution of an algorithm.

Definition 1.4. A *postcondition* is a condition that holds for the output after the execution of an algorithm.

Definition 1.5. The *signature of an algorithm* consists of its input, output, preconditions, postconditions, and how input parameters violating preconditions are handled.

1.3 Proof of correctness and simplification of algorithms

Definition 1.6. An *invariant* is a condition that holds during the execution of an algorithm.

Definition 1.7. A variable is *temporary or derived* for a loop if it is initialized inside the loop. A variable is *persistent or primary* for a loop if it is initialized before the loop and its value changes across different iterations.

Exercise 1.8. What are the invariants in Algorithm 1.1? Which quantities do a, b, c, h, u, v, w represent? Which of them are primary? Which of these variables are temporary? Draw pictures to illustrate the life spans of these variables.

Algorithm 1.9. A simplified bisection algorithm.

Input: $f : [a, b] \to \mathbb{R},\ a \in \mathbb{R},\ b \in \mathbb{R},$
$M \in \mathbb{N},\ \delta \in \mathbb{R}^+,\ \epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}[a, b],$
$\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$
Output: c, h, k
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

```
h ← b − a
u ← f(a)
for k = 0 : M do
    h ← h/2
    c ← a + h
    if |h| < δ or k = M then break
    w ← f(c)
    if |w| < ϵ then
        break
    else if sgn(w) = sgn(u) then
        a ← c
    end
end
```

1.4 Q-order convergence

Definition 1.10 (Q-order convergence)**.** A convergent sequence $\{x_n\}$ is said to *converge* to L with *Q-order* p ($p \geq 1$) if

$$\lim_{n\to\infty} \frac{|x_{n+1} - L|}{|x_n - L|^p} = c > 0; \tag{1.1}$$

the constant c is called the *asymptotic factor*. In particular, $\{x_n\}$ has *Q-linear convergence* if $p = 1$ and *Q-quadratic convergence* if $p = 2$.

Definition 1.11. A sequence of iterates $\{x_n\}$ is said to *converge linearly* to L if

$$\exists c \in (0,1), \exists d > 0, \text{ s.t. } \forall n \in \mathbb{N}, \ |x_n - L| \leq c^n d. \tag{1.2}$$

For a sequence $\{x_n\}$ that convergs to L, its *order of convergence* is the maximum $p \in \mathbb{R}^+$ satisfying

$$\exists c > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, \ |x_{n+1} - L| \leq c|x_n - L|^p. \tag{1.3}$$

In particular, $\{x_n\}$ *converges quadratically* if $p = 2$.

Theorem 1.12 (Monotonic sequence theorem)**.** Every bounded monotonic sequence is convergent.

Theorem 1.13 (Convergence of the bisection method)**.** For a continuous function $f : [a_0, b_0] \to \mathbb{R}$ satisfying $\operatorname{sgn}(f(a_0)) \neq \operatorname{sgn}(f(b_0))$, the sequence of iterates in the bisection method converges linearly with asymptotic factor $\frac{1}{2}$,

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \alpha, \tag{1.4}$$

$$f(\alpha) = 0, \tag{1.5}$$

$$|c_n - \alpha| \leq 2^{-(n+1)}(b_0 - a_0), \tag{1.6}$$

where $[a_n, b_n]$ is the interval in the nth iteration and $c_n = \frac{1}{2}(a_n + b_n)$ is the solution returned by the bisection method.

Proof. It follows from the bisection method that

$$\begin{aligned} a_0 \leq a_1 \leq a_2 \leq \cdots \leq b_0, \\ b_0 \geq b_1 \geq b_2 \geq \cdots \geq a_0, \\ b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n). \end{aligned}$$

In the rest of this proof, "lim" is a shorthand for "$\lim_{n\to\infty}$." By Theorem 1.12, both $\{a_n\}$ and $\{b_n\}$ converge. Also, $\lim(b_n - a_n) = \lim \frac{1}{2^n}(b_0 - a_0) = 0$, hence $\lim b_n = \lim a_n = \alpha$. By the given condition and the algorithm, the invariant $f(a_n)f(b_n) \leq 0$ always holds. Since f is continuous, $\lim f(a_n)f(b_n) = f(\lim a_n)f(\lim b_n)$, then $f^2(\alpha) \leq 0$ implies $f(\alpha) = 0$. (1.6) is another important invariant that can be proven by induction. Comparing (1.6) to (1.2) yields convergence of the bisection method. Also, the convergence is linear with asymptotic factor as $c = \frac{1}{2}$. □

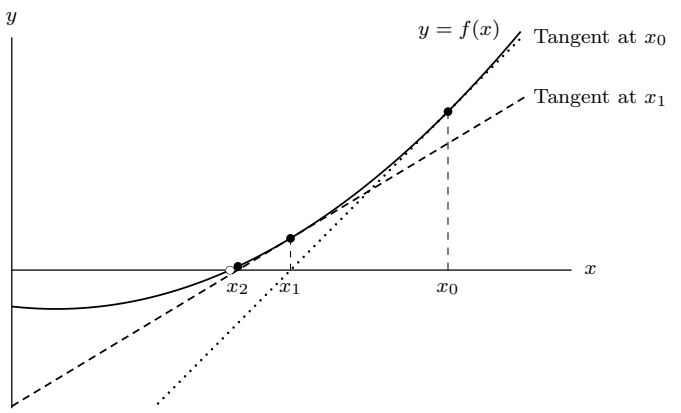
1.5 Newton's method

Algorithm 1.14. *Newton's method* finds the root of $f : \mathbb{R} \to \mathbb{R}$ near an initial guess x_0 by the iteration formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n \in \mathbb{N}. \tag{1.7}$$

Input: $f : \mathbb{R} \to \mathbb{R}$, f', $x_0 \in \mathbb{R}$, $M \in \mathbb{N}^+$, $\epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}^2$ and x_0 is sufficiently close to a root of f
Output: x, k
Postconditions: $|f(x)| < \epsilon$ or $k = M$

```
x ← x_0
for k = 0 : M do
    u ← f(x)
    if |u| < ε then break
    x ← x − u/f'(x)
end
```



Theorem 1.15 (Convergence of Newton's method)**.** Consider a $\mathcal{C}^2$ function $f : \mathcal{B} \to \mathbb{R}$ on $\mathcal{B} = [\alpha - \delta, \alpha + \delta]$ satisfying $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. If x_0 is chosen sufficiently close to α, then the sequence of iterates $\{x_n\}$ in the Newton's method converges at least quadratically to the root α, i.e.

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^2} = -\frac{f''(\alpha)}{2f'(\alpha)}. \tag{1.8}$$

Proof. By Taylor's theorem (Theorem C.61) and the assumption $f \in \mathcal{C}^2$, we have

$$f(\alpha) = f(x_n) + (\alpha - x_n)f'(x_n) + \frac{(\alpha - x_n)^2}{2} f''(\xi)$$

where ξ is between α and x_n. $f(\alpha) = 0$ yields

$$-\alpha = -x_n + \frac{f(x_n)}{f'(x_n)} + \frac{(\alpha - x_n)^2}{2}\frac{f''(\xi)}{f'(x_n)}.$$

By (1.7), we have

$$(*): \quad x_{n+1} - \alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \alpha = (x_n - \alpha)^2 \frac{f''(\xi)}{2f'(x_n)}.$$

The continuity of f' and the assumption $f'(\alpha) \neq 0$ yield

$$\exists \delta_1 \in (0, \delta) \text{ s.t. } \forall x \in \mathcal{B}_1, \ f'(x) \neq 0$$

where $\mathcal{B}_1 = [\alpha - \delta_1, \alpha + \delta_1]$. Define

$$M = \frac{\max_{x\in\mathcal{B}_1} |f''(x)|}{2\min_{x\in\mathcal{B}_1} |f'(x)|}$$

and pick x_0 sufficiently close to α such that

(i) $|x_0 - \alpha| = \delta_0 < \delta_1$;

(ii) $M\delta_0 < 1$.

The definition of M and (*) imply

$$|x_{n+1} - \alpha| \le M|x_n - \alpha|^2.$$

Comparing the above to (1.3) implies that if $\{x_n\}$ converges, then the order of convergence is 2. We must still show that (a) it converges and (b) it converges to α.

By (i) and (ii), we have $M|x_0 - \alpha| < 1$. Then it is easy to obtain the following via induction,

$$|x_n - \alpha| \le \frac{1}{M}\left(M|x_0 - \alpha|\right)^{2^n},$$

which shows both (a) and (b). Finally, the convergence rate could be higher than 2 in the case of $f''(\alpha) = 0$. □

Theorem 1.16. A continuous function $f : [a, b] \to [c, d]$ is bijective if and only if it is strictly monotonic.

Theorem 1.17. If a $\mathcal{C}^2$ function $f : \mathbb{R} \to \mathbb{R}$ satisfies $f(\alpha) = 0$, $f' > 0$ and $f'' > 0$, then α is the only root of f and, $\forall x_0 \in \mathbb{R}$, the sequence of iterates $\{x_n\}$ in the Newton's method converges quadratically to α.

Proof. By Theorem 1.16, f is a bijection since f is continuous and strictly monotonic. With 0 in its range, f must have a unique root. When proving Theorem 1.15, we had

$$x_{n+1} - \alpha = (x_n - \alpha)^2 \frac{f''(\xi)}{2f'(x_n)}. \tag{1.9}$$

Then $f' > 0$ and $f'' > 0$ further imply $x_n > \alpha$ for all $n > 0$. f being strictly increasing implies that $f(x_n) > f(\alpha) = 0$ for all $n > 0$. By the definition of Newton's method, $x_{n+1} - \alpha = x_n - \alpha - \frac{f(x_n)}{f'(x_n)}$, hence the sequence $\{x_n - \alpha : n > 0\}$ is strictly monotonically decreasing with 0 as a lower bound. By Theorem 1.12 it converges.

Suppose $\lim_{n\to\infty} x_n = a$. Then take the limit of $n \to \infty$ on both sides of (1.7) and we have

$$a = a - \frac{f(a)}{f'(a)},$$

which implies $f(a) = 0$. By the uniqueness of the root of f, we have $a = \alpha$.

The quadratic convergence rate can be proved by an induction using (1.9), as in Theorem 1.15. □

Definition 1.18. Let $\mathcal{V}$ be a vector space. A subset $\mathcal{U} \subseteq \mathcal{V}$ is a *convex set* iff

$$\forall x, y \in \mathcal{U}, \forall t \in (0, 1), \qquad tx + (1 - t)y \in \mathcal{U}. \tag{1.10}$$

A function $f : \mathcal{U} \to \mathbb{R}$ is *convex* iff

$$\begin{aligned} &\forall x, y \in \mathcal{U}, \forall t \in (0, 1), \\ &f(tx + (1 - t)y) \le tf(x) + (1 - t)f(y). \end{aligned} \tag{1.11}$$

In particular, f is *strictly convex* if we replace "$\le$" with "$<$" in the above equation.

1.6 The secant method

Algorithm 1.19. The *secant method* finds a root of $f : \mathbb{R} \to \mathbb{R}$ near initial guesses x_0, x_1 by the iteration

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \qquad n \in \mathbb{N}^+. \tag{1.12}$$

Input: $f : \mathbb{R} \to \mathbb{R},\ x_0 \in \mathbb{R},\ x_1 \in \mathbb{R},$
$M \in \mathbb{N}^+,\ \delta \in \mathbb{R}^+,\ \epsilon \in \mathbb{R}^+$
Preconditions : $f \in \mathcal{C}^2$; x_0, x_1 are sufficiently close to a root of f
Output: x_n, x_{n-1}, k
Postconditions: $|f(x_n)| < \epsilon$ or $|x_n - x_{n-1}| < \delta$ or $k = M$

```
x_n ← x_1
x_{n-1} ← x_0
u ← f(x_n)
v ← f(x_{n-1})
for k = 2 : M do
    s ← (x_n − x_{n-1}) / (u − v)
    x_{n-1} ← x_n
    v ← u
    x_n ← x_n − u × s
    if |x_n − x_{n-1}| < δ then break
    u ← f(x_n)
    if |u| < ε then break
end
```

Definition 1.20. The sequence $\{F_n\}$ of *Fibonacci numbers* is defined as

$$F_0 = 0,\ F_1 = 1, \qquad F_{n+1} = F_n + F_{n-1}. \tag{1.13}$$

Theorem 1.21 (Binet's formula)**.** Denote the golden ratio by $r_0 = \frac{1+\sqrt{5}}{2}$ and let $r_1 = 1 - r_0 = \frac{1-\sqrt{5}}{2}$, then

$$F_n = \frac{r_0^n - r_1^n}{\sqrt{5}}. \tag{1.14}$$

Proof. By Definition 1.20, we have

$$\mathbf{u}_k := \begin{bmatrix} F_{k+1} \\ F_k \end{bmatrix} = A \begin{bmatrix} F_k \\ F_{k-1} \end{bmatrix}, \text{ where } A := \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Hence we have $\mathbf{u}_n = A^n \mathbf{u}_0$. It follows from

$$\det(A - \lambda I) = \lambda^2 - \lambda - 1 = 0$$

that the two eigenvalues of A are $\lambda = r_0, r_1$, with their eigenvectors as $\mathbf{x}_0 = (r_0, 1)^T$ and $\mathbf{x}_1 = (r_1, 1)^T$, respectively. Indeed, the two eigenpairs stem from $\lambda^2 = \lambda + 1$, a nice relation between multiplication and addition by 1. Finally, we express $\mathbf{u}_0$ as a linear combination of $\mathbf{x}_0$ and $\mathbf{x}_1$,

$$\mathbf{u}_0 = \frac{1}{r_0 - r_1}(\mathbf{x}_0 - \mathbf{x}_1),$$

which, together with $\mathbf{u}_n = A^n \mathbf{u}_0$, yields

$$\mathbf{u}_n = \frac{1}{r_0 - r_1}(r_0^n \mathbf{x}_0 - r_1^n \mathbf{x}_1),$$

the second equation of which yields (1.14). □

Corollary 1.22. The ratios r_0, r_1 in Theorem 1.21 satisfy

$$F_{n+1} = r_0 F_n + r_1^n. \tag{1.15}$$

Proof. This follows from (1.14) and values of r_0 and r_1. □

Lemma 1.23 (Error relation of the secant method)**.** For the secant method (1.12), there exist ξ_n between x_{n-1} and x_n and ζ_n between $\min(x_{n-1}, x_n, \alpha)$ and $\max(x_{n-1}, x_n, \alpha)$ such that

$$x_{n+1} - \alpha = (x_n - \alpha)(x_{n-1} - \alpha)\frac{f''(\zeta_n)}{2f'(\xi_n)}. \tag{1.16}$$

Proof. Define a divided difference as

$$f[a, b] = \frac{f(a) - f(b)}{a - b}. \tag{1.17}$$

Then it takes some algebra to show that the formula (1.12) is equivalent to

$$x_{n+1} - \alpha = (x_n - \alpha)(x_{n-1} - \alpha)\frac{\frac{f[x_{n-1}, x_n] - f[x_n, \alpha]}{x_{n-1} - \alpha}}{f[x_{n-1}, x_n]}. \tag{1.18}$$

By (1.17) and the mean value theorem (Theorem C.53), there exists ξ_n between x_{n-1} and x_n such that

$$f[x_{n-1}, x_n] = f'(\xi_n). \tag{1.19}$$

Define a function $g(x) := f[x, x_n]$, apply the mean value theorem to $g(x)$, and we have

$$\frac{f[x_{n-1}, x_n] - f[x_n, \alpha]}{x_{n-1} - \alpha} = g'(\beta) \tag{1.20}$$

for some β between x_{n-1} and α. Compute the derivative of $g'(\beta)$ from (1.17), use the Lagrangian remainder Theorem C.61, and we have

$$\frac{f[x_{n-1}, x_n] - f[x_n, \alpha]}{x_{n-1} - \alpha} = \frac{f''(\zeta_n)}{2} \tag{1.21}$$

for some ζ_n between $\min(x_{n-1}, x_n, \alpha)$ and $\max(x_{n-1}, x_n, \alpha)$. The proof is completed by substituting (1.19) and (1.21) into (1.18). □

Theorem 1.24 (Convergence of the secant method)**.** Consider a $\mathcal{C}^2$ function $f : \mathcal{B} \to \mathbb{R}$ on $\mathcal{B} = [\alpha - \delta, \alpha + \delta]$ satisfying $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. If both x_0 and x_1 are chosen sufficiently close to α and $f''(\alpha) \neq 0$, then the iterates $\{x_n\}$ in the secant method converges to the root α with order $p = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618$.

Proof. The continuity of f' and the assumption $f'(\alpha) \neq 0$ yield

$$\exists \delta_1 \in (0, \delta) \text{ s.t. } \forall x \in \mathcal{B}_1,\ f'(x) \neq 0$$

where $\mathcal{B}_1 = [\alpha - \delta_1, \alpha + \delta_1]$. Define $E_i = |x_i - \alpha|$,

$$M = \frac{\max_{x \in \mathcal{B}_1} |f''(x)|}{2 \min_{x \in \mathcal{B}_1} |f'(x)|},$$

and we have from Lemma 1.23

$$ME_{n+1} \leq ME_n ME_{n-1}.$$

Pick x_0, x_1 such that

(i) $E_0 < \delta$, $E_1 < \delta$;

(ii) $\max(ME_1, ME_0) = \eta < 1$,

then an induction by the above equation shows that $E_n < \delta$, $ME_n < \eta$. To prove convergence, we write $ME_0 \leq \eta$, $ME_1 \leq \eta$, $ME_2 \leq ME_1 ME_0 \leq \eta^2$, $ME_3 \leq ME_2 ME_1 \leq \eta^3$, $\cdots$, $ME_{n+1} \leq ME_n ME_{n-1} \leq \eta^{q_n + q_{n-1}} = \eta^{q_{n+1}}$, i.e.

$$E_n \leq B_n := \frac{1}{M}\eta^{q_n}.$$

$\{q_n\}$ is a Fibonacci sequence starting from $q_0 = 1, q_1 = 1$. By Theorem 1.21, as $n \to \infty$ we have $q_n \to \frac{1.618^{n+1}}{\sqrt{5}}$ since $|r_1| \approx 0.618 < 1$. Hence $\lim_{n\to\infty} E_n = 0$.

To guestimate the convergence rate, we first examine the rate at which the upper bounds $\{B_n\}$ decrease:

$$\frac{B_{n+1}}{B_n^{r_0}} = \frac{\frac{1}{M}\eta^{q_{n+1}}}{\left(\frac{1}{M}\right)^{r_0}\eta^{r_0 q_n}} = M^{r_0 - 1}\eta^{q_{n+1} - r_0 q_n} \leq M^{r_0 - 1}\eta^{-1}$$

where $q_{n+1} - r_0 q_n = r_1^{n+1} > -1$.

To prove the convergence rate, we define

$$m_n := \left|\frac{f''(\zeta_n)}{2f'(\xi_n)}\right|, \qquad m_\alpha := \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|, \tag{1.22}$$

where ζ_n and ξ_n are the same as those in Lemma 1.23. By induction, we have

$$E_n = E_1^{F_n} E_0^{F_{n-1}} m_1^{F_{n-1}} \cdots m_{n-1}^{F_1},$$
$$E_{n+1} = E_1^{F_{n+1}} E_0^{F_n} m_1^{F_n} \cdots m_{n-1}^{F_2} m_n^{F_1},$$

where F_n is a Fibonacci number as in Definition 1.20. Then

$$\begin{aligned}\frac{E_{n+1}}{E_n^{r_0}} =& E_1^{F_{n+1} - r_0 F_n} E_0^{F_n - r_0 F_{n-1}} m_1^{F_n - r_0 F_{n-1}} m_2^{F_{n-1} - r_0 F_{n-2}} \\ & \cdots m_{n-2}^{F_3 - r_0 F_2} m_{n-1}^{F_2 - r_0 F_1} m_n^{F_1} \\ =& E_1^{r_1^n} E_0^{r_1^{n-1}} m_1^{r_1^{n-1}} m_2^{r_1^{n-2}} \cdots m_{n-1}^{r_1^1} m_n^1, \end{aligned} \tag{1.23}$$

where the second step follows from Corollary 1.22. (1.22) and the convergence we just proved yield

$$\lim_{n \to +\infty} m_n = m_\alpha, \tag{1.24}$$

which means

$$\exists N \in \mathbb{N} \text{ s.t. } \forall n > N, m_n \in \left(\frac{1}{2}m_\alpha, 2m_\alpha\right). \tag{1.25}$$

We define

$$A := E_1^{r_1^n} \cdot E_0^{r_1^{n-1}} m_1^{r_1^{n-1}} \cdot m_2^{r_1^{n-2}} \cdots m_N^{r_1^{n-N}}$$
$$B := m_{N+1}^{r_1^{n-N-1}} \cdot m_{N+2}^{r_1^{n-N-2}} \cdots m_{n-1}^{r_1^1} \cdot m_n^1$$

so that $\frac{E_{n+1}}{E_n^{r_0}} = AB$. Since $|r_1| < 1$, we have $\lim_{n\to\infty} A = 1$. As for B, we have from (1.25)

$$\begin{aligned} B &\leq (2m_\alpha)^1 \left(\frac{1}{2}m_\alpha\right)^{r_1} \cdots m_{N+1}^{r_1^{n-N-1}} \\ &\leq 2^{1 - r_1 + r_1^2 - \cdots + (-r_1)^{n-N-1}} \cdot m_\alpha^{1 + r_1 + r_1^2 + \cdots r_1^{n-N-1}} \\ &\leq 2^{\sum_{i=0}^{n-N-1}(-r_1)^i} \cdot m_\alpha^{\sum_{i=0}^{n-N-1} r_1^i}, \end{aligned}$$

where the discrimination of the even-numbered factors in the first line comes from the fact of $r_1 < 0$. It follows that

$$\lim_{n\to\infty} \frac{E_{n+1}}{E_n^{r_0}} = \lim_{n\to\infty} A \lim_{n\to\infty} B = \lim_{n\to\infty} B \le 2^{\frac{1}{1+r_1}} m_\alpha^{\frac{1}{r_0}}.$$

The proof is then completed by Definition 1.10. □

Corollary 1.25. Consider solving $f(x) = 0$ near a root α. Let m and sm be the time to evaluate $f(x)$ and $f'(x)$ respectively. The minimum time to obtain the desired absolute accuracy ϵ with Newton's method and the secant method are respectively

$$T_N = (1+s)m\lceil \log_2 K\rceil, \tag{1.26}$$
$$T_S = m\lceil \log_{r_0} K\rceil + m, \tag{1.27}$$

where $r_0 = \frac{1+\sqrt{5}}{2}$, $c = \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|$,

$$K = \frac{\log c\epsilon}{\log c|x_0-\alpha|}, \tag{1.28}$$

and $\lceil\cdot\rceil$ denotes the rounding-up operator, i.e. it rounds towards $+\infty$.

Proof. We showed $|x_n-\alpha| \le \frac{1}{M}\left(M|x_0-\alpha|\right)^{2^n}$ in proving Theorem 1.15. Denote $E_n = |x_n-\alpha|$, we have

$$ME_n \le (ME_0)^{2^n}.$$

Let $i \in \mathbb{N}^+$ denote the smallest number of iterations such that the desired accuracy ϵ is satisfied, i.e. $(ME_0)^{2^i} \le M\epsilon$. When ϵ is sufficiently small, $M \to c$. Hence we have

$$i = \lceil \log_2 K\rceil.$$

For each iteration, Newton's method incurs one function evaluation and one derivative evaluation, which cost time m and sm, respectively. Therefore (1.26) holds.

For the secant method, assume $ME_0 \ge ME_1$. By the proof of Theorem 1.24, we have

$$ME_n \approx (ME_0)^{\frac{\sqrt{5}}{5}r_0^{n+1}},$$

where we have ignored the term $-\frac{\sqrt{5}}{5}r_1^{n+1}$ in the exponent. Let $j \in \mathbb{N}^+$ be the smallest number of iterations such that the desired accuracy ϵ is satisfied, i.e. $r_0^j \le \frac{\sqrt{5}}{r_0}K$. Hence

$$j = \left\lceil \log_{r_0} K + \log_{r_0}\frac{\sqrt{5}}{r_0}\right\rceil \le \lceil \log_{r_0} K\rceil + 1.$$

Since the first two values x_0 and x_1 are given in the secant method, the least number of iterations is $\lceil \log_{r_0} K\rceil$ (compare to Newton's method!). In addition, only the function value $f(x_n)$ needs to be evaluated per iteration because $f(x_{n-1})$ has already been evaluated in the previous iteration. Finally, the extra m is due to the first step where the function value $f(x_0)$ has to be evaluated. □

1.7 Fixed-point iterations

Definition 1.26. A *fixed point* of a function g is an independent parameter of g satisfying $g(\alpha) = \alpha$.

Example 1.27. A fixed point of $f(x) = x^2-3x+4$ is $x = 2$.

Lemma 1.28. If $g : [a,b] \to [a,b]$ is continuous, then g has at least one fixed point in $[a,b]$.

Proof. The function $f(x) = g(x) - x$ satisfies $f(a) \ge 0$ and $f(b) \le 0$. The proof is then completed by the intermediate value theorem (Theorem C.41). □

Exercise 1.29. Let $A = [-1,0)\cup(0,1]$. Give an example of a continuous function $g : A \to A$ that does not have a fixed point. Give an example of a continuous function $f : \mathbb{R} \to \mathbb{R}$ that does not have a fixed point.

Theorem 1.30 (Brouwer's fixed point)**.** Any continuous function $f : \mathbb{D}^n \to \mathbb{D}^n$ with

$$\mathbb{D}^n := \{\mathbf{x} \in \mathbb{R}^n : \|x\| \le 1\}$$

has a fixed point.

Example 1.31. Take a map of your country C and place it on the ground of your room. Let f be the function assigning to each point in your country the point on the map corresponding to it. Then f can be considered as a continuous function $C \to C$. If C is homeomorphic to $\mathbb{D}^2$, then there must exist a point on the map that corresponds exactly to the point on the ground directly beneath it.

Exercise 1.32. Take two pieces of the same-sized paper and lay one on top of the other. Every point on the top sheet of paper is associated with some point right below it on the bottom sheet. Crumple the top sheet into a ball without ripping it. Place the crumpled ball on top of (and simultaneously within the realm of) the bottom sheet of paper. Use Theorem 1.30 to prove that there always exists some point in the crumpled ball that sits above the same point it sat above prior to crumpling.

Definition 1.33. A *fixed-point iteration* is a method for finding a fixed point of g with a formula of the form

$$x_{n+1} = g(x_n), \qquad n \in \mathbb{N}. \tag{1.29}$$

Example 1.34. Newton's method is a fixed-point iteration.

Exercise 1.35. To calculate the square root of some positive real number a, we can formulate the problem as finding the root of $f(x) = x^2 - a$. For $a = 1$, the initial guess of $x_0 = 2$, and the three choices of $g_1(x) := x^2 + x - a$, $g_2(x) := \frac{a}{x}$, and $g_3(x) := \frac{1}{2}(x + \frac{a}{x})$, verify that g_1 diverges, g_2 oscillates, g_3 converges. The theorems in this section will explain why.

Definition 1.36. A function $f : [a,b] \to [a,b]$ is a *contraction* or *contractive mapping* on $[a,b]$ if

$$\exists\lambda \in [0,1) \text{ s.t. } \forall x,y \in [a,b],\ |f(x)-f(y)| \le \lambda|x-y|. \tag{1.30}$$

Example 1.37. Any linear function $f(x) = \lambda x + c$ with $0 \le \lambda < 1$ is a contraction.

Theorem 1.38 (Convergence of contractions)**.** If $g(x)$ is a continuous contraction on $[a,b]$, then it has a unique fixed point α in $[a,b]$. Furthermore, the fixed-point iteration (1.29) converges to α for any choice $x_0 \in [a,b]$ and

$$|x_n - \alpha| \le \frac{\lambda^n}{1-\lambda}|x_1 - x_0|. \tag{1.31}$$

Proof. By Lemma 1.28, g has at least one fixed point in $[a,b]$. Suppose there are two distinct fixed points α and β, then $|\alpha-\beta| = |g(\alpha)-g(\beta)| \le \lambda|\alpha-\beta|$, which implies $|\alpha-\beta| \le 0$, i.e. the two fixed points are identical.

By Definition 1.36, $x_{n+1} = g(x_n)$ implies that all x_n's stay in $[a,b]$. To prove convergence,

$$|x_{n+1}-\alpha| = |g(x_n)-g(\alpha)| \le \lambda|x_n-\alpha|.$$

By induction and the triangle inequality, we have

$$\begin{aligned}|x_n-\alpha| &\le \lambda^n|x_0-\alpha| \\ &\le \lambda^n(|x_1-x_0|+|x_1-\alpha|) \\ &\le \lambda^n(|x_1-x_0|+\lambda|x_0-\alpha|),\end{aligned}$$

which implies $|x_0-\alpha| \le \frac{1}{1-\lambda}|x_1-x_0|$ and (1.31). □

Theorem 1.39. Consider $g:[a,b]\to[a,b]$. If $g \in \mathcal{C}^1[a,b]$ and $\lambda = \max_{x\in[a,b]}|g'(x)| < 1$, then g has a unique fixed point α in $[a,b]$. Furthermore, the fixed-point iteration (1.29) converges to α for any choice $x_0 \in [a,b]$, the error bound (1.31) holds, and

$$\lim_{n\to\infty}\frac{x_{n+1}-\alpha}{x_n-\alpha} = g'(\alpha). \tag{1.32}$$

Proof. The mean value theorem (Theorem C.53) implies that, for all $x,y\in[a,b]$, $|g(x)-g(y)| \le \lambda|x-y|$. Theorem 1.38 yields all the results except (1.32), which follows from

$$x_{n+1}-\alpha = g(x_n)-g(\alpha) = g'(\xi)(x_n-\alpha),$$

$\lim x_n = \alpha$, and the fact that ξ is between x_n and α. □

Corollary 1.40. Let α be a fixed point of $g:\mathbb{R}\to\mathbb{R}$ with $|g'(\alpha)|<1$ and $g\in\mathcal{C}^1(\mathcal{B})$ on $\mathcal{B}=[\alpha-\delta,\alpha+\delta]$ with some $\delta>0$. If x_0 is chosen sufficiently close to α, then the results of Theorem 1.38 hold.

Proof. Choose λ so that $|g'(\alpha)|<\lambda<1$. Choose $\delta_0 \le \delta$ so that $\max_{x\in\mathcal{B}_0}|g'(x)| \le \lambda < 1$ on $\mathcal{B}_0 = [\alpha-\delta_0,\alpha+\delta_0]$. Then $g(\mathcal{B}_0)\subset\mathcal{B}_0$ and applying Theorem 1.39 completes the proof. □

Corollary 1.41. Consider $g:[a,b]\to[a,b]$ with a fixed point $g(\alpha)=\alpha\in[a,b]$. The fixed-point iteration (1.29) converges to α with pth-order accuracy ($p>1$, $p\in\mathbb{N}$) if

$$\begin{cases} g\in\mathcal{C}^p[a,b], \\ \forall k=1,2,\dots,p-1,\ g^{(k)}(\alpha)=0, \\ g^{(p)}(\alpha)\neq 0 \end{cases} \tag{1.33}$$

and if x_0 is chosen sufficiently close to α.

Proof. By Corollary 1.40, $g'(\alpha)=0$, and the continuity of g', there exists $\delta>0$ such that the fixed-point iteration converges uniquely to α inside $[\alpha-\delta,\alpha+\delta]$. By the Taylor expansion of g at α, we have

$$\begin{aligned}E_{\text{abs}}(x_{n+1}) &:= |x_{n+1}-\alpha| = |g(x_n)-g(\alpha)| \\ &= \left|\sum_{i=1}^{p-1}\frac{(x_n-\alpha)^i}{i!}g^{(i)}(\alpha) + \frac{(x_n-\alpha)^p}{p!}g^{(p)}(\xi)\right|\end{aligned}$$

for some $\xi\in[a,b]$. Since $g^{(p)}$ is continuous on $[a,b]$, Theorem C.50 implies that $g^{(p)}$ is bounded on $[a,b]$. Hence there exists a constant M such that $E_{\text{abs}}(x_{n+1}) < ME^p_{\text{abs}}(x_n)$. □

Example 1.42. The following method has third-order convergence for computing $\sqrt{R}$:

$$x_{n+1} = \frac{x_n(x_n^2+3R)}{3x_n^2+R}.$$

First, $\sqrt{R}$ is the fixed point of $F(x)=\frac{x(x^2+3R)}{3x^2+R}$:

$$F\left(\sqrt{R}\right) = \frac{\sqrt{R}(R+3R)}{3R+R} = \sqrt{R}.$$

Second, the derivatives of $F(x)$ are

n	$F^{(n)}(x)$	$F^{(n)}(\sqrt{R})$
1	$\frac{3(x^2-R)^2}{(3x^2+R)^2}$	0
2	$\frac{48Rx(x^2-R)}{(3x^2+R)^3}$	0
3	$\frac{-48R(9x^4-18Rx^2+R^2)}{(3x^2+R)^4}$	$\frac{-48R(-8R^2)}{(4R)^4} = \frac{3}{2R} \neq 0$

The rest follows from Corollary 1.41.

1.8 Problems

1.8.1 Theoretical questions

I. Consider the bisection method starting with the initial interval $[1.5, 3.5]$. In the following questions "the interval" refers to the bisection interval whose width changes across different loops.

- What is the width of the interval at the nth step?
- What is the supremum of the distance between the root r and the midpoint of the interval?

II. In using the bisection algorithm with its initial interval as $[a_0,b_0]$ with $a_0>0$, we want to determine the root with its *relative* error no greater than ϵ. Prove that this goal of accuracy is guaranteed by the following choice of the number of steps,

$$n \ge \frac{\log(b_0-a_0)-\log\epsilon-\log a_0}{\log 2} - 1.$$

III. Perform four iterations of Newton's method for the polynomial equation $p(x) = 4x^3-2x^2+3=0$ with the starting point $x_0=-1$. Use a hand calculator and organize results of the iterations in a table.

IV. Consider a variation of Newton's method in which only the derivative at x_0 is used,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}.$$

Find C and s such that

$$e_{n+1} = Ce_n^s,$$

where e_n is the error of Newton's method at step n, s is a constant, and C may depend on x_n, the true solution α, and the derivative of the function f.

V. Within $(-\frac{\pi}{2}, \frac{\pi}{2})$, will the iteration $x_{n+1} = \tan^{-1} x_n$ converge?

VI. Let $p > 1$. What is the value of the following continued fraction?

$$x = \cfrac{1}{p + \cfrac{1}{p + \cfrac{1}{p + \dots}}}$$

Prove that the sequence of values converges. (Hint: this can be interpreted as $x = \lim_{n\to\infty} x_n$, where $x_1 = \frac{1}{p}$, $x_2 = \cfrac{1}{p+\frac{1}{p}}$, $x_3 = \cfrac{1}{p+\cfrac{1}{p+\frac{1}{p}}}$, ..., and so forth. Formulate x as a fixed point of some function.)

VII. What happens in problem II if $a_0 < 0 < b_0$? Derive an inequality of the number of steps similar to that in II. In this case, is the relative error still an appropriate measure?

VIII. (∗) Consider solving $f(x) = 0$ ($f \in \mathcal{C}^{k+1}$) by Newton's method with the starting point x_0 close to a root of multiplicity k. Note that α is a zero of multiplicity k of the function f iff

$$f^{(k)}(\alpha) \neq 0; \qquad \forall i < k, \quad f^{(i)}(\alpha) = 0.$$

- How can a multiple zero be detected by examining the behavior of the points $(x_n, f(x_n))$?
- Prove that if r is a zero of multiplicity k of the function f, then quadratic convergence in Newton's iteration will be restored by making this modification:

$$x_{n+1} = x_n - k\frac{f(x_n)}{f'(x_n)}.$$

1.8.2 Programming assignments

A. Implement the bisection method, Newton's method, and the secant method in a C++ package. You should

(a) design an abstract base class `EquationSolver` with a pure virtual method `solve`,

(b) write a derived class of `EquationSolver` for each method to accomodate its particularities in the contract of solving nonlinear equations.

B. Test your implementation of the bisection method on the following functions and intervals.

- $x^{-1} - \tan x$ on $[0, \frac{\pi}{2}]$,
- $x^{-1} - 2^x$ on $[0, 1]$,
- $2^{-x} + e^x + 2\cos x - 6$ on $[1, 3]$,
- $(x^3 + 4x^2 + 3x + 5)/(2x^3 - 9x^2 + 18x - 2)$ on $[0, 4]$.

C. Test your implementation of Newton's method by solving $x = \tan x$. Find the roots near 4.5 and 7.7.

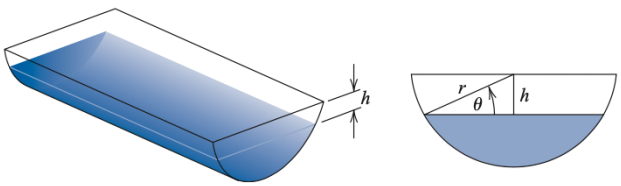
D. Test your implementation of the secant method by the following functions and initial values.

- $\sin(x/2) - 1$ with $x_0 = 0, x_1 = \frac{\pi}{2}$,
- $e^x - \tan x$ with $x_0 = 1, x_1 = 1.4$,
- $x^3 - 12x^2 + 3x + 1$ with $x_0 = 0, x_1 = -0.5$.

You should play with other initial values and (if you get different results) think about the reasons.

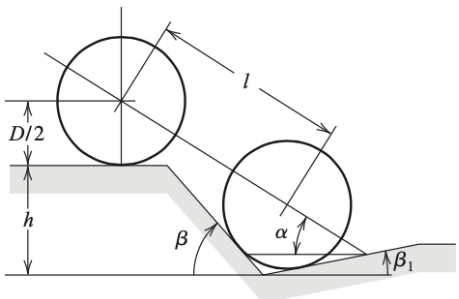
E. As shown below, a trough of length L has a cross section in the shape of a semi-circle with radius r. When filled to within a distance h of the top, the water has the volume

$$V = L\left[0.5\pi r^2 - r^2 \arcsin\frac{h}{r} - h(r^2 - h^2)^{\frac{1}{2}}\right].$$



Suppose $L = 10$ft, $r = 1$ft, and $V = 12.4$ft^3. Find the depth of water in the trough to within 0.01ft by each of the three implementations in A.

F. In the design of all-terrain vehicles, it is necessary to consider the failure of the vehicle when attempting to negotiate two types of obstacles. One type of failure is called *hang-up failure* and occurs when the vehicle attempts to cross an obstacle that causes the bottom of the vehicle to touch the ground. The other type of failure is called *nose-in failure* and occurs when the vehicle descends into a ditch and its nose touches the ground.



The above figure shows the components associated with the nose-in failure of a vehicle. The maximum angle α that can be negotiated by a vehicle when β is the

maximum angle at which hang-up failure does not occur satisfies the equation

$$A\sin\alpha\cos\alpha + B\sin^2\alpha - C\cos\alpha - E\sin\alpha = 0,$$

where

$$A = l\sin\beta_1,\ B = l\cos\beta_1,$$
$$C = (h + 0.5D)\sin\beta_1 - 0.5D\tan\beta_1,$$
$$E = (h + 0.5D)\cos\beta_1 - 0.5D.$$

(a) Use Newton's method to verify $\alpha \approx 33^\circ$ when $l = 89\,\text{in.}$, $h = 49\,\text{in.}$, $D = 55\,\text{in.}$ and $\beta_1 = 11.5^\circ$.

(b) Use Newton's method to find α with the initial guess 33° for the situation when l, h, β_1 are the same as in part (a) but $D = 30\,\text{in.}$.

(c) Use the secant method (with another initial value as far away as possbile from 33°) to find α. Show that you get a different result if the initial value is too far away from 33°; discuss the reasons.

Chapter 2

Polynomial Interpolation

Definition 2.1. *Interpolation* constructs new data points within the range of a discrete set of known data points, usually by generating an *interpolating function* whose graph goes through all known data points.

Example 2.2. The interpolating function may be piecewise constant, piecewise linear, polynomial, spline, or other non-polynomial functions.

2.1 The Vandermonde determinant

Definition 2.3. For n given parameters $t_1, t_2, \ldots, t_n$, the associated n-by-n *Vandermonde matrix* V has its (i,j)th element as $v_{i,j} = t_j^{i-1}$; in matrix norm, we have

$$V(t_1, t_2, \ldots, t_n) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{n-1} & t_2^{n-1} & \cdots & t_n^{n-1} \end{bmatrix}. \tag{2.1}$$

Lemma 2.4. The determinant of a $(n+1)$-by-$(n+1)$ Vandermonde matrix can be expressed as

$$\det V(x_0, x_1, \ldots, x_n) = \prod_{i>j} (x_i - x_j). \tag{2.2}$$

Proof. Consider the function

$$\begin{aligned} U(x) &= \det V(x_0, x_1, \ldots, x_{n-1}, x) \\ &= \begin{vmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_0 & x_1 & \cdots & x_{n-1} & x \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_0^n & x_1^n & \cdots & x_{n-1}^n & x^n \end{vmatrix}. \end{aligned} \tag{2.3}$$

Clearly, $U(x) \in \mathbb{P}_n$ and it vanishes at $x_0, x_1, \ldots, x_{n-1}$ since inserting these values in place of x yields two identical columns in the determinant. It follows that

$$U(x) = A \prod_{i=0}^{n-1} (x - x_i),$$

where A depends only on $x_0, x_1, \ldots, x_{n-1}$. Meanwhile, the expansion of $U(x)$ in (2.3) by minors of its last column implies that the coefficient of x^n is $\det V(x_0, x_1, \ldots, x_{n-1})$. Hence we have

$$U(x) = \det V(x_0, x_1, \ldots, x_{n-1}) \prod_{i=0}^{n-1} (x - x_i),$$

and consequently the recursion

$$\det V(x_0, x_1, \ldots, x_n) = \det V(x_0, x_1, \ldots, x_{n-1}) \prod_{i=0}^{n-1} (x_n - x_i).$$

An induction based on $U(x_0, x_1) = x_1 - x_0$ yields (2.2). □

Theorem 2.5 (Uniqueness of polynomial interpolation)**.** Given distinct points $x_0, x_1, \ldots, x_n \in \mathbb{C}$ and corresponding values $f_0, f_1, \ldots, f_n \in \mathbb{C}$. Denote by $\mathbb{P}_n$ the class of polynomials of degree at most n. There exists a unique polynomial $p_n(x) \in \mathbb{P}_n$ such that

$$\forall i = 0, 1, \ldots, n, \qquad p_n(x_i) = f_i. \tag{2.4}$$

Proof. Set up a polynomial $\sum_{i=0}^{n} a_i x^i$ with $n+1$ undetermined coefficients a_i. The condition (2.4) leads to the system of $n+1$ equations:

$$a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n = f_i,$$

where $i = 0, 1, \ldots, n$. By Lemma 2.4, the determinant of the system is $\prod_{i>j}(x_i - x_j)$. The proof is completed by the distinctness of the points and Cramer's rule. □

2.2 The Cauchy remainder

Theorem 2.6 (Generalized Rolle)**.** Let $n \geq 2$. Suppose that $f \in \mathcal{C}^{n-1}[a,b]$ and $f^{(n)}(x)$ exists at each point of (a,b). Suppose that $f(x_0) = f(x_1) = \cdots = f(x_n) = 0$ for $a \leq x_0 < x_1 < \cdots < x_n \leq b$. Then there is a point $\xi \in (x_0, x_n)$ such that $f^{(n)}(\xi) = 0$.

Proof. Applying Rolle's theorem (Theorem C.52) on the n intervals (x_i, x_{i+1}) yields n points ζ_i where $f'(\zeta_i) = 0$. Consider $f', f'', \ldots, f^{(n-1)}$ as new functions. Repeatedly applying the above arguments completes the proof. □

Theorem 2.7 (Cauchy remainder of polynomial interpolation)**.** Let $f \in \mathcal{C}^n[a,b]$ and suppose that $f^{(n+1)}(x)$ exists at each point of (a,b). Let $p_n(f;x)$ denote the unique polynomial in $\mathbb{P}_n$ that coincides with f at $x_0, x_1, \ldots, x_n$. Define

$$R_n(f;x) := f(x) - p_n(f;x) \tag{2.5}$$

as the *Cauchy remainder of the polynomial interpolation.* If $a \le x_0 < x_1 < \cdots < x_n \le b$, then there exists some $\xi \in (a,b)$ such that

$$R_n(f;x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{2.6}$$

where the value of ξ depends on $x, x_0, x_1, \ldots, x_n$, and f.

Proof. Since $f(x_k) = p_n(f;x_k)$, the remainder $R_n(f;x)$ vanishes at x_k's. Fix $x \neq x_0, x_1, \ldots, x_n$ and define

$$K(x) = \frac{f(x) - p_n(f;x)}{\prod_{i=0}^{n}(x - x_i)}$$

and a function of t

$$W(t) = f(t) - p_n(f;t) - K(x) \prod_{i=0}^{n} (t - x_i).$$

The function $W(t)$ vanishes at $t = x_0, x_1, \ldots, x_n$. In addition $W(x) = 0$. By Theorem 2.6, $W^{(n+1)}(\xi) = 0$ for some $\xi \in (a,b)$, i.e.

$$0 = W^{(n+1)}(\xi) = f^{(n+1)}(\xi) - (n+1)!K(x).$$

Hence $K(x) = f^{(n+1)}(\xi)/(n+1)!$ and (2.6) holds. □

Corollary 2.8. Suppose $f(x) \in \mathcal{C}^{n+1}[a,b]$. Then

$$|R_n(f;x)| \le \frac{M_{n+1}}{(n+1)!} \prod_{i=0}^{n} |x - x_i| \le \frac{M_{n+1}}{(n+1)!}(b-a)^{n+1}, \tag{2.7}$$

where $M_{n+1} = \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right|$.

Example 2.9. A value for $\arcsin(0.5335)$ is obtained by interpolating linearly between the values for $x = 0.5330$ and $x = 0.5340$. Estimate the error committed.

Let $f(x) = \arcsin(x)$. Then

$$f''(x) = x(1-x^2)^{-\frac{3}{2}}, \qquad f'''(x) = (1+2x^2)(1-x^2)^{-\frac{5}{2}}.$$

Since the third derivative is positive over $[0.5330, 0.5340]$. The maximum value of f'' occurs at 0.5340. By Corollary 2.8 we have $|R_1| \le 4.42 \times 10^{-7}$. The true error is about 1.10×10^{-7}.

2.3 The Lagrange formula

Definition 2.10. To interpolate given values $f_0, f_1, \ldots, f_n$ at distinct points $x_0, x_1, \ldots, x_n$, the *Lagrange formula* is

$$p_n(x) = \sum_{k=0}^{n} f_k \ell_k(x), \tag{2.8}$$

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

$$\ell_k(x) = \prod_{i \neq k; i=0}^{n} \frac{x - x_i}{x_k - x_i}. \tag{2.9}$$

In particular, for $n = 0$, $\ell_0 = 1$.

Example 2.11. For $i = 0, 1, 2$, we are given $x_i = 1, 2, 4$ and $f(x_i) = 8, 1, 5$, respectively. The Lagrangian formula generates $p_2(x) = 3x^2 - 16x + 21$.

Lemma 2.12. Define a symmetric polynomial

$$\pi_n(x) = \begin{cases} 1, & n = 0; \\ \prod_{i=0}^{n-1}(x - x_i), & n > 0. \end{cases} \tag{2.10}$$

Then for $n > 0$ the fundamental polynomial for pointwise interpolation can be expressed as

$$\forall x \neq x_k, \qquad \ell_k(x) = \frac{\pi_{n+1}(x)}{(x - x_k)\pi'_{n+1}(x_k)}. \tag{2.11}$$

Proof. By the chain rule, $\pi'_{n+1}(x)$ is the summation of $n+1$ terms, each of which is a product of n terms. When x is replaced with x_k, all of the $n+1$ terms vanish except one. □

Lemma 2.13 (Cauchy relations)**.** The fundamental polynomials $\ell_k(x)$ satisfy the Cauchy relations as follows.

$$\sum_{k=0}^{n} \ell_k(x) \equiv 1 \tag{2.12}$$

$$\forall j = 1, \ldots, n, \qquad \sum_{k=0}^{n} (x_k - x)^j \ell_k(x) \equiv 0 \tag{2.13}$$

Proof. By Theorems 2.5 and 2.7, for each $q(x) \in \mathbb{P}_n$ we have $p_n(q;x) \equiv q(x)$. Interpolating the constant function $f(x) \equiv 1$ with the Lagrange formula yields (2.12).

Similarly, define $q_j(u) := (u - x)^j$ where x is a free parameter. Then we have, for any $j = 1, \ldots, n$,

$$p_n(q_j;u) \equiv q_j(u) \;\Rightarrow\; \sum_{k=0}^{n} (x_k - x)^j \ell_k(u) \equiv (u - x)^j,$$

which yields (2.13) when we set $u = x$. □

2.4 The Newton formula

Definition 2.14 (Divided difference and the Newton formula)**.** The *Newton formula* for interpolating the values $f_0, f_1, \ldots, f_n$ at distinct points $x_0, x_1, \ldots, x_n$ is

$$p_n(x) = \sum_{k=0}^{n} a_k \pi_k(x), \tag{2.14}$$

where π_k is defined in (2.10) and the *kth divided difference* a_k is defined as the coefficient of x^k in $p_k(f;x)$ and is denoted by $f[x_0, x_1, \ldots, x_k]$ or $[x_0, x_1, \ldots, x_k]f$. In particular, $f[x_0] = f(x_0)$.

Corollary 2.15. Suppose $(i_0, i_1, i_2, \dots, i_k)$ is a permutation of $(0, 1, 2, \dots, k)$. Then

$$f[x_0, x_1, \dots, x_k] = f[x_{i_0}, x_{i_1}, \dots, x_{i_k}]. \tag{2.15}$$

Proof. The interpolating polynomial does not depend on the numbering of the interpolating nodes. The rest of the proof follows from the uniqueness of the interpolating polynomial in Theorem 2.5. □

Corollary 2.16. The kth divided difference can be expressed as

$$f[x_0, x_1, \dots, x_k] = \sum_{i=0}^{k} \frac{f_i}{\prod_{j\neq i; j=0}^{k}(x_i - x_j)} = \sum_{i=0}^{k} \frac{f_i}{\pi'_{k+1}(x_i)}, \tag{2.16}$$

where $\pi_{k+1}(x)$ is defined in (2.10).

Proof. The uniqueness of interpolating polynomials in Theorem 2.5 implies that the two polynomials in (2.8) and (2.14) are the same. Then the first equality follows from (2.9) and Definition 2.14, while the second equality follows from Lemma 2.12. □

Theorem 2.17. Divided differences satisfy the recursion

$$f[x_0, x_1, \dots, x_k] = \frac{f[x_1, x_2, \dots, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_0}. \tag{2.17}$$

Proof. By Definition 2.14, $f[x_1, x_2, \dots, x_k]$ is the coefficient of x^{k-1} in a degree-$(k-1)$ interpolating polynomial, say, $P_2(x)$. Similarly, let $P_1(x)$ be the interpolating polynomial whose coefficient of x^{k-1} is $f[x_0, x_1, \dots, x_{k-1}]$. Construct a polynomial

$$P(x) = P_1(x) + \frac{x - x_0}{x_k - x_0}\left(P_2(x) - P_1(x)\right).$$

Clearly $P(x_0) = P_1(x_0)$. Furthermore, the interpolation condition implies $P_2(x_i) = P_1(x_i)$ for $i = 1, 2, \dots, k-1$. Hence $P(x_i) = P_1(x_i)$ for $i = 1, 2, \dots, k-1$. Lastly, $P(x_k) = P_2(x_k)$. Therefore, $P(x)$ as above is the interpolating polynomial for the given values at the $k+1$ points. In particular, the term $f[x_0, x_1, \cdots, x_k]x^k$ in $P(x)$ is contained in $\frac{x}{x_k - x_0}(P_2(x) - P_1(x))$. The rest follows from Definition 2.14. □

Definition 2.18. The kth divided difference ($k \in \mathbb{N}^+$) on the *table of divided differences*

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

is calculated as the difference of the entry immediately to the left and the one above it, divided by the difference of the x-value horizontal to the left and the one corresponding to the f-value found by going diagonally up.

Example 2.19. Derive the interpolating polynomial via the Newton formula for the function f with given values as follows. Then estimate $f(\frac{3}{2})$.

x	0	1	2	3
$f(x)$	6	−3	−6	9

By Definition 2.18, we can construct the following table of divided difference,

$$\begin{array}{c|cccc} 0 & 6 & & & \\ 1 & -3 & -9 & & \\ 2 & -6 & -3 & 3 & \\ 3 & 9 & 15 & 9 & 2 \end{array} \tag{2.18}$$

By Definition 2.14, the interpolating polynomial is generated from the main diagonal and the first column of the above table as follows.

$$p_3 = 6 - 9x + 3x(x-1) + 2x(x-1)(x-2). \tag{2.19}$$

Hence $f(\frac{3}{2}) \approx p_3(\frac{3}{2}) = -6$.

Exercise 2.20. Redo Example 2.11 with the Newton formula.

Theorem 2.21. For distinct points $x_0, x_1, \dots, x_n$, and x, we have

$$\begin{aligned} f(x) = {} & f[x_0] + f[x_0, x_1](x - x_0) + \cdots \\ & + f[x_0, x_1, \cdots, x_n] \prod_{i=0}^{n-1}(x - x_i) \\ & + f[x_0, x_1, \cdots, x_n, x] \prod_{i=0}^{n}(x - x_i). \end{aligned} \tag{2.20}$$

Proof. Take another point $z \neq x_i$. The Newton formula applied to $x_0, x_1, \dots, x_n, z$ yields an interpolating polynomial

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

The interpolation condition $Q(z) = f(z)$ yields

$$\begin{aligned} f(z) = Q(z) = {} & f[x_0] + f[x_0, x_1](z - x_0) + \cdots \\ & + f[x_0, x_1, \cdots, x_n] \prod_{i=0}^{n-1}(z - x_i) \\ & + f[x_0, x_1, \cdots, x_n, z] \prod_{i=0}^{n}(z - x_i). \end{aligned}$$

Replacing the dummy variable z with x yields (2.20).

The above argument assumes $x \neq x_i$. We now consider the case of $x = x_j$ for some fixed j. Rewrite (2.20) as $f(x) = p_n(f; x) + R(x)$ where $R(x)$ is clearly the last term in (2.20). We need to show

$$\forall j = 0, 1, \cdots, n, \qquad p_n(f; x_j) + R(x_j) - f(x_j) = 0,$$

where $p_n(f; x_j)$ is the value of $p_n(f; x)$ at $x = x_j$; this clearly holds because $R(x_j) = 0$ and the interpolation condition at x_j dictates $p_n(f; x_j) = f(x_j)$. □

Corollary 2.22. Suppose $f \in \mathcal{C}^n[a,b]$ and $f^{(n+1)}(x)$ exists at each point of (a,b). If $a = x_0 < x_1 < \cdots < x_n = b$ and $x \in [a,b]$, then there exists $\xi(x) \in (a,b)$ such that

$$f[x_0, x_1, \cdots, x_n, x] = \frac{1}{(n+1)!} f^{(n+1)}(\xi(x)). \tag{2.21}$$

Proof. This follows from Theorems 2.21 and 2.7. □

Corollary 2.23. If $x_0 < x_1 < \cdots < x_n$ and $f \in \mathcal{C}^n[x_0, x_n]$, we have

$$\lim_{x_n \to x_0} f[x_0, x_1, \cdots, x_n] = \frac{1}{n!} f^{(n)}(x_0). \tag{2.22}$$

Proof. Set $x = x_{n+1}$ in Corollary 2.22, replace $n+1$ by n, and we have $\xi \to x_0$ as $x_n \to x_0$. □

Definition 2.24. A *bisequence* is a function $f : \mathbb{Z} \to \mathbb{R}$.

Definition 2.25. The *forward shift* E and the *backward shift* B are linear operators $V \mapsto V$ on the linear space V of bisequences given by

$$(Ef)(i) = f(i+1), \quad (Bf)(i) = f(i-1). \tag{2.23}$$

The *forward difference* Δ and the *backward difference* ∇ are linear operators $V \mapsto V$ given by

$$\Delta = E - I, \quad \nabla = I - B, \tag{2.24}$$

where I is the identity operator on V.

Example 2.26. With the notation $f_i := f(i)$ for a bisequence f, the nth forward difference and the nth backward difference are

$$\Delta^n f_i := (\Delta^n f)(i), \quad \nabla^n f_i := (\nabla^n f)(i). \tag{2.25}$$

In particular, for $n = 1$ we have

$$\Delta f_i = f_{i+1} - f_i, \quad \nabla f_i = f_i - f_{i-1}. \tag{2.26}$$

Theorem 2.27. The forward difference and backward difference are related as

$$\forall n \in \mathbb{N}^+, \qquad \Delta^n f_i = \nabla^n f_{i+n}. \tag{2.27}$$

Proof. An easy induction. □

Theorem 2.28. The forward difference can be expressed explicitly as

$$\Delta^n f_i = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k}. \tag{2.28}$$

Proof. For $n = 1$, (2.28) reduces to $\Delta f_i = f_{i+1} - f_i$. The rest of the proof is an induction utilizing the identity

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}. \tag{2.29}$$

Suppose (2.28) holds. For the inductive step, we have

$$\begin{aligned}
\Delta^{n+1} f_i =& \Delta \Delta^n f_i = \Delta \left(\sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k} \right) \\
=& \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k+1} - \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k} \\
=& \sum_{k=1}^{n} (-1)^{n+1-k} \binom{n}{k-1} f_{i+k} + f_{i+n+1} \\
&+ \sum_{k=1}^{n} (-1)^{n+1-k} \binom{n}{k} f_{i+k} + (-1)^{n+1} f_i \\
=& \sum_{k=0}^{n+1} (-1)^{n+1-k} \binom{n+1}{k} f_{i+k},
\end{aligned}$$

where the second line follows from (2.26), the third line from splitting one term out of each sum and replacing the dummy variable in the first sum, and the fourth line from (2.29) and the fact that $(-1)^{n+1} f_i$ and f_{i+n+1} contribute to the first and last terms, respectively. □

Theorem 2.29. On a grid $x_i = x_0 + ih$ with uniform spacing h, the sequence of values $f_i = f(x_i)$ satisfies

$$\forall n \in \mathbb{N}^+, \qquad f[x_0, x_1, \ldots, x_n] = \frac{\Delta^n f_0}{n! h^n}. \tag{2.30}$$

Proof. Of course (2.30) can be proven by induction. Here we provide a more informative proof. For $\pi_{n+1}(x)$ defined in (2.10), we have $\pi'_{n+1}(x_k) = \prod_{i=0, i\neq k}^{n} (x_k - x_i)$. It follows from $x_k - x_i = (k-i)h$ that

$$\pi'_{n+1}(x_k) = \prod_{i=0, i\neq k}^{n} (k-i)h = h^n k!(n-k)!(-1)^{n-k}. \tag{2.31}$$

Then we have

$$\begin{aligned}
f[x_0, x_1, \ldots, x_n] &= \sum_{k=0}^{n} \frac{f_k}{\pi'_{n+1}(x_k)} = \sum_{k=0}^{n} \frac{(-1)^{n-k} f_k}{h^n k!(n-k)!} \\
&= \frac{1}{h^n n!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_k = \frac{\Delta^n f_0}{h^n n!},
\end{aligned}$$

where the first step follows from Corollary 2.16, the second from (2.31), and the last from Theorem 2.28. □

Theorem 2.30 (Newton's forward difference formula)**.** Suppose $p_n(f;x) \in \mathbb{P}_n$ interpolates $f(x)$ on a uniform grid $x_i = x_0 + ih$ at $x_0, x_1, \ldots, x_n$ with $f_i = f(x_i)$. Then

$$\forall s \in \mathbb{R}, \qquad p_n(f; x_0 + sh) = \sum_{k=0}^{n} \binom{s}{k} \Delta^k f_0, \tag{2.32}$$

where $\Delta^0 f_0 = f_0$ and

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!}. \tag{2.33}$$

Proof. Set $f(x) = p_n(f; x)$ in Theorem 2.21, apply Theorem 2.29, and we have

$$p(x) = f_0 + \sum_{k=1}^{n} \frac{\Delta^k f_0}{k! h^k} \prod_{i=0}^{k-1} (x - x_i);$$

the remainder is zero because any $(n+1)$th divided difference applied to a degree n polynomial is zero. The proof is completed by $x = x_0 + sh$, $x_i = x_0 + ih$, and (2.33). □

2.5 The Neville-Aitken algorithm

Theorem 2.31. Denote $p_0^{[i]} = f(x_i)$ for $i = 0, 1, \ldots, n$. For all $k = 0, 1, \ldots, n-1$ and $i = 0, 1, \ldots, n-k-1$, define

$$p_{k+1}^{[i]}(x) = \frac{(x - x_i) p_k^{[i+1]}(x) - (x - x_{i+k+1}) p_k^{[i]}(x)}{x_{i+k+1} - x_i}. \quad (2.34)$$

Then each $p_k^{[i]}$ is the interpolating polynomial for the function f at the points $x_i, x_{i+1}, \ldots, x_{i+k}$. In particular, $p_n^{[0]}$ is the interpolating polynomial of degree n for the function f at the points $x_0, x_1, \ldots, x_n$.

Proof. The induction basis clearly holds for $k = 0$ because of the definition $p_0^{[i]} = f(x_i)$. Suppose that $p_k^{[i]}$ is the interpolating polynomial of degree k for the function f at the points $x_i, x_{i+1}, \ldots, x_{i+k}$. Then the interpolation conditions yield

$$\forall j = i+1, i+2, \ldots, i+k, \qquad p_k^{[i+1]}(x_j) = p_k^{[i]}(x_j) = f(x_j),$$

which, together with (2.34), implies

$$\forall j = i+1, i+2, \ldots, i+k, \qquad p_{k+1}^{[i]}(x_j) = f(x_j).$$

In addition, (2.34) and the induction hypothesis yield

$$p_{k+1}^{[i]}(x_i) = p_k^{[i]}(x_i) = f(x_i),$$
$$p_{k+1}^{[i]}(x_{i+k+1}) = p_k^{[i+1]}(x_{i+k+1}) = f(x_{i+k+1}).$$

The proof is completed by the last three equations and the uniqueness of interpolating polynomials. □

Example 2.32. To estimate $f(x)$ for $x = \frac{3}{2}$ directly from the table in Example 2.19, we construct a table by repeating (2.34) with $x_i = i$ for $i = 0, 1, 2, 3$.

x_i	$x - x_i$	$f(x_i)$	$p_1^{[i]}(x)$	$p_2^{[i]}(x)$	$p_3^{[i]}(x)$
0	$\frac{3}{2}$	6	$-\frac{15}{2}$	$-\frac{21}{4}$	-6
1	$\frac{1}{2}$	-3	$-\frac{9}{2}$	$-\frac{27}{4}$	
2	$-\frac{1}{2}$	-6	$-\frac{27}{2}$		
3	$-\frac{3}{2}$	9			

(2.35)

The result is the same as that in Example 2.19. In contrast, the calculation and layout of the two tables are distinct.

2.6 The Hermite interpolation

Definition 2.33. Given distinct points $x_0, x_1, \ldots, x_k$ in $[a, b]$, non-negative integers $m_0, m_1, \ldots, m_k$, and a function $f \in \mathcal{C}^M[a, b]$ where $M = \max_i m_i$, the *Hermite interpolation problem* seeks a polynomial p of the lowest degree such that

$$\forall i = 0, 1, \ldots, k, \ \forall \mu = 0, 1, \cdots, m_i, \ \ p^{(\mu)}(x_i) = f_i^{(\mu)}, \quad (2.36)$$

where $f_i^{(\mu)} = f^{(\mu)}(x_i)$ is the value of the μth derivative of f at x_i; in particular, $f_i^{(0)} = f(x_i)$.

Theorem 2.34. There exists a unique solution to the Hermite interpolation problem in Definition 2.33.

Proof. Write $N := k + \sum_{i=0}^{k} m_i$. The interpolation conditions in (2.36) lead to a linear system of $N+1$ equations, where the unknowns are coefficients of the monomials in the polynomial $p(x) = \sum_i a_i x^i$. It suffices to show that the matrix M of this linear system is invertible, i.e., $\ker M = \{\mathbf{0}\}$. To this end, consider the linear system $M\mathbf{a} = \mathbf{0}$. Since each x_i in (2.36) is a zero of $p(x)$ with its multiplicity no less than $(m_i + 1)$, the product $P(x) = \prod_{i=0}^{k} (x - x_i)^{m_i+1}$ must be a factor of $p(x)$. However, the degree of P is $N+1$ whereas that of p is N, thus we must have $p \equiv 0$. □

Definition 2.35 (Generalized divided difference)**.** Let $x_0, x_1, \ldots, x_k$ 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 *$N$th divided difference* associated with these points is the cofficient of the monomial x^N in the polynomial p that uniquely solves the Hermite interpolation problem in Definition 2.33.

Corollary 2.36. The nth divided difference at $n+1$ "confluent" (i.e. identical) points is

$$f[x_0, \cdots, x_0] = \frac{1}{n!} f^{(n)}(x_0), \quad (2.37)$$

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

Proof. For the Hermite interpolation problem with $k = 0$ and $m_0 = n$, the unique solution is a polynomial of degree n. By Definition 2.35, we have $p^{(n)} \equiv n! f[x_0, \ldots, x_0]$, then the interpolation conditions yield

$$f^{(n)}(x_0) = p^{(n)}(x_0) = n! f[x_0, \ldots, x_0]. \qquad \square$$

Theorem 2.37. For the Hermite interpolation problem in Definition 2.33, denote $N = k + \sum_i m_i$. Denote by $p_N(f; x)$ the unique element of $\mathbb{P}_N$ for which (2.36) holds. Suppose $f^{(N+1)}(x)$ exists in (a, b). Then there exists some $\xi \in (a, b)$ such that

$$f(x) - p_N(f; x) = \frac{f^{(N+1)}(\xi)}{(N+1)!} \prod_{i=0}^{k} (x - x_i)^{m_i+1}. \quad (2.38)$$

Proof. The proof is similar to that of Theorem 2.7. Pay attention to the difference caused by the multiple roots of the polynomial $\prod_{i=0}^{k} (x - x_i)^{m_i+1}$. □

Example 2.38. For the Hermite interpolation problem

$$p(x_0) = f_0,\ p'(x_0) = f_0',\ p''(x_0) = f_0'',$$

Newton's formula yields the interpolating polynomial as

$$p(x) = f_0 + f_0'(x - x_0) + \frac{1}{2}f_0''(x - x_0)^2,$$

which is exactly the Taylor polynomial of degree 2. Thus a Taylor polynomial is a special case of a Hermite interpolating polynomial. By Theorem 2.37, the Cauchy remainder of this interpolation is

$$R_2(f;x) = f(x) - p_2(f;x) = \frac{f^{(3)}(\xi)}{6}(x - x_0)^3,$$

which is Lagrange's formula of the remainder term in Taylor's formula; see Theorem C.61.

Example 2.39. For the Hermite interpolation problem

$$p(x_0) = f_0,\ p(x_1) = f_1,\ p'(x_1) = f_1',\ p(x_2) = f_2,$$

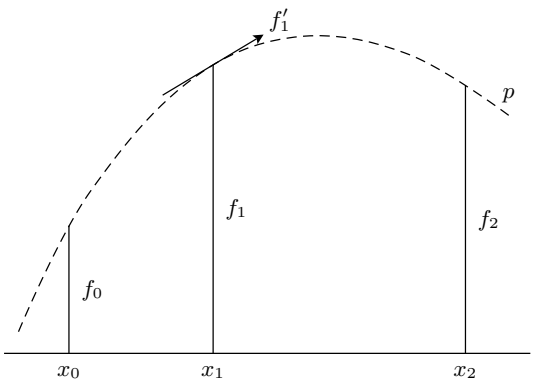
the table of divided differences has the form

$$\begin{array}{c|cccc} x_0 & f_0 & & & \\ x_1 & f_1 & f[x_0,x_1] & & \\ x_1 & f_1 & f_1' & f[x_0,x_1,x_1] & \\ x_2 & f_2 & f[x_1,x_2] & f[x_1,x_1,x_2] & f[x_0,x_1,x_1,x_2] \end{array}$$

and the interpolating polynomial follows from Newton's formula. By Theorem 2.37, the Cauchy remainder is

$$R_3(f;x) = \frac{f^{(4)}(\xi)}{4!}(x - x_0)(x - x_1)^2(x - x_2)$$

for some $\xi \in [x_0, x_2]$.

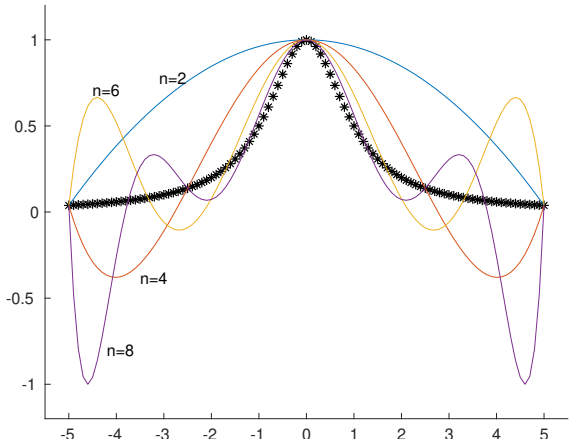


2.7 The Chebyshev polynomials

Example 2.40 (Runge phenomenon)**.** The points $x_0, x_1, \ldots, x_n$ in Theorem 2.5 are usually given *a priori*, e.g., as uniformly distributed over the interval $[x_0, x_n]$. As n increases, the degree of the interpolating polynomial also increases. Ideally we would like to have

$$\forall f \in \mathcal{C}[x_0, x_n], \forall x \in [x_0, x_n],\quad \lim_{n\to+\infty} p_n(f;x) = f(x). \tag{2.39}$$

However, this is not true for polynomial interpolation on equally spaced points. The famous Runge's example illustrates the violent oscillations at the end of the interval.



The above plot is created by interpolating

$$f(x) = \frac{1}{1 + x^2} \tag{2.40}$$

on $x_i = -5 + 10\frac{i}{n}$, $i = 0, 1, \ldots, n$ with $n = 2, 4, 6, 8$.

Definition 2.41. The *Chebyshev polynomial* of degree $n \in \mathbb{N}$ of the first kind is a polynomial $T_n : [-1, 1] \to [-1, 1]$,

$$T_n(x) = \cos(n \arccos x). \tag{2.41}$$

Theorem 2.42. The Chebyshev polynomials of the first kind satisfy the following recursive relations,

$$\forall n \in \mathbb{N}^+, \qquad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{2.42}$$

Proof. By trigonometric identities, we have

$$\cos(n+1)\theta = \cos n\theta \cos\theta - \sin n\theta \sin\theta,$$
$$\cos(n-1)\theta = \cos n\theta \cos\theta + \sin n\theta \sin\theta.$$

Adding up the two equations and setting $\cos\theta = x$ complete the proof. □

Corollary 2.43. The coefficient of x^n in T_n is 2^{n-1} for each $n > 0$.

Proof. Use (2.42) and $T_1 = x$ in an induction. □

Theorem 2.44. $T_n(x)$ has simple zeros at the n points

$$x_k = \cos\frac{2k-1}{2n}\pi, \tag{2.43}$$

where $k = 1, 2, \ldots, n$. For $x \in [-1, 1]$ and $n \in \mathbb{N}^+$, $T_n(x)$ has extreme values at the $n + 1$ points

$$x_k' = \cos\frac{k}{n}\pi, \qquad k = 0, 1, \ldots, n, \tag{2.44}$$

where it assumes the alternating values $(-1)^k$.

Proof. (2.41) and (2.43) yield

$$T_n(x_k) = \cos\left(n \arccos(\cos\frac{2k-1}{2n}\pi)\right) = \cos\left(\frac{2k-1}{2}\pi\right) = 0.$$

Differentiate (2.41) and we have

$$T_n'(x) = \frac{n}{\sqrt{1-x^2}} \sin(n \arccos x).$$

Then each x_k must be a simple zero since

$$T_n'(x_k) = \frac{n}{\sqrt{1-x_k^2}} \sin\left(\frac{2k-1}{2}\pi\right) \neq 0.$$

In contrast, $\forall k = 1, 2, \ldots, n-1$,

$$T_n'(x_k') = n\left(1 - \cos^2\frac{k\pi}{n}\right)^{-\frac{1}{2}} \sin(k\pi) = 0;$$

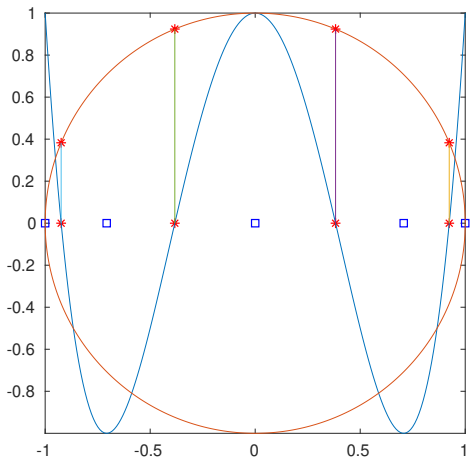
$$T_n''(x) = \frac{n^2 \cos(n \arccos(x))}{x^2-1} + \frac{n\,x \sin(n \arccos(x))}{(1-x^2)^{3/2}};$$

$$T_n''(x_k') \neq 0.$$

Hence a Taylor expansion of T_n yields

$$T_n(x_k' + \delta) = T_n(x_k') + \frac{1}{2}T_n''(x_k')\delta^2 + O(\delta^3),$$

and T_n must attain local extremes at each x_k'. For any $k = 0, 1, \ldots, n$, $T_n(x)$ attains its extreme values at x_k' since $T_n(x_0') = 1$, $T_n(x_1') = -1$, ..., and by (2.41) we have $|T_n(x)| \leq 1$. Clearly these are the only extrema of $T_n(x)$ on $[-1, 1]$. □



Exercise 2.45. Write a program to reproduce the above plot.

Theorem 2.46 (Chebyshev)**.** Denote by $\tilde{\mathbb{P}}_n$ the class of all polynomials of degree $n \in \mathbb{N}^+$ with leading coefficient 1. Then

$$\forall p \in \tilde{\mathbb{P}}_n, \qquad \max_{x\in[-1,1]} \left|\frac{T_n(x)}{2^{n-1}}\right| \leq \max_{x\in[-1,1]} |p(x)|. \tag{2.45}$$

Proof. By Theorem 2.44, $T_n(x)$ assumes its extrema $n+1$ times at the points x_k' defined in (2.44). Suppose (2.45) does not hold. Then Theorem 2.44 implies that

$$\exists p \in \tilde{\mathbb{P}}_n \text{ s.t. } \max_{x\in[-1,1]} |p(x)| < \frac{1}{2^{n-1}}. \tag{2.46}$$

Consider the polynomial $Q(x) = \frac{1}{2^{n-1}}T_n(x) - p(x)$.

$$Q(x_k') = \frac{(-1)^k}{2^{n-1}} - p(x_k'), \qquad k = 0, 1, \ldots, n.$$

By (2.46), $Q(x)$ has alternating signs at these $n+1$ points. Hence $Q(x)$ must have n zeros. However, by the construction of $Q(x)$, the degree of $Q(x)$ is at most $n-1$. Therefore, $Q(x) \equiv 0$ and $p(x) = \frac{1}{2^{n-1}}T_n(x)$, which implies $\max|p(x)| = \frac{1}{2^{n-1}}$. This is a contradiction to (2.46). □

Corollary 2.47. For $n \in \mathbb{N}^+$, we have

$$\max_{x\in[-1,1]} |x^n + a_1 x^{n-1} + \cdots + a_n| \geq \frac{1}{2^{n-1}}. \tag{2.47}$$

Corollary 2.48. Suppose polynomial interpolation is performed for f on the $n+1$ zeros of $T_{n+1}(x)$ as in Theorem 2.44. The Cauchy remainder in Theorem 2.7 satisfies

$$|R_n(f;x)| \leq \frac{1}{2^n(n+1)!} \max_{x\in[-1,1]} \left|f^{(n+1)}(x)\right|. \tag{2.48}$$

Proof. Theorem 2.7, Corollary 2.43, and Theorem 2.44 yield

$$|R_n(f;x)| = \frac{|f^{(n+1)}(\xi)|}{(n+1)!}\left|\prod_{i=0}^{n}(x-x_i)\right| = \frac{|f^{(n+1)}(\xi)|}{2^n(n+1)!}|T_{n+1}|.$$

Definition 2.41 completes the proof as $|T_{n+1}| \leq 1$. □

2.8 The Bernstein polynomials

Definition 2.49. The *Bernstein base polynomials* of degree $n \in \mathbb{N}^+$ relative to the unit interval $[0, 1]$ are

$$b_{n,k}(t) = \binom{n}{k} t^k (1-t)^{n-k} \tag{2.49}$$

where $k = 0, 1, \ldots, n$.

Lemma 2.50. The Bernstein base polynomials satisfy

$$\forall k = 0, 1, \ldots, n, \forall t \in (0, 1), \quad b_{n,k}(t) > 0 \tag{2.50a}$$

$$\sum_{k=0}^{n} b_{n,k}(t) = 1, \tag{2.50b}$$

$$\sum_{k=0}^{n} k b_{n,k}(t) = nt, \tag{2.50c}$$

$$\sum_{k=0}^{n} (k-nt)^2 b_{n,k}(t) = nt(1-t). \tag{2.50d}$$

Lemma 2.51. The Bernstein base polynomials of degree n form a basis of $\mathbb{P}_n$, the vector space of all polynomials with degree no more than n.

Proof. This follows from Definition 2.49. □

Definition 2.52. The *nth Bernstein polynomial of a map* $f \in \mathcal{C}[0, 1]$ is

$$(B_n f)(t) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_{n,k}(t), \tag{2.51}$$

where $b_{n,k}$ is a Bernstein base polynomial in (2.49).

Theorem 2.53 (Weierstrass approximation)**.** Every continuous function $f : [a, b] \to \mathbb{R}$ can be uniformly approximated as closely as desired by a polynomial function.

$$\forall f \in \mathcal{C}[a,b],\ \forall \epsilon > 0,\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > N,\\ \exists p_n \in \mathbb{P}_n \text{ s.t. } \forall x \in [a,b],\ |p_n(x) - f(x)| < \epsilon. \tag{2.52}$$

Proof. Without loss of generality, we assume $a = 0, b = 1$. Set $p_n = B_n f$ in (2.51). For any $\epsilon > 0$, there exist $\delta > 0$ and $n \in \mathbb{N}^+$ such that

$$\begin{aligned}
|(B_n f)(t) - f(t)| &= \left|(B_n f)(t) - f(t)\sum_{k=0}^{n} b_{n,k}(t)\right| \\
&\le \sum_{k=0}^{n}\left|f\left(\frac{k}{n}\right) - f(t)\right| b_{n,k}(t) \\
&= \left(\sum_{k:|\frac{k}{n}-t|<\delta} + \sum_{k:|\frac{k}{n}-t|\ge\delta}\right)\left|f\left(\frac{k}{n}\right) - f(t)\right| b_{n,k}(t) \\
&\le \sup_{|t-s|\le\delta}|f(t) - f(s)| + \frac{\|f\|_\infty}{2n\delta^2} \\
&\le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

where the case $|k - nt| < n\delta$ in the second inequality follows from (2.50a) and (2.50b), the other case $|k - nt| \ge n\delta$ in the second inequality follows from (2.50d) and

$$\begin{aligned}
\sum_{k:|\frac{k}{n}-t|\ge\delta} b_{n,k}(t) &\le \sum_{k:|\frac{k}{n}-t|\ge\delta} b_{n,k}(t)\frac{(k-nt)^2}{\delta^2 n^2} \\
&\le \sum_{k=0}^{n} b_{n,k}(t)\frac{(k-nt)^2}{\delta^2 n^2} = \frac{t(1-t)}{n\delta^2} \le \frac{1}{4n\delta^2},
\end{aligned}$$

and the last inequality follows from the uniform continuity of f (c.f. Theorem C.46) and the choice of $n > \frac{\|f\|_\infty}{\epsilon\delta^2}$. □

2.9 Problems

2.9.1 Theoretical questions

I. For $f \in \mathcal{C}^2[x_0, x_1]$ and $x \in (x_0, x_1)$, linear interpolation of f at x_0 and x_1 yields

$$f(x) - p_1(f; x) = \frac{f''(\xi(x))}{2}(x - x_0)(x - x_1).$$

Consider the case $f(x) = \frac{1}{x}$, $x_0 = 1$, $x_1 = 2$.

- Determine $\xi(x)$ explicitly.
- Extend the domain of ξ continuously from (x_0, x_1) to $[x_0, x_1]$. Find $\max \xi(x)$, $\min \xi(x)$, and $\max f''(\xi(x))$.

II. Let $\mathbb{P}_m^+$ be the set of all polynomials of degree $\le m$ that are non-negative on the real line,

$$\mathbb{P}_m^+ = \{p : p \in \mathbb{P}_m,\ \forall x \in \mathbb{R},\ p(x) \ge 0\}.$$

Find $p \in \mathbb{P}_{2n}^+$ such that $p(x_i) = f_i$ for $i = 0, 1, \ldots, n$ where $f_i \ge 0$ and x_i are distinct points on $\mathbb{R}$.

III. Consider $f(x) = e^x$.

- Prove by induction that

$$\forall t \in \mathbb{R}, \qquad f[t, t+1, \ldots, t+n] = \frac{(e-1)^n}{n!}e^t.$$

- From Corollary 2.22 we know

$$\exists \xi \in (0, n) \text{ s.t. } f[0, 1, \ldots, n] = \frac{1}{n!}f^{(n)}(\xi).$$

Determine ξ from the above two equations. Is ξ located to the left or to the right of the midpoint $n/2$?

IV. Consider $f(0) = 5$, $f(1) = 3$, $f(3) = 5$, $f(4) = 12$.

- Use the Newton formula to obtain $p_3(f; x)$;
- The data suggest that f has a minimum in $x \in (1, 3)$. Find an approximate value for the location $x_{\min}$ of the minimum.

V. Consider $f(x) = x^7$.

- Compute $f[0, 1, 1, 1, 2, 2]$.
- We know that this divided difference is expressible in terms of the 5th derivative of f evaluated at some $\xi \in (0, 2)$. Determine ξ.

VI. f is a function on $[0, 3]$ for which one knows that

$$f(0) = 1, f(1) = 2, f'(1) = -1, f(3) = f'(3) = 0.$$

- Estimate $f(2)$ using Hermite interpolation.
- Estimate the maximum possible error of the above answer if one knows, in addition, that $f \in \mathcal{C}^5[0, 3]$ and $|f^{(5)}(x)| \le M$ on $[0, 3]$. Express the answer in terms of M.

VII. Define forward difference by

$$\begin{aligned}
\Delta f(x) &= f(x + h) - f(x), \\
\Delta^{k+1} f(x) &= \Delta\Delta^k f(x) = \Delta^k f(x + h) - \Delta^k f(x)
\end{aligned}$$

and backward difference by

$$\begin{aligned}
\nabla f(x) &= f(x) - f(x - h), \\
\nabla^{k+1} f(x) &= \nabla\nabla^k f(x) = \nabla^k f(x) - \nabla^k f(x - h).
\end{aligned}$$

Prove

$$\begin{aligned}
\Delta^k f(x) &= k! h^k f[x_0, x_1, \ldots, x_k], \\
\nabla^k f(x) &= k! h^k f[x_0, x_{-1}, \ldots, x_{-k}],
\end{aligned}$$

where $x_j = x + jh$.

VIII. Assume f is differentiable at x_0. Prove

$$\frac{\partial}{\partial x_0} f[x_0, x_1, \ldots, x_n] = f[x_0, x_0, x_1, \ldots, x_n].$$

What about the partial derivative with respect to one of the other variables?

IX. A min-max problem.
For $n \in \mathbb{N}^+$, determine

$$\min \max_{x\in[a,b]} |a_0x^n + a_1x^{n-1} + \cdots + a_n|,$$

where $a_0 \neq 0$ is fixed and the minimum is taken over all $a_i \in \mathbb{R}$, $i = 1, 2, \cdots, n$.

X. Imitate the proof of Chebyshev Theorem.
Express the Chebyshev polynomial of degree $n \in \mathbb{N}$ as a polynomial T_n and change its domain from $[-1, 1]$ to $\mathbb{R}$. For a fixed $a > 1$, define $\mathbb{P}_n^a := \{p \in \mathbb{P}_n : p(a) = 1\}$ and a polynomial $\hat{p}_n(x) \in \mathbb{P}_n^a$,

$$\hat{p}_n(x) := \frac{T_n(x)}{T_n(a)}.$$

Prove

$$\forall p \in \mathbb{P}_n^a, \qquad \|\hat{p}_n\|_\infty \leq \|p\|_\infty$$

where the max-norm of a function $f : \mathbb{R} \to \mathbb{R}$ is defined as $\|f\|_\infty = \max_{x\in[-1,1]} |f(x)|$.

XI. Prove Lemma 2.50.

2.9.2 Programming assignments

A. Implement the Newton formula in a subroutine that produces the value of the interpolation polynomial $p_n(f; x_0, x_1, \ldots, x_n; x)$ at any real x, where $n \in \mathbb{N}^+$, x_i's are distinct, and f is a function assumed to be available in the form of a subroutine.

B. Run your routine on the function

$$f(x) = \frac{1}{1 + x^2}$$

for $x \in [-5, 5]$ using $x_i = -5 + 10\frac{i}{n}$, $i = 0, 1, \ldots, n$, and $n = 2, 4, 6, 8$. Plot the polynomials against the exact function to reproduce the plot in the notes that illustrate the Runge phenomenon.

C. Reuse your subroutine of Newton interpolation to perform Chebyshev interpolation for the function

$$f(x) = \frac{1}{1 + 25x^2}$$

for $x \in [-1, 1]$ on the zeros of Chebyshev polynomials T_n with $n = 5, 10, 15, 20$. Clearly the Runge function $f(x)$ is a scaled version of the function in B. Plot the interpolating polynomials against the exact function to observe that the Chebyshev interpolation is free of the wide oscillations in the previous assignment.

D. A car traveling along a straight road is clocked at a number of points. The data from the observations are given in the following table, where the time is in seconds, the displacement is in feet, and the velocity is in feet per second.

Time	0	3	5	8	13
displacement	0	225	383	623	993
velocity	75	77	80	74	72

(a) Use a Hermite polynomial to predict the position of the car and its speed for $t = 10$ seconds.

(b) Use the derivative of the Hermite polynomial to determine whether the car ever exceeds the speed limit of 55 miles per hour, i.e., 81 feet per second.

E. It is suspected that the high amounts of tannin in mature oak leaves inhibit the growth of the winter moth larvae that extensively damage these trees in certain years. The following table lists the average weight of two samples of larvae at times in the first 28 days after birth. The first sample was reared on young oak leaves, whereas the second sample was reared on mature leaves from the same tree.

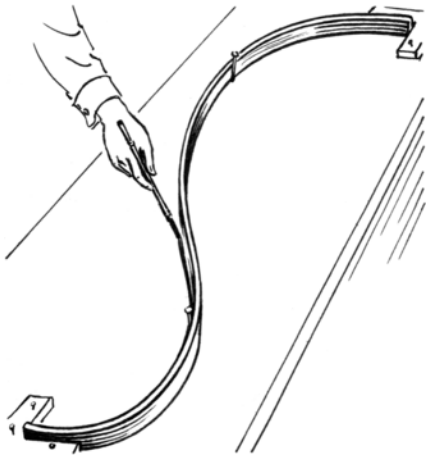
Day	0	6	10	13	17	20	28
Sp1	6.67	17.3	42.7	37.3	30.1	29.3	28.7
Sp2	6.67	16.1	18.9	15.0	10.6	9.44	8.89

(a) Use Newton's formula to approximate the average weight curve for each sample.

(b) Predict whether the two samples of larvae will die after another 15 days.

Chapter 3

Splines



3.1 Piecewise-polynomial splines

Definition 3.1. Given nonnegative integers n, k, and a strictly increasing sequence $\{x_i\}$ that partitions $[a,b]$,

$$a = x_1 < x_2 < \cdots < x_N = b, \tag{3.1}$$

the set of *spline functions of degree n and smoothness class k* relative to the partition $\{x_i\}$ is

$$\mathbb{S}_n^k = \big\{ s : \ s \in \mathcal{C}^k[a,b];\ \forall i \in [1, N-1],\ s|_{[x_i,x_{i+1}]} \in \mathbb{P}_n \big\}. \tag{3.2}$$

The x_i's are called *knot*s of the spline.

Notation 1. In Chapter 2 and this chapter, the polynomial degree is denoted by n for all methods. Here we use N to denote the number of knots for a spline.

Example 3.2. As an extreme, $\mathbb{S}_n^n = \mathbb{P}_n$, i.e. all the pieces of $s \in \mathbb{S}_n^n$ belong to a single polynomial. On the other end, $\mathbb{S}_1^0$ is the class of piecewise linear interpolating functions. The most popular splines are the cubic splines in $\mathbb{S}_3^2$.

Lemma 3.3. Denote $m_i = s'(f; x_i)$ for $s \in \mathbb{S}_3^2$. Then, for each $i = 2, 3, \ldots, N-1$, we have

$$\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = 3\mu_i f[x_i, x_{i+1}] + 3\lambda_i f[x_{i-1}, x_i], \tag{3.3}$$

where

$$\mu_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}, \qquad \lambda_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}. \tag{3.4}$$

Proof. Denote $p_i(x) = s|_{[x_i,x_{i+1}]}$ and $K_i = f[x_i, x_{i+1}]$. The table of divided difference for the Hermite interpolation problem $p_i(x_i) = f_i$, $p_i(x_{i+1}) = f_{i+1}$, $p_i'(x_i) = m_i$, $p_i'(x_{i+1}) = m_{i+1}$ is

$$\begin{array}{c|cccc} x_i & f_i & & & \\ x_i & f_i & m_i & & \\ x_{i+1} & f_{i+1} & K_i & \frac{K_i - m_i}{x_{i+1}-x_i} & \\ x_{i+1} & f_{i+1} & m_{i+1} & \frac{m_{i+1}-K_i}{x_{i+1}-x_i} & \frac{m_i + m_{i+1} - 2K_i}{(x_{i+1}-x_i)^2} \end{array}$$

Then the Newton formula yields

$$\begin{aligned} p_i(x) =& f_i + (x - x_i)m_i + (x - x_i)^2 \frac{K_i - m_i}{x_{i+1} - x_i} \\ &+ (x - x_i)^2(x - x_{i+1}) \frac{m_i + m_{i+1} - 2K_i}{(x_{i+1} - x_i)^2}, \end{aligned} \tag{3.5}$$

or equivalently

$$\left\{ \begin{array}{ll} p_i(x) & = c_{i,0} + c_{i,1}(x - x_i) + c_{i,2}(x - x_i)^2 + c_{i,3}(x - x_i)^3, \\ c_{i,0} & = f_i, \\ c_{i,1} & = m_i, \\ c_{i,2} & = \frac{3K_i - 2m_i - m_{i+1}}{x_{i+1} - x_i}, \\ c_{i,3} & = \frac{m_i + m_{i+1} - 2K_i}{(x_{i+1} - x_i)^2}. \end{array} \right. \tag{3.6}$$

$s \in \mathcal{C}^2$ implies that $p_{i-1}''(x_i) = p_i''(x_i)$, i.e.

$$3c_{i-1,3}(x_i - x_{i-1}) = c_{i,2} - c_{i-1,2}.$$

The substitution of the coefficients $c_{i,j}$ into the above equation yields (3.3). □

Lemma 3.4. Denote $M_i = s''(f; x_i)$ for $s \in \mathbb{S}_3^2$. Then, for each $i = 2, 3, \ldots, N-1$, we have

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = 6f[x_{i-1}, x_i, x_{i+1}] \tag{3.7}$$

where μ_i and λ_i are the same as those in (3.4).

Proof. Taylor expansion of $s(x)$ at x_i yields

$$s(x) = f_i + s'(x_i)(x - x_i) + \frac{M_i}{2}(x - x_i)^2 + \frac{s'''(x_i)}{6}(x - x_i)^3, \tag{3.8}$$

where $x \in [x_i, x_{i+1}]$ and the derivatives should be interpreted as the right-hand derivatives. Differentiate (3.8) twice, set $x = x_{i+1}$, and we have

$$s'''(x_i) = \frac{M_{i+1} - M_i}{x_{i+1} - x_i}. \tag{3.9}$$

Substitute (3.9) into (3.8), set $x = x_{i+1}$, and we have

$$s'(x_i) = f[x_i, x_{i+1}] - \frac{1}{6}(M_{i+1} + 2M_i)(x_{i+1} - x_i). \tag{3.10}$$

Similarly, differentiate (3.8) twice, set $x = x_{i-1}$, and we have $s'''(x_i) = \frac{M_{i-1} - M_i}{x_{i-1} - x_i}$. Its substitution into (3.8) yields

$$s'(x_i) = f[x_{i-1}, x_i] - \frac{1}{6}(M_{i-1} + 2M_i)(x_{i-1} - x_i). \tag{3.11}$$

The subtraction of (3.10) from (3.11) yields (3.7). □

Definition 3.5 (Boundary conditions of cubic splines)**.** Common cubic splines include the following.

- A *complete cubic spline* $s \in \mathbb{S}_3^2$ satisfies boundary conditions $s'(f; a) = f'(a)$ and $s'(f; b) = f'(b)$.
- A *cubic spline with specified second derivatives at its end points*: $s''(f; a) = f''(a)$ and $s''(f; b) = f''(b)$.
- A *natural cubic spline* $s \in \mathbb{S}_3^2$ satisfies boundary conditions $s''(f; a) = 0$ and $s''(f; b) = 0$.
- A *not-a-knot cubic spline* $s \in \mathbb{S}_3^2$ satisfies that $s'''(f; x)$ exists at $x = x_2$ and $x = x_{N-1}$.
- A *periodic cubic spline* $s \in \mathbb{S}_3^2$ is obtained from replacing $s(f; b) = f(b)$ with $s(f; b) = s(f; a)$, $s'(f; b) = s'(f; a)$, and $s''(f; b) = s''(f; a)$.

Lemma 3.6. For a complete cubic spline $s \in \mathbb{S}_3^2$, denote $M_i = s''(f; x_i)$ and we have

$$2M_1 + M_2 = 6f[x_1, x_1, x_2], \tag{3.12}$$
$$M_{N-1} + 2M_N = 6f[x_{N-1}, x_N, x_N]. \tag{3.13}$$

Proof. As for (3.12), the cubic polynomial on $[x_1, x_2]$ can be written as

$$\begin{aligned} s_1(x) =& f[x_1] + f[x_1, x_1](x - x_1) \\ &+ \frac{M_1}{2}(x - x_1)^2 + \frac{s_1'''(x_1)}{6}(x - x_1)^3. \end{aligned}$$

Differentiate the above equation twice, replace x with x_2, and we have $s_1'''(x_1) = \frac{M_2 - M_1}{x_2 - x_1}$, which implies

$$\begin{aligned} s_1(x) =& f[x_1] + f[x_1, x_1](x - x_1) \\ &+ \frac{M_1}{2}(x - x_1)^2 + \frac{M_2 - M_1}{6(x_2 - x_1)}(x - x_1)^3. \end{aligned} \tag{3.14}$$

Set $x = x_2$, divide both sides by $x_2 - x_1$, and we have

$$f[x_1, x_2] = f[x_1, x_1] + \left(\frac{M_1}{2} + \frac{M_2 - M_1}{6}\right)(x_2 - x_1),$$

which yields (3.12). (3.13) can be proven similarly. □

Theorem 3.7. For a given function $f : [a, b] \to \mathbb{R}$, there exists a unique complete/natural/periodic cubic spline $s(f; x)$ that interpolates f.

Proof. We only prove the case of complete cubic splines since the other cases are similar.

By the proof of Lemma 3.3, s is uniquely determined if all the m_i's are uniquely determined on all intervals. For a complete cubic spline we already have $m_1 = f'(a)$ and $m_N = f'(b)$. Assemble (3.3) into a linear system

$$\begin{bmatrix} 2 & \mu_2 & & & & & \\ \lambda_3 & 2 & \mu_3 & & & & \\ & & \ddots & & & & \\ & & \lambda_i & 2 & \mu_i & & \\ & & & & \ddots & & \\ & & & & \lambda_{N-2} & 2 & \mu_{N-2} \\ & & & & & \lambda_{N-1} & 2 \end{bmatrix} \begin{bmatrix} m_2 \\ m_3 \\ \vdots \\ m_i \\ \vdots \\ m_{N-2} \\ m_{N-1} \end{bmatrix} = \mathbf{b}, \tag{3.15}$$

where the vector $\mathbf{b}$ is determined from the known information. (3.4) implies that the matrix in the above equation is strictly diagonally dominant. Therefore its determinant is nonzero and the m_i's can be uniquely determined.

Alternatively, a complete cubic spline can be uniquely determined from Lemmas 3.4 and 3.6, following arguments similar to the above. □

Example 3.8. Construct a complete cubic spline $s(x)$ on points $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 6$ from the function values of $f(x) = \ln(x)$ and its derivatives at x_1 and x_5. Approximate $\ln(5)$ by $s(5)$.

From the given conditions, we set up the table of divided differences as follows.

x_i	$f[x_i]$		
1	0		
1	0	1	
2	0.6931	0.6931	−0.3069
3	1.0986	0.4055	−0.1438
4	1.3863	0.2877	−0.05889
6	1.7918	0.2027	−0.02831
6	1.7918	0.1667	−0.01803

All values of λ_i and μ_i are $\frac{1}{2}$ except that

$$\lambda_4 = \frac{2}{3},\ \mu_4 = \frac{1}{3}.$$

Then Lemma 3.4 and Lemma 3.6 yield a linear system

$$\begin{bmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & 1 & 4 & 1 & \\ & & 1 & 6 & 2 \\ & & & 1 & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \end{bmatrix} \approx \begin{bmatrix} -1.84112 \\ -1.72610 \\ -0.70670 \\ -0.50967 \\ -0.10820 \end{bmatrix},$$

where elements in the right-hand side (RHS) vector are obtained from the last column of the table of divided differences by multiplying $6, 12, 12, 18$, and 6. Why? Solve the linear system and we have all the M_i's. Then we derive an expression of the spline on the last interval following the procedures similar to those for (3.14). After this expression is obtained, we then evaluate it and obtain $s(5) \approx 1.60977$. In comparison, $\ln(5) \approx 1.60944$.

3.2 The minimum properties

Theorem 3.9 (Minimum bending energy)**.** For any function $g \in \mathcal{C}^2[a,b]$ that satisfies $g'(a) = f'(a)$, $g'(b) = f'(b)$, and $g(x_i) = f(x_i)$ for each $i = 1, 2, \ldots, N$, the complete cubic spline $s = s(f;x)$ satisfies

$$\int_a^b [s''(x)]^2 \mathrm{d}x \le \int_a^b [g''(x)]^2 \mathrm{d}x, \tag{3.16}$$

where the equality holds only when $g(x) = s(f;x)$.

Proof. Define $\eta(x) = g(x) - s(x)$. From the given conditions we have $\eta \in \mathcal{C}^2[a,b]$, $\eta'(a) = \eta'(b) = 0$, and $\forall i = 1, 2, \ldots, N$, $\eta(x_i) = 0$. Then

$$\begin{aligned}&\int_a^b [g''(x)]^2 \mathrm{d}x = \int_a^b [s''(x) + \eta''(x)]^2 \mathrm{d}x \\ =&\int_a^b [s''(x)]^2 \mathrm{d}x + \int_a^b [\eta''(x)]^2 \mathrm{d}x + 2\int_a^b s''(x)\eta''(x)\mathrm{d}x.\end{aligned}$$

From

$$\begin{aligned}&\int_a^b s''(x)\eta''(x)\mathrm{d}x = \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} s''(x)\mathrm{d}\eta' \\ =&\sum_{i=1}^{N-1} s''(x)\eta'(x)|_{x_i}^{x_{i+1}} - \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} \eta'(x)s'''(x)\mathrm{d}x \\ =&s''(b)\eta'(b) - s''(a)\eta'(a) - \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} s'''(x)\mathrm{d}\eta \\ =&-\sum_{i=1}^{N-1} s'''(x)\eta(x)|_{x_i}^{x_{i+1}} + \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} \eta(x)s^{(4)}(x)\mathrm{d}x \\ =&0,\end{aligned}$$

we have

$$\int_a^b [g''(x)]^2 \mathrm{d}x = \int_a^b [s''(x)]^2 \mathrm{d}x + \int_a^b [\eta''(x)]^2 \mathrm{d}x,$$

which completes the proof. □

Theorem 3.10 (Minimum bending energy)**.** For any function $g \in \mathcal{C}^2[a,b]$ with $g(x_i) = f(x_i)$ for each $i = 1, 2, \ldots, N$, the natural cubic spline $s = s(f;x)$ satisfies

$$\int_a^b [s''(x)]^2 \mathrm{d}x \le \int_a^b [g''(x)]^2 \mathrm{d}x, \tag{3.17}$$

where the equality holds only when $g(x) = s(f;x)$.

Proof. The proof is similar to that of Theorem 3.9. Although $\eta'(a) = \eta'(b) = 0$ does not hold, we do have $s''(a) = s''(b) = 0$. □

Lemma 3.11. Suppose a $\mathcal{C}^2$ function $f : [a,b] \to \mathbb{R}$ is interpolated by a complete cubic spline or a cubic spline with specified second derivatives at its end points. Then

$$\forall x \in [a,b], \qquad |s''(x)| \le 3 \max_{x\in[a,b]} |f''(x)|. \tag{3.18}$$

Proof. Since $s''(x)$ is linear on $[x_i, x_{i+1}]$, $|s''(x)|$ attains its maximum at x_j for some j. If $j = 2, \ldots, N-1$, it follows from Lemma 3.4 and Corollary 2.22 that

$$\begin{aligned}&2M_j = 6f[x_{j-1}, x_j, x_{j+1}] - \mu_j M_{j-1} - \lambda_j M_{j+1} \\ \Rightarrow\ &2|M_j| \le 6\,|f[x_{j-1}, x_j, x_{j+1}]| + (\mu_j + \lambda_j)|M_j| \\ \Rightarrow\ &\exists \xi \in (x_{j-1}, x_{j+1}) \ \text{ s.t. } |M_j| \le 3\,|f''(\xi)| \\ \Rightarrow\ &|s''(x)| \le 3 \max_{x\in[a,b]} |f''(x)|.\end{aligned} \tag{3.19}$$

If $|s''(x)|$ attains its maximum at x_1 or x_N, (3.19) clearly holds at these end points for a cubic spline with specified second derivatives. After all, $s''(a) = f''(a)$ and $s''(b) = f''(b)$. As for the complete cubic spline, it suffices to prove (3.19) when $|s''(x)|$ attains its maximum at x_1. Since the first derivative $f'(a) = f[x_1, x_1]$ is specified, $f[x_1, x_1, x_2]$ is a constant. By (3.12), we have

$$2|M_1| \le 6|f[x_1, x_1, x_2]| + |M_2| \le 6|f[x_1, x_1, x_2]| + |M_1|$$

which, together with Corollary 2.22, implies

$$\exists \xi \in (x_1, x_2) \text{ s.t. } |M_1| \le 3|f''(\xi)|.$$

This completes the proof. □

3.3 Error analysis

Theorem 3.12. Suppose a $\mathcal{C}^4$ function $f : [a,b] \to \mathbb{R}$ is interpolated by a complete cubic spline or a cubic spline with specified second derivatives at its end points. Then

$$\forall j = 0, 1, 2, \quad \left|f^{(j)}(x) - s^{(j)}(x)\right| \le c_j h^{4-j} \max_{x\in[a,b]} \left|f^{(4)}(x)\right|, \tag{3.20}$$

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

Proof. Our plan is to first prove the case of $j = 2$, then utilize the conclusion to study other cases.

Consider an auxiliary function $\hat{s} \in \mathcal{C}^2[a,b]$ that satisfies

$$\forall i = 1, 2, \ldots, N-1, \qquad \begin{cases} \hat{s}|_{[x_i, x_{i+1}]} \in \mathbb{P}_3; \\ \hat{s}''(x_i) = f''(x_i) \end{cases}$$

and $\hat{s}''(x_N) = f''(x_N)$. We can obtain such an $\hat{s}$ by interpolating $f''(x)$ with some $\tilde{s} \in \mathbb{S}_1^0$ and integrating $\tilde{s}$ twice. Then the theorem of Cauchy remainder (Theorem 2.7) implies

$$\begin{aligned}&\exists \xi_i \in [x_i, x_{i+1}], \ \text{ s.t. } \forall x \in [x_i, x_{i+1}], \\ &|f''(x) - \tilde{s}(x)| \le \frac{1}{2}\left|f^{(4)}(\xi_i)\right| |(x - x_i)(x - x_{i+1})|,\end{aligned}$$

hence we have

$$|f''(x) - \hat{s}''(x)|_{x\in[x_i,x_{i+1}]} \le \frac{1}{8} \max_{x\in[x_i,x_{i+1}]} \left|f^{(4)}(x)\right| (x_{i+1} - x_i)^2$$

and thus

$$|f''(x) - \hat{s}''(x)| \le \frac{h^2}{8} \max_{x\in[a,b]} |f^{(4)}(x)|. \tag{3.21}$$

Now consider interpolating $f(x) - \hat{s}(x)$ with a cubic spline. Since $\hat{s}(x) \in \mathbb{S}_3^2$, the interpolant must be $s(x) - \hat{s}(x)$. Then Lemma 3.11 yields

$$\forall x \in [a,b], \qquad |s''(x) - \hat{s}''(x)| \le 3 \max_{x\in[a,b]} |f''(x) - \hat{s}''(x)|,$$

which, together with (3.21), leads to (3.20) for $j = 2$:

$$\begin{aligned}|f''(x) - s''(x)| &\le |f''(x) - \hat{s}''(x)| + |\hat{s}''(x) - s''(x)| \\ &\le 4 \max_{x\in[a,b]} |f''(x) - \hat{s}''(x)| \\ &\le \frac{1}{2}h^2 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|. \end{aligned} \tag{3.22}$$

For $j = 0$, we have $f(x) - s(x) = 0$ for $x = x_i, x_{i+1}$. Then Rolle's theorem C.52 implies $f'(\xi_i) - s'(\xi_i) = 0$ for some $\xi_i \in [x_i, x_{i+1}]$. It follows from the second fundamental theorem of calculus (Theorem C.74) that

$$\forall x \in [x_i, x_{i+1}], \qquad f'(x) - s'(x) = \int_{\xi_i}^{x} (f''(t) - s''(t))\,\mathrm{d}t,$$

which, together with the integral mean value theorem C.72 and (3.22), yields

$$\begin{aligned}|f'(x) - s'(x)|_{x\in[x_i,x_{i+1}]} &= |x - \xi_i|\,|f''(\eta_i) - s''(\eta_i)| \\ &\le \frac{1}{2}h^3 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|. \end{aligned}$$

This proves (3.20) for $j = 1$. Finally, consider interpolating $f(x)-s(x)$ with some linear spline $\bar{s} \in \mathbb{S}_1^0$. The interpolation conditions dictate $\forall x \in [a, b]$, $\bar{s}(x) \equiv 0$. Hence

$$\begin{aligned}&|f(x) - s(x)|_{x\in[x_i,x_{i+1}]} = |f(x) - s(x) - \bar{s}|_{x\in[x_i,x_{i+1}]} \\ \le&\frac{1}{8}(x_{i+1} - x_i)^2 \max_{x\in[x_i,x_{i+1}]} |f''(x) - s''(x)| \\ \le&\frac{1}{16}h^4 \max_{x\in[a,b]} |f^{(4)}(x)|,\end{aligned}$$

where the second step follows from Theorem 2.7 and the third step from (3.22). □

Exercise 3.13. Verify Theorem 3.12 using the results in Example 3.8.

3.4 B-splines

Notation 2. In the notation $\mathbb{S}_n^{n-1}(t_1, t_2, \cdots, t_N)$, t_i's in the parentheses represent knots of a spline. When there is no danger of ambiguity, we also use the shorthand notation $\mathbb{S}_{n,N}^{n-1} := \mathbb{S}_n^{n-1}(t_1, t_2, \cdots, t_N)$ or simply $\mathbb{S}_n^{n-1}$.

Theorem 3.14. The set of splines $\mathbb{S}_n^{n-1}(t_1, t_2, \cdots, t_N)$ is a linear space with dimension $n + N - 1$.

Proof. It is easy to verify from (3.2) and Definition B.2 that $\mathbb{S}_n^{n-1}(t_1, t_2, \cdots, t_N)$ is indeed a linear space. Note that the additive identity is the zero function not the number zero. One polynomial of degree n is determined by $n + 1$ coefficients. The $N - 1$ intervals lead to $(N - 1)(n + 1)$ coefficients. At each of the $N - 2$ interval knots, the smoothness condition requires that the 0th, 1st, ..., $(n - 1)$th derivatives of adjacent polynomials match. Hence the dimension is $(N - 1)(n + 1) - n(N - 2) = n + N - 1$. □

Example 3.15. The cubic splines in Definition 3.5, have $n = 3$ and hence the dimension of $\mathbb{S}_3^2$ is $N + 2$. Apart from the N interpolation conditions at the knots, we need to impose two other conditions at the ends of the interpolating interval to obtain a unique spline, this leads to different types of cubic splines in Definition 3.5.

3.4.1 Truncated power functions

Definition 3.16. The *truncated power function* with exponent n is defined as

$$x_+^n = \begin{cases} x^n & \text{if } x \ge 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{3.23}$$

Example 3.17. According to Definition 3.16, we have

$$\forall t \in [a, b], \quad \int_a^b (t - x)_+^n \mathrm{d}x = \int_a^t (t - x)^n \mathrm{d}x = \frac{(t - a)^{n+1}}{n + 1}. \tag{3.24}$$

Lemma 3.18. The following is a basis of $\mathbb{S}_n^{n-1}(t_1, \ldots, t_N)$,

$$1, x, x^2, \ldots, x^n, (x-t_2)_+^n, (x-t_3)_+^n, \ldots, (x-t_{N-1})_+^n. \tag{3.25}$$

Proof. $\forall i = 2, 3, \ldots, N - 1$, $(x - t_i)_+^n \in \mathbb{S}_{n,N}^{n-1}$. Also, $\forall i = 0, 1, \ldots, n$, $x^i \in \mathbb{S}_{n,N}^{n-1}$. Suppose

$$\sum_{i=0}^{n} a_i x^i + \sum_{j=2}^{N-1} a_{n+j}(x - t_j)_+^n = \mathbf{0}(x). \tag{3.26}$$

To satisfy (3.26) for all $x < t_2$, a_i must be 0 for each $i = 0, 1, \cdots, n$. To satisfy (3.26) for all $x \in (t_2, t_3)$, a_{n+2} must be 0. Similarly, all a_{n+j}'s must be zero. Hence, the functions in (3.25) are linearly independent by Definition B.25. The proof is completed by Theorem 3.14, Lemma B.41, and the fact that there are $n + N - 1$ functions in (3.25). □

Corollary 3.19. Any $s \in \mathbb{S}_{n,N}^{n-1}$ can be expressed as

$$s(x) = \sum_{i=0}^{n} a_i (x - t_1)^i + \sum_{j=2}^{N-1} a_{n+j}(x - t_j)_+^n, \qquad x \in [t_1, t_N]. \tag{3.27}$$

Proof. By Lemma 3.18, it suffices to point out that

$$\operatorname{span}\{1, x, \ldots, x^n\} = \operatorname{span}\{1, (x - t_1), \ldots, (x - t_1)^n\}. \quad \square$$

Example 3.20. (3.27) with $n = 1$ is the linear spline interpolation. Imagine a plastic rod that is initially straight. Place one of its end at (t_1, f_1) and let it go through (t_2, f_2). In general (t_3, f_3) will be off the rod, but we can bend the rod at (t_2, f_2) to make the rod go through (t_3, f_3). This "bending" process corresponds to adding the first truncated power function in (3.27).

3.4.2 The local support of B-splines

Definition 3.21. The *hat function* at t_i is

$$\hat{B}_i(x) = \begin{cases} \frac{x-t_{i-1}}{t_i-t_{i-1}} & x \in (t_{i-1}, t_i], \\ \frac{t_{i+1}-x}{t_{i+1}-t_i} & x \in (t_i, t_{i+1}], \\ 0 & \text{otherwise.} \end{cases} \tag{3.28}$$

Theorem 3.22. The hat functions form a basis of $\mathbb{S}_1^0$.

Proof. By Definition 3.21, we have

$$\hat{B}_i(t_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{3.29}$$

Suppose $\sum_{i=1}^N c_i \hat{B}_i(x) = \mathbf{0}(x)$. Then we have $c_i = 0$ for each $i = 1, 2, \cdots, N$ by setting $x = t_j$ and applying (3.29). Hence by Definition B.25 these N hat functions are linearly independent. The proof is completed by Theorem 3.14 and Lemma B.41. □

Definition 3.23. *B-splines* are defined recursively by

$$B_i^{n+1}(x) = \frac{x - t_{i-1}}{t_{i+n} - t_{i-1}} B_i^n(x) + \frac{t_{i+n+1} - x}{t_{i+n+1} - t_i} B_{i+1}^n(x). \tag{3.30}$$

The recursion base is the B-spline of degree zero,

$$B_i^0(x) = \begin{cases} 1 & \text{if } x \in (t_{i-1}, t_i], \\ 0 & \text{otherwise.} \end{cases} \tag{3.31}$$

Example 3.24. The hat functions in Definition 3.21 are clearly the B-splines of degree one:

$$B_i^1 = \hat{B}_i. \tag{3.32}$$

In (3.30), B-splines of higher degrees are defined by generalizing the idea of hat functions.

Example 3.25. The quadratic B-splines $B_i^2(x) =$

$$\begin{cases} \frac{(x-t_{i-1})^2}{(t_{i+1}-t_{i-1})(t_i-t_{i-1})}, & x \in (t_{i-1}, t_i]; \\ \frac{(x-t_{i-1})(t_{i+1}-x)}{(t_{i+1}-t_{i-1})(t_{i+1}-t_i)} + \frac{(t_{i+2}-x)(x-t_i)}{(t_{i+2}-t_i)(t_{i+1}-t_i)}, & x \in (t_i, t_{i+1}]; \\ \frac{(t_{i+2}-x)^2}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & x \in (t_{i+1}, t_{i+2}]; \\ 0, & \text{otherwise.} \end{cases} \tag{3.33}$$

Definition 3.26. The *support* of a function $f : X \to \mathbb{R}$ is

$$\text{supp}(f) = \text{closure}\{x \in X \mid f(x) \neq 0\}. \tag{3.34}$$

Lemma 3.27 (Support of B-splines)**.** For $n \in \mathbb{N}^+$, the interval of support of B_i^n is $[t_{i-1}, t_{i+n}]$ where

$$\forall x \in (t_{i-1}, t_{i+n}),\ B_i^n(x) > 0. \tag{3.35}$$

Proof. This is an easy induction by (3.31) and (3.30). □

Definition 3.28. Let X be a vector space. For each $x \in X$ we associate a unique real (or complex) number $L(x)$. If $\forall x, y \in X$ and $\forall \alpha, \beta \in \mathbb{R}$ (or $\mathbb{C}$), we have

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y), \tag{3.36}$$

then L is called a *linear functional* over X.

Example 3.29. $X = \mathcal{C}[a, b]$, then the elements of X are functions continuous over $[a, b]$.

$$L(f) = \int_a^b f(x)\mathrm{d}x, \qquad L(f) = \int_a^b x^2 f(x)\mathrm{d}x$$

are both linear functionals over X.

Notation 3. We have used the notation $f[x_0, \ldots, x_k]$ for the kth divided difference of f, inline with considering $f[x_0, \ldots, x_k]$ as a generalization of the Taylor expansion. Hereafter, for analyzing B-splines, it is both semantically and syntactically better to use the notation $[x_0, \ldots, x_k]f$, inline with considering the *procedures* of a divided difference as a linear functional over $\mathcal{C}[x_0, x_k]$.

Theorem 3.30 (Leibniz formula)**.** For $k \in \mathbb{N}$, the kth divided difference of a product of two functions satisfies

$$[x_0, \ldots, x_k]fg = \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_i, \ldots, x_k]g. \tag{3.37}$$

Proof. The induction basis $k = 0$ holds because (3.37) reduces to $[x_0]fg = f(x_0)g(x_0)$. Now suppose (3.37) holds. For the induction step, we have from Theorem 2.17 that

$$[x_0, \ldots, x_{k+1}]fg = \frac{[x_1, \ldots, x_{k+1}]fg - [x_0, \ldots, x_k]fg}{x_{k+1} - x_0}.$$

By the induction hypothesis, we have

$$\begin{aligned} &[x_1, \ldots, x_{k+1}]fg = \sum_{i=0}^k [x_1, \ldots, x_{i+1}]f \cdot [x_{i+1}, \ldots, x_{k+1}]g \\ =&S_1 + \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_{i+1}, \ldots, x_{k+1}]g, \text{ where} \\ &S_1 = \sum_{i=0}^k (x_{i+1} - x_0) \cdot [x_0, \ldots, x_{i+1}]f \cdot [x_{i+1}, \ldots, x_{k+1}]g \\ &\quad = \sum_{i=1}^{k+1} (x_i - x_0) \cdot [x_0, \ldots, x_i]f \cdot [x_i, \ldots, x_{k+1}]g. \end{aligned}$$

$$\begin{aligned} &[x_0, \ldots, x_k]fg = \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_i, \ldots, x_k]g \\ =&- S_2 + \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_{i+1}, \ldots, x_{k+1}]g, \quad \text{where} \\ &S_2 = \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot (x_{k+1} - x_i) \cdot [x_i, \ldots, x_{k+1}]g. \end{aligned}$$

In the above derivation, we have applied Theorem 2.17 to go from the kth divided difference to the $(k+1)$th. Then

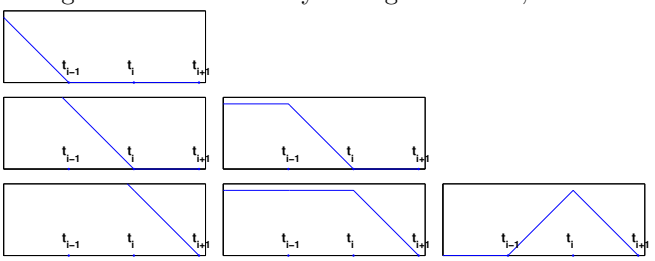
$$\begin{aligned} [x_0, \ldots, x_{k+1}]fg &= \frac{S_1 + S_2}{x_{k+1} - x_0} \\ &= \sum_{i=0}^{k+1} [x_0, \ldots, x_i]f \cdot [x_i, \ldots, x_{k+1}]g, \end{aligned}$$

which completes the inductive proof. □

Example 3.31. There exists a relation between B-splines and truncated power functions, e.g.,

$$\begin{aligned}
&(t_{i+1}-t_{i-1})[t_{i-1},t_i,t_{i+1}](t-x)_+\\
=&[t_i,t_{i+1}](t-x)_+ - [t_{i-1},t_i](t-x)_+\\
=&\frac{(t_{i+1}-x)_+-(t_i-x)_+}{t_{i+1}-t_i} - \frac{(t_i-x)_+-(t_{i-1}-x)_+}{t_i-t_{i-1}}\\
=&B_i^1 = \begin{cases} \frac{x-t_{i-1}}{t_i-t_{i-1}} & x\in(t_{i-1},t_i],\\ \frac{t_{i+1}-x}{t_{i+1}-t_i} & x\in(t_i,t_{i+1}],\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}$$

The algebra is illustrated by the figures below,



The significance is that, by applying divided difference to truncated power functions we can "cure" their drawback of non-local support. This idea is made precise in the next Theorem.

Theorem 3.32 (B-splines as divided difference of truncated power functions)**.** For any $n\in\mathbb{N}$, we have

$$B_i^n(x) = (t_{i+n}-t_{i-1})\cdot[t_{i-1},\dots,t_{i+n}](t-x)_+^n. \tag{3.38}$$

Proof. For $n=0$, (3.38) reduces to

$$\begin{aligned}
B_i^0(x) &= (t_i-t_{i-1})\cdot[t_{i-1},t_i](t-x)_+^0\\
&= (t_i-x)_+^0 - (t_{i-1}-x)_+^0\\
&= \begin{cases} 0 & \text{if } x\in(-\infty,t_{i-1}],\\ 1 & \text{if } x\in(t_{i-1},t_i],\\ 0 & \text{if } x\in(t_i,+\infty),\end{cases}
\end{aligned}$$

which is the same as (3.31). Hence the induction basis holds. Now assume the induction hypothesis (3.38) hold.

By Definition 3.16, $(t-x)_+^{n+1} = (t-x)(t-x)_+^n$. Then the application of the Leibniz formula (Theorem 3.30) with $f=(t-x)$ and $g=(t-x)_+^n$ yields

$$\begin{aligned}
&[t_{i-1},\dots,t_{i+n}](t-x)_+^{n+1}\\
=&(t_{i-1}-x)\cdot[t_{i-1},\dots,t_{i+n}](t-x)_+^n\\
&+[t_i,\dots,t_{i+n}](t-x)_+^n.
\end{aligned} \tag{3.39}$$

Definition 3.23 and the induction hypothesis yield

$$\begin{aligned}
&B_i^{n+1}(x) = \beta(x)+\gamma(x), \text{ with}\\
&\beta(x) = \frac{x-t_{i-1}}{t_{i+n}-t_{i-1}}B_i^n(x)\\
&\quad= (x-t_{i-1})\cdot[t_{i-1},\dots,t_{i+n}](t-x)_+^n\\
&\quad= [t_i,\dots,t_{i+n}](t-x)_+^n - [t_{i-1},\dots,t_{i+n}](t-x)_+^{n+1},
\end{aligned}$$

where the last step follows from (3.39). Similarly,

$$\begin{aligned}
\gamma(x) =&\frac{t_{i+n+1}-x}{t_{i+n+1}-t_i}B_{i+1}^n(x)\\
=&(t_{i+n+1}-x)\cdot[t_i,\dots,t_{i+n+1}](t-x)_+^n\\
=&(t_{i+n+1}-t_i)\cdot[t_i,\dots,t_{i+n+1}](t-x)_+^n\\
&+(t_i-x)\cdot[t_i,\dots,t_{i+n+1}](t-x)_+^n\\
=&[t_{i+1},\dots,t_{i+n+1}](t-x)_+^n - [t_i,\dots,t_{i+n}](t-x)_+^n\\
&+[t_i,\dots,t_{i+n+1}](t-x)_+^{n+1}\\
&-[t_{i+1},\dots,t_{i+n+1}](t-x)_+^n\\
=&[t_i,\dots,t_{i+n+1}](t-x)_+^{n+1} - [t_i,\dots,t_{i+n}](t-x)_+^n,
\end{aligned}$$

where the second last step follows from Theorem 2.17 and (3.39). The above arguments yield

$$\begin{aligned}
B_i^{n+1}(x) =&[t_i,\dots,t_{i+n+1}](t-x)_+^{n+1}\\
&-[t_{i-1},\dots,t_{i+n}](t-x)_+^{n+1}\\
=&(t_{i+n+1}-t_{i-1})\cdot[t_{i-1},\dots,t_{i+n+1}](t-x)_+^{n+1},
\end{aligned}$$

which completes the inductive proof. □

3.4.3 Integrals and derivatives

Corollary 3.33 (Integrals of B-splines)**.** The average of a B-spline over its support only depends on its degree,

$$\frac{1}{t_{i+n}-t_{i-1}}\int_{t_{i-1}}^{t_{i+n}} B_i^n(x)\mathrm{d}x = \frac{1}{n+1}. \tag{3.40}$$

Proof. The left-hand side (LHS) of (3.40) is

$$\begin{aligned}
&\frac{1}{t_{i+n}-t_{i-1}}\int_{t_{i-1}}^{t_{i+n}} B_i^n(x)\mathrm{d}x\\
=&\int_{t_{i-1}}^{t_{i+n}}[t_{i-1},\dots,t_{i+n}](t-x)_+^n\mathrm{d}x\\
=&[t_{i-1},\dots,t_{i+n}]\int_{t_{i-1}}^{t_{i+n}}(t-x)_+^n\mathrm{d}x\\
=&[t_{i-1},\dots,t_{i+n}]\frac{(t-t_{i-1})^{n+1}}{n+1}\\
=&\frac{1}{n+1},
\end{aligned}$$

where the first step follows from Theorem 3.32, the second step from the commutativity of integration and taking divided difference, the third step from (3.24), and the last step from Corollary 2.22. □

Theorem 3.34 (Derivatives of B-splines)**.** For $n\geq 2$, we have, $\forall x\in\mathbb{R}$,

$$\frac{\mathrm{d}}{\mathrm{d}x}B_i^n(x) = \frac{nB_i^{n-1}(x)}{t_{i+n-1}-t_{i-1}} - \frac{nB_{i+1}^{n-1}(x)}{t_{i+n}-t_i}. \tag{3.41}$$

For $n=1$, (3.41) holds for all x except at the three knots t_{i-1}, t_i, and t_{i+1}, where the derivative of B_i^1 is not defined.

Proof. We first show that (3.41) holds for all x except at the knots t_j. By (3.32), (3.28), and (3.31), we have

$$\begin{aligned}
&\forall x\in\mathbb{R}\setminus\{t_{i-1},t_i,t_{i+1}\},\\
&\quad\frac{\mathrm{d}}{\mathrm{d}x}B_i^1(x) = \frac{1}{t_i-t_{i-1}}B_i^0(x) - \frac{1}{t_{i+1}-t_i}B_{i+1}^0(x).
\end{aligned}$$

Hence the induction basis holds. Now suppose (3.41) holds $\forall x \in \mathbb{R} \setminus \{t_{i-1}, \dots, t_{i+n}\}$. Differentiate (3.30), apply the induction hypothesis (3.41), and we have

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^{n+1}(x) = \frac{B_i^n(x)}{t_{i+n} - t_{i-1}} - \frac{B_{i+1}^n(x)}{t_{i+n+1} - t_i} + nC(x), \quad (3.42)$$

where $C(x)$ is

$$\begin{aligned}
&\frac{x - t_{i-1}}{t_{i+n} - t_{i-1}} \left[\frac{B_i^{n-1}(x)}{t_{i+n-1} - t_{i-1}} - \frac{B_{i+1}^{n-1}(x)}{t_{i+n} - t_i} \right] \\
+&\frac{t_{i+n+1} - x}{t_{i+n+1} - t_i} \left[\frac{B_{i+1}^{n-1}(x)}{t_{i+n} - t_i} - \frac{B_{i+2}^{n-1}(x)}{t_{i+n+1} - t_{i+1}} \right] \\
=&\frac{1}{t_{i+n} - t_{i-1}} \left[\frac{(x - t_{i-1}) B_i^{n-1}(x)}{t_{i+n-1} - t_{i-1}} + \frac{(t_{i+n} - x) B_{i+1}^{n-1}(x)}{t_{i+n} - t_i} \right] \\
-&\frac{1}{t_{i+n+1} - t_i} \left[\frac{(x - t_i) B_{i+1}^{n-1}(x)}{t_{i+n} - t_i} + \frac{(t_{i+n+1} - x) B_{i+2}^{n-1}(x)}{t_{i+n+1} - t_{i+1}} \right] \\
=&\frac{B_i^n(x)}{t_{i+n} - t_{i-1}} - \frac{B_{i+1}^n(x)}{t_{i+n+1} - t_i},
\end{aligned}$$

where the last step follows from (3.30). Then (3.42) can be written as

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^{n+1}(x) = \frac{(n+1) B_i^n(x)}{t_{i+n} - t_{i-1}} - \frac{(n+1) B_{i+1}^n(x)}{t_{i+n+1} - t_i},$$

which completes the inductive proof of (3.41) except at the knots. Since $B_i^1 = \hat{B}_i$ is continuous, an easy induction with (3.30) shows that B_i^n is continuous for all $n \geq 1$. Hence the RHS of (3.41) is continuous for all $n \geq 2$. Therefore, if $n \geq 2$, $\frac{\mathrm{d}}{\mathrm{d}x} B_i^n(x)$ exists for all $x \in \mathbb{R}$. This completes the proof. □

Corollary 3.35 (Smoothness of B-splines)**.** $B_i^n \in \mathbb{S}_n^{n-1}$.

Proof. For $n = 1$, the induction basis $B_i^1(x) \in \mathbb{S}_1^0$ holds because of (3.32). The rest of the proof follows from (3.30) and Theorem 3.34 via an easy induction. □

3.4.4 Marsden's identity

Theorem 3.36 (Marsden's identity)**.** For any $n \in \mathbb{N}$,

$$(t - x)^n = \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n-1}) B_i^n(x), \quad (3.43)$$

where the product $(t - t_i) \cdots (t - t_{i+n-1})$ is defined as 1 for $n = 0$.

Proof. For $n = 0$, (3.43) follows from Definition 3.23. Now suppose (3.43) holds. A linear interpolation of the linear function $f(t) = t - x$ is the function itself,

$$t - x = \frac{t - t_{i+n}}{t_{i-1} - t_{i+n}} (t_{i-1} - x) + \frac{t - t_{i-1}}{t_{i+n} - t_{i-1}} (t_{i+n} - x). \quad (3.44)$$

Hence for the inductive step we have

$$\begin{aligned}
(t - x)^{n+1} &= (t - x) \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n-1}) B_i^n(x) \\
&= \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) \frac{t_{i-1} - x}{t_{i-1} - t_{i+n}} B_i^n(x) \\
&\quad + \sum_{i=-\infty}^{+\infty} (t - t_{i-1}) \cdots (t - t_{i+n-1}) \frac{t_{i+n} - x}{t_{i+n} - t_{i-1}} B_i^n(x) \\
&= \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) \frac{x - t_{i-1}}{t_{i+n} - t_{i-1}} B_i^n(x) \\
&\quad + \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) \frac{t_{i+n+1} - x}{t_{i+n+1} - t_i} B_{i+1}^n(x) \\
&= \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) B_i^{n+1}(x),
\end{aligned}$$

where the first step follows from the induction hypothesis, the second step from (3.44), the third step from replacing i with $i + 1$ in the second summation, and the last step from (3.30). □

Corollary 3.37 (Truncated power functions as linear combinations of B-splines)**.** For any $j \in \mathbb{Z}$ and $n \in \mathbb{N}$,

$$(t_j - x)_+^n = \sum_{i=-\infty}^{j-n} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x). \quad (3.45)$$

Proof. We need to show that the RHS is $(t_j - x)^n$ if $x \leq t_j$ and 0 otherwise. Set $t = t_j$ in (3.43) and we have

$$(t_j - x)^n = \sum_{i=-\infty}^{+\infty} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x).$$

For each $i = j - n + 1, \dots, j$, the corresponding term in the summation is zero regardless of x; for each $i \geq j + 1$, Lemma 3.27 implies that $B_i^n(x) = 0$ for all $x \leq t_j$. Hence

$$x \leq t_j \;\Rightarrow\; \sum_{i=-\infty}^{j-n} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x) = (t_j - x)^n.$$

Otherwise $x > t_j$, then Lemma 3.27 implies $B_i^n(x) = 0$ for each $i \leq j - n$. This completes the proof. □

3.4.5 Symmetric polynomials

Definition 3.38. The *elementary symmetric polynomial* of degree k in n variables is the sum of all products of k distinct variables chosen from the n variables,

$$\sigma_k(x_1, \dots, x_n) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k}. \quad (3.46)$$

In particular, $\sigma_0(x_1, \dots, x_n) = 1$ and

$$\forall k > n, \qquad \sigma_k(x_1, \dots, x_n) = 0.$$

If the distinctiveness condition is dropped, we have the *complete symmetric polynomial* of degree k in n variables,

$$\tau_k(x_1, \dots, x_n) = \sum_{1 \leq i_1 \leq \cdots \leq i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k}. \quad (3.47)$$

Example 3.39. $\sigma_2(x_1,x_2,x_3) = x_1x_2 + x_1x_3 + x_2x_3$. In comparison, $\tau_2(x_1,x_2,x_3) = \sigma_2(x_1,x_2,x_3) + x_1^2 + x_2^2 + x_3^2$.

Lemma 3.40. For $k \le n$, the elementary symmetric polynomials satisfy a recursion,

$$\begin{aligned}&\sigma_{k+1}(x_1,\dots,x_n,x_{n+1})\\ =&\sigma_{k+1}(x_1,\dots,x_n) + x_{n+1}\sigma_k(x_1,\dots,x_n).\end{aligned} \tag{3.48}$$

Proof. The terms in $\sigma_{k+1}(x_1,\dots,x_n,x_{n+1})$ can be assorted into two groups: (a) those that contain the factor x_{n+1} and (b) those that do not. By the symmetry in (3.46), group (a) must be $x_{n+1}\sigma_k(x_1,\dots,x_n)$ and group (b) must be $\sigma_{k+1}(x_1,\dots,x_n)$. □

Example 3.41. $\sigma_2(x_1,x_2,x_3) = x_1x_2 + x_3(x_1+x_2)$.

Definition 3.42. The *generating function for the elementary symmetric polynomials* is

$$g_{\sigma,n}(z) = \prod_{i=1}^{n}(1+x_iz) = (1+x_1z)\cdots(1+x_nz) \tag{3.49}$$

while that for the complete symmetric polynomials is

$$g_{\tau,n}(z) = \prod_{i=1}^{n}\frac{1}{1-x_iz} = \frac{1}{1-x_1z}\cdots\frac{1}{1-x_nz}. \tag{3.50}$$

Lemma 3.43 (Generating elementary and complete symmetric polynomials)**.** The elementary and complete symmetric polynomials are related to their generating functions as

$$g_{\sigma,n}(z) = \sum_{k=0}^{n}\sigma_k(x_1,\dots,x_n)z^k. \tag{3.51}$$

$$g_{\tau,n}(z) = \sum_{k=0}^{+\infty}\tau_k(x_1,\dots,x_n)z^k. \tag{3.52}$$

Proof. With Lemma 3.40, we can prove (3.51) by an easy induction. For (3.52), (3.50) and the identity

$$\frac{1}{1-x} = \sum_{k=0}^{+\infty}x^k \tag{3.53}$$

yield

$$\begin{aligned}g_{\tau,n}(z) =&\prod_{i=1}^{n}\sum_{k=0}^{+\infty}x_i^kz^k\\ =&(1+x_1z+x_1^2z^2+\cdots)(1+x_2z+x_2^2z^2+\cdots)\\ &\cdots(1+x_nz+x_n^2z^2+\cdots).\end{aligned}$$

The coefficient of the monomial z^k, is the sum of all possible products of k variables from $x_1,x_2,\dots,x_n$. Definition 3.38 then completes the proof. □

Example 3.44.

$$\begin{aligned}&(1+x_1z)(1+x_2z)(1+x_3z)\\ =&1+(x_1+x_2+x_3)z\\ &+(x_1x_2+x_1x_3+x_2x_3)z^2+x_1x_2x_3z^3.\end{aligned}$$

Lemma 3.45 (Recursive relations of complete symmetric polynomials)**.** The complete symmetric polynomials satisfy a recursion,

$$\begin{aligned}&\tau_{k+1}(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_1,\dots,x_n) + x_{n+1}\tau_k(x_1,\dots,x_n,x_{n+1}).\end{aligned} \tag{3.54}$$

Proof. (3.50) implies

$$g_{\tau,n+1} = g_{\tau,n} + x_{n+1}zg_{\tau,n+1}. \tag{3.55}$$

The proof is completed by requiring that the coefficient of z^{k+1} on the LHS equal that of z^{k+1} on the RHS. □

Theorem 3.46 (Complete symmetric polynomials as divided difference of monomials)**.** The complete symmetric polynomial of degree $m-n$ in $n+1$ variables is the nth divided difference of the monomial x^m, i.e.

$$\begin{aligned}&\forall m\in\mathbb{N}^+,\ \forall i\in\mathbb{N},\ \forall n = 0,1,\dots,m,\\ &\tau_{m-n}(x_i,\dots,x_{i+n}) = [x_i,\dots,x_{i+n}]x^m.\end{aligned} \tag{3.56}$$

Proof. By Lemma 3.45, we have

$$\begin{aligned}&(x_{n+1}-x_1)\tau_k(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_1,\dots,x_n,x_{n+1}) - \tau_{k+1}(x_1,\dots,x_n)\\ &-x_1\tau_k(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_2,\dots,x_n,x_{n+1}) + x_1\tau_k(x_1,\dots,x_n,x_{n+1})\\ &-\tau_{k+1}(x_1,\dots,x_n) - x_1\tau_k(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_2,\dots,x_n,x_{n+1}) - \tau_{k+1}(x_1,\dots,x_n).\end{aligned} \tag{3.57}$$

The rest of the proof is an induction on n. For $n=0$, (3.56) reduces to

$$\tau_m(x_i) = [x_i]x^m,$$

which is trivially true. Now suppose (3.56) holds for a nonnegative integer $n < m$. Then (3.57) and the induction hypothesis yield

$$\begin{aligned}&\tau_{m-n-1}(x_i,\dots,x_{i+n+1})\\ =&\frac{\tau_{m-n}(x_{i+1},\dots,x_{i+n+1}) - \tau_{m-n}(x_i,\dots,x_{i+n})}{x_{i+n+1}-x_i}\\ =&\frac{[x_{i+1},\dots,x_{i+n+1}]x^m - [x_i,\dots,x_{i+n}]x^m}{x_{i+n+1}-x_i}\\ =&[x_i,\dots,x_{i+n+1}]x^m,\end{aligned}$$

which completes the proof. □

3.4.6 B-splines indeed form a basis

Theorem 3.47. Given any $k\in\mathbb{N}$, the monomial x^k can be expressed as a linear combination of B-splines for any fixed $n \ge k$, in the form

$$\binom{n}{k}x^k = \sum_{i=-\infty}^{+\infty}\sigma_k(t_i,\dots,t_{i+n-1})B_i^n(x), \tag{3.58}$$

where $\sigma_k(t_i,\dots,t_{i+n-1})$ is the elementary symmetric polynomial of degree k in the n variables $t_i,\dots,t_{i+n-1}$.

Proof. Lemma 3.43 yields

$$(1+t_i x)\cdots(1+t_{i+n-1}x)=\sum_{k=0}^{n}\sigma_k(t_i,\ldots,t_{i+n-1})x^k.$$

Replace x with $-1/t$, multiply both sides with t^n, and we have

$$(t-t_i)\cdots(t-t_{i+n-1})=\sum_{k=0}^{n}\sigma_k(t_i,\ldots,t_{i+n-1})(-1)^k t^{n-k}.$$

Substituting the above into (3.43) yields

$$\begin{aligned}(t-x)^n &= \sum_{i=-\infty}^{+\infty}\sum_{k=0}^{n}\sigma_k(t_i,\ldots,t_{i+n-1})(-1)^k t^{n-k}B_i^n(x)\\ &= \sum_{k=0}^{n}\left\{t^{n-k}(-1)^k\sum_{i=-\infty}^{+\infty}\sigma_k(t_i,\ldots,t_{i+n-1})B_i^n(x)\right\}.\end{aligned}$$

On the other hand, the binomial theorem states that

$$(t-x)^n=\sum_{k=0}^{n}\binom{n}{k}t^{n-k}(-x)^k=\sum_{k=0}^{n}t^{n-k}(-1)^k\binom{n}{k}x^k.$$

Comparing the last two equations completes the proof. □

Corollary 3.48 (Partition of Unity)**.**

$$\forall n\in\mathbb{N},\qquad \sum_{i=-\infty}^{+\infty}B_i^n=1. \tag{3.59}$$

Proof. Setting $k=0$ in Theorem 3.47 yields (3.59). □

Theorem 3.49. The following list of B-splines is a basis of $\mathbb{S}_n^{n-1}(t_1,t_2,\ldots,t_N)$,

$$B_{2-n}^n(x),B_{3-n}^n(x),\ldots,B_N^n(x). \tag{3.60}$$

Proof. It is easy to verify that

$$\forall t_i\in\mathbb{R},\quad (x-t_i)_+^n=(x-t_i)^n-(-1)^n(t_i-x)_+^n. \tag{3.61}$$

Then it follows from Theorem 3.36 and Corollary 3.37 that each truncated power function $(x-t_i)_+^n$ can be expressed as a linear combination of B-splines. By Lemma 3.18, each element in $\mathbb{S}_n^{n-1}(t_1,t_2,\ldots,t_N)$ can be expressed as a linear combination of

$$1,x,x^2,\ldots,x^n,(x-t_2)_+^n,(x-t_3)_+^n,\ldots,(x-t_{N-1})_+^n.$$

Theorem 3.47 states that each monomial x^j can also be expressed as a linear combination of B-splines. Since the domain is restricted to $[t_1,t_N]$, we know from Lemma 3.27 that only those B-splines in the list of (3.60) appear in the linear combination. Therefore, these B-splines form a spanning list of $\mathbb{S}_n^{n-1}(t_1,t_2,\ldots,t_N)$. The proof is completed by Lemma B.40, Theorem 3.14, and the fact that the length of the list (3.60) is also $n+N-1$. □

3.4.7 Cardinal B-splines

Definition 3.50. The *cardinal B-spline* of degree n, denoted by $B_{i,\mathbb{Z}}^n$, is the B-spline in Definition 3.23 on the knot set $\mathbb{Z}$.

Corollary 3.51. Cardinal B-splines of the same degree are translates of one another, i.e.

$$\forall x\in\mathbb{R},\qquad B_{i,\mathbb{Z}}^n(x)=B_{i+1,\mathbb{Z}}^n(x+1). \tag{3.62}$$

Proof. The recurrence relation (3.30) reduces to

$$B_{i,\mathbb{Z}}^{n+1}(x)=\frac{x-i+1}{n+1}B_{i,\mathbb{Z}}^n(x)+\frac{i+n+1-x}{n+1}B_{i+1,\mathbb{Z}}^n(x). \tag{3.63}$$

The rest of the proof is an easy induction on n. □

Corollary 3.52. A cardinal B-spline is symmetric about the center of its interval of support, i.e.

$$\forall n>0,\ \forall x\in\mathbb{R},\qquad B_{i,\mathbb{Z}}^n(x)=B_{i,\mathbb{Z}}^n(2i+n-1-x). \tag{3.64}$$

Proof. The proof is similar with that of Corollary 3.51. □

Example 3.53. For $t_i=i$, the quadratic B-spline in Example 3.25 simplifies to

$$B_{i,\mathbb{Z}}^2(x)=\begin{cases}\frac{(x-i+1)^2}{2} & \text{if } x\in(i-1,i];\\ \frac{3}{4}-\left(x-(i+\frac{1}{2})\right)^2 & \text{if } x\in(i,i+1];\\ \frac{(i+2-x)^2}{2} & \text{if } x\in(i+1,i+2];\\ 0 & \text{otherwise.}\end{cases} \tag{3.65}$$

It is straightforward to verify Corollaries 3.51 and 3.52. It also follows from (3.65) that

$$B_{i,\mathbb{Z}}^2(j)=\begin{cases}\frac{1}{2} & \text{if } j\in\{i,i+1\};\\ 0 & \text{if } j\in\mathbb{Z}\setminus\{i,i+1\}.\end{cases} \tag{3.66}$$

Example 3.54. For $t_i=i$, the cubic cardinal B-spline is

$$B_{i,\mathbb{Z}}^3(x)=\begin{cases}\frac{(x-i+1)^3}{6} & \text{if } x\in(i-1,i];\\ \frac{2}{3}-\frac{1}{2}(x-i+1)(i+1-x)^2 & \text{if } x\in(i,i+1];\\ B_{i,\mathbb{Z}}^3(2i+2-x) & \text{if } x\in(i+1,i+3);\\ 0 & \text{otherwise.}\end{cases} \tag{3.67}$$

It follows that

$$B_{i,\mathbb{Z}}^3(j)=\begin{cases}\frac{1}{6} & \text{if } j\in\{i,i+2\};\\ \frac{2}{3} & \text{if } j=i+1;\\ 0 & \text{if } j\in\mathbb{Z}\setminus\{i,i+1,i+2\}.\end{cases} \tag{3.68}$$

This illustrates Corollary 3.51 that cardinal B-splines have the same shape, i.e., they are invariant under integer translations.

Theorem 3.55. The cardinal B-spline of degree n can be explicitly expressed as

$$B_{i,\mathbb{Z}}^n(x)=\frac{1}{n!}\sum_{k=-1}^{n}(-1)^{n-k}\binom{n+1}{k+1}(k+i-x)_+^n. \tag{3.69}$$

Proof. Theorems 3.32, 2.29, and 2.28 yield

$$\begin{aligned} B^n_{i,\mathbb{Z}}(x) &= (n+1)[i-1,\dots,i+n](t-x)^n_+ \\ &= \frac{n+1}{(n+1)!}\Delta^{n+1}(i-1-x)^n_+ \\ &= \frac{1}{n!}\sum_{k=0}^{n+1}(-1)^{n+1-k}\binom{n+1}{k}(i-1+k-x)^n_+. \end{aligned}$$

Replacing k with $k+1$ and accordingly changing the summation bounds complete the proof. □

Corollary 3.56. The value of a cardinal B-spline at an integer j is

$$B^n_{i,\mathbb{Z}}(j) = \frac{1}{n!}\sum_{k=j-i+1}^{n}(-1)^{n-k}\binom{n+1}{k+1}(k+i-j)^n \quad (3.70)$$

for $j \in [i, n+i)$ and is zero otherwise.

Proof. This follows directly from Theorem 3.55 and Definition 3.16. □

Theorem 3.57 (Unique interpolation by complete cubic cardinal B-splines)**.** There is a unique B-spline $S(x) \in \mathbb{S}^2_3$ that interpolates $f(x)$ at $1, 2, \dots, N$ with $S'(1) = f'(1)$ and $S'(N) = f'(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=-1}^{N} a_i B^3_{i,\mathbb{Z}}(x), \quad (3.71)$$

where

$$a_{-1} = a_1 - 2f'(1), \qquad a_N = a_{N-2} + 2f'(N), \quad (3.72)$$

and $\mathbf{a}^T = [a_0, \dots, a_{N-1}]$ is the solution of the linear system $M\mathbf{a} = \mathbf{b}$ with

$$\begin{aligned} \mathbf{b}^T =& [3f(1) + f'(1), 6f(2), \\ & \dots, 6f(N-1), 3f(N) - f'(N)], \\ M =& \begin{bmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{bmatrix}. \end{aligned}$$

Proof. By Theorem 3.49 and Lemma 3.27, we have, at each interpolation site $i = 1, 2, \dots, N$,

$$f(i) = a_{i-2}B^3_{i-2,\mathbb{Z}}(i) + a_{i-1}B^3_{i-1,\mathbb{Z}}(i) + a_i B^3_{i,\mathbb{Z}}(i).$$

Then (3.68) yields

$$\forall i = 1, 2, \dots, N, \qquad a_{i-2} + 4a_{i-1} + a_i = 6f(i), \quad (3.73)$$

which proves the middle $N-2$ equations of $M\mathbf{a} = \mathbf{b}$. By Theorem 3.34, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}B^n_{i,\mathbb{Z}}(x) = B^{n-1}_{i,\mathbb{Z}}(x) - B^{n-1}_{i+1,\mathbb{Z}}(x). \quad (3.74)$$

Differentiate (3.71), apply (3.74), set $x = 1$, apply (3.66) and we have the first identity in (3.72), which, together with (3.73), yields

$$2a_0 + a_1 = f'(1) + 3f(1);$$

this proves the first equation of $M\mathbf{a} = \mathbf{b}$. The last equation $M\mathbf{a} = \mathbf{b}$ and the second identity in (3.72) can be shown similarly. The strictly diagonal dominance of M implies a nonzero determinant of M and therefore $\mathbf{a}$ is uniquely determined. The uniqueness of $S(x)$ then follows from (3.72). □

Theorem 3.58. There is a unique B-spline $S(x) \in \mathbb{S}^1_2$ that interpolates $f(x)$ at $t_i = i + \frac{1}{2}$ for each $i = 1, 2, \dots, N-1$ with end conditions $S(1) = f(1)$ and $S(N) = f(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=0}^{N} a_i B^2_{i,\mathbb{Z}}(x), \quad (3.75)$$

where

$$a_0 = 2f(1) - a_1, \qquad a_N = 2f(N) - a_{N-1}, \quad (3.76)$$

and $\mathbf{a}^T = [a_1, \dots, a_{N-1}]$ is the solution of the linear system $M\mathbf{a} = \mathbf{b}$ with

$$\begin{aligned} \mathbf{b}^T =& \left[8f\left(\frac{3}{2}\right) - 2f(1), 8f\left(\frac{5}{2}\right),\right. \\ & \left. \dots, 8f\left(N - \frac{3}{2}\right), 8f\left(N - \frac{1}{2}\right) - 2f(N)\right], \\ M =& \begin{bmatrix} 5 & 1 & & & \\ 1 & 6 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 6 & 1 \\ & & & 1 & 5 \end{bmatrix}. \end{aligned}$$

Proof. It follows from Lemma 3.27 and Definition 3.50 that there are three quadratic cardinal B-splines, namely $B^2_{i-1,\mathbb{Z}}$, $B^2_{i,\mathbb{Z}}$, and $B^2_{i+1,\mathbb{Z}}$, that have nonzero values at each interpolation site $t_i = i + \frac{1}{2}$. Hence we have

$$f(t_i) = a_{i-1}B^2_{i-1,\mathbb{Z}}(t_i) + a_i B^2_{i,\mathbb{Z}}(t_i) + a_{i+1}B^2_{i+1,\mathbb{Z}}(t_i). \quad (3.77)$$

Hence the dimension of the space of relevant cardinal B-splines is $N - 1 + 2 = N + 1$, which is different from that in the proof of Theorem 3.49! By Theorem 3.55, we can calculate the values of B-splines as:

$$\begin{aligned} B^2_{0,\mathbb{Z}}(x) &= \frac{1}{2}\sum_{k=-1}^{2}(-1)^{2-k}\binom{3}{k+1}(k-x)^2_+, \\ B^2_{0,\mathbb{Z}}\left(\frac{1}{2}\right) &= \frac{3}{4}, \\ B^2_{0,\mathbb{Z}}\left(-\frac{1}{2}\right) = B^2_{0,\mathbb{Z}}\left(\frac{3}{2}\right) &= \frac{1}{8}, \end{aligned}$$

where for $B^2_{0,\mathbb{Z}}\left(-\frac{1}{2}\right)$ we have used Corollary 3.52. Then Corollary 3.51 and (3.77) yield

$$a_{i-1} + 6a_i + a_{i+1} = 8f(t_i), \quad (3.78)$$

which proves the middle $N-3$ equations in $M\mathbf{a}=\mathbf{b}$. At the end point $x=1$, only two quadratic cardinal B-splines, $B^2_{0,\mathbb{Z}}(x)$ and $B^2_{1,\mathbb{Z}}$, are nonzero. Then Example 3.25 yields

$$\frac{1}{2}a_0+\frac{1}{2}a_1=f(1)$$

and this proves the first identity in (3.76). Also, the above equation and (3.78) with $i=1$ yield

$$5a_1+a_2=8f\left(\frac{3}{2}\right)-2f(1),$$

which proves the first equation in $M\mathbf{a}=\mathbf{b}$. The last equation in $M\mathbf{a}=\mathbf{b}$ can be proven similarly. □

3.5 Curve fitting via splines

Definition 3.59. An open *curve* is (the image of) a continuous map $\gamma:(\alpha,\beta)\to\mathbb{R}^n$ for some α,β with $-\infty\le\alpha<\beta\le+\infty$. It is *simple* if the map γ is injective.

Definition 3.60. The *tangent vector of a curve* γ is its first derivative

$$\gamma':=\frac{\mathrm{d}\gamma}{\mathrm{d}t} \tag{3.79}$$

and the *unit tangent vector* of γ, denoted by $\mathbf{t}$, is the normalization of its tangent vector.

Definition 3.61. A *unit-speed curve* is a curve whose tangent vector has unit length at each of its points.

Definition 3.62. A point $\gamma(t_0)$ is a *regular point* of γ if $\mathbf{t}(t_0)$ exists and $\mathbf{t}(t_0)\neq\mathbf{0}$ holds; a curve is *regular* if all of its points are regular.

Definition 3.63. The *arc-length* of a curve starting at the point $\gamma(t_0)$ is defined as

$$s_\gamma(t)=\int_{t_0}^{t}\|\gamma'(u)\|_2\mathrm{d}u. \tag{3.80}$$

Definition 3.64. A map $X\mapsto Y$ is a *homeomorphism* if it is continuous and bijective and its inverse is also continuous; then the two sets X and Y are said to be *homeomorphic*.

Definition 3.65. A curve $\tilde\gamma(\tilde\alpha,\tilde\beta)\to\mathbb{R}^n$ is a *reparametrization* of another curve $\gamma(\alpha,\beta)\to\mathbb{R}^n$ if there exists a homeomorphism $\phi:(\tilde\alpha,\tilde\beta)\to(\alpha,\beta)$ such that $\tilde\gamma(\tilde t)=\gamma\left(\phi(\tilde t)\right)$ for each $\tilde t\in(\tilde\alpha,\tilde\beta)$.

Lemma 3.66. A reparametrization of a regular curve is unit-speed if and only if it is based on the arc-length.

Example 3.67. The spiral $\gamma:\mathbb{R}\to\mathbb{R}^2$ given by

$$\gamma(t)=(e^t\cos t,e^t\sin t) \tag{3.81}$$

is a curve. Its tangent vector is

$$\gamma'(t)=(e^t(\cos t-\sin t),e^t(\cos t+\sin t)) \tag{3.82}$$

and thus the modulus of the tangent vector is

$$\|\gamma'(t)\|_2=\sqrt{2}e^t.$$

Consequently, the arc length of the spiral is

$$s(t)=\int_0^t e^\tau\sqrt{2}\mathrm{d}\tau=\sqrt{2}(e^t-1) \tag{3.83}$$

and we have $t=\ln(\frac{s}{\sqrt{2}}+1)$. According to Lemma 3.66, the spiral can be expressed as a unit-speed curve

$$\gamma(s)=\left(\frac{s}{\sqrt{2}}+1\right)\left(\cos\left(\ln(\frac{s}{\sqrt{2}}+1)\right),\sin\left(\ln(\frac{s}{\sqrt{2}}+1)\right)\right). \tag{3.84}$$

Despite of its complicated form, the parametrization of the spiral in (3.84) makes the curve unit-speed; this is a prominent advantage over the parametrization in (3.81).

Definition 3.68. A *closed curve* is (the image of) a continuous map $\mathring\gamma:[0,1]\to\mathbb{R}^2$ that satisfies $\mathring\gamma(0)=\mathring\gamma(1)$. If the restriction of $\mathring\gamma$ to $[0,1)$ is further injective, then the closed curve is a *simple closed curve* or *Jordan curve*.

Definition 3.69. The *signed unit normal* of a curve, denoted by $\mathbf{n}_s$, is the unit vector obtained by rotating its unit tangent vector counterclockwise by $\frac{\pi}{2}$.

Definition 3.70. For a unit-speed curve γ, its *signed curvature* is defined as

$$\kappa_s:=\gamma''\cdot\mathbf{n}_s. \tag{3.85}$$

Definition 3.71. The *cumulative chordal length*s associated with a sequence of n points

$$\{\mathbf{x}_i\in\mathbb{R}^{\mathrm{D}}:i=1,2,\dots,n\} \tag{3.86}$$

are the n real numbers,

$$t_i=\begin{cases}0, & i=1;\\ t_{i-1}+\|\mathbf{x}_i-\mathbf{x}_{i-1}\|_2, & i>1,\end{cases} \tag{3.87}$$

where $\|\cdot\|_2$ denotes the Euclidean 2-norm.

Algorithm 3.72. A curve $\gamma:(0,1)\to\mathbb{R}^{\mathrm{D}}$ can be approximated by fitting D splines constructed from n characteristic points (3.86), each of which is on γ.

(a) Compute the cumulative chordal lengths.

(b) Fit a spline for each coordinate of γ with the independent parameter as the cumulative chordal lengths.

3.6 Problems

3.6.1 Theoretical questions

I. Consider $s\in\mathbb{S}^2_3$ on $[0,2]$:

$$s(x)=\begin{cases}p(x) & \text{if } x\in[0,1],\\ (2-x)^3 & \text{if } x\in[1,2].\end{cases}$$

Determine $p\in\mathbb{P}_3$ such that $s(0)=0$. Is $s(x)$ a natural cubic spline?

II. Given $f_i = f(x_i)$ of some scalar function at points $a = x_1 < x_2 < \cdots < x_n = b$, we consider interpolating f on $[a, b]$ with a quadratic spline $s \in \mathbb{S}_2^1$.

(a) Why is an additional condition needed to determine s uniquely?

(b) Define $m_i = s'(x_i)$ and $p_i = s|_{[x_i, x_{i+1}]}$.
Determine p_i in terms of f_i, f_{i+1}, and m_i for $i = 1, 2, \ldots, n-1$.

(c) Suppose $m_1 = f'(a)$ is given. Show how $m_2, m_3, \ldots, m_{n-1}$ can be computed.

III. Let $s_1(x) = 1 + c(x+1)^3$ where $x \in [-1, 0]$ and $c \in \mathbb{R}$. Determine $s_2(x)$ on $[0, 1]$ such that

$$s(x) = \begin{cases} s_1(x) & \text{if } x \in [-1, 0], \\ s_2(x) & \text{if } x \in [0, 1] \end{cases}$$

is a natural cubic spline on $[-1, 1]$ with knots $-1, 0, 1$. How must c be chosen if one wants $s(1) = -1$?

IV. Consider $f(x) = \cos\left(\frac{\pi}{2}x\right)$ with $x \in [-1, 1]$.

(a) Determine the natural cubic spline interpolant to f on knots $-1, 0, 1$.

(b) As discussed in the class, natural cubic splines have the minimal total bending energy. Verify this by taking $g(x)$ be (i) the quadratic polynomial that interpolates f at $-1, 0, 1$, and (ii) $f(x)$.

V. The quadratic B-spline $B_i^2(x)$.

(a) Derive the same explicit expression of $B_i^2(x)$ as that in the notes from the recursive definition of B-splines and the hat function.

(b) Verify that $\frac{\mathrm{d}}{\mathrm{d}x} B_i^2(x)$ is continuous at t_i and t_{i+1}.

(c) Show that only one $x^* \in (t_{i-1}, t_{i+1})$ satisfies $\frac{\mathrm{d}}{\mathrm{d}x} B_i^2(x^*) = 0$. Express x^* in terms of the knots within the interval of support.

(d) Consequently, show $B_i^2(x) \in [0, 1)$.

(e) Plot $B_i^2(x)$ for $t_i = i$.

VI. Verify Theorem 3.32 algebraically for the case of $n = 2$, i.e.

$$(t_{i+2} - t_{i-1})[t_{i-1}, t_i, t_{i+1}, t_{i+2}](t - x)_+^2 = B_i^2.$$

VII. Scaled integral of B-splines.
Deduce from the Theorem on derivatives of B-splines that the scaled integral of a B-spline $B_i^n(x)$ over its support is independent of its index i even if the spacing of the knots is not uniform.

VIII. Symmetric Polynomials.
We have a theorem on expressing complete symmetric polynomials as divided difference of monomials.

(a) Verify this theorem for $m = 4$ and $n = 2$ by working out the table of divided difference and comparing the result to the definition of complete symmetric polynomials.

(b) Prove this theorem by the lemma on the recursive relation on complete symmetric polynomials.

3.6.2 Programming assignments

A. Write a program for cubic-spline interpolation of the function

$$f(x) = \frac{1}{1 + 25x^2}$$

on evenly spaced nodes within the interval $[-1, 1]$ with $N = 6, 11, 21, 41, 81$. Compute for each N the max-norm of the interpolation error vector at mid-points of the subintervals and report the errors and convergence rates with respect to the number of subintervals.

Your algorithm should follow the example of interpolating the natural logarithm in the notes and your program must use an implementation of `lapack`.

Plot the interpolating spline against the exact function to observe that spline interpolation is free of the wide oscillations in the Runge phenomenon.

B. Let $f : \mathbb{R} \to \mathbb{R}$ be a given function. Implement two subroutines to interpolate f by the quadratic and cubic cardinal B-splines, which corresponds to Theorems 3.57 and 3.58, respectively.

C. Run your subroutines on the function

$$f(x) = \frac{1}{1 + x^2}, \qquad x \in [-5, 5],$$

using $t_i = -6 + i$, $i = 1, \ldots, 11$ for Theorem 3.57 and $t_i = i - \frac{11}{2}$, $i = 1, \ldots, 10$ for Corollary 3.58, respectively. Plot the polynomials against the exact function.

D. Define $E_S(x) = |S(x) - f(x)|$ as the interpolation error. For the two cardinal B-spline interpolants, output values of $E_S(x)$ at the sites

$$x = -3.5, -3, -0.5, 0, 0.5, 3, 3.5.$$

Output these values by a program. Why are some of the errors close to machine precision? Which of the two B-splines is more accurate?

E. The roots of the following equation constitute a closed planar curve in the shape of a heart:

$$x^2 + \left(\frac{3}{2}y - \sqrt{|x|}\right)^2 = 3. \tag{3.88}$$

Write a program to plot the heart. The parameter of the curve should be the *cumulative chordal length* defined in (3.87). Choose $n = 10, 40, 160$ and produce three plots of the heart function. (*Hints*: Your knots should include the characteristic points and you should think about (i) how many pieces of splines to use? (ii) what boundary conditions are appropriate?)

F. (*) Write a program to illustrate (3.38) by plotting the truncated power functions for $n = 1, 2$ and build a table of divided difference where the entries are figures instead of numbers. The pictures you generated for $n = 1$ should be the same as those in Example 3.31.

Chapter 4

Computer Arithmetic and Conditioning

4.1 Floating-point number systems

Definition 4.1. The *base or radix of a positional numeral system* is the number of unique symbols used to represent numbers.

Example 4.2. The *binary numeral system* consists of two digits: "0" and "1," and thus its base is 2. The *decimal system* consists of ten digits: "0" – "9," and thus its base is 10.

Definition 4.3. A *bit* is the basic unit of information in computing; it can have only one of two values 0 and 1.

Definition 4.4. A *byte* is a unit of information in computing that commonly consists of 8 bits; it is the smallest addressable unit of memory in many computers.

Definition 4.5. A *word* is a group of bits with fixed size that are handled as a unit by the instruction set architecture (ISA) and/or hardware of the processor. The *word size/width/length* is the number of bits in a word and is an important characteristic of processor or computer architecture.

Example 4.6. 32-bit and 64-bit computers are mostly common these days. A 32-bit register can store 2^{32} values, hence a processor with 32-bit memory address can directly access 4GB byte-addressable memory.

Definition 4.7 (Floating point numbers)**.** A *floating point number* (FPN) is a number of the form

$$x = \pm m \times \beta^e, \tag{4.1}$$

where β is the base or radix, $e \in [L, U]$, and the *significand* (or *mantissa*) m is a number of the form

$$m = \left(d_0 + \frac{d_1}{\beta} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right), \tag{4.2}$$

where the integer d_i satisfies $\forall i \in [0, p-1]$, $d_i \in [0, \beta-1]$. d_0 and d_{p-1} are called the *most significant digit* and the *least significant digit*, respectively. The *string of digits* of m is the string $d_0.d_1d_2\cdots d_{p-1}$, of which the portion $.d_1d_2\cdots d_{p-1}$ is called the *fraction* of m.

Algorithm 4.8. A decimal integer can be converted to a binary number via the following method:

- divide by 2 and record the remainder,
- repeat until you reach 0,
- concatenate the remainder backwards.

A decimal fraction can be converted to a binary number via the following method:

- multiply by 2 and check whether the integer part is no less than 1: if so record 1; otherwise record 0,
- repeat until you reach 0,
- concatenate the recorded bits forward.

Combine the above two methods and we can convert any decimal number to its binary counterpart.

Example 4.9. Convert 156 to binary number:

$$156 = (10011100)_2.$$

Example 4.10. What is the normalized binary form of $\frac{2}{3}$?

$$\begin{aligned} \frac{2}{3} &= (0.a_1a_2a_3\cdots)_2 = (0.1010\cdots)_2 \\ &= (1.0101010\cdots)_2 \times 2^{-1}. \end{aligned}$$

Definition 4.11 (FPN systems)**.** A *floating point number system* $\mathcal{F}$ is a proper subset of the rational numbers $\mathbb{Q}$, and it is characterized by a 4-tuple (β, p, L, U) with

- the *base* (or radix) β;
- the *precision* (or significand digits) p;
- the *exponent range* $[L, U]$.

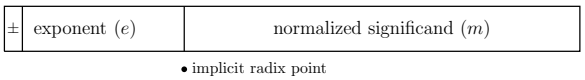
Definition 4.12. An FPN is *normalized* if its mantissa satisfies $1 \le m < \beta$.

Definition 4.13. The *subnormal* or *denormalized* numbers are FPNs of the form (4.1) with $e = L$ and $m \in (0, 1)$. A normalized FPN system can be *extended* by including the subnormal numbers.

Definition 4.14 (IEEE standard 754-2019)**.** The *single precision and double precision* FPNs of current IEEE (Institute of Electrical and Electronics Engineers) standard 754 published in 2019 are normalized FPN systems with three binary formats (32, 64, and 128 bits) and two decimal formats (64 and 128 bits).

$$\beta = 2,\ p = 23 + 1,\ e \in [-126, 127]; \tag{4.3a}$$
$$\beta = 2,\ p = 52 + 1,\ e \in [-1022, 1023]; \tag{4.3b}$$
$$\beta = 2,\ p = 112 + 1,\ e \in [-16382, 16383]; \tag{4.3c}$$
$$\beta = 10,\ p = 16,\ e \in [-1022, 1023]; \tag{4.3d}$$
$$\beta = 10,\ p = 34,\ e \in [-6143, 6144]. \tag{4.3e}$$

Example 4.15. In the IEEE 754 standard, there are some further details on the representation specifications of FPNs.



For example, some major representation specifications of the 32-bit FPNs are as follows.

(a) Out of the 32 bits, 1 is reserved for the sign, 8 for the exponents, 23 for the significand (see the plot above for the locations and the implicit radix point).

(b) The precision is 24 because we can choose $d_0 = 1$ for normalized binary floating point numbers and get away with never storing d_0.

(c) The exponent has $2^8 = 256$ possibilities. If we assign $1, 2, \ldots, 256$ to these possibilities, it would not be possible to represent numbers whose magnitudes are smaller than one. Hence we subtract $1, 2, \ldots, 256$ by 128 to shift the exponents to $-127, -126, \ldots, 0, \ldots, 127, 128$. Out of these numbers , $\pm m \times \beta^{-127}$ is reserved for ± 0 and subnormal numbers while $\pm m \times \beta^{128}$ is reserved for $\pm\infty$ and `NaN`s including `qNaN` (quiet) and `sNaN` (signaling).

Definition 4.16. The *machine precision* of a normalized FPN system $\mathcal{F}$ is the distance between 1.0 and the next larger FPN in $\mathcal{F}$,

$$\epsilon_M := \beta^{1-p}. \tag{4.4}$$

Definition 4.17. The underflow limit (UFL) and the overflow limit (OFL) of a normalized FPN system $\mathcal{F}$ are respectively

$$\mathrm{UFL}(\mathcal{F}) := \min |\mathcal{F} \setminus \{0\}| = \beta^L, \tag{4.5}$$
$$\mathrm{OFL}(\mathcal{F}) := \max |\mathcal{F}| = \beta^U(\beta - \beta^{1-p}). \tag{4.6}$$

Example 4.18. By default Matlab adopts IEEE 754 double precision arithmetic. Three characterizing constants are

- `eps` is the machine precision

$$\epsilon_M = \beta^{1-p} = 2^{1-(52+1)} = 2^{-52} \approx 2.22 \times 10^{-16},$$

- `realmin` is $\mathrm{UFL}(\mathcal{F})$

$$\min |\mathcal{F} \setminus \{0\}| = \beta^L = 2^{-1022} \approx 2.22 \times 10^{-308},$$

- `realmax` is $\mathrm{OFL}(\mathcal{F})$

$$\max |\mathcal{F}| = \beta^U(\beta - \beta^{1-p}) \approx 1.80 \times 10^{308}.$$

In `C/C++`, these constants are defined in `<cfloat>` and `float.h` by macros `DBL_EPSILON`, `DBL_MIN`, and `DBL_MAX`.

Corollary 4.19 (Cardinality of $\mathcal{F}$)**.** For a normalized binary FPN system $\mathcal{F}$,

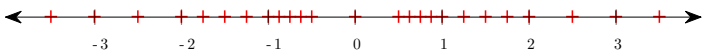
$$\#\mathcal{F} = 2^p(U - L + 1) + 1. \tag{4.7}$$

Proof. The cardinality can be proved by Axiom A.21. The factor 2^p comes from the sign bit and the mantissa. By Example 4.15, $U - L + 1$ is the number of exponents represented in $\mathcal{F}$. The trailing "+1" in (4.7) accounts for the number 0. □

Definition 4.20. The *range* of a normalized FPN system is a subset of $\mathbb{R}$ that consists of two intervals,

$$\mathcal{R}(\mathcal{F}) := \{x : x \in \mathbb{R}, \mathrm{UFL}(\mathcal{F}) \le |x| \le \mathrm{OFL}(\mathcal{F})\}. \tag{4.8}$$

Example 4.21. Consider a normalized FPN system with the characterization $\beta = 2, p = 3, L = -1, U = +1$.



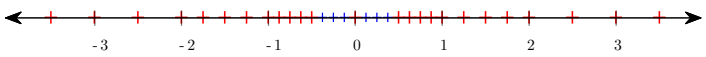
The four FPNs

$$1.00 \times 2^0,\ 1.01 \times 2^0,\ 1.10 \times 2^0,\ 1.11 \times 2^0$$

correspond to the four ticks in the plot starting at 1 while

$$1.00 \times 2^1,\ 1.01 \times 2^1,\ 1.10 \times 2^1,\ 1.11 \times 2^1$$

correspond to the four ticks starting at 2. Add subnormal FPNs and we have the following plot.



Definition 4.22. Two normalized FPNs a, b are *adjacent* to each other in $\mathcal{F}$ iff

$$\forall c \in \mathcal{F} \setminus \{a, b\},\quad |a - b| < |a - c| + |c - b|. \tag{4.9}$$

Lemma 4.23. Let a, b be two adjacent normalized FPNs satisfying $|a| < |b|$ and $ab > 0$. Then

$$\beta^{-1}\epsilon_M|a| < |a - b| \le \epsilon_M|a|. \tag{4.10}$$

Proof. Consider $a > 0$, then $\Delta a := b - a > 0$. By Definitions 4.7 and 4.12, $a = m \times \beta^e$ with $1.0 \le m < \beta$. a and b only differ from each other at the least significant digit, hence $\Delta a = \epsilon_M\beta^e$. Since $\frac{\epsilon_M}{\beta} < \frac{\epsilon_M}{m} \le \epsilon_M$, we have $\frac{\Delta a}{a} \in (\beta^{-1}\epsilon_M, \epsilon_M]$. The other case is similar. □

4.2 Rounding error analysis

4.2.1 Rounding a single number

Definition 4.24 (Rounding)**.** *Rounding* is a map $\mathrm{fl}: \mathbb{R} \to \mathcal{F} \cup \{+\infty, -\infty, \mathrm{NaN}\}$. The default rounding mode is *round to nearest*, i.e. $\mathrm{fl}(x)$ is chosen to minimize $|\mathrm{fl}(x) - x|$ for $x \in \mathcal{R}(\mathcal{F})$. In the case of a tie, $\mathrm{fl}(x)$ is chosen by *round to even*, i.e. $\mathrm{fl}(x)$ is the one with an even last digit d_{p-1}.

Definition 4.25. A rounded number $\mathrm{fl}(x)$ *overflows* if $|x| > \mathrm{OFL}(\mathcal{F})$, in which case $\mathrm{fl}(x)$=NaN, or *underflows* if $0 < |x| < \mathrm{UFL}(\mathcal{F})$, in which case $\mathrm{fl}(x)$=0. An underflow of an extended FPN system is called a *gradual underflow*.

Definition 4.26. The *unit roundoff* of $\mathcal{F}$ is the number

$$\epsilon_u := \frac{1}{2}\epsilon_M = \frac{1}{2}\beta^{1-p}. \tag{4.11}$$

Theorem 4.27 (Range of round-off errors)**.** For $x \in \mathcal{R}(\mathcal{F})$ as in (4.8), we have

$$\mathrm{fl}(x) = x(1+\delta), \qquad |\delta| < \epsilon_u. \tag{4.12}$$

Proof. By Definition A.32, $\mathcal{R}(\mathcal{F})$ is a subset of $\mathbb{R}$ and is thus a chain. Therefore $\forall x \in \mathcal{R}(\mathcal{F})$, $\exists x_L, x_R \in \mathcal{F}$ s.t.

- x_L and x_R are adjacent,
- $x_L \le x \le x_R$.

If $x = x_L$ or x_R, then $\mathrm{fl}(x) - x = 0$ and (4.12) clearly holds. Otherwise $x_L < x < x_R$. Then Lemma 4.23 and Definitions 4.22 and 4.24 yield

$$|\mathrm{fl}(x) - x| \le \frac{1}{2}|x_R - x_L| \le \epsilon_u \min(|x_L|, |x_R|) < \epsilon_u |x|. \tag{4.13}$$

Hence $-\epsilon_u|x| < \mathrm{fl}(x) - x < \epsilon_u|x|$, which yields (4.12). □

Theorem 4.28. For $x \in \mathcal{R}(\mathcal{F})$, we have

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \qquad |\delta| \le \epsilon_u. \tag{4.14}$$

Proof. The proof is the same as that of Theorem 4.27, except that we replace the last inequality "$< \epsilon_u|x|$" in (4.13) by "$\le \epsilon_u|\mathrm{fl}(x)|$." Consequently, the equality in (4.14) holds when $x = \frac{1}{2}(x_L + x_R)$ and $|\mathrm{fl}(x)| = \min(|x_L|, |x_R|)$ has its significand as $m = 1.0$. □

Example 4.29. Find x_L, x_R of $x = \frac{2}{3}$ in normalized single-precision IEEE 754 standard, which of them is $\mathrm{fl}(x)$?

By Example 4.10, we have

$$\frac{2}{3} = (0.1010\cdots)_2 = (1.0101010\cdots)_2 \times 2^{-1}.$$
$$x_L = (1.010\cdots10)_2 \times 2^{-1};$$
$$x_R = (1.010\cdots11)_2 \times 2^{-1},$$

where the last bit of x_L must be 0 because the IEEE 754 standard states that 23 bits are reserved for the mantissa. It follows that

$$x - x_L = \frac{2}{3} \times 2^{-24};$$
$$x_R - x_L = 2^{-24},$$
$$x_R - x = (x_R - x_L) - (x - x_L) = \frac{1}{3} \times 2^{-24}.$$

Thus Definition 4.24 implies $\mathrm{fl}(x) = x_R$.

4.2.2 Binary floating-point operations

Definition 4.30 (Addition/subtraction of two FPNs)**.** Express $a, b \in \mathcal{F}$ as $a = M_a \times \beta^{e_a}$ and $b = M_b \times \beta^{e_b}$ where $M_a = \pm m_a$ and $M_b = \pm m_b$. With the assumption $|a| \ge |b|$, the sum $c := \mathrm{fl}(a+b) \in \mathcal{F}$ is calculated in a register of precision at least $2p$ as follows.

(i) Exponent comparison:

- If $e_a - e_b > p + 1$, set $c = a$ and return c;
- otherwise set $e_c \leftarrow e_a$ and $M_b \leftarrow M_b/\beta^{e_a - e_b}$.

(ii) Perform the addition $M_c \leftarrow M_a + M_b$ in the register with rounding to nearest.

(iii) Normalization:

- If $|M_c| = 0$, return 0.
- If $|M_c| \in (\beta, \beta^2)$, set $M_c \leftarrow M_c/\beta$ and $e_c \leftarrow e_c + 1$.
- If $|M_c| \in (0, \beta - \epsilon_u(p))$, repeat $M_c \leftarrow M_c\beta$, $e_c \leftarrow e_c - 1$ until $|M_c| \in [1, \beta)$.
- If $|M_c| \in [\beta - \epsilon_u(p), \beta]$, set $|M_c| \leftarrow 1.0$ and $e_c \leftarrow e_c + 1$.

(iv) Check range:

- return NaN if e_c overflows,
- return 0 if e_c underflows.

(v) Round M_c (to nearest) to precision p.

(vi) Set $c \leftarrow M_c \times \beta^{e_c}$.

Here $\epsilon_u(p)$ in step (iii) is the unit round-off for FPNs with precision p, c.f. Definition 4.26.

Example 4.31. Consider the calculation of $c := \mathrm{fl}(a+b)$ with $a = 1.234 \times 10^4$ and $b = 5.678 \times 10^0$ in an FPN system $\mathcal{F} : (10, 4, -7, 8)$.

(i) $b \leftarrow 0.0005678 \times 10^4$; $e_c \leftarrow 4$.

(ii) $m_c \leftarrow 1.2345678$.

(iii) do nothing.

(iv) do nothing.

(v) $m_c \leftarrow 1.235$.

(vi) $c = 1.235 \times 10^4$.

For $b = 5.678 \times 10^{-2}$, $c = a$ would be returned in step (i).

Example 4.32. Consider the calculation of $c := \mathrm{fl}(a+b)$ with $a = 1.000 \times 10^0$ and $b = -9.000 \times 10^{-5}$ in an FPN system $\mathcal{F} : (10, 4, -7, 8)$.

(i) $b \leftarrow -0.0000900 \times 10^0$; $e_c \leftarrow 0$.

(ii) $m_c \leftarrow 0.9999100$.

(iii) $e_c \leftarrow e_c - 1$; $m_c \leftarrow 9.9991000$.

(iv) do nothing.

(v) $m_c \leftarrow 9.999$.

(vi) $c = 9.999 \times 10^{-1}$.

For $b = -9.000 \times 10^{-6}$, $c = a$ would be returned in step (i).

Exercise 4.33. Repeat Example 4.31 with $b = 8.769 \times 10^4$, $b = -5.678 \times 10^0$, and $b = -5.678 \times 10^3$.

Lemma 4.34. For $a, b \in \mathcal{F}$, $a + b \in \mathcal{R}(\mathcal{F})$ implies

$$\mathrm{fl}(a + b) = (a + b)(1 + \delta), \qquad |\delta| < \epsilon_u. \tag{4.15}$$

Proof. The round-off error in step (v) always dominates those in step (i) and step (ii), both of which, because of the $2p$ precision, are nonzero only in the case of $e_a - e_b = p + 1$. Then (4.15) follows from Theorem 4.27. □

Definition 4.35 (Multiplication of two FPNs)**.** Express $a, b \in \mathcal{F}$ as $a = M_a \times \beta^{e_a}$ and $b = M_b \times \beta^{e_b}$ where $M_a = \pm m_a$ and $M_b = \pm m_b$. The product $c := \mathrm{fl}(ab) \in \mathcal{F}$ is calculated in a register of precision at least $2p$ as follows.

(i) Exponent sum: $e_c \leftarrow e_a + e_b$.

(ii) Perform the multiplication $M_c \leftarrow M_a M_b$ in the register.

(iii) Normalization:

- If $|M_c| \in (\beta, \beta^2)$, set $M_c \leftarrow M_c/\beta$ and $e_c \leftarrow e_c + 1$.
- If $|M_c| \in [\beta - \epsilon_u(p), \beta]$, set $|M_c| \leftarrow 1.0$ and $e_c \leftarrow e_c + 1$.

(iv) Check range:

- return NaN if e_c overflows,
- return 0 if e_c underflows.

(v) Round M_c (to nearest) to precision p.

(vi) Set $c \leftarrow M_c \times \beta^{e_c}$.

Here $\epsilon_u(p)$ in step (iii) is the unit round-off for FPNs with precision p, c.f. Definition 4.26.

Example 4.36. Consider the calculation of $c := \mathrm{fl}(ab)$ with $a = 2.345 \times 10^4$ and $b = 6.789 \times 10^0$ in an FPN system $\mathcal{F} : (10, 4, -7, 8)$.

(i) $e_c \leftarrow 4$.

(ii) $M_c \leftarrow 15.920205$.

(iii) $m_c \leftarrow 1.5920205$, $e_c \leftarrow 5$.

(iv) do nothing.

(v) $m_c \leftarrow 1.592$.

(vi) $c = 1.592 \times 10^5$.

Lemma 4.37. For $a, b \in \mathcal{F}$, $|ab| \in \mathcal{R}(\mathcal{F})$ implies

$$\mathrm{fl}(ab) = (ab)(1 + \delta), \qquad |\delta| < \epsilon_u. \tag{4.16}$$

Proof. The error only comes from the round-off in steps (v). Then (4.16) follows from Theorem 4.27. □

Definition 4.38 (Division of two FPNs)**.** Express $a, b \in \mathcal{F}$ as $a = M_a \times \beta^{e_a}$ and $b = M_b \times \beta^{e_b}$ where $M_a = \pm m_a$ and $M_b = \pm m_b$. The quotient $c = \mathrm{fl}\left(\frac{a}{b}\right) \in \mathcal{F}$ is calculated in a register of precision at least $2p + 1$ as follows.

(i) If $m_b = 0$, return NaN; otherwise set $e_c \leftarrow e_a - e_b$.

(ii) Perform the division $M_c \leftarrow M_a/M_b$ in the register with rounding to nearest.

(iii) Normalization:

- If $|M_c| < 1$, set $M_c \leftarrow M_c\beta$, $e_c \leftarrow e_c - 1$.

(iv) Check range:

- return NaN if e_c overflows,
- return 0 if e_c underflows.

(v) Round M_c (to nearest) to precision p.

(vi) Set $c \leftarrow M_c \times \beta^{e_c}$.

Lemma 4.39. For $a, b \in \mathcal{F}$, $\frac{a}{b} \in \mathcal{R}(\mathcal{F})$ implies

$$\mathrm{fl}\left(\frac{a}{b}\right) = \frac{a}{b}(1 + \delta), \qquad |\delta| < \epsilon_u. \tag{4.17}$$

Proof. In the case of $|M_a| = |M_b|$, there is no rounding error in Definition 4.38 and (4.17) clearly holds. Hereafter we denote by M_{c1} and M_{c2} the results of steps (ii) and (v) in Definition 4.38, respectively.

In the case of $|M_a| > |M_b|$, the condition $a, b \in \mathcal{F}$, Definition 4.16, and $|M_a|, |M_b| \in [1, \beta)$ imply

$$\left|\frac{M_a}{M_b}\right| \geq \frac{\beta - \epsilon_M}{\beta - 2\epsilon_M} > 1 + \beta^{-1}\epsilon_M, \tag{4.18}$$

which further implies that the normalization step (iii) in Definition 4.38 is not invoked. By Definitions 4.24, 4.16, and 4.26, the unit roundoff of a register with precision $p + k$ is

$$\frac{1}{2}\beta^{1-p-k} = \frac{1}{2}\beta^{1-p}\beta^{1-p}\beta^{p-1-k} = \beta^{p-1-k}\epsilon_u\epsilon_M,$$

and hence the unit roundoff of the register in Definition 4.38 is $\beta^{-2}\epsilon_u\epsilon_M$. Therefore we have

$$\begin{aligned} M_{c2} &= M_{c1} + \delta_2, \qquad |\delta_2| < \epsilon_u \\ &= \frac{M_a}{M_b} + \delta_1 + \delta_2, \qquad |\delta_1| < \beta^{-2}\epsilon_u\epsilon_M \\ &= \frac{M_a}{M_b}(1 + \delta); \\ |\delta| &= \left|\frac{\delta_1 + \delta_2}{M_a/M_b}\right| < \frac{\epsilon_u\left(1 + \beta^{-2}\epsilon_M\right)}{1 + \beta^{-1}\epsilon_M} < \epsilon_u, \end{aligned}$$

where we have applied (4.18) and the triangular inequality in deriving the first inequality of the last line.

Consider the last case $|M_a| < |M_b|$. It is impossible to have $|M_{c1}| = 1$ in step (ii) because

$$\frac{|M_a|}{|M_b|} \leq \frac{\beta - 2\epsilon_M}{\beta - \epsilon_M} = 1 - \frac{\epsilon_M}{\beta - \epsilon_M} < 1 - \beta^{-1}\epsilon_M$$

and the precision of the register is greater than $p+1$. Therefore $|M_{c1}| < 1$ must hold and in Definition 4.38 step (iii) is invoked to yield

$$\begin{aligned} M_{c1} &= \frac{M_a}{M_b} + \delta_1, \qquad |\delta_1| < \beta^{-2}\epsilon_u\epsilon_M; \\ M_{c2} &= \beta M_{c1} + \delta_2, \qquad |\delta_2| < \epsilon_u \\ &= \beta\frac{M_a}{M_b}\left(1 + \frac{\beta\delta_1 + \delta_2}{\beta M_a/M_b}\right), \end{aligned}$$

where the denominator in the parentheses satisfies

$$\beta\left|\frac{M_a}{M_b}\right| \geq \frac{\beta}{\beta - \epsilon_M} = 1 + \frac{\epsilon_M}{\beta - \epsilon_M} > 1 + \beta^{-1}\epsilon_M.$$

Hence we have

$$|\delta| = \left|\frac{\beta\delta_1 + \delta_2}{\beta M_a/M_b}\right| < \frac{\beta^{-1}\epsilon_u\epsilon_M + \epsilon_u}{1 + \beta^{-1}\epsilon_M} = \epsilon_u. \qquad \square$$

Theorem 4.40 (Model of machine arithmetic)**.** Denote by $\mathcal{F}$ a normalized FPN system with precision p. For each arithmetic operation $\odot = +, -, \times, /$, we have

$$\forall a, b \in \mathcal{F},\ a \odot b \in \mathcal{R}(\mathcal{F}) \ \Rightarrow\ \mathrm{fl}(a \odot b) = (a \odot b)(1+\delta) \quad (4.19)$$

where $|\delta| < \epsilon_u$ if and only if these binary operations are performed in a register with precision $2p + 1$.

Proof. This follows from Lemmas 4.34, 4.37, and 4.39. $\square$

4.2.3 The propagation of rounding errors

Theorem 4.41. If $\forall i = 0, 1, \cdots, n$, $a_i \in \mathcal{F}$, $a_i > 0$, then

$$\mathrm{fl}\left(\sum_{i=0}^{n} a_i\right) = (1 + \delta_n)\sum_{i=0}^{n} a_i, \quad (4.20)$$

where $|\delta_n| < (1 + \epsilon_u)^n - 1 \approx n\epsilon_u$.

Proof. Define $s_k := \sum_{i=0}^{k} a_i$,

$$\begin{cases} s_0 & := a_0; \\ s_{k+1} & := s_k + a_{k+1}, \end{cases} \qquad \begin{cases} s_0^* & := a_0; \\ s_{k+1}^* & := \mathrm{fl}(s_k^* + a_{k+1}), \end{cases}$$

$$\delta_k := \frac{s_k^* - s_k}{s_k}, \qquad \epsilon_k := \frac{s_{k+1}^* - (s_k^* + a_{k+1})}{s_k^* + a_{k+1}}.$$

In words, δ_k is the accumulated rounding error and ϵ_k is the rounding error at the kth step. Then we have

$$\begin{aligned} \delta_{k+1} &= \frac{s_{k+1}^* - s_{k+1}}{s_{k+1}} = \frac{(s_k^* + a_{k+1})(1+\epsilon_k) - s_{k+1}}{s_{k+1}} \\ &= \frac{(s_k(1+\delta_k) + a_{k+1})(1+\epsilon_k) - s_k - a_{k+1}}{s_{k+1}} \\ &= \frac{(\epsilon_k + \delta_k + \epsilon_k\delta_k)s_k + \epsilon_k a_{k+1}}{s_{k+1}} \\ &= \frac{\epsilon_k s_{k+1} + \delta_k(1+\epsilon_k)s_k}{s_{k+1}} = \epsilon_k + \delta_k(1+\epsilon_k)\frac{s_k}{s_{k+1}}. \end{aligned}$$

The condition of a_i's being positive implies $s_k < s_{k+1}$, and Theorem 4.27 states $|\epsilon_k| < \epsilon_u$. Hence we have

$$|\delta_{k+1}| < |\epsilon_k| + |\delta_k|(1+\epsilon_u) < \epsilon_u + |\delta_k|(1+\epsilon_u).$$

An easy induction then shows that

$$\begin{aligned} \forall k \in \mathbb{N},\ |\delta_{k+1}| &< \epsilon_u \sum_{i=0}^{k}(1+\epsilon_u)^i \quad (4.21) \\ &= \epsilon_u\frac{(1+\epsilon_u)^{k+1} - 1}{1 + \epsilon_u - 1} = (1+\epsilon_u)^{k+1} - 1, \end{aligned}$$

where the second step follows from the summation formula of geometric series. The proof is completed by the binomial theorem. $\square$

Exercise 4.42. If we sort the positive numbers $a_i > 0$ according to their magnitudes and carry out the additions in this ascending order, we can minimize the rounding error term δ in Theorem 4.41. Can you give some examples?

Exercise 4.43. Derive $\mathrm{fl}(a_1b_1 + a_2b_2 + a_3b_3)$ for $a_i, b_i \in \mathcal{F}$ and make some observations on the corresponding derivation of $\mathrm{fl}(\sum_i \prod_j a_{i,j})$.

Theorem 4.44. For given $\mu \in \mathbb{R}^+$ and a positive integer $n \leq \lfloor \frac{\ln 2}{\mu} \rfloor$, suppose $|\delta_i| \leq \mu$ for each $i = 1, 2, \ldots, n$. Then

$$1 - n\mu \leq \prod_{i=1}^{n}(1+\delta_i) \leq 1 + n\mu + (n\mu)^2, \quad (4.22)$$

or equivalently, for $I_n := [-\frac{1}{1+n\mu}, 1]$,

$$\exists \theta \in I_n \text{ s.t. } \prod_{i=1}^{n}(1+\delta_i) = 1 + \theta(n\mu + n^2\mu^2). \quad (4.23)$$

Proof. The condition $|\delta_i| \leq \mu$ implies

$$(1-\mu)^n \leq \prod_{i=1}^{n}(1+\delta_i) \leq (1+\mu)^n.$$

Taylor expansion of $f(\mu) = (1-\mu)^n$ at $\mu = 0$ with Lagrangian remainder yields

$$(1-\mu)^n \geq 1 - n\mu,$$

which implies the first inequality in (4.22). On the other hand, the Taylor series of e^x for $x \in \mathbb{R}^+$ satisfies

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ &= 1 + x + \frac{x^2}{2!}\left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right) \\ &\leq 1 + x + \frac{x^2}{2}e^x. \end{aligned}$$

Set $x = n\mu$ in the above inequality, apply the condition $n\mu \leq \ln 2$, and we have

$$e^{n\mu} \leq 1 + n\mu + (n\mu)^2,$$

which, together with the inequality $(1+\mu)^n \leq e^{n\mu}$, yields the second inequality in (4.22).

Finally, (4.22) implies that $\prod_{i=1}^{n}(1+\delta_i)$ is in the range of the continuous function $f(\tau) = 1 + \tau(1+n\mu)n\mu$ on I_n. The rest of the proof follows from the intermediate value theorem. $\square$

4.3 Accuracy and stability

4.3.1 Avoiding catastrophic cancellation

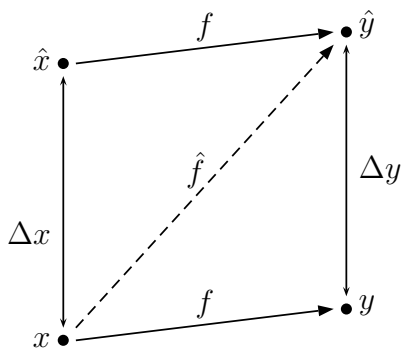
Definition 4.45. Let $\hat{x}$ be an approximation to $x \in \mathbb{R}$. The accuracy of $\hat{x}$ can be measured by its *absolute error*

$$E_{\text{abs}}(\hat{x}) = |\hat{x} - x| \tag{4.24}$$

and/or its *relative error*

$$E_{\text{rel}}(\hat{x}) = \frac{|\hat{x} - x|}{|x|}. \tag{4.25}$$

Definition 4.46. For an approximation $\hat{y}$ to $y = f(x)$ computed by $\hat{y} = \hat{f}(x)$, the *forward error* is the relative error of $\hat{y}$ in approximating y and the *backward error* is the smallest relative error in approximating x by an $\hat{x}$ that satisfies $f(\hat{x}) = \hat{f}(x)$, assuming such an $\hat{x}$ exists.



Definition 4.47 (Accuracy)**.** An algorithm $\hat{y} = \hat{f}(x)$ for computing the function $y = f(x)$ is *accurate* if its forward error is small for all x, i.e. $\forall x \in \text{dom}(f)$, $E_{\text{rel}}(\hat{f}(x)) \le c\epsilon_u$ where c is a small constant.

Example 4.48 (Catastrophic cancellation)**.** For two real numbers $x, y \in \mathcal{R}(\mathcal{F})$, Theorems 4.27 and 4.40 imply

$$\begin{aligned}\text{fl}(\text{fl}(x) \odot \text{fl}(y)) &= (\text{fl}(x) \odot \text{fl}(y))(1+\delta_3) \\ &= (x(1+\delta_1) \odot y(1+\delta_2))(1+\delta_3)\end{aligned}$$

where $|\delta_i| \le \epsilon_u$. From Theorems 4.40 and 4.44, we know that *multiplication is accurate*:

$$\begin{aligned}\text{fl}(\text{fl}(x) \times \text{fl}(y)) &= xy(1+\delta_1)(1+\delta_2)(1+\delta_3) \\ &= xy[1 + \theta(3\epsilon_u + 3\epsilon_u^2 + \epsilon_u^3)],\end{aligned}$$

where $\theta \in [-1, 1]$. Similarly, *division is also accurate*:

$$\begin{aligned}\text{fl}(\text{fl}(x)/\text{fl}(y)) &= \frac{x(1+\delta_1)}{y(1+\delta_2)}(1+\delta_3) \\ =&\frac{x}{y}(1+\delta_1)(1-\delta_2+\delta_2^2-\cdots)(1+\delta_3) \\ \approx&\frac{x}{y}(1+\delta_1)(1-\delta_2)(1+\delta_3).\end{aligned}$$

However, *addition and subtraction might not be accurate*:

$$\begin{aligned}&\text{fl}(\text{fl}(x) + \text{fl}(y)) = (x(1+\delta_1) + y(1+\delta_2))(1+\delta_3) \\ =&(x + y + x\delta_1 + y\delta_2)(1+\delta_3) \\ =&(x+y)\left(1 + \delta_3 + \frac{x\delta_1 + y\delta_2}{x+y} + \delta_3\frac{x\delta_1 + y\delta_2}{x+y}\right).\end{aligned}$$

In other words, the relative error of addition or subtraction can be arbitrarily large when $x + y \to 0$.

Theorem 4.49 (Loss of most significant digits)**.** Suppose $x, y \in \mathcal{F}$, $x > y > 0$, and

$$\beta^{-t} \le 1 - \frac{y}{x} \le \beta^{-s}. \tag{4.26}$$

Then the number of most significant digits that are lost in the subtraction $x - y$ is at most t and at least s.

Proof. Rewrite $x = m_x \times \beta^n$ and $y = m_y \times \beta^m$ with $1 \le m_x, m_y < \beta$. Definition 4.30 and the condition $x > y$ imply that m_y, the significand of y, is shifted so that y has the same exponent as x before $m_x - m_y$ is performed in the register. Then

$$\begin{aligned}&y = (m_y \times \beta^{m-n}) \times \beta^n \\ \Rightarrow\ &x - y = (m_x - m_y \times \beta^{m-n}) \times \beta^n \\ \Rightarrow\ &m_{x-y} = m_x\left(1 - \frac{m_y \times \beta^m}{m_x \times \beta^n}\right) = m_x\left(1 - \frac{y}{x}\right) \\ \Rightarrow\ &\beta^{-t} \le m_{x-y} < \beta^{1-s}.\end{aligned}$$

To normalize m_{x-y} into the interval $[1, \beta)$, it should be multiplied by at least β^s and at most β^t. In other words, m_{x-y} should be shifted to the left for at least s times and at most t times. Therefore the conclusion on the number of lost significant digits follows. □

Rule 4.50. Catastrophic cancellation should be avoided whenever possible.

Example 4.51. Calculate $y = f(x) = x - \sin x$ for $x \to 0$. When x is small, a straightforward calculation would result in a catastrophic cancellation because $x \approx \sin x$. The solution is to use the Taylor series

$$\begin{aligned}x - \sin x &= x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right) \\ &= \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots\end{aligned}$$

4.3.2 Backward stability and numerical stability

Definition 4.52 (Backward stability)**.** An algorithm $\hat{f}(x)$ for computing $y = f(x)$ is *backward stable* if its backward error is small for all x, i.e.

$$\begin{aligned}&\forall x \in \text{dom}(f),\ \exists \hat{x} \in \text{dom}(f),\ \text{ s.t.} \\ &\hat{f}(x) = f(\hat{x})\ \Rightarrow\ E_{\text{rel}}(\hat{x}) \le c\epsilon_u,\end{aligned} \tag{4.27}$$

where c is a small constant.

Definition 4.53. An algorithm $\hat{f}(x_1, x_2)$ for computing $y = f(x_1, x_2)$ is *backward stable* if

$$\begin{aligned}&\forall (x_1, x_2) \in \text{dom}(f),\ \exists (\hat{x}_1, \hat{x}_2) \in \text{dom}(f) \text{ s.t.} \\ &\hat{f}(x_1, x_2) = f(\hat{x}_1, \hat{x}_2)\ \Rightarrow\ \begin{cases} E_{\text{rel}}(\hat{x}_1) \le c_1\epsilon_u, \\ E_{\text{rel}}(\hat{x}_2) \le c_2\epsilon_u, \end{cases}\end{aligned} \tag{4.28}$$

where c_1, c_2 are two small constants.

Lemma 4.54. For $f(x_1, x_2) = x_1 - x_2$, $x_1, x_2 \in \mathcal{R}(\mathcal{F})$, the algorithm $\hat{f}(x_1, x_2) = \text{fl}(\text{fl}(x_1) - \text{fl}(x_2))$ is backward stable.

Proof. We have $\hat{f}(x_1, x_2) = (\text{fl}(x_1) - \text{fl}(x_2))(1 + \delta_3)$ from Theorem 4.40. Then Theorem 4.27 implies

$$\begin{aligned}\hat{f}(x_1, x_2) &= (x_1(1 + \delta_1) - x_2(1 + \delta_2))(1 + \delta_3) \\ &= x_1(1 + \delta_1 + \delta_3 + \delta_1\delta_3) - x_2(1 + \delta_2 + \delta_3 + \delta_2\delta_3).\end{aligned}$$

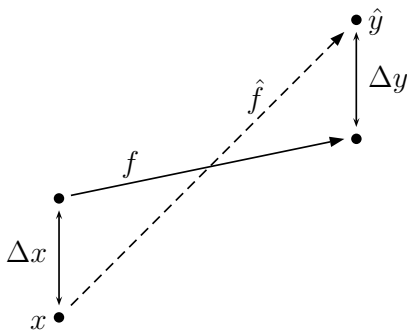
Take $\hat{x}_1$ and $\hat{x}_2$ to be the two terms in the above line and we have

$$\begin{aligned}E_{\text{rel}}(\hat{x}_1) &= |\delta_1 + \delta_3 + \delta_1\delta_3|, \\ E_{\text{rel}}(\hat{x}_2) &= |\delta_2 + \delta_3 + \delta_2\delta_3|.\end{aligned}$$

Then Definition 4.53 completes the proof. □

Example 4.55. For $f(x) = 1 + x$, $x \in (0, \text{OFL})$, the algorithm $\hat{f}(x) = \text{fl}(1.0 + \text{fl}(x))$ is not backward stable.

We prove a stronger statement that implies the negation of (4.27). For each $x \in (0, \epsilon_u)$, Definition 4.24 yields $\hat{f}(x) = 1.0$. Then $\hat{f}(x) = f(\hat{x})$ implies $\hat{x} = 0$, which further implies $E_{\text{rel}}(\hat{x}) = 1$.



Definition 4.56. An algorithm $\hat{f}(x)$ for computing $y = f(x)$ is *stable* or *numerically stable* iff

$$\forall x \in \text{dom}(f),\ \exists \hat{x} \in \text{dom}(f) \text{ s.t. } \begin{cases} \left|\frac{\hat{f}(x) - f(\hat{x})}{f(\hat{x})}\right| \le c_f \epsilon_u, \\ E_{\text{rel}}(\hat{x}) \le c\epsilon_u, \end{cases} \tag{4.29}$$

where c_f, c are two small constants.

Lemma 4.57. If an algorithm is backward stable, then it is numerically stable.

Proof. By Definition 4.52, $f(\hat{x}) = \hat{f}(x)$, hence $c_f = 0$. The other condition also follows trivially. □

Example 4.58. For $f(x) = 1 + x$, $x \in (0, \text{OFL})$, we show that the algorithm $\hat{f}(x) = \text{fl}(1.0 + \text{fl}(x))$ is stable.

If $|x| < \epsilon_u$, then $\hat{f}(x) = 1.0$. Choose $\hat{x} = x$, then $f(\hat{x}) - x = \hat{f}(x)$ and $\left|\frac{\hat{f}(x) - f(\hat{x})}{f(\hat{x})}\right| = \left|\frac{x}{1+x}\right| < 2\epsilon_u$.

Otherwise $|x| \ge \epsilon_u$. The definitions of the range and unit roundoff (Definitions 4.26 and 4.20) yield $x \in \mathcal{R}(\mathcal{F})$; here we have assumed $\epsilon_u >$UFL, which holds for all realworld FPN systems. By Theorem 4.27, $\hat{f}(x) = (1 + x(1 + \delta_1))(1 + \delta_2)$, i.e., $\hat{f}(x) = 1 + \delta_2 + x(1 + \delta_1 + \delta_2 + \delta_1\delta_2)$, where $|\delta_1|, |\delta_2| < \epsilon_u$.

Choose $\hat{x} = x(1 + \delta_1 + \delta_2 + \delta_1\delta_2)$ and we have

$$\begin{aligned}&E_{\text{rel}}(\hat{x}) = |\delta_1 + \delta_2 + \delta_1\delta_2| < 3\epsilon_u, \\ \Rightarrow &\left|\frac{\hat{f}(x) - f(\hat{x})}{f(\hat{x})}\right| = \left|\frac{\delta_2}{1 + \delta_2 + x(1 + \delta_1 + \delta_2 + \delta_1\delta_2)}\right| \le \epsilon_u,\end{aligned}$$

where the denominator is never close to zero since $x > 0$.

4.3.3 Condition numbers: scalar functions

Definition 4.59. The (relative) *condition number of a function* $y = f(x)$ is a measure of the relative change in the output for a small change in the input,

$$C_f(x) = \left|\frac{xf'(x)}{f(x)}\right|. \tag{4.30}$$

Definition 4.60. A problem with a low condition number is said to be *well-conditioned*. A problem with a high condition number is said to be *ill-conditioned*.

Example 4.61. Definition 4.59 yields

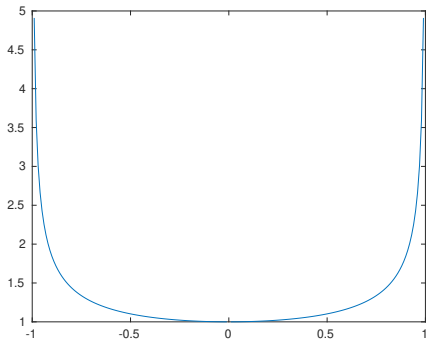
$$E_{\text{rel}}(\hat{y}) \lessapprox C_f E_{\text{rel}}(\hat{x}). \tag{4.31}$$

The approximation mark "$\lessapprox$" refers to the fact that the quadratic term $(\Delta x)^2$ has been ignored. As one way to interpret (4.31) and to understand Definition 4.59, *the computed solution to an ill-conditioned problem may have a large forward error.*

Example 4.62. For the function $f(x) = \arcsin(x)$, its condition number, according to Definition 4.59, is

$$C_f(x) = \left|\frac{xf'(x)}{f(x)}\right| = \frac{x}{\sqrt{1 - x^2}\arcsin x}.$$

Hence $C_f(x) \to +\infty$ as $x \to \pm 1$.



Lemma 4.63. Consider solving the equation $f(x) = 0$ near a simple root r, i.e. $f(r) = 0$ and $f'(r) \ne 0$. Suppose we perturb the function f to $F = f + \epsilon g$ where $f, g \in \mathcal{C}^2$, $g(r) \ne 0$, and $|\epsilon g'(r)| \ll |f'(r)|$. Then the root of F is $r + h$ where

$$h \approx -\epsilon\frac{g(r)}{f'(r)}. \tag{4.32}$$

Proof. Suppose $r + h$ is the new root, i.e. $F(r + h) = 0$, or,

$$f(r + h) + \epsilon g(r + h) = 0.$$

Taylor's expansion of $F(r + h)$ yields

$$f(r) + hf'(r) + \epsilon[g(r) + hg'(r)] = O(h^2)$$

and we have

$$h \approx -\epsilon\frac{g(r)}{f'(r) + \epsilon g'(r)} \approx -\epsilon\frac{g(r)}{f'(r)}.$$ □

Example 4.64 (Wilkinson)**.** Define

$$f(x) := \prod_{k=1}^{p}(x-k),$$
$$g(x) := x^p.$$

How is the root $x = p$ affected by perturbing f to $f + \epsilon g$?

By Lemma 4.63, the answer is

$$h \approx -\epsilon \frac{g(p)}{f'(p)} = -\epsilon \frac{p^p}{(p-1)!}.$$

For $p = 20, 30, 40$, the value of $\frac{p^p}{(p-1)!}$ is about 8.6×10^8, 2.3×10^{13}, 5.9×10^{17}, respectively. Hence a small change of the coefficient in the monomial x^p would cause a large change of the root. Consequently, the problem of root finding for polynomials with very high degrees is hopeless.

4.3.4 Condition numbers: vector functions

Definition 4.65. The *condition number of a vector function* $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$ is

$$\operatorname{cond}_{\mathbf{f}}(\mathbf{x}) = \frac{\|\mathbf{x}\| \, \|\nabla \mathbf{f}\|}{\|\mathbf{f}(\mathbf{x})\|}, \tag{4.33}$$

where $\|\cdot\|$ denotes a Euclidean norm such as the 1-, 2-, and ∞-norms.

Example 4.66. In solving the linear system $A\mathbf{u} = \mathbf{b}$, the algorithm can be viewed as taking the input $\mathbf{b}$ and returning the output $A^{-1}\mathbf{b}$, i.e. $\mathbf{f}(\mathbf{b}) = A^{-1}\mathbf{b}$. Clearly $\nabla \mathbf{f} = A^{-1}$. Definition 4.65 yields

$$\operatorname{cond}_{\mathbf{f}}(\mathbf{x}) = \frac{\|\mathbf{b}\| \left\|A^{-1}\right\|}{\|\mathbf{u}\|} = \frac{\|A\mathbf{u}\| \left\|A^{-1}\right\|}{\|\mathbf{u}\|}.$$

In practice the input $\mathbf{b}$ can take any value, hence we have

$$\max \operatorname{cond}_{\mathbf{f}}(\mathbf{x}) = \max_{\mathbf{u}\neq\mathbf{0}} \frac{\|A\mathbf{u}\| \left\|A^{-1}\right\|}{\|\mathbf{u}\|} = \|A\| \left\|A^{-1}\right\|,$$

where we have used the common definition

$$\|A\| := \max_{\mathbf{u}\neq\mathbf{0}} \frac{\|A\mathbf{u}\|}{\|\mathbf{u}\|}. \tag{4.34}$$

This explains Definition 4.67.

Definition 4.67. The *condition number of a nonsingular square matrix* A is

$$\operatorname{cond} A := \left\|A\right\| \left\|A^{-1}\right\|. \tag{4.35}$$

Lemma 4.68. Based on the 2-norm, the condition number of a nonsingular square matrix A is

$$\operatorname{cond}_2 A := \left\|A\right\|_2 \left\|A^{-1}\right\|_2 = \frac{\sigma_{\max}}{\sigma_{\min}},$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are respectively the largest and the smallest singular values of A. If A is also normal, we have

$$\operatorname{cond}_2 A = \frac{|\lambda_{\max}|}{|\lambda_{\min}|},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are eigenvalues of A with the largest and the smallest moduli, respectively. Furthermore, if A is unitary, we have $\operatorname{cond}_2 A = 1$.

Example 4.69. For the matrix

$$A = \begin{bmatrix} 1 & 1-\delta \\ 1 & 1+\delta \end{bmatrix} \tag{4.36}$$

and $\delta = 10^{-8}$, we have $\operatorname{cond}_2 A = 199999999.137258$.

Theorem 4.70. Let A be an invertible matrix, $\hat{A}$ a perturbation of A satisfying

$$\|A^{-1}\|\|\hat{A} - A\| < 1, \tag{4.37}$$

and $\mathbf{x}$ and $\hat{\mathbf{x}}$ solutions to the linear systems

$$A\mathbf{x} = \mathbf{b}, \quad \hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}}. \tag{4.38}$$

Then the relative error of $\hat{\mathbf{x}}$ approximating $\mathbf{x}$ is bounded as

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\operatorname{cond} A}{1 - (\operatorname{cond} A)\frac{\|\hat{A}-A\|}{\|A\|}} \left(\frac{\|\hat{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\hat{A} - A\|}{\|A\|} \right). \tag{4.39}$$

Proof. Write $\hat{A} = A[I + A^{-1}(\hat{A} - A)]$ and we have from Theorem E.141 and (4.37) that $\hat{A}$ is invertible and

$$\hat{A}^{-1} = [I + A^{-1}(\hat{A} - A)]^{-1} A^{-1}.$$

Then Theorem E.141 and the triangular inequality yield

$$\|\hat{A}^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\hat{A} - A\|}. \tag{4.40}$$

By (4.38) we can write $\hat{\mathbf{x}} - \mathbf{x} = \hat{A}^{-1}[\hat{\mathbf{b}} - \mathbf{b} - (\hat{A} - A)\mathbf{x}]$, which, together with (4.40) and the triangular inequality, yields

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{\operatorname{cond} A}{1 - \|\hat{A} - A\|\|A^{-1}\|} \left(\frac{\|\hat{\mathbf{b}} - \mathbf{b}\|}{\|A\|\|\mathbf{x}\|} + \frac{\|\hat{A} - A\|}{\|A\|} \right).$$

Then (4.39) follows from $\|A\|\|\mathbf{x}\| \ge \|\mathbf{b}\|$. □

Definition 4.71. The *componentwise condition number* of a vector function $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$ is

$$\operatorname{cond}_{\mathbf{f}}(\mathbf{x}) = \|A(\mathbf{x})\|, \tag{4.41}$$

where the matrix $A(\mathbf{x}) = [a_{ij}(\mathbf{x})]$ and each component is

$$a_{ij}(\mathbf{x}) = \left| \frac{x_j \frac{\partial f_i}{\partial x_j}}{f_i(\mathbf{x})} \right|. \tag{4.42}$$

Example 4.72. For the vector function

$$\mathbf{f}(\mathbf{x}) := \begin{bmatrix} \frac{1}{x_1} + \frac{1}{x_2} \\ \frac{1}{x_1} - \frac{1}{x_2} \end{bmatrix},$$

its Jacobian matrix is

$$\nabla \mathbf{f} = -\frac{1}{x_1^2 x_2^2} \begin{bmatrix} x_2^2 & x_1^2 \\ x_2^2 & -x_1^2 \end{bmatrix}.$$

The condition number based on Definition 4.71 clearly captures the fact that $x_1 \pm x_2 \approx 0$ leads to ill-conditioning,

$$C_c = \begin{bmatrix} \left|\frac{x_2}{x_1+x_2}\right| & \left|\frac{x_1}{x_1+x_2}\right| \\ \left|\frac{x_2}{x_1-x_2}\right| & \left|\frac{x_1}{x_1-x_2}\right| \end{bmatrix},$$

while that based on 1-norm of Definition 4.65 fails to capture the ill-conditioning,

$$C_1 = \frac{\|\mathbf{x}\|_1 \|\nabla \mathbf{f}\|_1}{\|\mathbf{f}\|_1} = \frac{|x_1|+|x_2|}{|x_1 x_2|} \frac{2\max(x_1^2, x_2^2)}{|x_1+x_2|+|x_1-x_2|},$$

in that the condition $x_1 \pm x_2 \approx 0$ yields $C_1 \approx 2$. Note that we have used the well-known formula

$$\forall A \in \mathbb{R}^{n\times n}, \qquad \|A\|_1 = \max_j \sum_i |a_{ij}|.$$

Example 4.73. If the parameters $t_1, t_2, \ldots, t_n$ are equally spaced in $[-1, 1]$, the condition number of the Vandermonde matrices in Definition 2.3 based on the ∞-norm is

$$\mathrm{cond}_\infty V_n \sim \frac{1}{\pi} e^{-\pi/4} e^{n/4(\pi+2\ln 2)},$$

which is 9.86×10^8 for $n = 20$.

Definition 4.74. The *Hilbert matrix* $H_n \in \mathbb{R}^{n\times n}$ is

$$h_{i,j} = \frac{1}{i+j-1}. \tag{4.43}$$

Example 4.75. The condition number of Hilbert matrices based on the 2-norm is

$$\mathrm{cond}_2 H_n \sim \frac{(\sqrt{2}+1)^{4n+4}}{2^{15/4}\sqrt{\pi n}},$$

which is 9.22×10^{14} for $n = 10$. See Example 5.36.

4.3.5 Condition numbers: algorithms

Definition 4.76. Consider approximating a function $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$ with an algorithm $\mathbf{f}_A : \mathcal{F}^m \to \mathcal{F}^n$. Assume

$$\forall \mathbf{x} \in \mathcal{F}^m, \ \exists \mathbf{x}_A \in \mathbb{R}^m \text{ s.t. } \mathbf{f}_A(\mathbf{x}) = \mathbf{f}(\mathbf{x}_A), \tag{4.44}$$

the *condition number of the algorithm* $\mathbf{f}_A$ is defined as

$$\mathrm{cond}_A(\mathbf{x}) = \frac{1}{\epsilon_u} \inf_{\{\mathbf{x}_A\}} \frac{\|\mathbf{x}_A - \mathbf{x}\|}{\|\mathbf{x}\|}. \tag{4.45}$$

Example 4.77. Consider an algorithm A for calculating $y = \ln x$. Suppose that, for any positive number x, this program produces a y_A satisfying $y_A = (1+\delta)\ln x$ where $|\delta| \le 5\epsilon_u$. What is the condition number of the algorithm?

We clearly have

$$y_A = \ln x_A \text{ where } x_A = x^{1+\delta},$$

and consequently

$$E_{\mathrm{rel}}(x_A) = \left|\frac{x^{1+\delta} - x}{x}\right| = |x^\delta - 1| = |e^{\delta \ln x} - 1|$$
$$\approx |\delta \ln x| \le 5|\ln x|\epsilon_u.$$

Hence A is well conditioned except when $x \to 0^+$.

Theorem 4.78. Suppose a smooth function $f : \mathbb{R} \to \mathbb{R}$ is approximated by an algorithm $A : \mathcal{F} \to \mathcal{F}$, producing $f_A(x) = f(x)(1+\delta(x))$ where $|\delta(x)| \le \varphi(x)\epsilon_u$. If $\mathrm{cond}_f(x)$ is bounded and nonzero, then we have

$$\forall x \in \mathcal{F}, \quad \mathrm{cond}_A(x) \le \frac{\varphi(x)}{\mathrm{cond}_f(x)}. \tag{4.46}$$

Proof. Assume $\forall x$, $\exists x_A$ such that $f(x_A) = f_A(x)$. Write $x_A = x(1+\epsilon_A)$ and we have

$$f(x)(1+\delta) = f(x_A) = f(x(1+\epsilon_A)) = f(x + x\epsilon_A)$$
$$= f(x) + x\epsilon_A f'(x) + O(\epsilon_A^2).$$

Neglecting the quadratic term yields

$$x\epsilon_A f'(x) = f(x)\delta$$
$$\Rightarrow \left|\frac{x_A - x}{x}\right| = |\epsilon_A| = \left|\frac{f(x)}{x f'(x)}\right| |\delta(x)|.$$

Dividing both sides by ϵ_u yields

$$\frac{1}{\epsilon_u}\left|\frac{x_A - x}{x}\right| = \frac{\delta(x)}{\epsilon_u \mathrm{cond}_f(x)}.$$

Take inf with respect to all x_A's, apply the condition $|\delta(x)| \le \varphi(x)\epsilon_u$, and we have (4.46). □

Example 4.79. Assume that $\sin x$ and $\cos x$ are computed with relative error within machine roundoff (this can be satisfied easily by truncating the Taylor series). Apply Theorem 4.78 to analyze the conditioning of the algorithm

$$f_A = \mathrm{fl}\left[\frac{\mathrm{fl}\big(1 - \mathrm{fl}(\cos x)\big)}{\mathrm{fl}(\sin x)}\right] \tag{4.47}$$

that computes $f(x) = \frac{1-\cos x}{\sin x}$ for $x \in (0, \pi/2)$.

By Definition 4.59, it is easy to compute that

$$\mathrm{cond}_f(x) = \frac{x}{\sin x}.$$

Furthermore, by Theorem 4.40 and the assumptions on $\sin x$ and $\cos x$, we have

$$f_A(x) = \frac{(1 - (\cos x)(1+\delta_1))(1+\delta_2)}{(\sin x)(1+\delta_3)}(1+\delta_4),$$

where $|\delta_i| \le \epsilon_u$ for $i = 1, 2, 3, 4$. Neglecting the quadratic terms of $O(\delta_i^2)$, the above equation is equivalent to

$$f_A(x) = \frac{1-\cos x}{\sin x}\left\{1 + \delta_2 + \delta_4 - \delta_3 - \delta_1 \frac{\cos x}{1-\cos x}\right\},$$

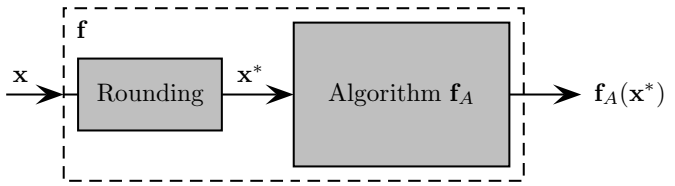
hence we have $\varphi(x) = 3 + \frac{\cos x}{1-\cos x}$ and

$$\mathrm{cond}_A(x) \le \frac{\sin x}{x}\left(3 + \frac{\cos x}{1-\cos x}\right).$$

Hence, $\mathrm{cond}_A(x)$ may be unbounded as $x \to 0$. On the other hand, $\mathrm{cond}_A(x)$ is controlled by $\frac{6}{\pi}$ as $x \to \frac{\pi}{2}$.

Exercise 4.80. Repeat Example 4.79 for $f(x) = \frac{\sin x}{1+\cos x}$ on the same interval.

4.3.6 Overall error of a computer solution



Theorem 4.81. Consider using normalized FPN arithmetics to solve a math problem

$$\mathbf{f}:\mathbb{R}^m\to\mathbb{R}^n,\qquad \mathbf{y}=\mathbf{f}(\mathbf{x}). \tag{4.48}$$

Denote the computer input and output as

$$\mathbf{x}^*\approx\mathbf{x},\qquad \mathbf{y}_A^*=\mathbf{f}_A(\mathbf{x}^*), \tag{4.49}$$

where $\mathbf{f}_A$ is the algorithm that approximates $\mathbf{f}$. The relative error of approximating $\mathbf{y}$ with $\mathbf{y}_A^*$ can be bounded as

$$E_{\text{rel}}(\mathbf{y}_A^*)\lessapprox E_{\text{rel}}(\mathbf{x}^*)\text{cond}_{\mathbf{f}}(\mathbf{x})+\epsilon_u\text{cond}_{\mathbf{f}}(\mathbf{x}^*)\,\text{cond}_A(\mathbf{x}^*), \tag{4.50}$$

where the relative error is defined in (4.25).

Proof. By the triangle inequality, we have

$$\begin{aligned}\frac{\|\mathbf{y}_A^*-\mathbf{y}\|}{\|\mathbf{y}\|}&=\frac{\|\mathbf{f}_A(\mathbf{x}^*)-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}\\&\le\frac{\|\mathbf{f}(\mathbf{x}^*)-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}+\frac{\|\mathbf{f}_A(\mathbf{x}^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x})\|}.\end{aligned}$$

By (4.31), the first term is

$$\begin{aligned}\frac{\|\mathbf{f}(\mathbf{x}^*)-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}&\lessapprox\text{cond}_{\mathbf{f}}(\mathbf{x})\frac{\|\mathbf{x}^*-\mathbf{x}\|}{\|\mathbf{x}\|}\\&=E_{\text{rel}}(\mathbf{x}^*)\text{cond}_{\mathbf{f}}(\mathbf{x}).\end{aligned}$$

By (4.31) and Definition 4.76, the second term is

$$\begin{aligned}\frac{\|\mathbf{f}_A(\mathbf{x}^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x})\|}=\frac{\|\mathbf{f}(\mathbf{x}_A^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x})\|}&\approx\frac{\|\mathbf{f}(\mathbf{x}_A^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x}^*)\|}\\&\le\text{cond}_{\mathbf{f}}(\mathbf{x}^*)\frac{\|\mathbf{x}_A^*-\mathbf{x}^*\|}{\|\mathbf{x}^*\|}\\&=\epsilon_u\text{cond}_A(\mathbf{x}^*)\,\text{cond}_{\mathbf{f}}(\mathbf{x}^*),\end{aligned}$$

where the last step follows from the fact that we only consider the $\mathbf{x}_A^*$ that is the least dangerous. □

4.4 Problems

4.4.1 Theoretical questions

I. Convert the decimal integer 477 to a normalized FPN with $\beta=2$.

II. Convert the decimal fraction 3/5 to a normalized FPN with $\beta=2$.

III. Let $x=\beta^e$, $e\in\mathbb{Z}$, $L<e<U$ be a normalized FPN in $\mathbb{F}$ and $x_L,x_R\in\mathbb{F}$ the two normalized FPNs adjacent to x such that $x_L<x<x_R$. Prove $x_R-x=\beta(x-x_L)$.

IV. By reusing your result of II, find out the two normalized FPNs adjacent to $x=3/5$ under the IEEE 754 single-precision protocol. What is $\text{fl}(x)$ and the relative roundoff error?

V. If the IEEE 754 single-precision protocol did not round off numbers to the nearest, but simply dropped excess bits, what would the unit roundoff be?

VI. How many bits of precision are lost in the subtraction $1-\cos x$ when $x=\frac{1}{4}$?

VII. Suggest at least two ways to compute $1-\cos x$ to avoid catastrophic cancellation caused by subtraction.

VIII. What are the condition numbers of the following functions? Where are they large?

- $(x-1)^\alpha$,
- $\ln x$,
- e^x,
- $\arccos x$.

IX. Consider the function $f(x)=1-e^{-x}$ for $x\in[0,1]$.

- Show that $\text{cond}_f(x)\le 1$ for $x\in[0,1]$.
- Let A be the algorithm that evaluates $f(x)$ for the machine number $x\in\mathbb{F}$. Assume that the exponential function is computed with relative error within machine roundoff. Estimate $\text{cond}_A(x)$ for $x\in[0,1]$.
- Plot $\text{cond}_f(x)$ and the estimated upper bound of $\text{cond}_A(x)$ as a function of x on $[0,1]$. Discuss your results.

X. Prove Lemma 4.68.

XI. The math problem of root finding for a polynomial

$$q(x)=\sum_{i=0}^n a_ix^i,\qquad a_n=1,a_0\neq 0,a_i\in\mathbb{R} \tag{4.51}$$

can be considered as a vector function $f:\mathbb{R}^n\to\mathbb{C}$:

$$r=f(a_0,a_1,\ldots,a_{n-1}).$$

Derive the componentwise condition number of f based on the 1-norm. For the Wilkinson example, compute your condition number, and compare your result with that in the Wilkinson Example. What does the comparison tell you?

XII. Suppose the division of two FPNs is calculated in a register of precision $2p$. Give an example that contradicts the conclusion of the model of machine arithmetic.

XIII. If the bisection method is used in single precision FPNs of IEEE 754 starting with the interval $[128,129]$, can we compute the root with absolute accuracy $<10^{-6}$? Why?

XIV. In fitting a curve by cubic splines, one gets inaccurate results when the distance between two adjacent points is much smaller than those of other adjacent pairs. Use the condition number of a matrix to explain this phenomenon.

4.4.2 Programming assignments

A. Print values of the functions in (4.52) at 101 equally spaced points covering the interval $[0.99, 1.01]$. Calculate each function in a straightforward way without rearranging or factoring. Note that the three functions are theoretically the same, but the computed values might be very different. Plot these functions near 1.0 using a magnified scale for the function values to see the variations involved. Discuss what you see. Which one is the most accurate? Why?

B. Consider a normalized FPN system $\mathbb{F}$ with the characterization $\beta = 2, p = 3, L = -1, U = +1$.

- compute UFL($\mathbb{F}$) and OFL($\mathbb{F}$) and output them as decimal numbers;
- enumerate all numbers in $\mathbb{F}$ and verify the corollary on the cardinality of $\mathbb{F}$ in the summary handout;
- plot $\mathbb{F}$ on the real axis;
- enumerate all the subnormal numbers of $\mathbb{F}$;
- plot the *extended* $\mathbb{F}$ on the real axis.

$$f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1 \tag{4.52a}$$

$$g(x) = (((((((x - 8)x + 28)x - 56)x + 70)x - 56)x + 28)x - 8)x + 1 \tag{4.52b}$$

$$h(x) = (x - 1)^8 \tag{4.52c}$$

Chapter 5

Best Approximation and Least Squares

Definition 5.1. Given a normed vector space Y of functions and its subspace $X \subset Y$. A function $\hat{\varphi} \in X$ is called the *best approximation* to $f \in Y$ from X with respect to the norm $\|\cdot\|$ iff

$$\forall \varphi \in X, \qquad \|f - \hat{\varphi}\| \leq \|f - \varphi\|. \tag{5.1}$$

Example 5.2. The Chebyshev Theorem 2.46 can be restated in the format of Definition 5.1 as follows. As in Example B.24, denote by $\mathbb{P}_n(\mathbb{R})$ the set of all polynomials with coefficients in $\mathbb{R}$ and degree at most n. For $Y = \mathbb{P}_n(\mathbb{R})$, and $X = \mathbb{P}_{n-1}(\mathbb{R})$, the best approximation to $f(x) = -x^n$ in Y from X with respect to the max-norm $\|\cdot\|_\infty$

$$\|g\|_\infty = \max_{x \in [-1,1]} |g(x)| \tag{5.2}$$

is $\hat{\varphi} = \frac{T_n}{2^{n-1}} - x^n$, where T_n is the Chebyshev polynomial of degree n. Clearly $\hat{\varphi}$ satisfies (5.1).

Definition 5.3. The *fundamental problem of linear approximation* is to find the best approximation $\hat{\varphi} = \sum_{i=1}^n a_i u_i$ to $f \in Y$ from n elements $u_1, u_2, \ldots, u_n \in X \subset Y$ that are linearly independent and given a priori.

Example 5.4. For $f(x) = e^x$ in $\mathcal{C}^\infty[-1,1]$, seeking its best approximation of the form $\hat{\varphi} = \sum_{i=1}^n a_i u_i$ in the subspace $X = \text{span}\{1, x, x^2, \ldots\}$ is a problem of linear approximation, where n can be any positive integer and the norm can be the max-norm (5.2), the 1-norm

$$\|g\|_1 := \int_{-1}^{+1} |g(x)| \mathrm{d}x, \tag{5.3}$$

or the 2-norm

$$\|g\|_2 := \left(\int_{-1}^{+1} |g(x)|^2 \mathrm{d}x \right)^{\frac{1}{2}}. \tag{5.4}$$

The three different norms are motivated differently: the max-norm corresponds to the min-max error, the 1-norm is related to the area bounded between $g(x)$ and the x-axis, and the 2-norm is related to the Euclidean distance, c.f. Section 5.4.

Example 5.5. For a simple closed curve $\gamma : [0,1) \to \mathbb{R}^2$ and n points $\mathbf{x}_i \in \gamma$, consider a spline approximation $p : [0,1) \to \mathbb{R}^2$ with its knots at $\mathbf{x}_i$'s and a scaled cumulative chordal length as in Definition 3.71. Denote by $\text{Int}(\gamma)$ as the complement of γ that always lies at the left of an observer who travels γ according to its parametrization. Then the area difference between $\mathcal{S}_1 := \text{Int}(\gamma)$ and $\mathcal{S}_2 := \text{Int}(p)$ can be defined as

$$\|\mathcal{S}_1 \oplus \mathcal{S}_2\|_1 := \int_{\mathcal{S}_1 \oplus \mathcal{S}_2} \mathrm{d}\mathbf{x},$$

where

$$\mathcal{S}_1 \oplus \mathcal{S}_2 := \mathcal{S}_1 \cup \mathcal{S}_2 \setminus (\mathcal{S}_1 \cap \mathcal{S}_2)$$

is the exclusive disjunction of S_1 and S_2.

The minimization of this area difference can be formulated by a best approximation problem based on the 1-norm.

Theorem 5.6. Suppose X is a finite-dimensional subspace of a normed vector space $(Y, \|\cdot\|)$. Then we have

$$\forall y \in Y,\ \exists \hat{\varphi} \in X \text{ s.t. } \forall \varphi \in X,\ \|\hat{\varphi} - y\| \leq \|\varphi - y\|. \tag{5.5}$$

Proof. For a given $y \in Y$, define a closed ball

$$B_y := \{x \in X : \|x\| \leq 2\|y\|\}.$$

Clearly $0 \in B_y$, and the distance from y to B_y is

$$\text{dist}(y, B_y) := \inf_{x \in B_y} \|y - x\| \leq \|y - 0\| = \|y\|.$$

By definition, any $z \in X$, $z \notin B_y$ must satisfy $\|z\| > 2\|y\|$, and thus

$$\|z - y\| \geq \|z\| - \|y\| > \|y\|.$$

Therefore, if a best approximation to y exists, it must be in B_y. As a subspace of X, B_y is finite dimensional, closed, and bounded, hence B_y is compact. The extreme value theorem states that a continuous scalar function attains its minimum and maximum on a compact set. A norm is a continuous function, hence the function $d : B_y \to \mathbb{R}^+ \cup \{0\}$ given by $d(x) = \|x - y\|$ must attain its minimum on B_y. □

Theorem 5.7. The set $\mathcal{C}[a,b]$ of continuous functions over $[a,b]$ is an inner-product space over $\mathbb{C}$ with its inner product as

$$\langle u, v \rangle := \int_a^b \rho(t) u(t) \overline{v(t)} \mathrm{d}t, \tag{5.6}$$

where $\overline{v(t)}$ is the complex conjugate of $v(t)$ and the *weight function* $\rho(x) \in \mathcal{C}[a,b]$ satisfies $\rho(x) > 0$ for all $x \in (a,b)$. In addition, $\mathcal{C}[a,b]$ with

$$\|u\|_2 := \left(\int_a^b \rho(t)|u(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}} \tag{5.7}$$

is a normed vector space over $\mathbb{R}$.

Proof. This follows from Definitions B.2, B.148, and B.153. □

Definition 5.8. The *least-square approximation* on $\mathcal{C}[a,b]$ is a best approximation problem with the norm in (5.1) set to that in (5.7).

5.1 Orthonormal systems

Definition 5.9. A subset S of an inner product space X is called *orthogonal* iff $\forall u, v \in S$, $\langle u, v\rangle > 0$ if $u = v$ and $\langle u, v\rangle = 0$ otherwise. The subset S is *orthonormal* if it is both orthogonal and satisfies $\forall u \in S, \langle u, u\rangle = 1$.

Example 5.10. The standard basis vectors in $\mathbb{R}^n$ are orthonormal.

Example 5.11. The Chebyshev polynomials of the first kind as in Definition 2.41 are orthogonal with respect to (5.6) where $a = -1, b = 1, \rho = \frac{1}{\sqrt{1-x^2}}$. However, they are not orthonormal; see Definition 5.9.

Theorem 5.12. Any finite set of nonzero orthogonal elements $u_1, u_2, \ldots, u_n$ is linearly independent.

Proof. This is easily proven by contradiction using Definitions B.25 and 5.9. □

Definition 5.13. The *Gram-Schmidt process* takes in a finite or infinite independent list $(u_1, u_2, \ldots)$ and output two other lists $(v_1, v_2, \ldots)$ and $(u_1^*, u_2^*, \ldots)$ by

$$v_{n+1} = u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u_k^* \rangle \, u_k^*, \tag{5.8a}$$

$$u_{n+1}^* = v_{n+1} / \|v_{n+1}\|, \tag{5.8b}$$

with the recursion basis as $v_1 = u_1,\ u_1^* = v_1/\|v_1\|$.

Theorem 5.14. For a finite or infinite independent list $(u_1, u_2, \ldots)$, the Gram-Schmidt process yields constants

$$\begin{matrix} a_{11} & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \\ \vdots & & \end{matrix}$$

such that $a_{kk} = \frac{1}{\|v_k\|} > 0$ and the elements $u_1^*, u_2^*, \ldots$

$$\begin{aligned} u_1^* &= a_{11} u_1 \\ u_2^* &= a_{21} u_1 + a_{22} u_2 \\ u_3^* &= a_{31} u_1 + a_{32} u_2 + a_{33} u_3 \\ &\vdots \end{aligned} \tag{5.9}$$

are orthonormal.

Proof. Definition 5.13 implies that the formulae (5.8) can be rewritten in the form of (5.9); this can be proven by induction. The induction basis is the recursion basis $u_1^* = u_1/\|u_1\|$ and the inductive step follows from (5.8) as

$$u_n^* = \frac{1}{\|v_n\|} \left(u_n - \sum_{k=1}^{n-1} \langle u_n, u_k^* \rangle \, u_k^* \right),$$

where each u_k^* is a linear combination of $u_1, u_2, \ldots, u_k$ and a_{nn}, the coefficient of u_n in (5.9), is clearly $\frac{1}{\|v_n\|}$. By (5.8b), u_{n+1}^* is normal. We show by induction that u_{n+1}^* is orthogonal to u_n^*, u_{n-1}^*, …, u_1^*. The induction base holds because

$$\begin{aligned} \langle v_2, u_1^* \rangle &= \langle u_2 - \langle u_2, u_1^* \rangle \, u_1^*, u_1^* \rangle \\ &= \langle u_2, u_1^* \rangle - \langle u_2, u_1^* \rangle \, \langle u_1^*, u_1^* \rangle = 0, \end{aligned}$$

where the second step follows from (IP-3) in Definition B.148 and the third step from u_1^* being normal. The inductive step also holds because for any $j < n+1$ we have

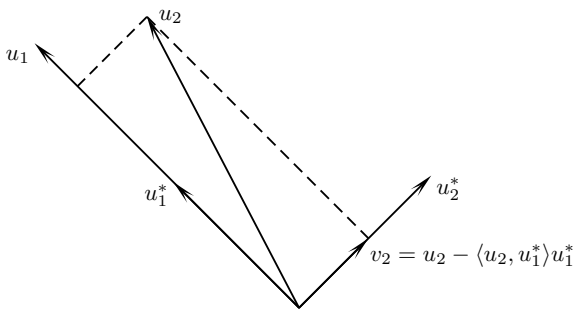
$$\begin{aligned} \left\langle v_{n+1}, u_j^* \right\rangle &= \left\langle u_{n+1} - \textstyle\sum_{k=1}^{n} \langle u_{n+1}, u_k^* \rangle \, u_k^*, u_j^* \right\rangle \\ &= \left\langle u_{n+1}, u_j^* \right\rangle - \textstyle\sum_{k=1}^{n} \langle u_{n+1}, u_k^* \rangle \left\langle u_k^*, u_j^* \right\rangle \\ &= \left\langle u_{n+1}, u_j^* \right\rangle - \left\langle u_{n+1}, u_j^* \right\rangle = 0, \end{aligned} \tag{5.10}$$

where the third step follows from the induction hypothesis and (5.8b), i.e.,

$$\left\langle u_k^*, u_j^* \right\rangle = \begin{cases} 1 & \text{if } k = j; \\ 0 & \text{otherwise.} \end{cases} \tag{5.11}$$

□

Exercise 5.15. Prove $a_{kk} = \frac{1}{\|v_k\|}$ by using $\|u_n^*\| = 1$.



Corollary 5.16. For a finite or infinite independent list $(u_1, u_2, \ldots)$, we can find constants

$$\begin{matrix} b_{11} & & \\ b_{21} & b_{22} & \\ b_{31} & b_{32} & b_{33} \\ \vdots & & \end{matrix}$$

and an orthonormal list $(u_1^*, u_2^*, \ldots)$ such that $b_{ii} > 0$ and

$$\begin{aligned} u_1 &= b_{11} u_1^* \\ u_2 &= b_{21} u_1^* + b_{22} u_2^* \\ u_3 &= b_{31} u_1^* + b_{32} u_2^* + b_{33} u_3^* \\ &\vdots \end{aligned} \tag{5.12}$$

Proof. This follows from (5.9) and that a lower-triangular matrix with positive diagonal elements is invertible. □

Corollary 5.17. In Theorem 5.14, we have $\langle u_n^*, u_i\rangle = 0$ for each $i = 1, 2, \ldots, n-1$.

Proof. By Corollary 5.16, each u_i can be expressed as

$$u_i = \sum_{k=1}^{i} b_{ik} u_k^*.$$

Inner product the above equation with u_n^*, apply the orthogonal conditions, and we reach the conclusion. □

Definition 5.18. Using the Gram-Schmidt orthonormalizing process with the inner product (5.6), we obtain from the independent list of monomials $(1, x, x^2 \ldots)$ the following *classic orthonormal polynomials*:

	a	b	$\rho(x)$
Chebyshev polynomials of the first kind	-1	1	$\frac{1}{\sqrt{1-x^2}}$
Chebyshev polynomials of the second kind	-1	1	$\sqrt{1-x^2}$
Legendre polynomials	-1	1	1
Jacobi polynomials	-1	1	$(1-x)^\alpha(1+x)^\beta$
Laguerre polynomials	0	$+\infty$	$x^\alpha e^{-x}$
Hermite polynomials	$-\infty$	$+\infty$	e^{-x^2}

where $\alpha, \beta > -1$ for Jacobi polynomials and $\alpha > -1$ for Laguerre polynomials.

Example 5.19. We compute the first 3 Legendre polynomials using the Gram-Schmidt process.

$$u_1 = 1,\ v_1 = 1, \|v_1\|^2 = \int_{-1}^{+1} \mathrm{d}x = 2,\ \ u_1^* = \frac{1}{\sqrt{2}}.$$

$$u_2 = x,\ v_2 = x - \left\langle x, \frac{1}{\sqrt{2}}\right\rangle \frac{1}{\sqrt{2}} = x,\ \|v_2\|^2 = \frac{2}{3},$$

$$u_2^* = \sqrt{\frac{3}{2}}x.$$

$$v_3 = x^2 - \left\langle x^2, \sqrt{\frac{3}{2}}x\right\rangle \sqrt{\frac{3}{2}}x - \left\langle x^2, \frac{1}{\sqrt{2}}\right\rangle \frac{1}{\sqrt{2}} = x^2 - \frac{1}{3},$$

$$\|v_3\|^2 = \int_{-1}^{+1} \left(x^2 - \frac{1}{3}\right)^2 \mathrm{d}x = \frac{8}{45},$$

$$u_3^* = \frac{3}{4}\sqrt{10}\left(x^2 - \frac{1}{3}\right).$$

5.2 Fourier expansions

Definition 5.20. Let $(u_1^*, u_2^*, \ldots)$ be a finite or infinite orthonormal list. The *orthogonal expansion* or *Fourier expansion* for an arbitrary w is the series

$$\sum_{n=1}^{m} \langle w, u_n^*\rangle\, u_n^*, \tag{5.13}$$

where the constants $\langle w, u_n^*\rangle$ are known as the *Fourier coefficients* of w and the term $\langle w, u_n^*\rangle\, u_n^*$ the *projection* of w on u_n^*. The *error of the Fourier expansion* of w with respect to $(u_1^*, u_2^*, \ldots)$ is simply $\sum_n \langle w, u_n^*\rangle\, u_n^* - w$.

Example 5.21. With the Euclidean inner product in Definition B.152, we select orthonormal vectors in $\mathbb{R}^3$ as

$$u_1^* = (1,0,0)^T,\ u_2^* = (0,1,0)^T,\ u_3^* = (0,0,1)^T.$$

For the vector $w = (a,b,c)^T$, the Fourier coefficients are

$$\langle w, u_1^*\rangle = a,\ \langle w, u_2^*\rangle = b,\ \langle w, u_3^*\rangle = c,$$

and the projections of w onto u_1^* and u_2^* are

$$\langle w, u_1^*\rangle\, u_1^* = (a,0,0)^T,\quad \langle w, u_2^*\rangle\, u_2^* = (0,b,0)^T.$$

The Fourier expansion of w is

$$w = \langle w, u_1^*\rangle\, u_1^* + \langle w, u_2^*\rangle\, u_2^* + \langle w, u_3^*\rangle\, u_3^*,$$

with the error of Fourier expansion as 0; see Theorem 5.23.

Exercise 5.22. For the orthonormal list in $L^2_{\rho=1}[-\pi, \pi]$,

$$\frac{1}{\sqrt{2\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos x}{\sqrt{\pi}}, \ldots, \frac{\sin(nx)}{\sqrt{\pi}}, \frac{\cos(nx)}{\sqrt{\pi}}, \ldots, \tag{5.14}$$

derive the *Fourier series* of a function $f(x)$ as

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{+\infty}(a_k \cos kx + b_k \sin kx), \tag{5.15}$$

where the coefficients are

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx \mathrm{d}x,\ b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx \mathrm{d}x.$$

Theorem 5.23. Let $u_1, u_2, \ldots, u_n$ be linearly independent and let u_i^* be the u_i's orthonormalized by the Gram-Schmidt process. If $w = \sum_{i=1}^n a_i u_i$, then

$$w = \sum_{i=1}^{n} \langle w, u_i^*\rangle\, u_i^*, \tag{5.16}$$

i.e. w is equal to its Fourier expansion.

Proof. By the condition $w = \sum_{i=1}^n a_i u_i$ and Corollary 5.16, we can express w as a linear combination of u_i^*'s,

$$w = \sum_{i=1}^{n} c_i u_i^*.$$

Then the orthonormality of u_i^*'s implies

$$\forall k = 1, 2, \cdots, n, \qquad \langle w, u_k^*\rangle = c_k,$$

which completes the proof. □

Theorem 5.24 (Minimum properties of Fourier expansions)**.** Let $u_1^*, u_2^*, \ldots$ be an orthonormal system and let w be arbitrary. Then

$$\left\| w - \sum_{i=1}^{N} \langle w, u_i^*\rangle\, u_i^* \right\| \le \left\| w - \sum_{i=1}^{N} a_i u_i^* \right\|, \tag{5.17}$$

for any selection of constants $a_1, a_2, \cdots, a_N$.

Proof. With the shorthand notation $\sum_i = \sum_{i=1}^N$, we deduce from the definition and properties of inner products

$$\begin{aligned}
&\left\| w - \sum_i a_i u_i^* \right\|^2 = \left\langle w - \sum_i a_i u_i^*, w - \sum_i a_i u_i^* \right\rangle \\
&= \langle w, w \rangle - \left\langle w, \sum_i a_i u_i^* \right\rangle - \left\langle \sum_i a_i u_i^*, w \right\rangle \\
&\quad + \left\langle \sum_i a_i u_i^*, \sum_i a_i u_i^* \right\rangle \\
&= \langle w, w \rangle - \sum_i \overline{a_i} \langle w, u_i^* \rangle - \sum_i a_i \langle u_i^*, w \rangle \\
&\quad + \sum_i \sum_j a_i \overline{a_j} \langle u_i^*, u_j^* \rangle \\
&= \langle w, w \rangle - \sum_i \overline{a_i} \langle w, u_i^* \rangle - \sum_i a_i \langle u_i^*, w \rangle + \sum_i |a_i|^2 \\
&\quad - \sum_i \langle u_i^*, w \rangle \langle w, u_i^* \rangle + \sum_i \langle u_i^*, w \rangle \langle w, u_i^* \rangle \\
&= \|w\|^2 - \sum_i |\langle w, u_i^* \rangle|^2 + \sum_i |a_i - \langle w, u_i^* \rangle|^2 , \qquad (5.18)
\end{aligned}$$

where "$|\cdot|$" denotes the modulus of a complex number. The first two terms are independent of a_i. Therefore $\|w - \sum_i a_i u_i^*\|^2$ is minimized only when $a_i = \langle w, u_i^* \rangle$. □

Corollary 5.25. Let $(u_1, u_2, \ldots, u_n)$ be an independent list. The fundamental problem of linearly approximating an arbitrary vector w is solved by the best approximation $\hat{\varphi} = \sum_k \langle w, u_k^* \rangle u_k^*$ where u_k^*'s are the u_k's orthonormalized by the Gram-Schmidt process. The error norm is

$$\|w - \hat{\varphi}\|^2 := \min_{a_k} \left\| w - \sum_{k=1}^n a_k u_k \right\|^2 = \|w\|^2 - \sum_{k=1}^n |\langle w, u_k^* \rangle|^2 . \qquad (5.19)$$

Proof. This follows directly from (5.18). □

Corollary 5.26 (Bessel inequality)**.** If $u_1^*, u_2^*, \ldots, u_N^*$ are orthonormal, then, for an arbitrary w,

$$\sum_{i=1}^N |\langle w, u_i^* \rangle|^2 \le \|w\|^2. \qquad (5.20)$$

Proof. This follows directly from Corollary 5.25 and the real positivity of a norm. □

Corollary 5.27. The Gram-Schmidt process in Definition 5.13 satisfies

$$\forall n \in \mathbb{N}^+, \quad \|v_{n+1}\|^2 = \|u_{n+1}\|^2 - \sum_{k=1}^n |\langle u_{n+1}, u_k^* \rangle|^2 . \qquad (5.21)$$

Proof. By (5.8a), each v_{n+1} can be regarded as the error of Fourier expansion of u_{n+1} with respect to the orthonormal list $(u_1^*, u_2^*, \ldots, u_n^*)$. In Corollary 5.25, identifying w with u_{n+1} completes the proof. □

Example 5.28. Consider the problem in Example 5.4 in the sense of least square approximation with the weight function $\rho = 1$. It is equivalent to

$$\min_{a_i} \int_{-1}^{+1} \left(e^x - \sum_{i=0}^n a_i x^i \right)^2 \mathrm{d}x. \qquad (5.22)$$

For $n = 1, 2$, use the Legendre polynomials derived in Example 5.19:

$$u_1^* = \frac{1}{\sqrt{2}}, \ u_2^* = \sqrt{\frac{3}{2}} x, \ u_3^* = \frac{1}{4}\sqrt{10}(3x^2 - 1),$$

and we have the Fourier coefficients of e^x as

$$\begin{aligned}
b_1 &= \int_{-1}^{+1} \frac{1}{\sqrt{2}} e^x \mathrm{d}x = \frac{1}{\sqrt{2}} \left(e - \frac{1}{e} \right), \\
b_2 &= \int_{-1}^{+1} \sqrt{\frac{3}{2}} x e^x \mathrm{d}x = \sqrt{6} e^{-1}, \\
b_3 &= \int_{-1}^{+1} \frac{1}{4}\sqrt{10}(3x^2 - 1) e^x \mathrm{d}x = \frac{\sqrt{10}}{2} \left(e - \frac{7}{e} \right).
\end{aligned}$$

The minimizing polynomials are thus

$$\hat{\varphi}_n = \begin{cases} \frac{1}{2e}(e^2 - 1) + \frac{3}{e} x & \text{if } n = 1; \\ \hat{\varphi}_1 + \frac{5}{4e}(e^2 - 7)(3x^2 - 1) & \text{if } n = 2. \end{cases} \qquad (5.23)$$

5.3 The normal equations

Theorem 5.29. Let $u_1, u_2, \ldots, u_n \in X$ be linearly independent and let u_i^* be the u_i's orthonormalized by the Gram-Schmidt process. Then, for any element w,

$$\forall j = 1, 2, \ldots, n, \qquad \left(w - \sum_{k=1}^n \langle w, u_k^* \rangle u_k^* \right) \perp u_j^*, \qquad (5.24)$$

where "$\perp$" denotes orthogonality.

Proof. If $w \in X$, we have $w - \sum_{k=1}^n \langle w, u_k^* \rangle u_k^* = \mathbf{0}$ and thus (5.24) holds trivially. For the other case of $w \notin X$, set $w = v_{n+1}$, apply (5.11), and we have (5.24). □

Corollary 5.30. Let $u_1, u_2, \ldots, u_n \in X$ be linearly independent. If $\hat{\varphi} = \sum_{k=1}^n a_k u_k$ is the best linear approximant to w, then

$$\forall j = 1, 2, \ldots, n, \qquad (w - \hat{\varphi}) \perp u_j. \qquad (5.25)$$

Proof. Since $\hat{\varphi} = \sum_{k=1}^n a_k u_k$ is the best linear approximant to w, Theorem 5.24 implies that

$$\sum_{k=1}^n a_k u_k = \sum_{k=1}^n \langle w, u_k^* \rangle u_k^*.$$

Corollary 5.16 and Theorem 5.29 complete the proof. □

Definition 5.31. Let $u_1, u_2, \ldots, u_n$ be a sequence of elements in an inner product space. The $n \times n$ matrix

$$G = G(u_1, u_2, \cdots, u_n) = (\langle u_i, u_j \rangle) = \begin{bmatrix} \langle u_1, u_1 \rangle & \langle u_1, u_2 \rangle & \ldots & \langle u_1, u_n \rangle \\ \langle u_2, u_1 \rangle & \langle u_2, u_2 \rangle & \ldots & \langle u_2, u_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, u_1 \rangle & \langle u_n, u_2 \rangle & \ldots & \langle u_n, u_n \rangle \end{bmatrix} \tag{5.26}$$

is the *Gram matrix* of $u_1, u_2, \ldots, u_n$. Its determinant

$$g = g(u_1, u_2, \ldots, u_n) = \det(\langle u_i, u_j \rangle) \tag{5.27}$$

is the *Gram determinant*.

Lemma 5.32. Let $w_i = \sum_{j=1}^n a_{ij} u_j$ for $i = 1, 2, \ldots, n$. Let $A = (a_{ij})$ and its conjugate transpose $A^H = (\overline{a_{ji}})$. Then we have

$$G(w_1, w_2, \ldots, w_n) = AG(u_1, u_2, \ldots, u_n)A^H \tag{5.28}$$

and

$$g(w_1, w_2, \ldots, w_n) = |\det A|^2 g(u_1, u_2, \ldots, u_n). \tag{5.29}$$

Proof. The inner product of u_i and w_j yields

$$\begin{aligned} &\begin{bmatrix} \langle u_1, w_1 \rangle & \langle u_1, w_2 \rangle & \ldots & \langle u_1, w_n \rangle \\ \langle u_2, w_1 \rangle & \langle u_2, w_2 \rangle & \ldots & \langle u_2, w_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, w_1 \rangle & \langle u_n, w_2 \rangle & \ldots & \langle u_n, w_n \rangle \end{bmatrix} \\ =& \begin{bmatrix} \langle u_1, u_1 \rangle & \langle u_1, u_2 \rangle & \ldots & \langle u_1, u_n \rangle \\ \langle u_2, u_1 \rangle & \langle u_2, u_2 \rangle & \ldots & \langle u_2, u_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, u_1 \rangle & \langle u_n, u_2 \rangle & \ldots & \langle u_n, u_n \rangle \end{bmatrix} \begin{bmatrix} \overline{a_{11}} & \ldots & \overline{a_{n1}} \\ \overline{a_{12}} & \ldots & \overline{a_{n2}} \\ \vdots & \ddots & \vdots \\ \overline{a_{1n}} & \ldots & \overline{a_{nn}} \end{bmatrix} \\ =& G(u_1, u_2, \ldots, u_n)A^H. \end{aligned}$$

Therefore (5.28) holds since

$$\begin{aligned} G(w_1, w_2, \ldots, w_n) =& \begin{bmatrix} \langle w_1, w_1 \rangle & \langle w_1, w_2 \rangle & \ldots & \langle w_1, w_n \rangle \\ \langle w_2, w_1 \rangle & \langle w_2, w_2 \rangle & \ldots & \langle w_2, w_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle w_n, w_1 \rangle & \langle w_n, w_2 \rangle & \ldots & \langle w_n, w_n \rangle \end{bmatrix} \\ =& \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ a_{21} & \ldots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} \langle u_1, w_1 \rangle & \langle u_1, w_2 \rangle & \ldots & \langle u_1, w_n \rangle \\ \langle u_2, w_1 \rangle & \langle u_2, w_2 \rangle & \ldots & \langle u_2, w_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, w_1 \rangle & \langle u_n, w_2 \rangle & \ldots & \langle u_n, w_n \rangle \end{bmatrix} \\ =& AG(u_1, u_2, \ldots, u_n)A^H. \end{aligned}$$

The following properties of complex conjugate are well known:

$$\overline{z + w} = \overline{z} + \overline{w}, \qquad \overline{zw} = \overline{z}\,\overline{w}.$$

Then the identity $\det(A) = \det(A^T)$ and the Leibniz formula of determinants (Definition B.227) yield

$$\overline{\det A} = \overline{\det A^T} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^n \overline{a_{i,\sigma_i}} = \det A^H.$$

Finally, (5.29) follows from the determinant of (5.28) and the identity $\det(AB) = \det(A)\det(B)$. □

Theorem 5.33. For nonzero elements $u_1, u_2, \ldots, u_n \in X$, we have

$$0 \le g(u_1, u_2, \ldots, u_n) \le \prod_{k=1}^n \|u_k\|^2, \tag{5.30}$$

where the lower equality holds if and only if $u_1, u_2, \ldots, u_n$ are linearly dependent and the upper equality holds if and only if they are orthogonal.

Proof. Suppose $u_1, u_2, \ldots, u_n$ are linearly dependent. Then we can find constants $c_1, c_2, \ldots, c_n$ satisfying $\sum_{i=1}^n c_i u_i = \mathbf{0}$ with at least one c_j being nonzero. Construct vectors

$$w_k = \begin{cases} \sum_{i=1}^n c_i u_i = \mathbf{0}, & k = j; \\ u_k, & k \ne j. \end{cases}$$

We have $g(w_1, w_2, \ldots, w_n) = 0$ because $\langle w_j, w_k \rangle = 0$ for each k. By the Laplace theorem (Theorem B.232), we can expand the determinant of $C = (c_{ij})$ according to minors of its jth row:

$$\begin{aligned} \det(C) = \det &\begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ c_1 & c_2 & \cdots & c_j & \cdots & c_n \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \\ &= 0 + \cdots + 0 + c_j + 0 + \cdots + 0 = c_j \ne 0, \end{aligned}$$

where the determinant of each minor matrix M_i of c_i with $i \ne j$ is zero because the i_0th row of each M_i is a row of all zeros, with $i_0 = i$ if $i < j$ and $i_0 = i - 1$ if $i > j$. Then Lemma 5.32 yields $g(u_1, u_2, \ldots, u_n) = 0$.

Now suppose $u_1, u_2, \ldots, u_n$ are linearly independent. Theorem 5.14 yields constants a_{ij} such that $a_{kk} > 0$ and the following vectors are orthonormal:

$$u_k^* = \sum_{i=1}^k a_{ki} u_i.$$

Then Definition 5.31 implies $g(u_1^*, u_2^*, \ldots, u_n^*) = 1$. Also, we have $\det(a_{ij}) = \prod_{k=1}^n a_{kk}$ because the matrix (a_{ij}) is triangular. It then follows from Lemma 5.32 that

$$g(u_1, u_2, \ldots, u_n) = \prod_{k=1}^n \frac{1}{a_{kk}^2} > 0. \tag{5.31}$$

Since the list of vectors $(u_1, u_2, \ldots, u_n)$ is either dependent or independent, the arguments so far show that $g(u_1, u_2, \ldots, u_n) = 0$ if and only if $u_1, u_2, \ldots, u_n$ are linearly dependent.

Suppose $u_1, u_2, \ldots, u_n$ are orthogonal. By Definition 5.31, $G(u_1, u_2, \ldots, u_n)$ is a diagonal matrix with $\|u_k\|^2$ on the diagonals. Hence the orthogonality of u_k's implies

$$g(u_1, u_2, \ldots, u_n) = \prod_{k=1}^n \|u_k\|^2. \tag{5.32}$$

For the converse statement, suppose (5.32) holds. Then $u_1, u_2, \ldots, u_n$ must be independent because otherwise it would contradict the lower equality of (5.30) proved as above. Apply the Gram-Schmidt process to $(u_1, u_2, \ldots, u_n)$ and we know from Theorem 5.14 that $\frac{1}{a_{kk}} = \|v_k\|$. Set the length of the list in Theorem 5.14 to $1, 2, \ldots, n$ and we know from (5.31) and (5.32) that

$$\forall k = 1, 2, \ldots, n, \qquad \|u_k\|^2 = \|v_k\|^2. \tag{5.33}$$

Then Corollary 5.27 and (5.33) imply

$$\forall k = 1, 2, \ldots, n, \qquad \sum_{j=1}^{k-1} \left|\langle u_k, u_j^* \rangle\right|^2 = 0,$$

which further implies

$$\forall k = 1, 2, \ldots, n,\ \forall j = 1, 2, \ldots, k-1, \qquad \langle u_k, u_j^* \rangle = 0,$$

which, together with Corollary 5.16, implies the orthogonality of u_k's. Finally, we remark that the maximum of $g(u_1, u_2, \ldots, u_n)$ is indeed $\prod_{k=1}^n \|u_k\|^2$ because of (5.31), $\frac{1}{a_{kk}} = \|v_k\|$, and Corollary 5.27. □

Theorem 5.34. Let $\hat{\varphi} = \sum_{i=1}^n a_i u_i$ be the best approximation to w constructed from the list of independent vectors $(u_1, u_2, \ldots, u_n)$. Then the coefficients

$$\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$$

are uniquely determined from the linear system of *normal equations*,

$$G(u_1, u_2, \ldots, u_n)^T \mathbf{a} = \mathbf{c}, \tag{5.34}$$

where $\mathbf{c} = [\langle w, u_1 \rangle, \langle w, u_2 \rangle, \ldots, \langle w, u_n \rangle]^T$.

Proof. Take inner product of $\hat{\varphi} = \sum_{i=1}^n a_i u_i$ with u_j, apply Corollary 5.30, and we have

$$\langle w, u_j \rangle = \sum_{k=1}^n a_k \langle u_k, u_j \rangle,$$

which is simply the jth equation of (5.34). The uniqueness of the coefficients follows from Theorem 5.33 and Cramer's rule. □

Example 5.35. Solve Example 5.28 by normal equations.

To find the best approximation $\hat{\varphi} = a_0 + a_1 x + a_2 x^2$ to e^x from the linearly independent list $(1, x, x^2)$, we first construct the Gram matrix from (5.26), (5.6), and $\rho = 1$:

$$G(1, x, x^2) = \begin{bmatrix} \langle 1, 1 \rangle & \langle 1, x \rangle & \langle 1, x^2 \rangle \\ \langle x, 1 \rangle & \langle x, x \rangle & \langle x, x^2 \rangle \\ \langle x^2, 1 \rangle & \langle x^2, x \rangle & \langle x^2, x^2 \rangle \end{bmatrix} = \begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix}.$$

We then calculate the vector

$$\mathbf{c} = \begin{bmatrix} \langle e^x, 1 \rangle \\ \langle e^x, x \rangle \\ \langle e^x, x^2 \rangle \end{bmatrix} = \begin{bmatrix} e - 1/e \\ 2/e \\ e - 5/e \end{bmatrix}.$$

The normal equations then yields

$$a_0 = \frac{3(11 - e^2)}{4e},\ a_1 = \frac{3}{e},\ a_2 = \frac{15(e^2 - 7)}{4e}.$$

With these values, it is easily verified that the best approximation $\hat{\varphi} = a_0 + a_1 x + a_2 x^2$ equals that in (5.23).

Example 5.36 (The Hilbert matrix as the Gram matrix)**.** Consider the least-square approximation of a given function $f \in \mathcal{C}[0, 1]$ by a polynomial $p(x) = \sum_{k=0}^n \alpha_k x^k$. If the monomials are used as the basis of the subspace, their Gram matrix is the Hilbert matrix as in Definition 4.74 because

$$\langle x^j, x^k \rangle = \int_0^1 x^j x^k \mathrm{d}x = \frac{1}{j + k + 1}.$$

The coefficients α_k's are determined by the normal equations

$$\forall j = 0, 1, \ldots, n, \quad \sum_{k=0}^n \frac{\alpha_k}{j + k + 1} = \int_0^1 f(x) x^j \mathrm{d}x. \tag{5.35}$$

For the special case of $f(x) = \frac{1}{1+x}$, the RHS

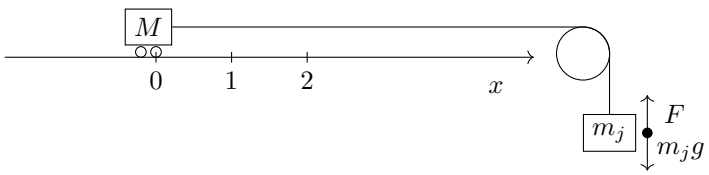
$$r_j := \int_0^1 \frac{x^j}{1 + x} \mathrm{d}x$$

can be derived as

$$\forall j = 0, 1, \ldots, n, \quad r_j = (-1)^j \ln 2 + \sum_{i=1}^j (-1)^{i+j} \frac{1}{i}. \tag{5.36}$$

5.4 Discrete least squares (DLS)

Example 5.37 (Processing experimental data on Newton's second law using discrete least squares)**.** A cart with mass M is pulled along a horizontal track by a cable attached to a weight of mass m_j through a pulley.



Neglect the friction of the track and the pulley system, assume Newton's second law holds and we have

$$m_j g = (m_j + M) a_j = (m_j + M) \frac{\mathrm{d}^2 x}{\mathrm{d}t^2}.$$

The following experiments verify Newton's second law.

(i) For fixed M and m_j, we measure a number of data points (t_i, x_i) by recording the position of the cart with a high-speed camera.

(ii) Fit a quadratic polynomial $p(t) = c_0 + c_1 t + c_2 t^2$ by minimizing the total length squared,

$$\min \sum_i (x_i - p(t_i))^2.$$

(iii) Take $a_j = 2c_2$ as the experimental result of acceleration for the force $F_j = m_j(g - a_j)$.

(iv) Change the weight m_j and repeat steps (i)-(iii) a number of times to get data points (a_j, F_j).

(v) Fit a linear polynomial $f(x) = c_0 + c_1 x$ by minimizing the total length squared,

$$\min \sum_j (F_j - f(a_j))^2,$$

where the expressions in steps (ii) and (v) justify the name "least squares."

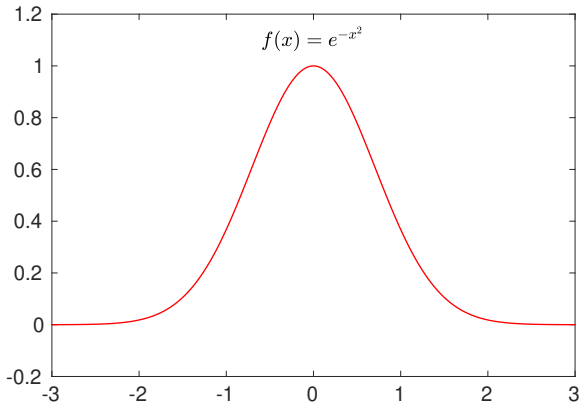
One verifies Newton's second law by showing that the data fitting result c_1 is very close to M. This verification process have neither derived nor proved Newton's second law. But we claim that we did not find any self-inconsistency of Newton's second law, nor did we find any discrepancy between the experimental results and Newton's second law.

5.4.1 Gaussian and Dirac delta functions

Definition 5.38. A *Gaussian function*, or a *Gaussian*, is a function of the form

$$f(x) = a \exp\left(-\frac{(x-b)^2}{2c^2}\right), \tag{5.37}$$

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 RMS (root mean square) width*.



Lemma 5.39. The integral of a Gaussian is

$$\int_{-\infty}^{+\infty} a e^{-\frac{(x-b)^2}{2c^2}} \mathrm{d}x = ac\sqrt{2\pi}. \tag{5.38}$$

Proof. By the trick of combining two one-dimensional Gaussians and the Polar coordinate transformation, we have

$$\begin{aligned}
\int_{-\infty}^{+\infty} e^{-x^2} \mathrm{d}x &= \sqrt{\left(\int_{-\infty}^{+\infty} e^{-x^2} \mathrm{d}x\right)\left(\int_{-\infty}^{+\infty} e^{-y^2} \mathrm{d}y\right)} \\
&= \sqrt{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)} \, \mathrm{d}x\mathrm{d}y} \\
&= \sqrt{\int_0^{2\pi}\int_0^{+\infty} e^{-r^2} r \, \mathrm{d}r\mathrm{d}\theta} \\
&= \sqrt{2\pi \cdot \left(-\frac{1}{2}e^{-r^2}\right)\Big|_0^{+\infty}} = \sqrt{\pi},
\end{aligned}$$

and hence

$$\int_{-\infty}^{+\infty} a e^{-\frac{(x-b)^2}{2c^2}} \mathrm{d}x = \sqrt{2}ac \int_{-\infty}^{+\infty} e^{-y^2} \mathrm{d}y = ac\sqrt{2\pi},$$

where it follows from the transformation of $x = b + \sqrt{2}cy$. □

Definition 5.40. A *normal distribution* or *Gaussian distribution* is a continuous probability distribution of the form

$$f_{\mu,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{5.39}$$

where μ is the *mean* or *expectation* and σ is the *standard deviation*.

Definition 5.41. The *Dirac delta function* $\delta(x-\bar{x})$ centered at $\bar{x}$ is

$$\delta(x-\bar{x}) = \lim_{\epsilon\to 0} \phi_\epsilon(x-\bar{x}) \tag{5.40}$$

where $\phi_\epsilon(x-\bar{x}) = f_{\bar{x},\epsilon}$ is a normal distribution with its mean at $\bar{x}$ and its standard deviation as ϵ.

Lemma 5.42. The Dirac delta function satisfies

$$\delta(x-\bar{x}) = \begin{cases} +\infty, & x = \bar{x}, \\ 0, & x \neq \bar{x}; \end{cases} \tag{5.41a}$$

$$\int_{-\infty}^{+\infty} \delta(x-\bar{x})\mathrm{d}x = 1. \tag{5.41b}$$

Proof. These follow directly from Definitions 5.40 and 5.41 and Lemma 5.39. □

Lemma 5.43 (Sifting property of δ)**.** If $f : \mathbb{R} \to \mathbb{R}$ is continuous, then

$$\int_{-\infty}^{+\infty} \delta(x-\bar{x}) f(x) \mathrm{d}x = f(\bar{x}). \tag{5.42}$$

Proof. Since $I_\epsilon := [\bar{x}-\epsilon, \bar{x}+\epsilon]$ is a compact interval and $f(x)$ is continuous over I_ϵ, $f(x)$ is bounded over I_ϵ, say, $f(x) \in [m, M]$. The nonnegativeness of ϕ_ϵ and the integral mean value theorem C.72 imply that

$$(*): \qquad \int_{-\infty}^{+\infty} \delta(x-\bar{x}) f(x) \mathrm{d}x = \int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon} \delta(x-\bar{x}) f(x) \mathrm{d}x$$

is bounded within the interval

$$\left[m \int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon} \delta(x-\bar{x})\mathrm{d}x, \; M \int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon} \delta(x-\bar{x})\mathrm{d}x\right].$$

It follows that

$$\lim_{\epsilon\to 0} \int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon} \delta(x-\bar{x}) f(x)\mathrm{d}x = f(\bar{x}) \lim_{\epsilon\to 0} \int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon} \delta(x-\bar{x})\mathrm{d}x = f(\bar{x}).$$

Apply $\lim_{\epsilon\to 0}$ to $(*)$ and we have (5.42). □

Definition 5.44. The *Heaviside function* or *step function* is

$$H(x) := \begin{cases} 0 & \text{if } x < 0; \\ 1 & \text{if } x \geq 0. \end{cases} \tag{5.43}$$

Lemma 5.45. The Dirac delta function and the Heaviside function are related as

$$\int_{-\infty}^{x} \delta(t)\mathrm{d}t = H(x). \tag{5.44}$$

Proof. This follows from Definitions 5.41 and 5.44 and Lemma 5.42. □

5.4.2 Reusing the formalism

Definition 5.46. Define a function $\lambda : \mathbb{R} \to \mathbb{R}$

$$\lambda(t) = \begin{cases} 0 & \text{if } t \in (-\infty, a), \\ \int_a^t \rho(\tau)\mathrm{d}\tau & \text{if } t \in [a, b], \\ \int_a^b \rho(\tau)\mathrm{d}\tau & \text{if } t \in (b, +\infty). \end{cases} \tag{5.45}$$

Then a corresponding *continuous measure* $\mathrm{d}\lambda$ can be defined as

$$\mathrm{d}\lambda = \begin{cases} \rho(t)\mathrm{d}t & \text{if } t \in [a, b], \\ 0 & \text{otherwise}, \end{cases} \tag{5.46}$$

where the *support of the continuous measure* $\mathrm{d}\lambda$ is the interval $[a, b]$.

Definition 5.47. The *discrete measure* or the *Dirac measure* associated with the point set $\{t_1, t_2, \ldots, t_N\}$ is a measure $\mathrm{d}\lambda$ that is nonzero only at the points t_i and has the value ρ_i there. The *support of the discrete measure* is the set $\{t_1, t_2, \ldots, t_N\}$.

Lemma 5.48. For a function $u : \mathbb{R} \to \mathbb{R}$, define

$$\lambda(t) = \sum_{i=1}^{N} \rho_i H(t - t_i), \tag{5.47}$$

and we have

$$\int_{\mathbb{R}} u(t)\mathrm{d}\lambda = \sum_{i=1}^{N} \rho_i u(t_i). \tag{5.48}$$

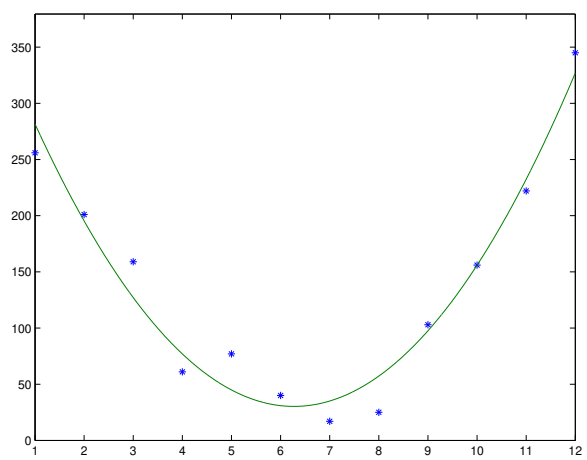
Proof. By (5.47) and Lemmas 5.43 and 5.45, we have

$$\int_{\mathbb{R}} u(t)\mathrm{d}\lambda = \int_{\mathbb{R}} \sum_{i=1}^{N} \rho_i \delta(t - t_i) u(t)\mathrm{d}t = \sum_{i=1}^{N} \rho_i u(t_i). \qquad \square$$

5.4.3 DLS via normal equations

Example 5.49. Consider a table of sales record.

x	1	2	3	4	5	6
y	256	201	159	61	77	40
x	7	8	9	10	11	12
y	17	25	103	156	222	345



From the plot of the discrete data, it appears that a quadratic polynomial would be a good fit. Hence we formulate the least square problem as finding the coefficients of a quadratic polynomial to minimize

$$\sum_{i=1}^{12} \left(y_i - \sum_{j=0}^{2} a_j x_i^j \right)^2 .$$

Reusing the procedures in Example 5.35, we have

$$G(1, x, x^2) = \begin{bmatrix} \langle 1,1\rangle & \langle 1,x\rangle & \langle 1,x^2\rangle \\ \langle x,1\rangle & \langle x,x\rangle & \langle x,x^2\rangle \\ \langle x^2,1\rangle & \langle x^2,x\rangle & \langle x^2,x^2\rangle \end{bmatrix} = \begin{bmatrix} 12 & 78 & 650 \\ 78 & 650 & 6084 \\ 650 & 6084 & 60710 \end{bmatrix},$$

$$\mathbf{c} = \begin{bmatrix} \langle y,1\rangle \\ \langle y,x\rangle \\ \langle y,x^2\rangle \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{12} y_i \\ \sum_{i=1}^{12} y_i x_i \\ \sum_{i=1}^{12} y_i x_i^2 \end{bmatrix} = \begin{bmatrix} 1662 \\ 11392 \\ 109750 \end{bmatrix}.$$

Then the normal equations yield

$$\mathbf{a} = G^{-1}\mathbf{c} = \begin{bmatrix} 386.00 \\ -113.43 \\ 9.04 \end{bmatrix}.$$

The corresponding polynomial is shown in the above plot.

5.4.4 DLS via QR decomposition

Definition 5.50. A matrix $A \in \mathbb{R}^{n\times n}$ is *orthogonal* iff $A^T A = I$.

Definition 5.51. A matrix A is *upper triangular* iff

$$\forall i, j, \qquad i > j \;\Rightarrow\; a_{i,j} = 0.$$

Similarly, a matrix A is *lower triangular* iff

$$\forall i, j, \qquad i < j \;\Rightarrow\; a_{i,j} = 0.$$

Theorem 5.52 (QR factorization)**.** For any $A \in \mathbb{R}^{m\times n}$, there exists an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m\times n}$ so that $A = QR$.

Proof. Rewrite $A = [\xi_1, \xi_2, \cdots, \xi_n] \in \mathbb{R}^{m\times n}$ and denote by r the column rank of A. Construct a rank-r matrix

$$A_r = [u_1, u_2, \ldots, u_r]$$

by the following steps.

(S-1) Set $u_1 = \xi_{k_1}$ where k_1 satisfies $\xi_{k_1} \neq \mathbf{0}$ and $\forall \ell < k_1$, $\xi_\ell = \mathbf{0}$.

(S-2) For each $j = 2, \ldots, r$, set $u_j = \xi_{k_j}$ where k_j satisfies that $K_j = (\xi_{k_1}, \ldots, \xi_{k_j})$ is a list of independent column vectors and, $\forall \ell \in R_j := \{k_{j-1} + 1, \ldots, k_j - 1\}$, ξ_ℓ can be expressed as a linear combination of the column vectors in K_{j-1}.

In plain words, (S-1) means that we jump over the lead zero vectors and (S-2) states that, starting from u_{j-1}, we pick the first vector as u_j that is not in $\text{span}(u_1, u_2, \ldots, u_{j-1})$.

By Corollary 5.16, the Gram-Schmidt process determines a unique orthogonal matrix $A_r^* = [u_1^*,\ u_2^*,\ \ldots,\ u_r^*] \in \mathbb{R}^{m\times r}$ and a unique upper triangular matrix such that

$$A_r = A_r^* \begin{bmatrix} b_{11} & b_{21} & \ldots & b_{r1} \\ & b_{22} & \ldots & b_{r2} \\ & & \ddots & \vdots \\ & & & b_{rr} \end{bmatrix}. \tag{5.49}$$

By definition of the column rank of a matrix, we have $r \le m$.

In the rest of this proof, we insert each column vector in $X = \{\xi_1, \xi_2, \ldots, \xi_n\} \backslash \{u_1, u_2, \ldots, u_r\}$ back into (5.49) and show that the QR form of (5.49) is maintained. For those zero column vectors in (S-1), we have

$$\begin{aligned} A_\xi =& [\xi_1\ \ldots\ \xi_{k_1-1}\ u_1\ u_2\ \ldots\ u_r] \\ =& A_r^* \begin{bmatrix} 0 & \ldots & 0 & b_{11} & b_{21} & \ldots & b_{r1} \\ 0 & \ldots & 0 & & b_{22} & \ldots & b_{r2} \\ \vdots & \ddots & \vdots & & & \ddots & \vdots \\ 0 & \ldots & 0 & & & & b_{rr} \end{bmatrix}. \end{aligned} \tag{5.50}$$

For each ξ_ℓ with $\ell \in R_j$ in (S-2), we have

$$\begin{aligned} &[u_1,\ u_2,\ \ldots,\ u_{j-1},\ \xi_\ell] \\ =& \begin{bmatrix} u_1^*,\ u_2^*,\ \ldots,\ u_{j-1}^* \end{bmatrix} \begin{bmatrix} b_{11} & \ldots & b_{j-1,1} & c_{\ell,1} \\ & \ddots & \vdots & \vdots \\ & & b_{j-1,j-1} & c_{\ell,j-1} \end{bmatrix}, \end{aligned} \tag{5.51}$$

where $\xi_\ell = c_{\ell,1}u_1^* + \ldots + c_{\ell,j-1}u_{j-1}^*$. With (5.50) as the induction basis and (5.51) as the inductive step, it is straightforward to prove by induction that we have $A = A_r^* R$ where R is an upper triangular matrix.

If $r = m$, Definitions 5.50 and 5.9 complete the proof. Otherwise $r < m$ and the proof is completed by the well-known fact in linear algebra that a list of orthonormal vectors can be extended to an orthonormal basis. □

Lemma 5.53. An orthogonal matrix preserves the 2-norm of the vectors it acts on.

Proof. Definition 5.50 yields

$$\forall \mathbf{x} \in \text{dom}(Q), \qquad \|Q\mathbf{x}\|_2^2 = \mathbf{x}^T Q^T Q \mathbf{x} = \mathbf{x}^T \mathbf{x} = \|x\|_2^2. \quad \square$$

Theorem 5.54. Consider an over-determined linear system $A\mathbf{x} = \mathbf{b}$ where $A \in \mathbb{R}^{m\times n}$ and $m \ge n$. The discrete linear least square problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|Ax - b\|_2^2$$

is solved by $\mathbf{x}^*$ satisfying

$$R_1 \mathbf{x}^* = \mathbf{c}, \tag{5.52}$$

where $R_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{c} \in \mathbb{R}^n$ result from the QR factorization of A:

$$Q^T A = R = \begin{bmatrix} R_1 \\ \mathbf{0} \end{bmatrix}, \qquad Q^T \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{r} \end{bmatrix}. \tag{5.53}$$

Furthermore, the minimum is $\|\mathbf{r}\|_2^2$.

Proof. For any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|Q^T A\mathbf{x} - Q^T\mathbf{b}\|_2^2 = \|R_1\mathbf{x} - \mathbf{c}\|_2^2 + \|\mathbf{r}\|_2^2,$$

where the first step follows from Lemma 5.53. □

5.5 Solving ill-posed linear systems

Definition 5.55. A problem in mathematical physics is *well-posed* iff (i) it admits a solution, (ii) the solution is unique, and (iii) the problem has a small condition number; otherwise it is *ill-posed*.

Lemma 5.56 (Solvability of a linear system)**.** A linear system $A\mathbf{x} = \mathbf{b}$ with $A \in \mathbb{C}^{m\times n}$ admits a solution if and only if $\mathbf{b}$ is in the range of A, or equivalently, $\mathbf{b}$ is perpendicular to the null space of the adjoint of A, i.e.,

$$\forall \mathbf{z} \in \{\mathbf{y} \in \mathbb{C}^m : A^*\mathbf{y} = \mathbf{0}\}, \quad \langle \mathbf{b}, \mathbf{z}\rangle = 0, \tag{5.54}$$

in which case a particular solution of the linear system is

$$\mathbf{x}^\oplus := \sum_{j=1}^r \frac{1}{\sigma_j} \langle \mathbf{b}, \mathbf{v}_j\rangle \mathbf{u}_j, \tag{5.55}$$

where r is the rank of A and $(\sigma_j, \mathbf{u}_j, \mathbf{v}_j)$ is the singular system of A, c.f. Definition B.214.

Proof. Suppose the linear system has a solution $\mathbf{x}$. Then

$$\langle \mathbf{b}, \mathbf{z}\rangle = \langle A\mathbf{x}, \mathbf{z}\rangle = \langle \mathbf{x}, A^*\mathbf{z}\rangle = 0.$$

As for the sufficiency, suppose (5.54) holds. Then we have

$$\mathbf{b} = \sum_{j=1}^r \langle \mathbf{b}, \mathbf{v}_j\rangle \mathbf{v}_j = \sum_{j=1}^r \frac{1}{\sigma_j} \langle \mathbf{b}, \mathbf{v}_j\rangle A\mathbf{u}_j = A\mathbf{x}^\oplus,$$

where the first step follows from (B.81b), the second from (B.81a), and the last from (5.55). □

5.5.1 No solutions or multiple solutions

Theorem 5.57. Suppose the linear system $A\mathbf{x} = \mathbf{b}$ in Lemma 5.56 has no solutions. Then $\mathbf{x}^\oplus$ in (5.55) solves the problem of minimizing the 2-norm of the residual, i.e.,

$$\mathbf{x}^\oplus = \arg\min_{\mathbf{x}\in\mathbb{C}^n} \|A\mathbf{x} - \mathbf{b}\|_2. \tag{5.56}$$

Proof. The Fourier expansion of $\mathbf{b}$ is

$$\mathbf{b} = \sum_{i=1}^m \langle \mathbf{b}, \mathbf{v}_j\rangle \mathbf{v}_j.$$

It follows from (5.55) that

$$\forall \mathbf{b} \notin \text{range}A, \quad A\mathbf{x}^\oplus - \mathbf{b} \in \text{span}(\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_m).$$

On the other hand, the rank of A being r implies

$$\forall \mathbf{x} \in \mathbb{C}^n, \quad A\mathbf{x} - A\mathbf{x}^\oplus \in \text{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r).$$

Hence $\langle A\mathbf{x}^\oplus - \mathbf{b}, A\mathbf{x} - A\mathbf{x}^\oplus\rangle = 0$ and the Pythagorean theorem B.170 yields

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|A\mathbf{x} - A\mathbf{x}^\oplus\|_2^2 + \|A\mathbf{x}^\oplus - \mathbf{b}\|_2^2,$$

which completes the proof. □

Theorem 5.58. Suppose a linear system $A\mathbf{x} = \mathbf{b}$ has multiple solutions. Then its general solution is $\mathbf{x}^{\oplus} + \mathbf{y}$ where $\mathbf{x}^{\oplus}$ is given by (5.55) and $\mathbf{y}$ is any vector in the null space of A. Furthermore, $\mathbf{x}^{\oplus}$ is the solution with the minimum 2-norm.

Proof. By Lemma 5.56, we have $\mathbf{x}^{\oplus} \in \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r)$ and $\operatorname{null} A = \operatorname{span}(\mathbf{u}_{r+1}, \mathbf{u}_{r+1}, \dots, \mathbf{u}_n)$. Hence $\forall \mathbf{y} \in \operatorname{null} A$, $\langle \mathbf{x}^{\oplus}, \mathbf{y} \rangle = 0$ and the Pythagorean theorem B.170 yields

$$\|\mathbf{x}^{\oplus} + \mathbf{y}\|_2 = \|\mathbf{x}^{\oplus}\|_2 + \|\mathbf{y}\|_2.$$

The proof is completed by the fact that all solutions of the linear system are of the form $\mathbf{x}^{\oplus} + \mathbf{y}$. □

5.5.2 The Moore-Penrose inverse

Definition 5.59. The *Moore-Penrose inverse* or *pseudo-inverse* or *generalized inverse* of a matrix $A \in \mathbb{F}^{m\times n}$ where $\mathbb{F} = \mathbb{R}, \mathbb{C}$ is the matrix $A^+ \in \mathbb{F}^{n\times m}$ given by

$$A^+\mathbf{y} = \sum_{j=1}^{r} \frac{1}{\sigma_j} \langle \mathbf{y}, \mathbf{v}_j \rangle \mathbf{u}_j, \tag{5.57}$$

where r is the rank of A and $(\sigma_j, \mathbf{u}_j, \mathbf{v}_j)$ is the singular system of A, c.f. Definition B.214.

Theorem 5.60. The pseudo-inverse in Definition 5.59 has the following properties.

(PDI-1) AA^+ maps all columns of A to themselves: $AA^+A = A$;

(PDI-2) A^+ acts like a weak inverse: $A^+AA^+ = A^+$;

(PDI-3) Both AA^+ and A^+A are Hermitian: $(AA^+)^* = AA^+$, $(A^+A)^* = A^+A$.

Lemma 5.61. If A has linearly independent columns, then A^*A is invertible and we have

$$A^+ = (A^*A)^{-1}A^*, \tag{5.58}$$

which is then a *left inverse*, i.e. $A^+A = I$. Similarly, if A has linearly independent rows, then AA^* is invertible and

$$A^+ = A^*(AA^*)^{-1}, \tag{5.59}$$

which is then a *right inverse*, i.e. $AA^+ = I$.

5.5.3 Spectral cutoff for ill-conditioning

Example 5.62. The condition number of solving the linear system $\mathbf{Ax} = \mathbf{b}$ is dominated by the small singular values. If we perturb $\mathbf{b}$ to $\hat{\mathbf{b}} = \mathbf{b} + \epsilon\mathbf{v}_j$ for some $\epsilon \in \mathbb{C}$, by Lemma 5.56 the solution would change from $\mathbf{x}$ to $\hat{\mathbf{x}} = \mathbf{x} + \frac{\epsilon}{\sigma_j}\mathbf{u}_j$. Hence the condition number based on *absolute* errors is

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_2}{\|\hat{\mathbf{b}} - \mathbf{b}\|_2} = \frac{1}{\sigma_j}.$$

Thus a solvable linear system is ill-conditioned if its matrix has small singular values.

Definition 5.63. The *spectral cutoff* is a method to stabilize an ill-conditioned linear system by neglecting the terms with small singular values in the solution (5.55).

Definition 5.64. Given a perturbed RHS $\hat{\mathbf{b}}$ to a linear system $A\mathbf{x} = \mathbf{b}$, the *discrepancy principle* is a strategy to come up with an approximate solution $\hat{\mathbf{x}}$ so that

$$\|A\hat{\mathbf{x}} - \hat{\mathbf{b}}\|_2 = C\|\mathbf{b} - \hat{\mathbf{b}}\|_2, \tag{5.60}$$

where $C \geq 1$ is some suitable constant.

Example 5.65. By the discrepancy principle for the spectral cutoff, we approximate the solution (5.55) by

$$\mathbf{x}_p := \begin{cases} \mathbf{0} & \text{if } p = 0; \\ \sum_{j=1}^{p} \frac{1}{\sigma_j} \left\langle \hat{\mathbf{b}}, \mathbf{v}_j \right\rangle \mathbf{u}_j & \text{if } p = 1, \dots, r, \end{cases} \tag{5.61}$$

where p is determined by the perturbation error $\|\hat{\mathbf{b}} - \mathbf{b}\|_2$ so that (5.60) is satisfied. A common choice is $C = 1$.

Theorem 5.66. Suppose the linear system $A\mathbf{x} = \mathbf{b}$ with $A \in \mathbb{C}^{m\times n}$ is solvable and a perturbation $\hat{\mathbf{b}}$ to $\mathbf{b}$ satisfies

$$\|\hat{\mathbf{b}} - \mathbf{b}\|_2 \leq \epsilon \leq \|\hat{\mathbf{b}}\|_2. \tag{5.62}$$

Then there exists a smallest integer $p = p(\epsilon)$ such that

$$\|A\mathbf{x}_p - \hat{\mathbf{b}}\|_2 \leq \epsilon, \tag{5.63}$$

where $\mathbf{x}_p$ is defined in (5.61). The discrepancy principle for the spectral cutoff is *convergent* in the sense that

$$\lim_{\epsilon\to 0} \mathbf{x}_p = A^+\mathbf{b}. \tag{5.64}$$

Proof. The function $f : \{0, 1, \dots, r\} \to \mathbb{R}$ given by

$$f(p) := \|A\mathbf{x}_p - \hat{\mathbf{b}}\|_2 - \epsilon^2$$

can be expressed as

$$f(p) = \sum_{j=p+1}^{m} \left| \left\langle \hat{\mathbf{b}}, \mathbf{v}_j \right\rangle \right|^2 - \epsilon^2, \tag{5.65}$$

where $\mathbf{v}_j$ is a right singular vector of A. Then (5.62) yields $f(0) = \|\hat{\mathbf{b}}\|_2 - \epsilon^2 \geq 0$ and $f(r) = -\epsilon^2 < 0$ if the rank r of A equals m. If $r < m$, the solvability condition of the linear system and Lemma 5.56 yield

$$\forall j = r + 1, \cdots, m, \quad \langle \mathbf{b}, \mathbf{v}_j \rangle = 0.$$

Then we have

$$f(r) = \sum_{j=r+1}^{m} \left| \left\langle (\hat{\mathbf{b}} - \mathbf{b}), \mathbf{v}_j \right\rangle \right|^2 - \epsilon^2 \leq \|\hat{\mathbf{b}} - \mathbf{b}\|_2^2 - \epsilon^2 \leq 0$$

where the last step follows from (5.62). By (5.65), f is monotonically decreasing, and the first conclusion concerning (5.63) follows from $f(0) \geq 0$ and $f(r) \leq 0$.

As for the second conclusion, (5.63) yields

$$\|A\mathbf{x}_p - \mathbf{b}\|_2 \leq \|A\mathbf{x}_p - \hat{\mathbf{b}}\|_2 + \|\hat{\mathbf{b}} - \mathbf{b}\|_2 \leq 2\epsilon.$$

Then (5.64) follows from

$$\forall \mathbf{v} \in \operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_r), \quad A^+A\mathbf{v} = \mathbf{v},$$

which is a consequence of Lemma 5.61. □

5.5.4 Tikhonov regularization

Definition 5.67. *Tikhonov regularization* is a method to approximate the solution of an ill-conditioned linear system $A\mathbf{x} = \mathbf{b}$ by that of a regularized linear system

$$\alpha\mathbf{x} + A^*A\mathbf{x} = A^*\mathbf{b}, \tag{5.66}$$

where $\alpha > 0$ is called the *regularization parameter*.

Theorem 5.68. Suppose a matrix $A \in \mathbb{C}^{m\times n}$ has rank r. Then for any $\mathbf{b} \in \mathbb{C}^m$ the Tikhonov regularized linear system (5.66) admits the unique solution

$$\mathbf{x}_\alpha := \sum_{j=1}^{r} \frac{\sigma_j}{\alpha + \sigma_j^2} \langle \mathbf{b}, \mathbf{v}_j\rangle \, \mathbf{u}_j, \tag{5.67}$$

where $(\sigma_j, \mathbf{u}_j, \mathbf{v}_j)$ is the singular system of A, c.f. Definition B.214. Furthermore, the Tikhonov regularization satisfies

$$\lim_{\alpha\to 0}(\alpha I + A^*A)^{-1}A^*\mathbf{b} = A^+\mathbf{b}. \tag{5.68}$$

Proof. The matrix $B := I + A^*A$ is strictly positive definite and thus nonsingular. Furthermore, $B\mathbf{u}_j = (\alpha + \sigma_j^2)\mathbf{u}_j$ and $B^* = B$ imply that $(\alpha + \sigma_j^2, \mathbf{u}_j, \mathbf{u}_j)$ is a singular system of B. Apply Lemma 5.56 to (5.66) and we have

$$\mathbf{x} = \sum_{j=1}^{r} \frac{1}{\alpha + \sigma_j^2} \langle A^*\mathbf{b}, \mathbf{u}_j\rangle \, \mathbf{u}_j = \sum_{j=1}^{r} \frac{1}{\alpha + \sigma_j^2} \langle \mathbf{b}, A\mathbf{u}_j\rangle \, \mathbf{u}_j = \mathbf{x}_\alpha,$$

where the last step follows from (B.81a) and (5.67). Finally, (5.68) follows directly from (5.57) and (5.67). □

Theorem 5.69. For any $\mathbf{b} \in \mathbb{C}^m$, the solution $\mathbf{x}_\alpha$ in (5.67) to the Tikhonov regularized linear system (5.66) is equivalent to the unique minimizer of $\|A\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$, i.e.,

$$\|A\mathbf{x}_\alpha - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}_\alpha\|_2^2 = \inf_{\mathbf{x}\in\mathbb{C}^n} \left\{\|A\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2\right\}. \tag{5.69}$$

Proof. Define $\mathbf{z}_\alpha := \alpha\mathbf{x}_\alpha + A^*A\mathbf{x}_\alpha - A^*\mathbf{b}$. The law of cosines (Theorem B.158) implies

$$\begin{aligned}&\|A\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2 = 2\mathrm{Re}\,\langle \mathbf{x} - \mathbf{x}_\alpha, \mathbf{z}_\alpha\rangle \\ &+\|A\mathbf{x}_\alpha - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}_\alpha\|_2^2 + \|A\mathbf{x} - A\mathbf{x}_\alpha\|_2^2 + \alpha\|\mathbf{x} - \mathbf{x}_\alpha\|_2^2.\end{aligned} \tag{5.70}$$

Hence the solution $\mathbf{x}_\alpha$ of (5.66) satisfies $\mathbf{z}_\alpha = \mathbf{0}$ and consequently (5.69) holds.

Conversely, let $\mathbf{x}_\alpha$ be the solution of (5.69) and suppose $\mathbf{z}_\alpha \neq \mathbf{0}$. Then for $\mathbf{x} := \mathbf{x}_\alpha - \epsilon\mathbf{z}_\alpha$ with $\epsilon \in \mathbb{R}$, (5.70) yields

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2 = \|A\mathbf{x}_\alpha - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}_\alpha\|_2^2 - 2\epsilon a + \epsilon^2 b,$$

where $a := \|\mathbf{z}_\alpha\|_2^2$ and $b := \|A\mathbf{z}_\alpha\|_2^2 + \alpha\|\mathbf{z}_\alpha\|_2^2$. The choice of $\epsilon = \frac{a}{b}$ gives

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2 < \|A\mathbf{x}_\alpha - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}_\alpha\|_2^2,$$

which contradicts the starting point. □

Example 5.70. In practical applications, the RHS side of a linear system is often perturbed with some uncertainties, say, $\|\hat{\mathbf{b}} - \mathbf{b}\|_2 < \epsilon$. Then the Tikhonov regularization approximates the solution $\mathbf{x} = A^+\mathbf{b}$ by $\mathbf{x}_\alpha$ that satisfies

$$B\mathbf{x}_\alpha = A^*\hat{\mathbf{b}},$$

where $B := \alpha I + A^*A$. Hence the total absolute error is

$$\mathbf{x}_\alpha - \mathbf{x} = B^{-1}A^*(\hat{\mathbf{b}} - \mathbf{b}) + (B^{-1}A^*\mathbf{b} - A^+\mathbf{b}),$$

where the first RHS term is the error caused by the RHS data and the second is the error caused by the regularization. On the one hand, α should be small to keep the regularization error small. On the other hand, Lemma 4.68 implies

$$\mathrm{cond}_2 B = \frac{\alpha + \sigma_1^2}{\alpha + \sigma_n^2}.$$

Thus α should be large to make the data error small; otherwise the conditioning of the linear system $B\mathbf{x} = A^*(\hat{\mathbf{b}} - \mathbf{b})$ would be large. Note that this conclusion remains valid even if the condition number is based on absolute errors as in Example 5.62. To sum up, the choice of α have to be made through a compromise between accuracy and stability.

Theorem 5.71. Suppose the linear system $A\mathbf{x} = \mathbf{b}$ with $A \in \mathbb{C}^{m\times n}$ is solvable and a perturbation $\hat{\mathbf{b}}$ to $\mathbf{b}$ satisfies

$$\|\hat{\mathbf{b}} - \mathbf{b}\|_2 \le \epsilon \le \|\hat{\mathbf{b}}\|_2. \tag{5.71}$$

Then there exists a unique $\alpha = \alpha(\epsilon)$ such that

$$\|A\mathbf{x}_\alpha - \hat{\mathbf{b}}\|_2 = \epsilon, \tag{5.72}$$

where $\mathbf{x}_\alpha$ is the unique solution of

$$(\alpha I + A^*A)\mathbf{x} = A^*\hat{\mathbf{b}}. \tag{5.73}$$

In addition, the discrepancy principle for Tikhonov regularization is *convergent*, i.e.,

$$\lim_{\epsilon\to 0} \mathbf{x}_\alpha = A^+\mathbf{b}. \tag{5.74}$$

Proof. By (5.67), the function $f : (0, +\infty) \to \mathbb{R}$ given by

$$f(\alpha) := \|A\mathbf{x}_\alpha - \hat{\mathbf{b}}\|_2^2 - \epsilon^2$$

can be expressed as

$$f(\alpha) = \sum_{j=1}^{m} \left| \frac{\alpha^2}{(\alpha + \sigma^2)^2} \left\langle \hat{\mathbf{b}}, \mathbf{v}_j \right\rangle \right|^2 - \epsilon^2, \tag{5.75}$$

where $\mathbf{v}_j$ is a right singular vector of A. Being continuous and strictly monotonically increasing, f also satisfies

$$\lim_{\alpha\to 0} f(\alpha) = -\epsilon^2 < 0, \quad \lim_{\alpha\to+\infty} f(\alpha) = \|\hat{\mathbf{b}}\|_2^2 - \epsilon^2 \ge 0,$$

hence f must have exactly one root α that satisfies (5.72).

Then (5.72) and the triangle inequality yield

$$\|\hat{\mathbf{b}}\|_2 - \epsilon = \|\hat{\mathbf{b}}\|_2 - \|A\mathbf{x}_\alpha - \hat{\mathbf{b}}\|_2 \le \|A\mathbf{x}_\alpha\|_2.$$

On the other hand, $\mathbf{x}_\alpha$ being a solution of (5.73) gives

$$\alpha\|A\mathbf{x}_\alpha\|_2 = \|AA^*(\hat{\mathbf{b}} - A\mathbf{x}_\alpha)\|_2 \le \|AA^*\|_2\epsilon.$$

By (5.71), we have $\|\hat{\mathbf{b}}\|_2 \ge \|\mathbf{b}\|_2 - \epsilon$, which, together with the above two inequalities, gives

$$\alpha \le \frac{\|AA^*\|_2\epsilon}{\|\mathbf{b}\|_2 - 2\epsilon}.$$

Hence $\lim_{\epsilon\to 0}\alpha = 0$. Then (5.74) follows from (5.71), (5.57), and (5.67). □

5.6 Problems

5.6.1 Theoretical questions

I. Give a detailed proof of Theorem 5.7.

II. Consider the Chebyshev polynomials of the first kind.

(a) Show that they are orthogonal on $[-1, 1]$ with respect to the inner product in Theorem 5.7 with the weight function $\rho(x) = \frac{1}{\sqrt{1-x^2}}$.

(b) Normalize the first three Chebyshev polynomials to arrive at an orthonormal system.

III. Least-square approximation of a continuous function. Approximate the circular arc given by the equation $y(x) = \sqrt{1-x^2}$ for $x \in [-1, 1]$ by a quadratic polynomial with respect to the inner product in Theorem 5.7.

(a) $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ with Fourier expansion,

(b) $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ with normal equations.

IV. Discrete least square via orthonormal polynomials. Consider the example on the table of sales record in Example 5.49.

(a) Starting from the independent list $(1, x, x^2)$, construct orthonormal polynomials by the Gram-Schmidt process using

$$\langle u(t), v(t)\rangle = \sum_{i=1}^{N}\rho(t_i)u(t_i)v(t_i) \tag{5.76}$$

as the inner product with $N = 12$ and $\rho(x) = 1$.

(b) Find the best approximation $\hat{\varphi} = \sum_{i=0}^{2} a_i x^i$ such that $\|y - \hat{\varphi}\| \le \|y - \sum_{i=0}^{2} b_i x^i\|$ for all $b_i \in \mathbb{R}$. Verify that $\hat{\varphi}$ is the same as that of the example on the table of sales record in the notes.

(c) Suppose there are other tables of sales record in the same format as that in the example. Values of N and x_i's are the same, but the values of y_i's are different. Which of the above calculations can be reused? Which cannot be reused? What advantage of orthonormal polynomials over normal equations does this reuse imply?

V. Prove Theorem 5.60 and Lemma 5.61.

5.6.2 Programming assignments

A. Write a program to perform discrete least square via normal equations. Your subroutine should take two arrays x and y as the input and output three coefficients a_0, a_1, a_2 that determines a quadratic polynomial as the best fitting polynomial in the sense of least squares with the weight function $\rho = 1$.

Run your subroutine on the following data.

x	0.0	0.5	1.0	1.5	2.0	2.5	3.0
y	2.9	2.7	4.8	5.3	7.1	7.6	7.7
x	3.5	4.0	4.5	5.0	5.5	6.0	6.5
y	7.6	9.4	9.0	9.6	10.0	10.2	9.7
x	7.0	7.5	8.0	8.5	9.0	9.5	10.0
y	8.3	8.4	9.0	8.3	6.6	6.7	4.1

B. Write a program to solve the previous discrete least square problem via QR factorization. Report the condition number based on the 2-norm of the matrix G in the normal-equation approach and that of the matrix R_1 in the QR-factorization approach, verifying that the former is much larger than the latter.

C. First, prove that the solution of (5.35) is of the form

$$\forall j = 0, 1, \ldots, n, \quad \alpha_j = \beta_j \ln 2 + \gamma_j,$$

where β_j and γ_j are rational numbers. Hint: derive (5.36) from $r_0 = \ln 2$ and the geometric series

$$\sum_{i=1}^{j}(-1)^{i-1}x^{i-1} = \frac{1-(-1)^j x^j}{1+x}.$$

Second, write a program to perform Gauss elimination on the Hilbert matrix in terms of rational numbers to obtain β_j and γ_j.

Third, for $n = 1, 2, \ldots, 6$, compute α_j's by approximating $\ln 2$ to machine precision and present your results in a table to show convergence of $\alpha_j = (-1)^j$ to the Taylor series of $f(x) = \frac{1}{1+x}$.

Fourth, compute the RHS r_j's in (5.36) by floating-point arithmetic with $\ln 2 \approx 0.69315$ (the relative error is 10^{-5}) and then solve (5.35) to obtain α_j's. Can you still obtain convergence? Why? Your explanation should contain at least the condition numbers of the Hilbert matrix for $n = 2, 3, \ldots, 6$.

Finally, use Tikhonov regularization with $\alpha = 10^{-10}$ to solve (5.35); show that you can recover convergence at least for α_0 and α_1.

Chapter 6

Numerical Integration and Differentiation

Definition 6.1. A *weighted quadrature formula* $I_n(f)$ is a linear functional

$$I_n(f) := \sum_{k=1}^{n} w_k f(x_k) \tag{6.1}$$

that approximates the integral of a function $f \in \mathcal{C}[a,b]$,

$$I(f) := \int_a^b f(x)\rho(x)\mathrm{d}x, \tag{6.2}$$

where the weight function $\rho \in \mathcal{C}[a,b]$ satisfies $\forall x \in (a,b)$, $\rho(x) > 0$. The points x_k's at which the integrand f is evaluated are called *nodes* or *abscissas*, and the multiplier w_k's are called *weights* or *coefficients*.

Example 6.2. If a and/or b are infinite, $I(f)$ and $I_n(f)$ in (6.1) may still be well defined if the *moment of weight function*

$$\mu_j := \int_a^b x^j \rho(x)\mathrm{d}x \tag{6.3}$$

exists and is finite for all $j \in \mathbb{N}$.

6.1 Accuracy and convergence

Definition 6.3. The *remainder*, or *error*, of $I_n(f)$ is

$$E_n(f) := I(f) - I_n(f). \tag{6.4}$$

$I_n(f)$ is said to be *convergent* for $\mathcal{C}[a,b]$ iff

$$\forall f \in \mathcal{C}[a,b], \qquad \lim_{n\to+\infty} I_n(f) = I(f). \tag{6.5}$$

Definition 6.4. A subset $\mathbb{V} \subset \mathcal{C}[a,b]$ is *dense* in $\mathcal{C}[a,b]$ iff

$$\forall f \in \mathcal{C}[a,b], \forall \epsilon > 0, \exists f_\epsilon \in \mathbb{V}, \text{ s.t. } \max_{x\in[a,b]} |f(x) - f_\epsilon(x)| \le \epsilon. \tag{6.6}$$

Theorem 6.5. Let $\{I_n(f) : n \in \mathbb{N}^+\}$ be a sequence of quadrature formulas that approximate $I(f)$, where I_n and $I(f)$ are defined in (6.1) and (6.2). Let $\mathbb{V}$ be a dense subset of $\mathcal{C}[a,b]$. $I_n(f)$ is convergent for $\mathcal{C}[a,b]$ if and only if

(a) $\forall f \in \mathbb{V}$, $\lim_{n\to+\infty} I_n(f) = I(f)$,

(b) $\exists B \in \mathbb{R}$ s.t. $\forall n \in \mathbb{N}^+, W_n := \sum_{k=1}^n |w_k| < B$.

Proof. For sufficiency, we need to prove that for any given f we have $\lim_{n\to+\infty} I_n(f) = I(f)$. To this end, we find $f_\epsilon \in \mathbb{V}$ such that (6.6) holds, define $K := \max_{x\in[a,b]} |f(x) - f_\epsilon(x)|$. Then we have

$$\begin{aligned}
|E_n(f)| \le& |I(f) - I(f_\epsilon)| + |I(f_\epsilon) - I_n(f_\epsilon)| + |I_n(f_\epsilon) - I_n(f)| \\
=& \left| \int_a^b [f(x) - f_\epsilon(x)]\,\rho(x)\mathrm{d}x \right| \\
&+ |I(f_\epsilon) - I_n(f_\epsilon)| + \left| \sum_{k=1}^n w_k \left[f(x_k) - f_\epsilon(x_k) \right] \right| \\
\le& K \left[\int_a^b \rho(x)\mathrm{d}x + \sum_{k=1}^n |w_k| \right] + |I(f_\epsilon) - I_n(f_\epsilon)|,
\end{aligned}$$

where the first step follows from the triangular inequality, the second from Definition 6.1, and the third from the definition of K. The terms inside the brackets is bounded because of $\rho \in \mathcal{C}[a,b]$ and condition (b). By condition (a), $|I(f_\epsilon) - I_n(f_\epsilon)|$ can be made arbitrarily small. Since K can also be arbitrarily small, we have (6.5).

For necessity, it is trivial to deduce (a) from (6.5). In contrast, it is nontrivial to deduce (b) from (6.5) as the process involves some key theorems in functional analysis. A reader not familiar with the principle of uniform boundedness may skip the rest of the proof.

The numerical quadrature formula $I_n : \mathcal{C}[a,b] \to \mathbb{R}$ is a linear functional and is continuous at $f = \mathbf{0}$ because of Definition E.58 and the fact that

$$\begin{aligned}
&\forall \epsilon > 0,\ \exists \delta = \frac{\epsilon}{2\sum_k |w_k|},\ \text{s.t. } \forall f \in \mathcal{C}[a,b], \\
&\|f - \mathbf{0}\|_\infty < \delta \ \Rightarrow\ |I_n(f) - I_n(\mathbf{0})| = \left| \sum_k w_k f(x_k) \right| \\
&\qquad \le \sum_k |w_k||f(x_k)| \le \delta \sum_k |w_k| < \epsilon.
\end{aligned}$$

By Theorem E.96, I_n is continuous and for each $n \in \mathbb{N}^+$ we have $I_n \in \mathcal{CL}(\mathcal{C}[a,b], \mathbb{R})$ and the convergence (6.5) implies

$$\forall f \in \mathcal{C}[a,b], \ \sup_{n\in\mathbb{N}^+} |I_n(f)| < +\infty.$$

Then Theorem E.85 and the principle of uniform boundedness (Theorem E.148) yield (b). Note that the operator norm of I_n, by Lemma E.109, equals W_n. □

Definition 6.6. A weighted quadrature formula (6.1) has (polynomial) *degree of exactness* d_E iff

$$\begin{cases} \forall f \in \mathbb{P}_{d_E}, & E_n(f) = 0, \\ \exists g \in \mathbb{P}_{d_E+1}, \text{ s.t.} & E_n(g) \neq 0, \end{cases} \tag{6.7}$$

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

Example 6.7. By Definition 6.6, $d_E \geq 0$ implies that $\sum_k w_k$ is bounded since $I_n(c) = c\int_a^b \rho(x)\mathrm{d}x$ holds for any constant $c \in \mathbb{R}$.

Lemma 6.8. Let $x_1, \ldots, x_n$ be given as distinct nodes of $I_n(f)$. If $d_E \geq n-1$, then its weights can be deduced as

$$\forall k = 1, \ldots, n, \qquad w_k = \int_a^b \rho(x)\ell_k(x)\mathrm{d}x, \tag{6.8}$$

where $\ell_k(x)$ is the fundamental polynomial for pointwise interpolation in (2.9) applied to the given nodes,

$$\ell_k(x) := \prod_{i\neq k; i=1}^{n} \frac{x - x_i}{x_k - x_i}. \tag{6.9}$$

Proof. Let $p_{n-1}(f;x)$ be the unique polynomial that interpolates f at the distinct nodes, as in the theorem on the uniqueness of polynomial interpolation (Theorem 2.5). Then we have

$$\begin{aligned} &\sum_{k=1}^{n} w_k p_{n-1}(x_k) = \int_a^b p_{n-1}(f;x)\rho(x)\mathrm{d}x \\ &= \int_a^b \sum_{k=1}^{n} \{\ell_k(x) f(x_k)\}\, \rho(x)\mathrm{d}x = \sum_{k=1}^{n} w_k f(x_k), \end{aligned}$$

where the first step follows from $d_E \geq n-1$ and the second step from the interpolation conditions (2.4), the Lagrange formula, and the uniqueness of $p_{n-1}(f;x)$. The proof is completed by setting f to be the hat function $\hat{B}_k(x)$ (see Definition 3.21) for each x_k. □

6.2 Newton-Cotes formulas

Definition 6.9. A *Newton-Cotes formula* is a formula (6.1) based on approximating $f(x)$ by interpolating it on uniformly spaced nodes $x_1, \ldots, x_n \in [a, b]$.

Definition 6.10. *The trapezoidal rule* is a formula (6.1) based on approximating $f(x)$ by the straight line that connects the points $(a, f(a))^T$ and $(b, f(b))^T$. In particular, for $\rho(x) \equiv 1$, it is simply

$$I^T(f) = \frac{b-a}{2}[f(a) + f(b)]. \tag{6.10}$$

Example 6.11. Derive the trapezoidal rule for the weight function $\rho(x) = x^{-1/2}$ on the interval $[0, 1]$. Note that one cannot apply (6.10) to $\rho(x)f(x)$ because $\rho(0) = \infty$. (6.8) yields

$$\begin{aligned} w_1 &= \int_0^1 x^{-1/2}(1-x)\mathrm{d}x = \frac{4}{3}, \\ w_2 &= \int_0^1 x^{-1/2} x \mathrm{d}x = \frac{2}{3}. \end{aligned}$$

Hence the formula is

$$I^T(f) = \frac{2}{3}[2f(0) + f(1)]. \tag{6.11}$$

Theorem 6.12. For $f \in \mathcal{C}^2[a, b]$ with weight function $\rho(x) \equiv 1$, the remainder of the trapezoidal rule satisfies

$$\exists \zeta \in [a, b] \text{ s.t. } E^T(f) = -\frac{(b-a)^3}{12} f''(\zeta). \tag{6.12}$$

Proof. By Theorem 2.5, the interpolating polynomial $p_1(f;x)$ is unique. Then we have

$$\begin{aligned} E^T(f) &= -\int_a^b \frac{f''(\xi(x))}{2}(x-a)(b-x)\mathrm{d}x \\ &= -\frac{f''(\zeta)}{2}\int_a^b (x-a)(b-x)\mathrm{d}x = -\frac{(b-a)^3}{12} f''(\zeta), \end{aligned}$$

where the first step follows from Theorem 2.7 and the second step from the integral mean value theorem (Theorem C.72). Here we can apply Theorem C.72 because

$$w(x) = (x-a)(b-x)$$

is always positive on (a, b). Also note that ξ is a function of x while ζ is a constant depending only on f, a, and b. □

Definition 6.13. *The midpoint rule* is a formula (6.1) based on approximating $f(x)$ by the constant $f(\frac{a+b}{2})$. In particular, for $\rho(x) \equiv 1$, it is simply

$$I^M(f) = (b-a) f\left(\frac{b+a}{2}\right). \tag{6.13}$$

Definition 6.14. *Simpson's rule* is a formula (6.1) based on approximating $f(x)$ by a quadratic polynomial that goes through the points $(a, f(a))^T$, $(b, f(b))^T$, and $(\frac{a+b}{2}, f(\frac{a+b}{2}))^T$. For $\rho(x) \equiv 1$, it is simply

$$I^S(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]. \tag{6.14}$$

Theorem 6.15. For $f \in \mathcal{C}^4[a, b]$ with weight $\rho(x) \equiv 1$, the remainder of Simpson's rule satisfies

$$\exists \zeta \in (a, b) \text{ s.t. } E^S(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta). \tag{6.15}$$

Proof. It is difficult to imitate the proof of Theorem 6.12, since $(x-a)(x-b)(x-\frac{a+b}{2})$ changes sign over $[a, b]$ and the integral mean value theorem is not applicable. To overcome this difficulty, we can formulate the interpolation via a Hermite problem so that Theorem C.72 can be applied. See problem I in Section 6.6 for the main steps. □

Example 6.16. Consider the integral

$$I = \int_{-4}^{4} \frac{\mathrm{d}x}{1+x^2} = 2\tan^{-1}(4) = 2.6516\cdots \tag{6.16}$$

Let $n-1$ be the number of sub-intervals that partition $[a,b]$ in Definition 6.9. As shown below, the Newton-Cotes formula appears to be non-convergent.

$n-1$	2	4	6	8	10
I_{n-1}	5.4902	2.2776	3.3288	1.9411	3.5956

For equally spaced nodes, the interpolating polynomials have wilder and wilder oscillations as the degree increases. Consequently, condition (b) of Theorem 6.5 does not hold. Hence Newton-Cotes formulas are not convergent even for well-behaved functions in $\mathcal{C}[a,b]$. In practice, Newton-Cotes formula with $n > 8$ is seldom used.

6.3 Composite formulas

Definition 6.17. The *composite trapezoidal rule* for approximating $I(f)$ in (6.2) with $\rho(x) \equiv 1$ is

$$I_n^T(f) = \frac{h}{2}f(x_0) + h\sum_{k=1}^{n-1} f(x_k) + \frac{h}{2}f(x_n), \tag{6.17}$$

where $h = \frac{b-a}{n}$ and $x_k = a + kh$.

Theorem 6.18. For $f \in \mathcal{C}^2[a,b]$, the remainder of the composite trapezoidal rule satisfies

$$\exists \xi \in (a,b) \text{ s.t. } E_n^T(f) = -\frac{b-a}{12}h^2 f''(\xi). \tag{6.18}$$

Proof. Apply Theorem 6.12 to the subintervals, sum up the errors, and we have

$$E_n^T(f) = -\frac{b-a}{12}h^2\left[\frac{1}{n}\sum_{k=0}^{n-1} f''(\xi_k)\right]. \tag{6.19}$$

The proof is completed by (6.19), the intermediate value Theorem C.41, and the fact $f \in \mathcal{C}^2[a,b] \Rightarrow f'' \in \mathcal{C}[a,b]$. □

Definition 6.19. The *composite Simpson's rule* for approximating $I(f)$ in (6.2) with $\rho(x) \equiv 1$ is

$$\begin{aligned} I_n^S(f) = \frac{h}{3}\big[&f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) \\ &+ \cdots + 4f(x_{n-1}) + f(x_n)\big], \end{aligned} \tag{6.20}$$

where $h = \frac{b-a}{n}$, $x_k = a + kh$, and n is even.

Theorem 6.20. For $f \in \mathcal{C}^4[a,b]$ and $n \in 2\mathbb{N}^+$, the remainder of the composite Simpson's rule satisfies

$$\exists \xi \in (a,b) \text{ s.t. } E_n^S(f) = -\frac{b-a}{180}h^4 f^{(4)}(\xi). \tag{6.21}$$

Proof. Exercise. □

Lemma 6.21. The trapezoidal rule satisfies

$$\forall f \in \mathbb{T}_{n-1}[0,2\pi], \qquad E_n^T(f) = 0, \tag{6.22}$$

where $\mathbb{T}_n[0,2\pi]$ is the class of trigonometric polynomials of degree at most n,

$$\mathbb{T}_n[0,2\pi] := \operatorname{span}\{1, \cos x, \sin x, \ldots, \cos(nx), \sin(nx)\}.$$

Proof. It suffices to verify that (6.22) holds for the complex exponential $e_m(x) := e^{imx} = \cos mx + i\sin mx$, $m \in \mathbb{N}$, i.e.

$$\begin{aligned} E_n^T(e_m) &= \int_0^{2\pi} e_m(x)\mathrm{d}x \\ &\quad - \frac{2\pi}{n}\left[\frac{e_m(0)+e_m(2\pi)}{2} + \sum_{k=1}^{n-1} e_m\left(\frac{2k\pi}{n}\right)\right] \\ &= \int_0^{2\pi} e^{imx}\mathrm{d}x - \frac{2\pi}{n}\sum_{k=0}^{n-1} e^{imk\cdot 2\pi/n}. \end{aligned}$$

Since $\int_0^{2\pi} e^{imx}\mathrm{d}x = (im)^{-1}\cdot e^{imx}\,|_0^{2\pi} = 0$, the geometric series yields

$$E_n^T(e_m) = \begin{cases} 0 & \text{if } m = 0; \\ -2\pi & \text{if } m = 0 \,(\mathrm{mod}\ n),\ m > 0; \\ -\frac{2\pi}{n}\frac{1-e^{imn\cdot 2\pi/n}}{1-e^{im\cdot 2\pi/n}} = 0 & \text{if } m \neq 0 \,(\mathrm{mod}\ n). \end{cases} \tag{6.23}$$

Hence (6.22) holds as $E_n^T(e_m) = 0$ for $m = 0, \ldots, n-1$. □

6.4 Gauss formulas

Lemma 6.22. Let $n, m \in \mathbb{N}^+$ and $m \leq n$. Given polynomials $p = \sum_{i=0}^{n+m} p_i x^i \in \mathbb{P}_{n+m}$ and $s = \sum_{i=0}^{n} s_i x^i \in \mathbb{P}_n$ satisfying $p_{n+m} \neq 0$ and $s_n \neq 0$, there exist unique polynomials $q \in \mathbb{P}_m$ and $r \in \mathbb{P}_{n-1}$ such that

$$p = qs + r. \tag{6.24}$$

Proof. Rewrite (6.24) as

$$\sum_{i=0}^{n+m} p_i x^i = \left(\sum_{i=0}^{m} q_i x^i\right)\left(\sum_{i=0}^{n} s_i x^i\right) + \sum_{i=0}^{n-1} r_i x^i. \tag{6.25}$$

Since monomials are linearly independent, (6.25) consists of $n+m+1$ equations, the last $m+1$ of which are

$$\begin{aligned} p_{n+m} &= q_m s_n, \\ p_{n+m-1} &= q_m s_{n-1} + q_{m-1} s_n, \\ &\cdots \\ p_n &= q_m s_{n-m} + \ldots + q_0 s_n, \end{aligned}$$

which can be written as $S\mathbf{q} = \mathbf{p}$ with S being a lower triangular matrix whose diagonal entries are $s_n \neq 0$. The coefficient vector $\mathbf{q}$ can be determined uniquely from coefficients of p and s. Then r can be determined uniquely by $p - qs$ from (6.25). □

Definition 6.23. The *node polynomial* associated with the nodes x_k's of a weighted quadrature formula is

$$v_n(x) = \prod_{k=1}^{n}(x - x_k). \tag{6.26}$$

Theorem 6.24. An *interpolatory formula*, i.e. a quadrature formula (6.1) with $d_E \geq n-1$, can be improved to have $d_E \geq n+j-1$ where $j \in (0,n]$ by and only by imposing the additional conditions on its node polynomial and weight function:

$$\forall p \in \mathbb{P}_{j-1}, \qquad \int_a^b v_n(x)p(x)\rho(x)\mathrm{d}x = 0. \tag{6.27}$$

Proof. For the necessity, we have

$$\int_a^b v_n(x)p(x)\rho(x)\mathrm{d}x = \sum_{k=1}^n w_k v_n(x_k)p(x_k) = 0,$$

where the first step follows from the facts $d_E \geq n+j-1$ and $v_n(x)p(x) \in \mathbb{P}_{n+j-1}$, and the second step from (6.26).

To prove the sufficiency, we must show that $E_n(p) = 0$ for any $p \in \mathbb{P}_{n+j-1}$. Lemma 6.22 yields

$$\forall p \in \mathbb{P}_{n+j-1},\ \exists! q \in \mathbb{P}_{j-1},\ \exists! r \in \mathbb{P}_{n-1},\ \text{s.t. } p = qv_n + r. \tag{6.28}$$

Consequently, we have

$$\begin{aligned}
&\int_a^b p(x)\rho(x)\mathrm{d}x = \int_a^b q(x)v_n(x)\rho(x)\mathrm{d}x + \int_a^b r(x)\rho(x)\mathrm{d}x \\
&= \int_a^b r(x)\rho(x)\mathrm{d}x = \sum_{k=1}^n w_k r(x_k) \\
&= \sum_{k=1}^n w_k[p(x_k) - q(x_k)v_n(x_k)] = \sum_{k=1}^n w_k p(x_k),
\end{aligned}$$

where the first step follows from (6.28), the second from (6.27), the third from the condition of $d_E \geq n-1$, the fourth from (6.28), and the last from (6.26). □

Definition 6.25. A *Gaussian quadrature formula* (or simply a *Gauss formula*) is an interpolatory formula that satisfies (6.27) for $j = n$ and some node polynomial v_n in (6.26); the nodes of the formula are exactly the roots of $v_n(x)$.

Corollary 6.26. A Gauss formula has $d_E = 2n-1$.

Proof. The index j in (6.27) cannot be $n+1$ because the node polynomial $v_n(x) \in \mathbb{P}_n$ cannot be orthogonal to itself. Therefore we know that $j = n$ in Theorem 6.24 is optimal: the formula (6.1) achieves the highest degree of exactness $2n-1$. From an algebraic viewpoint, the $2n$ degrees of freedom of nodes and weights in (6.1) determine a polynomial of degree at most $2n-1$. The proof is completed by Theorem 6.24 and Definition 6.6. □

Corollary 6.27. Weights of a Gauss formula $I_n(f)$ are

$$\forall k = 1, \cdots, n, \quad w_k = \int_a^b \frac{v_n(x)}{(x-x_k)v_n'(x_k)}\rho(x)\mathrm{d}x, \tag{6.29}$$

where $v_n(x)$ is the node polynomial that defines $I_n(f)$.

Proof. This follows from Lemma 6.8; also see (2.11). □

Example 6.28. Derive the Gauss formula of $n = 2$ for the weight function $\rho(x) = x^{-1/2}$ on the interval $[0,1]$.

We first construct an orthogonal polynomial

$$\pi(x) = c_0 - c_1 x + x^2$$

such that

$$\forall p \in \mathbb{P}_1, \quad \langle p(x), \pi(x)\rangle := \int_0^1 p(x)\pi(x)\rho(x)\mathrm{d}x = 0,$$

which is equivalent to $\langle 1, \pi(x)\rangle = 0$ and $\langle x, \pi(x)\rangle = 0$ because $\mathbb{P}_1 = \mathrm{span}(1, x)$. These two conditions yield

$$\begin{aligned}
\int_0^1 (c_0 - c_1 x + x^2)x^{-1/2}\mathrm{d}x &= \frac{2}{5} + 2c_0 - \frac{2}{3}c_1 = 0, \\
\int_0^1 x(c_0 - c_1 x + x^2)x^{-1/2}\mathrm{d}x &= \frac{2}{7} + \frac{2}{3}c_0 - \frac{2}{5}c_1 = 0.
\end{aligned}$$

Hence $c_1 = \frac{6}{7}$, $c_0 = \frac{3}{35}$, and the orthogonal polynomial is

$$\pi(x) = \frac{3}{35} - \frac{6}{7}x + x^2$$

with its zeros at

$$x_1 = \frac{1}{7}\left(3 - 2\sqrt{\frac{6}{5}}\right),\ x_2 = \frac{1}{7}\left(3 + 2\sqrt{\frac{6}{5}}\right).$$

To calculate w_1 and w_2, we could again use (6.8), but it is simpler to set up a linear system of equations by exploiting Corollary 6.26, i.e. Gauss quadrature is exactly for all constants and linear polynomials,

$$\begin{aligned}
w_1 + w_2 &= \int_0^1 x^{-1/2}\mathrm{d}x = 2, \\
x_1 w_1 + x_2 w_2 &= \int_0^1 x x^{-1/2}\mathrm{d}x = \frac{2}{3},
\end{aligned}$$

which yields

$$\begin{aligned}
w_1 &= \frac{-2x_2 + \frac{2}{3}}{x_1 - x_2} = 1 + \frac{1}{3}\sqrt{\frac{5}{6}}, \\
w_2 &= \frac{2x_1 - \frac{2}{3}}{x_1 - x_2} = 1 - \frac{1}{3}\sqrt{\frac{5}{6}}.
\end{aligned}$$

The desired two-point Gauss formula is thus

$$\begin{aligned}
I_2^G(f) = &\left(1 + \frac{1}{3}\sqrt{\frac{5}{6}}\right) f\left(\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}\right) \\
&+ \left(1 - \frac{1}{3}\sqrt{\frac{5}{6}}\right) f\left(\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}\right).
\end{aligned} \tag{6.30}$$

The degree of exactness of the trapezoidal rule is 1 while that of the two-point Gauss formula is 3. Hence we expect that the Gauss formula be much more accurate. Indeed, calculate errors of the two formulas (6.11) and (6.30) for $f(x) = \cos\left(\frac{1}{2}\pi x\right)$ and we have

$$\begin{aligned}
E^T &= 0.226453\ldots; \\
E_2^G &= 0.002197\ldots,
\end{aligned}$$

which can be verified by simple calculations.

Definition 6.29. A set of *orthogonal polynomials* is a set of polynomials $P = \{p_i : \deg(p_i) = i\}$ that satisfy

$$\forall p_i, p_j \in P, \quad i \neq j \;\Rightarrow\; \langle p_i, p_j \rangle = 0. \tag{6.31}$$

Example 6.30. In this chapter, the inner product in (6.31) is taken to be

$$\langle p_i, p_j \rangle = \int_a^b p_i(x) p_j(x) \rho(x) \mathrm{d}x,$$

where $[a, b]$ and ρ are the same as those in (6.2).

Theorem 6.31. Each zero of a real orthogonal polynomial over $[a, b]$ is real, simple, and inside (a, b).

Proof. For fixed $n \geq 1$, suppose $p_n(x)$ does not change sign in $[a, b]$. Then

$$\exists c \in \mathbb{R}^+ \text{ s.t. } \int_a^b \rho(x) p_n(x) \mathrm{d}x = c \langle p_n, p_0 \rangle \neq 0,$$

which contradicts the orthogonality of p_n and p_0. Hence there exists $x_1 \in [a, b]$ satisfying $p_n(x_1) = 0$.

Suppose there were a zero at x_1 which is multiple. Then $\frac{p_n(x)}{(x-x_1)^2}$ would be a polynomial of degree $n-2$. Hence $0 = \left\langle p_n(x), \frac{p_n(x)}{(x-x_1)^2} \right\rangle = \left\langle 1, \frac{p_n^2(x)}{(x-x_1)^2} \right\rangle > 0$, which is false. Therefore every zero is simple.

Suppose that only $j < n$ zeros of p_n, say $x_1, x_2, \ldots, x_j$, are inside (a, b) and all other zeros are out of (a, b). Let $v_j(x) = \prod_{i=1}^{j}(x - x_i) \in \mathbb{P}_j$. Then $p_n v_j = P_{n-j} v_j^2$ where P_{n-j} is a polynomial of degree $n - j$ that does not change sign on $[a, b]$. Hence $\left|\left\langle P_{n-j}, v_j^2 \right\rangle\right| > 0$, which contradicts the orthogonality of $p_n(x)$ and $v_j(x)$. □

Corollary 6.32. All nodes of a Gauss formula are real, distinct, and contained in (a, b).

Proof. This follows directly from Definition 6.25 and Theorem 6.31. □

Lemma 6.33. Gauss formulas have positive weights.

Proof. For each $k = 1, 2, \ldots, n$, the definition of $\ell_k(x)$ in (6.9) implies $\ell_k^2 \in \mathbb{P}_{2n-2}$, then we have

$$w_k = \sum_{j=1}^{n} w_j \ell_k^2(x_j) = \int_a^b \rho(x) \ell_k^2(x) \mathrm{d}x > 0,$$

where the first step follows from (6.9), second step from $d_E = 2n - 1$ and the last step from the conditions on ρ. □

Lemma 6.34. A Gauss formula satisfies

$$\sum_{k=1}^{n} w_k = \mu_0 \in (0, +\infty).$$

Proof. This follows from setting $j = 0$ in (6.3) and applying the condition on ρ in Definition 6.1. □

Theorem 6.35. Gauss formulas are convergent for $\mathcal{C}[a, b]$.

Proof. Denote by $\mathbb{P}$ the set of real polynomials. Theorem 2.53 states that $\mathbb{P}$ is dense in $\mathcal{C}[a, b]$, i.e. condition (a) in Theorem 6.5 holds. Condition (b) also holds because of Lemmas 6.34 and 6.33. Then the proof is completed by Theorem 6.5. □

Theorem 6.36. For $f \in \mathcal{C}^{2n}[a, b]$, the remainder of a Gauss formula $I_n(f)$ satisfies

$$\exists \xi \in [a, b] \text{ s.t. } E_n^G(f) = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b \rho(x) v_n^2(x) \mathrm{d}x, \tag{6.32}$$

where v_n is the node polynomial that defines I_n.

Proof. One proof is suggested in 1885 by Markov, a student of Chebyshev and famous for his work in probability theory on certain random processes now known as Markov chains. See exercise IV in Section 6.6.1. □

6.5 Numerical differentiation

Formula 6.37 (The method of undetermined coefficients)**.** A *general method to derive FD formulas* that approximate $u^{(k)}(\bar{x})$ is based on an arbitrary stencil of $n > k$ distinct points $x_1, x_2, \ldots, x_n$. Taylor expansions of u at each point x_i in the stencil about $u(\bar{x})$ yield

$$u(x_i) = u(\bar{x}) + (x_i - \bar{x}) u'(\bar{x}) + \cdots + \frac{1}{k!}(x_i - \bar{x})^k u^{(k)}(\bar{x}) + \cdots$$

for $i = 1, 2, \ldots, n$. This leads to a linear combination of point values that approximates $u^{(k)}(\bar{x})$,

$$u^{(k)}(\bar{x}) = c_1 u(x_1) + c_2 u(x_2) + \cdots + c_n u(x_n) + O(h^p),$$

where the c_j's are chosen to make p as large as possible:

$$\forall i = 0, \ldots, p-1, \qquad \frac{1}{i!} \sum_{j=1}^{n} c_j (x_j - \bar{x})^i = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{6.33}$$

Example 6.38. To approximate $u'(\bar{x})$ with an FD formula

$$D_2 u(\bar{x}) = a u(\bar{x}) + b u(\bar{x} - h) + c u(\bar{x} - 2h), \tag{6.34}$$

we determine the coefficients a, b, and c to give the best possible accuracy. Taylor expansions at $\bar{x}$ yield

$$\begin{aligned} D_2 u(\bar{x}) =& (a + b + c) u(\bar{x}) - (b + 2c) h u'(\bar{x}) \\ &+ \frac{1}{2}(b + 4c) h^2 u''(\bar{x}) - \frac{1}{6}(b + 8c) h^3 u'''(\bar{x}) \\ &+ O\left(h^4\right). \end{aligned}$$

Set $a + b + c = 0$, $b + 2c = -\frac{1}{h}$, and $b + 4c = 0$, solve

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{h} \\ 0 \end{bmatrix}, \tag{6.35}$$

and we get

$$a = \frac{3}{2h}, \quad b = -\frac{2}{h}, \quad c = \frac{1}{2h}. \tag{6.36}$$

Therefore the FD formula is determined as

$$D_2 u(\bar{x}) = \frac{1}{2h}[3u(\bar{x}) - 4u(\bar{x} - h) + u(\bar{x} - 2h)]. \tag{6.37}$$

Definition 6.39. In approximating the derivative of a smooth function, an FD formula is *p-th order accurate* if its error E has the form

$$E(h) = \Theta(h^p), \tag{6.38}$$

where h is the maximum distance of adjacent points in the stencil.

Example 6.40. Consider approximating $u'(x)$ at a point $\bar{x}$ using the nearby function values $u(\bar{x} \pm h)$. Three commonly used formulas are

$$D_+u(\bar{x}) := \frac{u(\bar{x}+h) - u(\bar{x})}{h}, \tag{6.39}$$

$$D_-u(\bar{x}) := \frac{u(\bar{x}) - u(\bar{x}-h)}{h}, \tag{6.40}$$

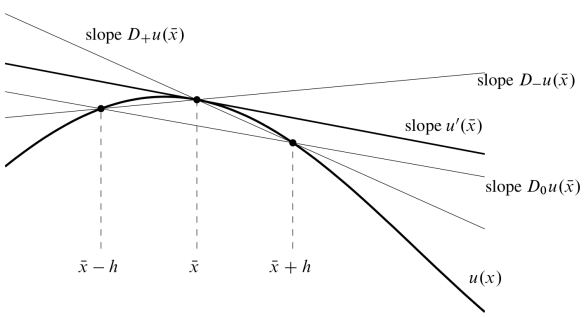
$$D_0u(\bar{x}) := \frac{u(\bar{x}+h) - u(\bar{x}-h)}{2h} = \frac{1}{2}(D_+ + D_-)u(\bar{x}). \tag{6.41}$$

For $u(x) = \sin(x)$ and $\bar{x} = 1$, we calculate the errors of the above three formulas in approximating $u'(1) = \cos(1) \approx 0.5403023$ with $h = 0.01$ and 0.005.

Define the error as $E := Du(\bar{x}) - u'(\bar{x})$ and the following table shows the values of E for the three formulas.

h	$D_+u(\bar{x})$	$D_-u(\bar{x})$	$D_0u(\bar{x})$
$h_1 =$ 1.0e-2	−4.2e-3	−4.2e-3	−9.00e-6
$h_2 =$ 5.0e-3	−2.1e-3	−2.1e-3	−2.25e-6
$p = \log_2 \frac{E(Du(\bar{x},h_1))}{E(Du(\bar{x},h_2))}$	1	1	2

The last row of the table shows that the error behaves like $E(h) \approx Ch^p$, which means reduction of h by a factor of r leads to error reduction by a factor of r^p. Geometric illustration of these FD formulas are shown below.



Exercise 6.41. Show that the FD formulas $D_+u(\bar{x})$ and $D_-u(\bar{x})$ are first-order accurate while $D_0u(\bar{x})$ is second-order accurate.

Exercise 6.42. Construct a table of divided difference (as in Definition 2.18) to derive a quadratic polynomial that agrees with $u(x)$ at $\bar{x}, \bar{x}-h$, and $\bar{x}-2h$. Then take derivative of this polynomial to obtain the FD formula (6.37).

Lemma 6.43. In approximating the second derivative of $u \in \mathcal{C}^4(\mathbb{R})$, the formula

$$D^2u(\bar{x}) = \frac{u(\bar{x}-h) - 2u(\bar{x}) + u(\bar{x}+h)}{h^2} \tag{6.42}$$

is second-order accurate. Furthermore, if the input function values $u(\bar{x}-h)$, $u(\bar{x})$, and $u(\bar{x}+h)$ are perturbed with random errors $\epsilon \in [-E, E]$, then there exists $\xi \in [\bar{x}-h, \bar{x}+h]$ such that

$$\left|u''(\bar{x}) - D^2u(\bar{x})\right| \le \frac{h^2}{12}|u^{(4)}(\xi)| + \frac{4E}{h^2}. \tag{6.43}$$

6.6 Problems

6.6.1 Theoretical questions

I. Simpson's rule.

(a) Show that Simpson's rule on $[-1, 1]$ can be obtained by

$$\int_{-1}^{1} y(t)\mathrm{d}t = \int_{-1}^{1} p_3(y; -1, 0, 0, 1; t)\mathrm{d}t + E^S(y),$$

where $y \in \mathcal{C}^4[-1, 1]$ and $p_3(y; -1, 0, 0, 1; t)$ is the interpolation polynomial of y that satisfies $p_3(-1) = y(-1)$, $p_3(0) = y(0)$, $p_3'(0) = y'(0)$, and $p_3(1) = y(1)$.

(b) Derive $E^S(y)$.

(c) Using (a), (b), and a change of variable, derive the composite Simpson's rule and prove the theorem on its error estimation.

II. Estimate the number of subintervals required to approximate $\int_0^1 e^{-x^2}\mathrm{d}x$ to six correct decimal places, i.e. the absolute error is less than 0.5×10^{-6},

(a) by the composite trapezoidal rule,

(b) by the composite Simpson's rule.

III. Gauss-Laguerre quadrature formula.

(a) Construct a polynomial $\pi_2(t) = t^2 + at + b$ that is orthogonal to $\mathbb{P}_1$ with respect to the weight function $\rho(t) = e^{-t}$, i.e.

$$\forall p \in \mathbb{P}_1, \qquad \int_0^{+\infty} p(t)\pi_2(t)\rho(t)\mathrm{d}t = 0.$$

(*hint*: $\int_0^{+\infty} t^m e^{-t}\mathrm{d}t = m!$)

(b) Derive the two-point Gauss-Laguerre quadrature formula

$$\int_0^{+\infty} f(t)e^{-t}\mathrm{d}t = w_1 f(t_1) + w_2 f(t_2) + E_2(f)$$

and express $E_2(f)$ in terms of $f^{(4)}(\tau)$ for some $\tau > 0$.

(c) Apply the formula in (b) to approximate

$$I = \int_0^{+\infty} \frac{1}{1+t} e^{-t}\mathrm{d}t.$$

Use the remainder to estimate the error and compare your estimate with the true error. With the true error, identify the unknown quantity τ contained in $E_2(f)$.
(*hint*: use the exact value $I = 0.596347361\cdots$)

IV. Remainder of Gauss formulas. Consider the Hermite interpolation problem: find $p \in \mathbb{P}_{2n-1}$ such that

$$\forall m = 1, 2, \ldots, n, \quad p(x_m) = f_m,\ p'(x_m) = f'_m. \tag{6.44}$$

There are *elementary Hermite interpolation polynomials* h_m, q_m such that the solution of (6.44) can be expressed in the form

$$p(t) = \sum_{m=1}^{n} [h_m(t) f_m + q_m(t) f'_m],$$

analogous to the Lagrange interpolation formula.

(a) Seek h_m and q_m in the form

$$h_m(t) = (a_m + b_m t)\ell_m^2(t),\ q_m(t) = (c_m + d_m t)\ell_m^2(t)$$

where ℓ_m is the elementary Lagrange polynomial in (2.9). Determine the constants a_m, b_m, c_m, d_m.

(b) Obtain the quadrature rule

$$I_n(f) = \sum_{k=1}^{n} [w_k f(x_k) + \mu_k f'(x_k)]$$

that satisfies $E_n(p) = 0$ for all $p \in \mathbb{P}_{2n-1}$.

(c) What conditions on the node polynomial or on the nodes x_k must be imposed so that $\mu_k = 0$ for each $k = 1, 2, \ldots, n$?

V. Prove Lemma 6.43. How do you choose h to minimize the error bound in (6.43)? Design a fourth-order accurate formula based on a symmetric stencil, derive its error bound, and minimize the error bound. What do you observe in comparing the second-order case and the fourth-order case?

Chapter 7

Finite Difference (FD) Methods for Boundary Value Problems (BVPs)

Definition 7.1. A *partial differential equation* (PDE) is an equation involving an unknown function of two or more variables and some of its partial derivatives.

Definition 7.2. *Laplace equation* is a second-order PDE of the form

$$\Delta u(\mathbf{x}) = 0, \tag{7.1}$$

where the unknown is a function $u : \overline{\Omega} \to \mathbb{R}$, $\overline{\Omega}$ is the closure of an open set $\Omega \subset \mathbb{R}^n$, and the *Laplacian operator* $\Delta : \mathcal{C}^2(\Omega) \to \mathcal{C}(\Omega)$ is

$$\Delta := \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}. \tag{7.2}$$

Example 7.3. *Potential flow* is a special type of flow where the velocity field $\mathbf{u}$ can be expressed as the gradient of a scalar function:

$$\mathbf{u} = \nabla\varphi,$$

where φ is called the *velocity potential*. For incompressible fluids with $\nabla\cdot\mathbf{u} = 0$, the velocity potential satisfies a Laplace equation $\Delta\varphi = 0$.

Definition 7.4. *Poisson's equation* is a second-order PDE of the form

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \tag{7.3}$$

where the unknown is a function $u : \overline{\Omega} \to \mathbb{R}$, $\overline{\Omega}$ is the closure of an open set $\Omega \subset \mathbb{R}^n$, and the RHS function $f : \Omega \to \mathbb{R}$ is given a priori.

Definition 7.5. A *boundary value problem* (BVP) is a differential equation together with a set of additional constraints, called the *boundary conditions*, that hold only on the domain boundary.

Definition 7.6. Common types of boundary conditions for a one-dimensional interval $\Omega = (a, b)$ are

- *Dirichlet conditions*: $u(a) = \alpha$ and $u(b) = \beta$;
- *Mixed conditions*: $u(a) = \alpha$ and $\left.\frac{\partial u}{\partial x}\right|_b = \beta$;
- *Neumann conditions*: $\left.\frac{\partial u}{\partial x}\right|_a = \alpha$ and $\left.\frac{\partial u}{\partial x}\right|_b = \beta$.

Theorem 7.7. Suppose f and g are two sufficiently smooth functions. Then there exists a unique solution (up to an additive constant) for the Neumann BVP

$$\Delta\phi = f \qquad \text{in } \Omega; \tag{7.4a}$$
$$\mathbf{n}\cdot\nabla\phi = g \qquad \text{on } \partial\Omega \tag{7.4b}$$

if and only if

$$\int_\Omega f \,\mathrm{d}V = \int_{\partial\Omega} g \,\mathrm{d}A. \tag{7.5}$$

Proof. See [Taylor, 2011, page 409]. □

Example 7.8. The fundamental theorem of vector calculus (Theorem **??**) states that a continuously differentiable vector field $\mathbf{v}^*$ can be uniquely decomposed into a divergence-free part and a curl-free part:

$$\begin{cases} \mathbf{v}^* &= \mathbf{v} + \nabla\phi, \\ \nabla\cdot\mathbf{v} &= 0, \qquad \nabla\times\nabla\phi = \mathbf{0}. \end{cases} \tag{7.6}$$

When a given boundary condition of $\mathbf{v}$ satisfies $\oint_{\partial\Omega} \mathbf{v}\cdot\mathbf{n} = 0$, the decomposition is realized by solving the Neumann BVP

$$\Delta\phi = \nabla\cdot\mathbf{v}^* \qquad \text{in } \Omega, \tag{7.7a}$$
$$\mathbf{n}\cdot\nabla\phi = \mathbf{n}\cdot(\mathbf{v}^* - \mathbf{v}) \qquad \text{on } \partial\Omega, \tag{7.7b}$$

for which the existence of the unique solution is guaranteed by Theorem 7.7 and

$$\int_\Omega \nabla\cdot\mathbf{v}^* \,\mathrm{d}V = \int_\Omega \nabla\cdot(\mathbf{v}^* - \mathbf{v}) \,\mathrm{d}V = \int_{\partial\Omega} \mathbf{n}\cdot(\mathbf{v}^* - \mathbf{v}) \,\mathrm{d}A.$$

7.1 The FD discretization

Formula 7.9. In solving a linear BVP, the general procedures of an FD method are as follows.

(FD-1) Discretize the problem domain by a grid.

(FD-2) Approximate each spatial derivative in the PDE with some finite difference formula at every grid point to get a system of linear equations $A\mathbf{U} = \mathbf{F}$

where the vector $\mathbf{U}$ approximates the unknown variable on the grid while the vector $\mathbf{F}$ contains given conditions of the BVP such as boundary conditions and derivatives of the unknown function.

(FD-3) Solve the system of algebraic equations.

Example 7.10 (An FD method for Poisson's equation in a unit interval)**.** Consider the one-dimensional BVP

$$-u''(x) = f(x) \text{ in } \Omega := (0,1) \tag{7.8}$$

with Dirichlet boundary conditions

$$u(0) = \alpha, \quad u(1) = \beta. \tag{7.9}$$

The general procedures of an FD method based on the central difference are as follows.

(a) Discretize Ω by a Cartesian grid with uniform spacing,

$$x_j = jh, \qquad h = \frac{1}{m+1}, \quad j = 0, 1, \ldots, m+1.$$

Set $U_0 = \alpha$, $U_{m+1} = \beta$ and we will compute m values $U_1, \ldots, U_m$ where each U_j approximates $u(x_j)$.

(b) Approximate the second derivative u'' with a centered difference

$$u''(x_j) = \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} + O(h^2) \tag{7.10}$$

and we get the following system of linear equations:

$$\begin{aligned} -\frac{\alpha - 2U_1 + U_2}{h^2} &= f(x_1), \\ -\frac{U_{j-1} - 2U_j + U_{j+1}}{h^2} &= f(x_j), \quad j = 2, \ldots, m-1, \\ -\frac{U_{m-1} - 2U_m + \beta}{h^2} &= f(x_m). \end{aligned}$$

These equations are written in the form

$$A\mathbf{U} = \mathbf{F}, \tag{7.11}$$

where

$$\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{m-1} \\ U_m \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} f(x_1) + \frac{\alpha}{h^2} \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) + \frac{\beta}{h^2} \end{bmatrix}, \tag{7.12}$$

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}. \tag{7.13}$$

(c) Solve the linear system (7.11).

7.2 Errors and consistency

Definition 7.11. The *global error* or *solution error* of an FD method in Formula 7.9 is

$$\mathbf{E} = \mathbf{U} - \hat{\mathbf{U}}, \tag{7.14}$$

where $\hat{\mathbf{U}} = [u(x_1), u(x_2), \cdots, u(x_m)]^T$ is the vector of true values and $\mathbf{U}$ the computed solution.

Definition 7.12. A *grid function* is a function $\mathbf{g} : \mathbf{X} \to \mathbb{R}$ on a discrete grid $\mathbf{X}$ that contains a finite number of points.

Definition 7.13. The *q-norm of a grid function* $\mathbf{g}$ on a one-dimensional grid $\mathbf{X} := \{x_1, x_2, \ldots, x_N\}$ is

$$\|\mathbf{g}\|_q = \left(h \sum_{i=1}^{N} |g_i|^q \right)^{\frac{1}{q}}, \tag{7.15}$$

where $\mathbf{g} = (g_1, g_2, \cdots, g_N)$. In particular, the *1-norm* is

$$\|\mathbf{g}\|_1 = h \sum_{i=1}^{N} |g_i| \tag{7.16}$$

and the *max-norm* is

$$\|\mathbf{g}\|_\infty = \max_{1 \le i \le N} |g_i|. \tag{7.17}$$

Exercise 7.14. Suppose a grid function $\mathbf{g} : \mathbf{X} \to \mathbb{R}$ has $\mathbf{X} := \{x_1, x_2, \ldots, x_N\}$, $g_1 = O(h)$, $g_N = O(h)$, and $g_j = O(h^2)$ for all $j = 2, \ldots, N-1$. Show that

$$\|\mathbf{g}\|_\infty = O(h), \ \|\mathbf{g}\|_1 = O(h^2), \ \|\mathbf{g}\|_2 = O(h^{\frac{3}{2}}). \tag{7.18}$$

As the main point of this exercise, the differences in the max-norm, 1-norm, and 2-norm of a grid function often reveal the percentage of components with large magnitude.

Definition 7.15. The *local truncation error* (LTE) of an FD method in Formula 7.9 is the error caused by replacing a continuous derivative with an FD formula.

Example 7.16. When we approximate Δu with

$$D^2 u(x_j) := \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2}, \tag{7.19}$$

the LTE of the FD method in Example 7.10 is

$$\tau_j = -D^2 u(x_j) - \big(-u''(x_j)\big) = -\frac{h^2}{12} u''''(x_j) + O(h^4).$$

Lemma 7.17. Let $A\mathbf{U} = \mathbf{F}$ be the linear system obtained by applying Formula 7.9 to a linear BVP $\mathcal{L}u = f(x)$ in $(0,1)$ with Dirichlet conditions. Then the LTE of this FD method is the error of calculating the RHS function $\mathbf{F}$ by replacing U_j with the exact solution $u(x_j)$, i.e.,

$$\boldsymbol{\tau} = A\hat{\mathbf{U}} - \mathbf{F}, \tag{7.20}$$

where $\hat{\mathbf{U}}$ is the vector of true solution values. In particular, this holds for the FD method in Example 7.10.

Proof. We only prove the case in Example 7.10 since other linear BVPs can be proven by similar arguments.

By (7.8) and (7.11), we have, $\forall j = 2, \dots, m-1$,

$$(A\hat{\mathbf{U}} - \mathbf{F})_j = -\frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} + \Delta u(x_j),$$

which also holds for $j = 1$ and $j = m$ since the boundary conditions yield $u(x_0) = u(0) = \alpha$ and $u(x_{m+1}) = u(1) = \beta$. Then the proof is completed by Definition 7.15. □

Lemma 7.18. The LTE and the global error are related as

$$A\mathbf{E} = -\boldsymbol{\tau}. \tag{7.21}$$

Proof. $A\mathbf{E} = A(\mathbf{U} - \hat{\mathbf{U}}) = \mathbf{F} - (\mathbf{F} + \boldsymbol{\tau}) = -\boldsymbol{\tau}$. □

Definition 7.19. An FD method in Formula 7.9 is said to be *consistent* with the BVP if

$$\lim_{h\to 0} \|\boldsymbol{\tau}^h\| = 0, \tag{7.22}$$

where $\boldsymbol{\tau}^h$ is the LTE.

7.3 Stability and convergence

Definition 7.20. An FD method is *convergent* if

$$\lim_{h\to 0} \left\|\mathbf{E}^h\right\| = 0, \tag{7.23}$$

where $\mathbf{E}^h$ is the solution error in Definition 7.11 and $\|\cdot\|$ is a q-norm in Definition 7.13.

Definition 7.21. An FD method in Formula 7.9 is *stable* if

(a) $\exists h_0 \in \mathbb{R}^+$ s.t. $\forall h \in (0, h_0)$, $\det(A) \neq 0$, where A is the matrix of the linear system for the grid size h;

(b) $\lim_{h\to 0} \left\|A^{-1}\right\| = O(1)$.

Theorem 7.22. A consistent and stable FD method is convergent.

Proof. Lemma 7.18 and Definitions 7.19 and 7.21 yield

$$\lim \left\|\mathbf{E}^h\right\| \le \lim \left\|(A^h)^{-1}\right\| \lim \left\|\boldsymbol{\tau}^h\right\| \le C \lim \left\|\boldsymbol{\tau}^h\right\| = 0,$$

where a norm is either a vector norm or an induced matrix norm. Then (7.23) follows from the observation that, by Definition 7.13, the q-norm $\|\cdot\|_q$ and the corresponding vector norm $\|\cdot\|$ are related by $\|\mathbf{E}\|_q = h^{\frac{1}{q}}\|\mathbf{E}\|$. □

7.3.1 Convergence in the 2-norm

Definition 7.23 (Matrix norms induced by vector norms)**.** The *norm* of a matrix $A \in \mathbb{R}^{n\times n}$ is defined by

$$\begin{aligned}\|A\| &= \sup\left\{\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}\right\} \\ &= \sup\left\{\|A\mathbf{x}\| : \|\mathbf{x}\| = 1\right\}.\end{aligned}$$

Example 7.24. Commonly used matrix norms include

$$\begin{cases} \|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|, \\ \|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^n |a_{ij}|, \\ \|A\|_2 = \sqrt{\rho(A^TA)}, \end{cases} \tag{7.24}$$

where $\rho(B) := \max_i |\lambda_i(B)|$ is the spectral radius of the matrix B, i.e. the maximum modulus of eigenvalues of B.

Lemma 7.25. The eigenvalues λ_k and eigenvectors $\mathbf{w}_k$ of the matrix A in (7.13) are

$$\lambda_k(A) = \frac{4}{h^2}\sin^2\frac{k\pi}{2(m+1)}, \tag{7.25}$$

$$w_{k,j} = \sin\frac{jk\pi}{m+1}, \tag{7.26}$$

where $j, k = 1, 2, \dots, m$.

Proof. It is straightforward to verify the conclusions using the trigonometric identity

$$\sin\alpha + \sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}.$$ □

Theorem 7.26. The FD method in Example 7.10 is second-order convergent in the 2-norm.

Proof. The symmetry of A gives $\|A\|_2 = \rho(A)$. Then Lemma 7.25 implies the stability, i.e., condition (b) in Definition 7.21:

$$\lim_{h\to 0} \|A^{-1}\|_2 = \lim_{h\to 0} \frac{1}{\min|\lambda_k(A)|} = \lim_{h\to 0} \frac{h^2}{4\sin^2\frac{\pi h}{2}} = \frac{1}{\pi^2}.$$

The rest of the proof follows from Example 7.16, Definition 7.19, and Theorem 7.22. □

7.3.2 Green's function

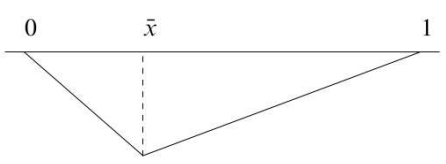
Definition 7.27. For any fixed $\bar{x} \in [0, 1]$, the *Green's function* $G(x; \bar{x})$ is the function of x that solves the BVP

$$\begin{cases} u''(x) = \delta(x - \bar{x}); \\ u(0) = u(1) = 0, \end{cases} \tag{7.27}$$

where $\delta(x - \bar{x})$ is the Dirac delta function in Definition 5.41.

Lemma 7.28. The Green's function $G(x; \bar{x})$ that solves (7.27) is

$$G(x;\bar{x}) = \begin{cases} (\bar{x}-1)x, & x \in [0, \bar{x}], \\ \bar{x}(x-1), & x \in [\bar{x}, 1]. \end{cases} \tag{7.28}$$



Proof. For any fixed ϵ, we have from (7.27) and (5.41b) that

$$\int_{x_0-\epsilon}^{x_0+\epsilon} G''(x)\mathrm{d}x = \int_{x_0-\epsilon}^{x_0+\epsilon} \delta(x-\bar{x})\mathrm{d}x = \begin{cases} 0, & \bar{x} \notin (x_0-\epsilon, x_0+\epsilon), \\ 1, & \bar{x} \in (x_0-\epsilon, x_0+\epsilon). \end{cases}$$

Take limit $\epsilon \to 0$ of the above equation and we deduce from the second fundamental theorem of calculus (Theorem C.74) that

$$\lim_{\epsilon\to 0} G'(x_0+\epsilon) - \lim_{\epsilon\to 0} G'(x_0-\epsilon) = \begin{cases} 0 & \text{if } x_0 \in (0,\bar{x}) \cup (\bar{x},1), \\ 1 & \text{if } x_0 = \bar{x}. \end{cases} \tag{7.29}$$

Substitute

$$G(x;\bar{x}) = \begin{cases} ax+b, & x \in [0,\bar{x}], \\ cx+d, & x \in [\bar{x},1] \end{cases}$$

into (7.29) and (7.27) and the continuity of $G(x;\bar{x})$ yields

$$\begin{cases} c = a+1 \\ b = 0 \\ c+d = 0 \\ a\bar{x}+b = c\bar{x}+d \end{cases} \Rightarrow \begin{cases} a = \bar{x}-1 \\ b = 0 \\ c = \bar{x} \\ d = -\bar{x} \end{cases},$$

which completes the proof. □

Corollary 7.29. The solution to the linear BVP

$$\begin{cases} u''(x) = c\delta(x-\bar{x}), \\ u(0) = u(1) = 0 \end{cases}$$

is

$$u(x) = cG(x;\bar{x}).$$

Proof. This follows directly from Lemma 7.28. □

7.3.3 Convergence in the max-norm

Lemma 7.30. For the matrix A in (7.13), any element of its inverse $B = A^{-1}$ is

$$b_{ij} = -hG(x_i;x_j) = \begin{cases} -h(x_j-1)x_i, & i \le j, \\ -hx_j(x_i-1), & i \ge j. \end{cases} \tag{7.30}$$

More explicitly, the matrix B is

$$B = -h\begin{bmatrix} x_1(x_1-1) & x_1(x_2-1) & \cdots & x_1(x_m-1) \\ x_1(x_2-1) & x_2(x_2-1) & \cdots & x_2(x_m-1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(x_m-1) & x_2(x_m-1) & \cdots & x_m(x_m-1) \end{bmatrix}.$$

Proof. To verify that B is indeed the inverse of A, it suffices to multiply the ith row of h^2A and the jth column of $-\frac{1}{h}B$,

$$\begin{bmatrix} 0,\cdots,0,-1,2,-1,0,\cdots,0 \end{bmatrix} \begin{bmatrix} x_1(x_j-1) \\ \vdots \\ x_{j-1}(x_j-1) \\ x_j(x_j-1) \\ x_j(x_{j+1}-1) \\ \vdots \\ x_j(x_m-1) \end{bmatrix},$$

the only nonzero case is when $i = j$:

$$2x_j(x_j-1) - x_{j-1}(x_j-1) - x_j(x_{j+1}-1) = -h. \qquad \square$$

Theorem 7.31. The max-norm of $B = A^{-1}$ satisfies

$$\|B\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{m} |b_{ij}| \le 1. \tag{7.31}$$

Proof. Lemma 7.30 yields

$$\begin{aligned} \sum_{j=1}^{m} |b_{ij}| &= \sum_{j=1}^{i} hx_j|x_i-1| + \sum_{j=i+1}^{m} hx_i|x_j-1| \\ &\le \sum_{j=1}^{i} h\left(\frac{m}{m+1}\right)^2 + \sum_{j=i+1}^{m} h\left(\frac{m}{m+1}\right)^2 \\ &= mh\left(\frac{m}{m+1}\right)^2 = \left(\frac{m}{m+1}\right)^3 \le 1. \end{aligned}$$ □

7.4 A solution via Green's function

Lemma 7.32. Suppose $\mathcal{L}$ is an invertible linear differential operator that satisfies

$$\mathcal{L}u(x) = f(x). \tag{7.32}$$

Then we have

$$u(x) = \int G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x}, \tag{7.33}$$

where G is the Green's function satisfying

$$\mathcal{L}G(x;\bar{x}) = \delta(x-\bar{x}). \tag{7.34}$$

Proof. Multiply (7.34) by $f(\bar{x})$, integrate w.r.t. $\bar{x}$, and we have

$$\int \mathcal{L}G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x} = \int \delta(x-\bar{x})f(\bar{x})\mathrm{d}\bar{x} = f(x),$$

where the second equality follows from the sifting property of the Dirac delta function (Lemma 5.43). Therefore

$$\mathcal{L}\int G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x} = f(x),$$

which further implies (7.33) since $\mathcal{L}$ is invertible. □

Theorem 7.33. The Dirichlet BVP

$$\begin{cases} u''(x) = f(x), \\ u(0) = \alpha, u(1) = \beta \end{cases} \tag{7.35}$$

is solved by

$$u(x) = \alpha G_0(x) + \beta G_1(x) + \hat{U}(x), \tag{7.36}$$

where $G_0(x), G_1(x)$ and $\hat{U}(x)$ are defined by BVPs as follows

$$\begin{cases} G_0''(x) = 0, \\ G_0(0) = 1, G_0(1) = 0 \end{cases} \Rightarrow \quad G_0(x) = 1-x, \tag{7.37a}$$

$$\begin{cases} G_1''(x) = 0, \\ G_1(0) = 0, G_1(1) = 1 \end{cases} \Rightarrow \quad G_1(x) = x, \tag{7.37b}$$

$$\begin{cases} \hat{U}''(x) = f(x), \\ \hat{U}(0) = 0, \hat{U}(1) = 0 \end{cases} \Rightarrow \quad \hat{U}(x) = \int_0^1 f(\bar{x})G(x;\bar{x})\mathrm{d}\bar{x}. \tag{7.37c}$$

Proof. This follows from the linearity of the BVP (7.35). □

7.5 Other boundary conditions

Example 7.34. Consider the second-order BVP

$$u''(x) = f(x) \quad \text{in } (0,1) \tag{7.38}$$

with mixed boundary conditions

$$u'(0) = \sigma, \quad u(1) = \beta. \tag{7.39}$$

As the crucial difference between this BVP and the BVP with pure Dirichlet conditions in Example 7.10, the value of $u(x)$ at $x = 0$ becomes an unknown to be solved for.

The first approach is to use a one-sided expression

$$\frac{U_1 - U_0}{h} = \sigma \tag{7.40}$$

to arrive at

$$A_E \mathbf{U}_E = \mathbf{F}_E \tag{7.41}$$

where

$$A_E = \frac{1}{h^2}\begin{bmatrix} -h & h & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & 0 & h^2 & \end{bmatrix},$$

$$\mathbf{U}_E = \begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_{m-1} \\ U_m \\ U_{m+1} \end{bmatrix}, \qquad \mathbf{F}_E = \begin{bmatrix} \sigma \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) \\ \beta \end{bmatrix}.$$

The LTE at $x_0 = 0$ is

$$\begin{aligned} \tau_0 &= \frac{1}{h^2}(hu(x_1) - hu(x_0)) - \sigma \\ &= u'(x_0) + \frac{1}{2}hu''(x_0) + O(h^2) - \sigma \\ &= \frac{1}{2}hu''(x_0) + O(h^2), \end{aligned}$$

which is only first order accurate.

The second approach is to extend the domain with a *ghost cell* $x_{-1} = -h$ and use a central difference to obtain

$$\frac{U_1 - U_{-1}}{2h} = \sigma \tag{7.42}$$

that is second-order accurate for the LTE. We do not have any information for U_{-1}, so we want to eliminate it by

$$\frac{1}{h^2}(U_{-1} - 2U_0 + U_1) = f(x_0). \tag{7.43}$$

(7.42) and (7.43) yield

$$\frac{1}{h}(-U_0 + U_1) = \sigma + \frac{h}{2}f(x_0). \tag{7.44}$$

The resulting matrix is the same as that of the first approach in (7.41) except that the first component of $\mathbf{F}_E$ has an additional term $\frac{h}{2}f(x_0)$.

The third approach is to use U_0, U_1, and U_2 to approximate $u'(0)$ and we can get a second-order FD formula, c.f. Example 6.38,

$$-\frac{1}{h}\left(\frac{3}{2}U_0 - 2U_1 + \frac{1}{2}U_2\right) = \sigma + O(h^2). \tag{7.45}$$

This results in the linear system

$$A_F \mathbf{U}_E = \mathbf{F}_E \tag{7.46}$$

where

$$A_F = \frac{1}{h^2}\begin{bmatrix} -\frac{3}{2}h & 2h & -\frac{1}{2}h & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & 0 & h^2 & \end{bmatrix}.$$

Exercise 7.35. Show that all elements of the first column of $B_E = A_E^{-1}$ are $O(1)$.

Example 7.36. Consider the second-order BVP

$$u''(x) = f(x) \text{ in } (0,1), \tag{7.47}$$

with pure Neumann conditions

$$u'(0) = \sigma_0, \quad u'(1) = \sigma_1. \tag{7.48}$$

To ensure the existence of a solution, the following compatibility condition on $f(x)$, σ_0, and σ_1 must be satisfied:

$$\int_0^1 f(x)\mathrm{d}x = \int_0^1 u''(x)\mathrm{d}x = u'(1) - u'(0) = \sigma_1 - \sigma_0. \tag{7.49}$$

In fact, if (7.49) holds, there are an infinite number of solutions: if v is a solution of (7.47), $v + \mathbb{R}$ are also solutions.

Using procedures similar to those in Example 7.34, we can discretize (7.47) and (7.48) as

$$A_F \mathbf{U}_E = \mathbf{F}_F, \tag{7.50}$$

where

$$A_F = \frac{1}{h^2}\begin{bmatrix} -h & h & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & h & -h & \end{bmatrix}, \tag{7.51}$$

$$\mathbf{F}_F = \begin{bmatrix} \sigma_0 + \frac{h}{2}f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) \\ -\sigma_1 + \frac{h}{2}f(x_{m+1}) \end{bmatrix}.$$

Lemma 7.37. The matrix A_F in (7.51) satisfies

$$\dim \mathcal{N}(A_F) = 1. \tag{7.52}$$

Proof. Clearly, $\mathbf{e} = (1, 1, \cdots, 1)^T$ is in the null space of A_F. The rest follows from the well-posedness of the BVP with mixed conditions. □

Theorem 7.38 (Solvability condition)**.** The linear system (7.50) has a solution if and only if

$$\frac{h}{2} f(x_0) + h \sum_{i=1}^{m} f(x_i) + \frac{h}{2} f(x_{m+1}) = \sigma_1 - \sigma_0. \tag{7.53}$$

Proof. The fundamental theorem of linear algebra (Theorem B.89) implies

$$\mathbb{R}^{m+2} = \mathcal{R}(A_F) \oplus \mathcal{N}(A_F^T) \tag{7.54}$$

and $\dim \mathcal{N}(A_F^T) = \dim \mathcal{N}(A_F)$. Lemma 7.37 further yields $\dim \mathcal{N}(A_F^T) = 1$. Then it is readily verified that

$$\mathcal{N}(A_F^T) = \operatorname{span}\left\{(1, h, h, \ldots, h, 1)^T\right\}.$$

For sufficiency, the above equation and (7.53) imply that $\mathbf{F}_F$ is orthogonal to $\mathcal{N}(A_F^T)$ and thus (7.54) yields $\mathbf{F}_F \in \mathcal{R}(A_F)$. Hence (7.50) must have a solution. As for necessity, the existence of a solution of (7.50) implies $\mathbf{F}_F \in \mathcal{R}(A_F)$, which, together with the above two equations, implies (7.53). □

7.6 BVPs in two dimensions

Example 7.39 (An FD method for Poisson's equation in a unit square)**.** Consider the two-dimensional BVP

$$-\frac{\partial^2}{\partial x^2} u(x, y) - \frac{\partial^2}{\partial y^2} u(x, y) = f(x, y) \tag{7.55}$$

in $\Omega := (0, 1)^2$ with homogeneous Dirichlet conditions

$$u(x, y)|_{\partial\Omega} = 0. \tag{7.56}$$

A uniform Cartesian grid can be generated with

$$x_i = ih, \quad y_j = jh, \quad i, j = 0, 1, \cdots, m, m+1, \tag{7.57}$$

where $h = \Delta x = \Delta y = \frac{1}{m+1}$ is the uniform grid size.

Approximate $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ separately and we have, $\forall i, j = 1, 2, \ldots, m$,

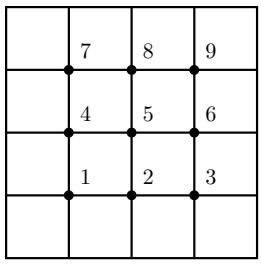
$$-\frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} - \frac{U_{i,j-1} - 2U_{ij} + U_{i,j+1}}{h^2} = f_{ij}. \tag{7.58}$$

These $m \times m$ equations organize into a single system

$$A_{2D}\mathbf{U} = \mathbf{F}. \tag{7.59}$$

Exercise 7.40. Show that the LTE $\boldsymbol{\tau}$ of the FD method in Example 7.39 is

$$\tau_{i,j} = -\frac{1}{12} h^2 \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right)\Big|_{(x_i, y_j)} + O(h^4). \tag{7.60}$$



Example 7.41. For $m = 3$ with ordering as shown above, we have

$$A_{2D} = \frac{1}{h^2} \begin{bmatrix} +4 & -1 & & -1 & & & & & \\ -1 & +4 & -1 & & -1 & & & & \\ & -1 & +4 & & & -1 & & & \\ -1 & & & +4 & -1 & & -1 & & \\ & -1 & & -1 & +4 & -1 & & -1 & \\ & & -1 & & -1 & +4 & & & -1 \\ & & & -1 & & & +4 & -1 & \\ & & & & -1 & & -1 & +4 & -1 \\ & & & & & -1 & & -1 & +4 \end{bmatrix}$$

$$= \frac{1}{h^2} \begin{bmatrix} T & -I & \\ -I & T & -I \\ & -I & T \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \\ U_{12} \\ U_{22} \\ U_{32} \\ U_{13} \\ U_{23} \\ U_{33} \end{bmatrix},$$

where

$$T = \begin{bmatrix} +4 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & +4 \end{bmatrix}. \tag{7.61}$$

where $\mathbf{U}$ is obtained by stacking the columns on top of each other.

Lemma 7.42. Let $\mathbf{E} = \mathbf{U} - \hat{\mathbf{U}}$ denote the global error of the linear system (7.59). Then the LTE (7.60) satisfies

$$A_{2D}\mathbf{E} = -\boldsymbol{\tau}. \tag{7.62}$$

Proof. The proof is the same as that of Lemma 7.18. □

7.6.1 Kronecker product

Definition 7.43. The *Kronecker product* of two matrices $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{p \times q}$ is another matrix $A \otimes B \in \mathbb{C}^{mp \times nq}$ given by

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}, \tag{7.63}$$

where a_{ij} is the (i, j)th element of A.

Example 7.44.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} 0 & 5 \\ 6 & 7 \end{bmatrix} = \begin{bmatrix} 0 & 5 & 0 & 10 \\ 6 & 7 & 12 & 14 \\ 0 & 15 & 0 & 20 \\ 18 & 21 & 24 & 28 \end{bmatrix}.$$

Definition 7.45. For $X \in \mathbb{C}^{m\times n}$, vec$(X)$ is defined to be a column vector of size mn made of the columns of X stacked on top of one another from left to right.

Lemma 7.46. Any $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{n\times n}$, and $X \in \mathbb{C}^{m\times n}$ satisfy

$$\text{vec}(AX) = (I_n \otimes A)\text{vec}(X), \tag{7.64}$$
$$\text{vec}(XB) = (B^T \otimes I_m)\text{vec}(X). \tag{7.65}$$

Proof. We have

$$\begin{aligned}\text{vec}(AX) &= \text{vec}([A\mathbf{X}_1, A\mathbf{X}_2, \cdots, A\mathbf{X}_n]) \\ &= \begin{bmatrix} A\mathbf{X}_1 \\ A\mathbf{X}_2 \\ \vdots \\ A\mathbf{X}_n \end{bmatrix} = \begin{bmatrix} A & & & \\ & A & & \\ & & \ddots & \\ & & & A \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \\ &= (I_n \otimes A) \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}.\end{aligned}$$

Let $Y = XB$, then

$$\mathbf{Y}_j = X\mathbf{b}_j \Rightarrow y_{kj} = \sum_{i=1}^{n} x_{ki}b_{ij}. \tag{7.66}$$

Let $C = B^T \otimes I_m$, then the (i,j)-th sub-block of C is

$$C_{ij} = b_{ji}I_m. \tag{7.67}$$

Let $D = C\text{vec}(X)$, then the j-th block of D is

$$\mathbf{D}_j = \sum_{i=1}^{n} C_{ji}\mathbf{X}_i = \sum_{i=1}^{n} b_{ij}I_m\mathbf{X}_i = \sum_{i=1}^{n} b_{ij}\mathbf{X}_i, \tag{7.68}$$

and the (k,j)-th entry of D is (Here we also use (k,j) to denote the scalar index corresponding to the multi-index (k,j).)

$$d_{(k,j)} = \sum_{i=1}^{n} b_{ij}x_{ki} = \sum_{i=1}^{n} x_{ki}b_{ij}. \tag{7.69}$$

Combining (7.66) and (7.69) yields (7.65). □

7.6.2 Convergence in the 2-norm

Lemma 7.47. The linear system (7.59) is equivalent to

$$AU_{m\times m} + U_{m\times m}A = F_{m\times m}, \tag{7.70}$$

where the (i,j)th element of $U_{m\times m}$ is the computed solution at the (i,j)th grid point, the (i,j)th element of $F_{m\times m}$ is

$$(F_{m\times m})_{ij} = f(ih, jh),$$

and A is the 1D discrete Laplacian in (7.13).

Proof. A direct computation gives

$$\begin{cases} (AU_{m\times m})_{ij} = \dfrac{1}{h^2}(-U_{i-1,j} + 2U_{ij} - U_{i+1,j}), \\ (U_{m\times m}A)_{ij} = \dfrac{1}{h^2}(-U_{i,j-1} + 2U_{ij} - U_{i,j+1}), \end{cases} \tag{7.71}$$

and the *homogeneous* Dirichlet condition yields

$$AU_{m\times m} + U_{m\times m}A = F_{m\times m}.$$ □

Lemma 7.48. The 1D discrete Laplacian A in (7.13) satisfies

$$\text{vec}(AU_{m\times m} + U_{m\times m}A) = (I_m \otimes A + A \otimes I_m)\text{vec}(U_{m\times m}).$$

Proof. By Lemma 7.46, we have

$$\text{vec}(AU_{m\times m}) = (I_m \otimes A)\text{vec}(U_{m\times m}),$$

and

$$\text{vec}(U_{m\times m}A) = (A^T \otimes I_m)\text{vec}(U_{m\times m}) = (A \otimes I_m)\text{vec}(U_{m\times m}),$$

where the second equality follows from the symmetry of A. Adding these two equations gives the desired result. □

Theorem 7.49. With matrix ordering, the linear system (7.59) can be written as

$$A_{2D} = I_m \otimes A + A \otimes I_m,\ \mathbf{U} = \text{vec}(U_{m\times m}),\ \mathbf{F} = \text{vec}(F_{m\times m}).$$

Proof. This follows from Lemma 7.47 and Lemma 7.48. □

Definition 7.50. The *discrete Laplacian* in n-dimensional space analogous to the 1D discrete Laplacian (7.13) is

$$A_{nD} = \sum_{j=0}^{n-1} \underbrace{I_m \otimes \cdots \otimes I_m}_{\#I_m = j} \otimes A \otimes \underbrace{I_m \otimes \cdots \otimes I_m}_{\#I_m = n-j-1}. \tag{7.72}$$

Example 7.51. For $n = 3$, we have

$$A_{3D} = A \otimes I_m \otimes I_m + I_m \otimes A \otimes I_m + I_m \otimes I_m \otimes A.$$

Theorem 7.52. The eigenpairs of A_{2D} are

$$\lambda_{ij} = \lambda_i + \lambda_j, \quad \mathbf{W}_{ij} = \text{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right), \tag{7.73}$$

where $i, j = 1, 2, \ldots, m$ and $(\lambda_i, \mathbf{w}_i)$ is an eigenpair of A in Lemma 7.25.

Proof. By Lemma 7.25, we have

$$\begin{aligned} A\mathbf{w}_i\mathbf{w}_j^T + \mathbf{w}_i\mathbf{w}_j^T A &= \lambda_i\mathbf{w}_i\mathbf{w}_j^T + \mathbf{w}_i\mathbf{w}_j^T A^T \\ &= \lambda_i\mathbf{w}_i\mathbf{w}_j^T + \mathbf{w}_i(A\mathbf{w}_j)^T \\ &= \lambda_i\mathbf{w}_i\mathbf{w}_j^T + \lambda_j\mathbf{w}_i\mathbf{w}_j^T \\ &= (\lambda_i + \lambda_j)\mathbf{w}_i\mathbf{w}_j^T. \end{aligned}$$

Then Theorem 7.49 and Lemma 7.48 yield

$$\begin{aligned} A_{2D}\text{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right) &= (I_m \otimes A + A \otimes I_m)\text{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right) \\ &= \text{vec}\left(A(\mathbf{w}_i\mathbf{w}_j^T) + (\mathbf{w}_i\mathbf{w}_j^T)A\right) \\ &= (\lambda_i + \lambda_j)\text{vec}\left(\mathbf{w}_i\mathbf{w}_j^T\right), \end{aligned}$$

and hence $\lambda_i + \lambda_j$ is an eigenvalue of A_{2D} with corresponding eigenvector $\mathbf{W}_{ij}$. □

Theorem 7.53. The FD method in Example 7.39 is second-order convergent in the 2-norm.

Proof. We have $\|A_{2D}\|_2 = \rho(A_{2D})$ since A_{2D} is symmetric. Then Theorem 7.52 yields

$$\lim_{h\to 0} \|A_{2D}^{-1}\|_2 = \lim_{h\to 0} \frac{1}{\min|\lambda_{ij}|} = \lim_{h\to 0} \frac{h^2}{8\sin^2\frac{\pi h}{2}} = \frac{1}{2\pi^2} = O(1).$$

By Definition 7.21, the method is stable. The proof is completed by (7.60), Definition 7.19, Theorem 7.22, and Lemma 7.42. □

7.6.3 Convergence in the max-norm via a discrete maximum principle

Theorem 7.54. The FD method in Example 7.39 is second-order convergent in the max-norm.

Proof. Let $\mathbf{X}_I$ be the grid obtained by removing from $\mathbf{X}$ in (7.57) those grid points satisfying $i = 0, m+1$ or $j = 0, m+1$. Define a linear map $\hat{A}_{2D} : \{\mathbf{X} \to \mathbb{R}\} \to \{\mathbf{X}_I \to \mathbb{R}\}$,

$$\begin{aligned}\hat{A}_{2D}U_{i,j} &:= (\hat{A}_{2D}\mathbf{U})_{i,j} \\ &= \frac{1}{h^2}(4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1}).\end{aligned} \tag{7.74}$$

The matrix of $\hat{A}_{2D}$ is different from A_{2D} in Example 7.39. (How?) Define a *comparison function* $\phi : \mathbb{R}^2 \to \mathbb{R}$,

$$\phi(x, y) := \left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2 \tag{7.75}$$

and write $\phi_{i,j} := \phi(ih, jh)$. (7.74) and (7.75) yield

$$\hat{A}_{2D}\phi_{i,j} = -4. \tag{7.76}$$

Let $E, \tau : \mathbf{X} \to \mathbb{R}$ be the solution error and the LTE, respectively; write $\tau_m := \max_{i,j} |\tau_{i,j}|$. Construct a grid function

$$\psi_{i,j} := E_{i,j} + \frac{1}{4}\tau_m\phi_{i,j} \tag{7.77}$$

and we have, from $\hat{A}_{2D}E_{i,j} = -\tau_{i,j}$ (why?) and (7.76),

$$\hat{A}_{2D}\psi_{i,j} = -\tau_{i,j} - \tau_m \le 0.$$

By (7.74), $\hat{A}_{2D}\psi_{i,j} \le 0$ dictates that $\psi_{i,j}$ be no greater than at least one of its neighbors $\psi_{i+1,j}$, $\psi_{i-1,j}$, $\psi_{i,j+1}$, and $\psi_{i,j-1}$. Therefore, *the maximum value of ψ must occur at a boundary point*, i.e., a point with $i = 0, m+1$ or $j = 0, m+1$. Consequently, there exists some constant $C > 0$ such that

$$E_{i,j} \le \psi_{i,j} \le \frac{1}{8}\tau_m < Ch^2,$$

where the first step follows from (7.77) and $\tau_m\phi_{i,j} \ge 0$, the second step from $E_{i,j}$ being zero at all boundary points and the fact that the maximum of ϕ is $\frac{1}{2}$ at the domain corners, and the last step from (7.60).

By similar arguments, the grid function

$$\chi_{i,j} := -E_{i,j} + \frac{1}{4}\tau_m\phi_{i,j} \tag{7.78}$$

satisfies $\hat{A}_{2D}\chi_{i,j} \le 0$ for all grid points and thus

$$-E_{i,j} \le \chi_{i,j} \le \frac{1}{8}\tau_m < Ch^2.$$

To sum up, we have $|E_{i,j}| = O(h^2)$ for all grid points. □

Notation 4. Consider discretizing a BVP on domain Ω. Denote by $\mathbf{X}_\Omega$ the set of *equation-discretization points* where the BVP is discretized and where values of the unknown function u are sought. Let $\mathbf{X}_{\partial\Omega} \subset \partial\Omega$ be the set of *boundary points* so that each point $Q \in \mathbf{X}_{\partial\Omega}$ satisfies

- the BVP is not discretized at Q,
- $u(Q)$ is prescribed by a Dirichlet condition,
- this Dirichlet condition at Q is involved in discretizing the BVP at some $P \in \mathbf{X}_\Omega$.

WLOG, we also assume $\mathbf{X} = \mathbf{X}_\Omega \cup \mathbf{X}_{\partial\Omega}$.

Example 7.55. A boundary point in Notation 4 must be on the domain boundary $\partial\Omega$, but an equation-discretization point P might also belong to $\partial\Omega$: if we have a Neumann condition for P, then the *value* of the unknown at P might still be needed. In general, a grid $\mathbf{X}$ that corresponds to a straightforward discretization of the problem domain might contain certain points that belong neither to $\mathbf{X}_\Omega$ nor to $\mathbf{X}_{\partial\Omega}$, e.g., the four corners of the square domain $(0, 1)^2$. In Notation 4, we have assumed $\mathbf{X} = \mathbf{X}_\Omega \cup \mathbf{X}_{\partial\Omega}$ to exclude these abnormal points so that hereafter the analysis can be relatively simple.

Lemma 7.56 (Discrete maximum principle)**.** Suppose that an FD discretization of a linear BVP yields

$$\forall P \in \mathbf{X}_\Omega, \quad L_hU_P - f_P + g_P = 0, \tag{7.79}$$

where f_P corresponds to the RHS of (7.3), g_P corresponds to all boundary data other than Dirichlet conditions, and L_h and $\mathbf{X}_\Omega$ satisfy

(DMP-1) for each equation-discretization point $P \in \mathbf{X}_\Omega$, L_h is of the form

$$L_hU_P = c_PU_P - \sum_{Q \in Q_P} c_QU_Q, \tag{7.80}$$

where $Q_P \subset \mathbf{X}_\Omega \cup \mathbf{X}_{\partial\Omega}$, $c_P > 0$ and each $c_Q > 0$. The set $\{P\} \cup Q_P$ in (7.80) is called the *P-stencil* or the *stencil of L_h at P*;

(DMP-2) $\forall P \in \mathbf{X}_\Omega$, $c_P \ge \sum_{Q \in Q_P} c_Q$;

(DMP-3) $\mathbf{X}_\Omega$ is *connected*, i.e.,

$$\begin{aligned}&\forall P_0, P_m \in \mathbf{X}_\Omega, \exists P_1,\ P_2,\ \ldots,\ P_{m-1} \text{ s.t.} \\ &\forall r = 1, 2, \ldots, m,\ P_r \text{ is in the } P_{r-1}\text{-stencil};\end{aligned} \tag{7.81}$$

(DMP-4) at least one equation (7.80) involves a boundary value U_Q given by a Dirichlet condition.

Then for any grid function $\psi : \mathbf{X} \to \mathbb{R}$ satisfying

$$\begin{cases}\max\limits_{P \in \mathbf{X}} \psi_P \ge 0, \\ \forall P \in \mathbf{X}_\Omega, \quad L_h\psi_P \le 0,\end{cases} \tag{7.82}$$

we have

$$\max_{P \in \mathbf{X}_\Omega} \psi_P \le \max_{Q \in \mathbf{X}_{\partial\Omega}} \psi_Q. \tag{7.83}$$

Proof. Suppose

$$M_\Omega := \max_{Q\in\mathbf{X}_\Omega} \psi_Q > M_{\partial\Omega} := \max_{Q\in\mathbf{X}_{\partial\Omega}} \psi_Q$$

and let P be the point where ψ attains M_Ω. Then

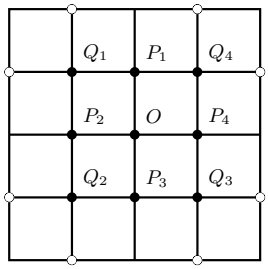
$$(*): \quad M_\Omega = \psi_P \le \frac{1}{c_P}\sum_{Q\in Q_P} c_Q\psi_Q \le \frac{1}{c_P}\sum_{Q\in Q_P} c_Q M_\Omega \le M_\Omega,$$

where the first inequality follows from (7.82) and (7.80), the second from the definition of M_Ω, and the third from (DMP-2) and $M_\Omega \ge 0$. For the second "$\le$" in $(*)$ to be "=," we must have $\psi_Q = \psi_P$ for each Q. By (DMP-3), ψ takes the same value M_Ω on all equation-discretization points. Then (DMP-4) implies $M_\Omega = M_{\partial\Omega}$, which contradicts the starting point $M_\Omega > M_{\partial\Omega}$. □

Example 7.57. Suppose the concept of connectedness in (DMP-3) were defined as

$$\forall P_0, P_m \in \mathbf{X}_\Omega, \exists P_1,\ P_2,\ \ldots,\ P_{m-1} \text{ s.t. } \forall r = 1, 2, \ldots, m, \text{ both } U_{P_r} \text{ and } U_{P_{r-1}} \text{ appear in some equation (7.79).} \tag{7.84}$$

Then the conclusion (7.83) of Lemma 7.56 would not hold. The following is a counter-example.



Let L_h be

$$\begin{aligned}
L_h\psi_{P_1} &= 3\psi_{P_1} - \psi_{P_2} - \psi_{P_3} - \psi_{P_4},\\
L_h\psi_{P_2} &= 3\psi_{P_2} - \psi_{P_1} - \psi_{P_3} - \psi_{P_4},\\
L_h\psi_{P_3} &= 3\psi_{P_3} - \psi_{P_1} - \psi_{P_2} - \psi_{P_4},\\
L_h\psi_{P_4} &= 3\psi_{P_4} - \psi_{P_1} - \psi_{P_2} - \psi_{P_3},\\
L_h\psi_O &= 4\psi_O - \psi_{P_1} - \psi_{P_2} - \psi_{P_3} - \psi_{P_4},
\end{aligned}$$

and the expressions of L_h at Q_i's be similar with that of L_h at O such that (DMP-1,2,4) hold. (7.84) also holds. The following distribution of the grid function ψ,

$$\psi_{\mathbf{x}} = \begin{cases} 10 & \text{if } \mathbf{x} = P_1, P_2, P_3, P_4;\\ 1 & \text{otherwise,}\end{cases}$$

satisfies (7.82), but the conclusion (7.83) does not hold.

The key point of this example is that, in order for two points P_0 and P_m to be called connected in the context of solving elliptic equations, there must simultaneously exist a path from P_0 to P_m and a path from P_m to P_0. The definition of connectedness in (7.84) is not satisfactory because it fails to capture the direction of the path. In contrast, (7.81) captures the dependence of path connectedness on the direction of the path.

Theorem 7.58. Suppose an FD discretization of a BVP satisfies the conditions (DMP-1,2,3,4) in Lemma 7.56. Then the solution error $E_P := U_P - u(P)$ of the FD method (7.79) is bounded by

$$\forall P \in \mathbf{X}, \quad |E_P| \le T_{\max}\left(\max_{Q\in\mathbf{X}_{\partial\Omega}} \phi(Q)\right), \tag{7.85}$$

where $T_{\max} = \max_{P\in\mathbf{X}_\Omega} |T_P|$, T_P is the LTE at P and $\phi : \mathbf{X} \to \mathbb{R}$ is a nonnegative grid function satisfying

$$\forall P \in \mathbf{X}_\Omega, \quad L_h\phi_P \le -1. \tag{7.86}$$

Proof. Lemma 7.18 implies $L_hE_P = -T_P$. Define

$$\psi_P := E_P + T_{\max}\phi_P$$

and we know from (7.86) that

$$L_h\psi_P \le -T_P - T_{\max} \le 0.$$

Furthermore, we have $\max_{P\in\mathbf{X}} \psi_P \ge 0$ because $\phi_P \ge 0$ and

$$(*): \quad \forall Q \in \mathbf{X}_{\partial\Omega},\ E_Q = 0,$$

c.f. Notation 4. Then we have

$$\begin{aligned}
E_P &\le \max_{P\in\mathbf{X}} (E_P + T_{\max}\phi_P)\\
&\le \max_{Q\in\mathbf{X}_{\partial\Omega}} (E_Q + T_{\max}\phi_Q) = T_{\max}\max_{Q\in\mathbf{X}_{\partial\Omega}} (\phi_Q),
\end{aligned}$$

where the first step follows from $T_{\max}\phi_P \ge 0$, the second from Lemma 7.56, and the third from $(*)$.

Repeat the above arguments with $\psi_P = -E_P + T_{\max}\phi_P$ and we have

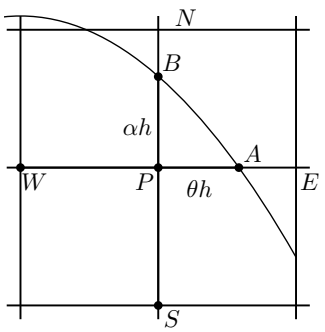
$$-E_P \le T_{\max}\max_{Q\in\mathbf{X}_{\partial\Omega}} (\phi_Q).$$ □

7.6.4 Convergence on irregular domains

Example 7.59 (An FD method for Poisson's equation in 2D irregular domains)**.** Consider the BVP

$$-\frac{\partial^2}{\partial x^2}u(x,y) - \frac{\partial^2}{\partial y^2}u(x,y) = f(x,y) \tag{7.87}$$

in a 2D irregular domain Ω with Dirichlet conditions.



An equation-discretization point is said to be *regular* if the standard 5-point stencil is applicable; otherwise it is *irregular*. For an irregular point, we modify the FD discretization in Example 7.39 to incorporate the info of local geometry and Dirichlet conditions. For example, in the above plot, the discrete operator becomes

$$L_h U_P := \frac{(1+\theta)U_P - U_A - \theta U_W}{\frac{1}{2}\theta(1+\theta)h^2} + \frac{(1+\alpha)U_P - U_B - \alpha U_S}{\frac{1}{2}\alpha(1+\alpha)h^2}. \tag{7.88}$$

In the resulting linear system $L_h U_P - f_P = 0$, the form of $L_h U_P$ as in (7.88) is different from that in (7.58) at a regular equation-discretization point. Consequently, a global analysis of this linear system is difficult.

Exercise 7.60. Show that, in Example 7.59, the LTE at an irregular equation-discretization point is $O(h)$ while the LTE at a regular equation-discretization point is $O(h^2)$.

Theorem 7.61. Suppose that, in the notation of Theorem 7.58, the set $\mathbf{X}_\Omega$ of equation-discretization points can be partitioned as

$$\mathbf{X}_\Omega = \mathbf{X}_1 \cup \mathbf{X}_2, \quad \mathbf{X}_1 \cap \mathbf{X}_2 = \emptyset,$$

the nonnegative function $\phi : \mathbf{X} \to \mathbb{R}$ satisfies

$$\forall P \in \mathbf{X}_1, \quad L_h\phi_P \le -C_1 < 0; \tag{7.89a}$$
$$\forall P \in \mathbf{X}_2, \quad L_h\phi_P \le -C_2 < 0, \tag{7.89b}$$

and the LTE of (7.79) satisfy

$$\forall P \in \mathbf{X}_1, \quad |T_P| < T_1; \tag{7.90a}$$
$$\forall P \in \mathbf{X}_2, \quad |T_P| < T_2. \tag{7.90b}$$

Then the solution error $E_P := U_P - u(P)$ of the FD method (7.79) is bounded by

$$\forall P \in \mathbf{X}, \quad |E_P| \le \left(\max_{Q \in \mathbf{X}_{\partial\Omega}} \phi(Q) \right) \max\left\{ \frac{T_1}{C_1}, \frac{T_2}{C_2} \right\}.$$

Exercise 7.62. Prove Theorem 7.61 by choosing a function ψ to which Lemma 7.56 applies.

Theorem 7.63. The FD method in Example 7.59 is second-order convergent in the max-norm.

Proof. Define a comparison function ϕ as

$$\phi(x,y) := \begin{cases} F_1\left[(x-p)^2 + (y-q)^2\right] & \text{if } (x,y) \in \mathbf{X}_\Omega; \\ F_1\left[(x-p)^2 + (y-q)^2\right] + F_2 & \text{if } (x,y) \in \mathbf{X}_{\partial\Omega}, \end{cases}$$

where (p,q) is the geometric center of Ω and $F_1, F_2 > 0$ are constants to be chosen later. Both regular points and irregular points belong to $\mathbf{X}_\Omega$. Their difference is that, for a regular point Q, we have

$$L_h\phi_Q = -4F_1$$

while for an irregular point P shown in Example 7.59 the coefficient of U_A is

$$-\frac{2}{\theta(1+\theta)h^2} < -\frac{1}{h^2}$$

because $\theta \in (0,1)$. Therefore,

$$L_h\phi_P < -4F_1 - \frac{1}{h^2}F_2 < -\frac{1}{h^2}F_2. \tag{7.91}$$

Note that the last upper bound $-\frac{1}{h^2}F_2$ can not be sharpened to $-\frac{2}{h^2}F_2$ or $-\frac{3}{h^2}F_2$ because the stencil of the irregular point p might as well only contains one point outside the domain.

By Exercise 7.60, we write the maximum LTEs on regular and irregular points as $T_1 = K_1h^2$ and $T_2 = K_2h$, respectively. Then Theorem 7.61 implies

$$|E_P| \le (F_1R^2 + F_2) \max\left\{ \frac{K_1h^2}{4F_1}, \frac{K_2h^3}{F_2} \right\}, \tag{7.92}$$

where R is the maximum distance of a point in Ω to the geometric center of Ω. The RHS of the above equation is minimized when we choose $\frac{F_1}{F_2} = \frac{K_1}{4K_2h}$ so that the two terms in max{} equal. It follows that

$$\forall P \in \mathbf{X}, \quad |E_P| \le \frac{1}{4}K_1R^2h^2 + K_2h^3. \qquad \square$$

7.7 Programming assignments

Write a C++ package to solve the two-dimensional Poisson equation in Definition 7.4 by the FD methods in Examples 7.39 and 7.59.

I. Your package must give the user the following options:

   (a) the problem domain: either $\Omega = (0,1)^2$ or $\Omega \setminus \mathbf{D}$ where $\mathbf{D}$ is a closed circular disk, of which the radius and the center, specified by the user, must keep $\Omega \setminus \mathbf{D}$ connected and $\mathbf{D}$ must cover at least four equation-discretization points in Notation 4 (you must check the validity of input parameters);

   (b) boundary conditions: Dirichlet, Neumann, or mixed (partly Dirichlet and partly Neumann).

   Use a direct method such as the LU factorization to solve the linear system, with your favorite implementation of a BLAS or LAPACK.

II. For the function

$$u(x,y) = \exp(y + \sin(x)), \tag{7.93}$$

derive the corresponding $f(x,y)$ and the boundary conditions. Test your solver for all combinations of (a,b) in I on grids with $n = 8, 16, 32, 64$ along each dimension, report the 1-, 2-, and ∞-norms of the errors and the corresponding convergence rates on the four grids. You should also design at least two of your own test functions and carry out the same process.

III. The user-specified parameters must be clearly listed in an input file with a key-value syntax; I recommend `json` (`https://www.json.org/json-en.html`), but you may use or write your own parser. The main program is supposed to read the input file, create objects from the input, call your BVP solver, generate test results, and report errors and convergence rates.

IV. The generation of pictures on solution and errors can be performed outside your program using another tool of your choice.

V. Write a GNU `makefile` under your root directory so that the command "`make run`" would trigger the compilation of your source code, the production of the executable, and the running of your tests.

VI. Write a report to summarize the main points of your numerical experiments, which should be designed to verify the analytic results in the notes. To some extent, your grade depends on the number of key conclusions in the notes confirmed by your numerical experiments.

Chapter 8

Basic Iterative Methods for Linear Systems

8.1 Jacobi, Gauss-Seidel, and SOR

Definition 8.1. A *fixed point iteration for solving a linear system*

$$A\mathbf{x} = \mathbf{b}, \tag{8.1}$$

is an iteration of the form

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + \mathbf{c}, \tag{8.2}$$

where T and $\mathbf{c}$ are functions of A and $\mathbf{b}$ such that $\mathbf{x} = T\mathbf{x}+\mathbf{c}$ and $\mathbf{x}^{(k)}$ is the kth iterate that approximates $\mathbf{x}$.

Definition 8.2. The *Jacobi iteration* determines the next iterate $\mathbf{x}^{(k+1)}$ by separately annihilating the ith component of the residual:

$$a_{ii}x_i^{(k+1)} = -\sum_{j\neq i;j=1}^{n} a_{ij}x_j^{(k)} + b_i, \tag{8.3}$$

where $x_i^{(k)}$ is the ith component of the kth iterate $\mathbf{x}^{(k)}$.

Lemma 8.3. The Jacobi iteration is a fixed point iteration for solving (8.1) with

$$\begin{cases} T_J = D^{-1}(L+U), \\ \mathbf{c} = D^{-1}\mathbf{b}, \end{cases} \tag{8.4}$$

where D, $-L$, and $-U$ are respectively the diagonal, lower triangular, and upper triangular part of A:

$$A = D - L - U. \tag{8.5}$$

Proof. This follows directly from (8.3). □

Example 8.4. For the linear system in (7.11), the iteration matrix of the Jacobi method $T_J = D^{-1}(L+U)$ is given by

$$t_{ij} = \begin{cases} \frac{1}{2} & \text{if } i-j = \pm 1; \\ 0 & \text{otherwise.} \end{cases}$$

Definition 8.5. The *Gauss-Seidel iteration* determines the next iterate $\mathbf{x}^{(k+1)}$ by annihilating the ith component of the residual in the order $i = 1, 2, \ldots, n$ using new components of $\mathbf{x}^{(k+1)}$ whenever possible:

$$a_{ii}x_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i. \tag{8.6}$$

Lemma 8.6. The Gauss-Seidel iteration is a fixed point iteration for solving (8.1) with

$$\begin{cases} T_{GS} = (D-L)^{-1}U, \\ \mathbf{c} = (D-L)^{-1}\mathbf{b}, \end{cases} \tag{8.7}$$

where D, $-L$, and $-U$ are the same as in Lemma 8.3.

Proof. This follows directly from (8.6). □

Definition 8.7. The *backward Gauss-Seidel iteration* is the same as the Gauss-Seidel iterate except that components of the next iterate $\mathbf{x}^{(k+1)}$ is updated in the reversed order $i = n, n-1, \ldots, 1$:

$$a_{ii}x_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k+1)} + b_i. \tag{8.8}$$

Lemma 8.8. The backward Gauss-Seidel iteration is a fixed point iteration for solving (8.1) with

$$\begin{cases} T_{BGS} = (D-U)^{-1}L, \\ \mathbf{c} = (D-U)^{-1}\mathbf{b}, \end{cases} \tag{8.9}$$

where D, $-L$, and $-U$ are the same as in Lemma 8.3.

Proof. This follows directly from (8.8). □

Definition 8.9. The *weighted Jacobi iteration* is a fixed point iteration of the form

$$\mathbf{x}_* = T_J\mathbf{x}^{(k)} + \mathbf{c} \tag{8.10a}$$

$$\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{x}^{(k)} + \omega\mathbf{x}_*, \tag{8.10b}$$

where T_J and $\mathbf{c}$ are given in (8.4).

Definition 8.10. The *successive over relaxation* (SOR) determines the next iterate $\mathbf{x}^{(k+1)}$ by annihilating components of the residual in the order $i = 1, 2, \ldots, n$ using a linear combination of the Gauss-Seidel iterate and the current one:

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega x_i^{GS}, \tag{8.11}$$

where x_i^{GS} equals $x_i^{(k+1)}$ in (8.6).

Lemma 8.11. The SOR is a fixed point iteration for solving (8.1) with

$$\begin{cases} T_{SOR} = (D-\omega L)^{-1}[\omega U + (1-\omega)D], \\ \mathbf{c} = \omega(D-\omega L)^{-1}\mathbf{b}, \end{cases} \tag{8.12}$$

where D, $-L$, and $-U$ are the same as in Lemma 8.3.

Proof. This follows from (8.11) and (8.6). □

Definition 8.12. The *backward SOR* determines the next iterate $\mathbf{x}^{(k+1)}$ by annihilating components of the residual in the reversed order $i = n, n-1, \ldots, 1$ using a linear combination of the backward Gauss-Seidel iterate and the current one:

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega x_i^{BGS}, \tag{8.13}$$

where x_i^{BGS} equals $x_i^{(k+1)}$ in (8.8).

Lemma 8.13. The backward SOR is a fixed point iteration for solving (8.1) with

$$\begin{cases} T_{BSOR} = (D-\omega U)^{-1}[\omega L + (1-\omega)D], \\ \mathbf{c} = \omega(D-\omega U)^{-1}\mathbf{b}, \end{cases} \tag{8.14}$$

where D, $-L$, and $-U$ are the same as in Lemma 8.3.

Proof. This follows from (8.13) and (8.8). □

Definition 8.14. Each step of the *symmetric SOR* (SSOR) consists of the SOR step followed by a backward SOR step,

$$(D-\omega L)\mathbf{x}_* = [(1-\omega)D + \omega U]\mathbf{x}^{(k)} + \omega\mathbf{b}, \tag{8.15a}$$
$$(D-\omega U)\mathbf{x}^{(k+1)} = [(1-\omega)D + \omega L]\mathbf{x}_* + \omega\mathbf{b}. \tag{8.15b}$$

Lemma 8.15. The SSOR is a fixed point iteration for solving (8.1) with

$$\begin{aligned} T_{SSOR} = \;& (D-\omega U)^{-1}[(1-\omega)D + \omega L] \\ & (D-\omega L)^{-1}[(1-\omega)D + \omega U], \\ \mathbf{c} = \;& \omega(2-\omega)(D-\omega U)^{-1}D(D-\omega L)^{-1}\mathbf{b}, \end{aligned} \tag{8.16}$$

where D, $-L$, and $-U$ are the same as in Lemma 8.3.

Exercise 8.16. Prove Lemma 8.15.

8.2 General convergence analysis

8.2.1 Similarity transformations

Definition 8.17. Two matrices $A, B \in \mathbb{C}^{n\times n}$ are *similar* if there exists a nonsingular matrix S such that

$$B = SAS^{-1}, \tag{8.17}$$

and the map $A \mapsto SAS^{-1}$ is a *similarity transformation*. In particular, B is *unitarily similar* to A if there exists a unitary matrix S as in Definition B.189 such that (8.17) holds.

Lemma 8.18. Two similar matrices A and B have the same set of eigenvalues.

Proof. Let $(\lambda, \mathbf{u})$ be an eigenpair of A, i.e., $A\mathbf{u} = \lambda\mathbf{u}$. Then (8.17) yields

$$BS\mathbf{u} = SA\mathbf{u} = \lambda S\mathbf{u},$$

and thus $(\lambda, S\mathbf{u})$ is an eigenpair of B. □

Theorem 8.19 (Jordan canonical form)**.** Every matrix $A \in \mathbb{C}^{n\times n}$ has a similarity transformation

$$A = RJR^{-1}, \tag{8.18}$$

where R is invertible and J is a block diagonal matrix of the form (B.52), and each $J(\lambda_i, k_i)$ is a Jordan block of order k_i, $\sum_{i=1}^{s} k_i = n$, and each k_i is no greater than the index of λ_i. Let m_a and m_g respectively denote the algebraic multiplicity and the geometric multiplicity of an eigenvalue λ of A. Then λ appears in m_g blocks and the sum of the orders of these blocks is m_a.

Proof. This is simply a restatement of Theorem B.147 using Definition B.121. □

8.2.2 Matrix powers

Theorem 8.20. Let A be a square matrix of finite size. The sequence $(A^n)_{n\in\mathbb{N}}$ converges to zero (in the sense of Definition E.22) if and only if the spectral radius $\rho(A) < 1$.

Proof. For necessity, let $\mathbf{x}$ be a unit eigenvector associated with an eigenvalue λ of maximum modulus. Then we have

$$A^n\mathbf{x} = \lambda^n\mathbf{x} \;\Rightarrow\; \lim_{n\to\infty} |\lambda^n| = \lim_{n\to\infty} \|A^n\mathbf{x}\|_2 = 0$$

from $\lim_{n\to\infty} A^n = \mathbf{0}$. This implies $\rho(A) = |\lambda| < 1$.

For sufficiency, we start with the Jordan canonical form (8.18). To prove $\lim_{n\to\infty} A^n = \mathbf{0}$, it suffices to show that $\lim_{n\to\infty} J^n = \mathbf{0}$. Since J^n has the same block form as that of J, it suffices to show that each Jordan block $J_i = \lambda_i I + S_i$ satisfies $\lim_{n\to\infty} J_i^n = \mathbf{0}$. Denote by ℓ_i the index of the nilpotent matrix S_i, i.e., $\forall m \geq \ell_i$, $S_i^m = \mathbf{0}$. Therefore,

$$J_i^m = \sum_{j=0}^{\ell_i - 1} \frac{m!}{j!(m-j)!} \lambda_i^{m-j} S_i^j.$$

The triangle inequality and the condition $\rho(A) < 1$ yield

$$\begin{aligned} \lim_{n\to\infty} \|J_i^n\| &= \lim_{n\to\infty} \left\| \textstyle\sum_{j=0}^{\ell_i-1} \frac{n!}{j!(n-j)!} \lambda_i^{n-j} S_i^j \right\| \\ &\leq \lim_{n\to\infty} \textstyle\sum_{j=0}^{\ell_i-1} \frac{n!}{j!(n-j)!} |\lambda_i|^{n-j} \left\| S_i^j \right\| = 0, \end{aligned}$$

where the last equality follows from $|\lambda_i| < 1$ and the fact that there are only a finite number of terms in the summation. Then Definition E.22 yields $\lim_{n\to\infty} J_i^n = \mathbf{0}$. □

Lemma 8.21. Let A be a square matrix of finite size. If $\rho(A) < 1$, then $I - A$ is nonsingular.

Exercise 8.22. Prove Lemma 8.21.

Theorem 8.23. Let A be a square matrix of finite size. The series $\sum_{k=0}^{+\infty} A^k$ converges if and only if $\rho(A) < 1$. Furthermore,

$$\sum_{k=0}^{+\infty} A^k = (I-A)^{-1}. \tag{8.19}$$

Proof. In finite-dimensional normed spaces, a Cauchy sequence is equivalent to a convergent sequence. If $\sum_{k=0}^{+\infty} A^k$ converges, the sequence $\left(\sum_{k=0}^{n} A^k\right)_{n\in\mathbb{N}^+}$ is Cauchy and $\lim_{n\to\infty}\|A^n\| = 0$. Then Theorem 8.20 yields $\rho(A) < 1$.

Conversely, $\rho(A) < 1$ and Lemma 8.21 imply that $I - A$ is invertible. The equality

$$I - A^{k+1} = (I - A)(I + A + \ldots + A^k)$$

implies

$$(I - A)^{-1} \lim_{n\to\infty}(I - A^{n+1}) = \lim_{n\to\infty}\sum_{k=0}^{n} A^k.$$

Then Theorem 8.20 yields (8.19). □

8.2.3 The spectral radius

Theorem 8.24 (Gelfand's formula)**.** For any matrix norm, we have

$$\lim_{n\to\infty}\|A^n\|^{\frac{1}{n}} = \rho(A). \tag{8.20}$$

Proof. Consider an arbitrary $\epsilon > 0$. For $A_b = \frac{1}{\rho(A)+\epsilon}A$, Theorem 8.20 yields

$$\begin{aligned}&\rho(A_b) < 1 \;\Rightarrow\; \lim_{n\to\infty} A_b^n = \mathbf{0}\\ &\Rightarrow\; \exists N_b \in \mathbb{N} \text{ s.t. } \forall m > N_b,\ \|A_b^m\| < 1\\ &\Rightarrow\; \exists N_b \in \mathbb{N} \text{ s.t. } \forall m > N_b,\ \|A^m\|^{\frac{1}{m}} < \rho(A) + \epsilon.\end{aligned}$$

Similarly, for $A_u = \frac{1}{\rho(A)-\epsilon}A$, we have

$$\begin{aligned}&\rho(A_u) > 1 \;\Rightarrow\; \exists N_u \in \mathbb{N} \text{ s.t. } \forall m > N_u,\ \|A_u^m\| > 1\\ &\Rightarrow\; \exists N_u \in \mathbb{N} \text{ s.t. } \forall m > N_u,\ \|A^m\|^{\frac{1}{m}} > \rho(A) - \epsilon.\end{aligned}$$

In summary, set $N = \max(N_b, N_u)$ and we have

$$\forall\epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall m > N,\ \left|\|A^m\|^{\frac{1}{m}} - \rho(A)\right| < \epsilon,$$

which, by Definition C.4, is equivalent to (8.20). □

Lemma 8.25. For any square matrix A and any matrix norm $\|\cdot\|$ induced from a vector norm, we have

$$\rho(A) \le \|A\|. \tag{8.21}$$

Proof. For each eigenvalue λ there exists an eigenvector $\mathbf{x}_0$ that is normalized with respect to the given vector norm. By Definition 7.23, we have

$$\|A\| = \sup_{\|\mathbf{x}\|=1}\|A\mathbf{x}\| \ge \|A\mathbf{x}_0\| = |\lambda|\|\mathbf{x}_0\| = |\lambda|. \qquad \square$$

Theorem 8.26. The spectral radius of a square matrix A is the infimum of all vector-induced matrix norms, i.e., for any $\epsilon > 0$, there exists some vector-induced matrix norm $\|\cdot\|$ satisfying $\|A\| \in [\rho(A), \rho(A) + \epsilon]$.

Proof. See [Isaacson and Keller, 1966, page 12]. □

8.2.4 General criteria for convergence

Theorem 8.27. The fixed point iteration in (8.2) converges if and only if the spectral radius $\rho(T) < 1$.

Proof. The exact solution $\mathbf{x}$ satisfies

$$\mathbf{x} = T\mathbf{x} + \mathbf{c}.$$

Subtract the above equation from (8.2) and we have

$$\begin{aligned}\mathbf{e}^{(k)} &= \mathbf{x} - \mathbf{x}^{(k)} = T(\mathbf{x} - \mathbf{x}^{(k-1)}) = \cdots = T^k(\mathbf{x} - \mathbf{x}^{(0)})\\ &= T^k\mathbf{e}^{(0)}.\end{aligned}$$

If $\rho(T) < 1$, Theorem 8.20 implies $\lim_{k\to\infty} T^k = \mathbf{0}$ and thus $\lim_{k\to\infty}\mathbf{e}^{(k)} = \mathbf{0}$. Conversely, if $\lim_{k\to\infty}\mathbf{e}^{(k)} = \mathbf{0}$, we have $\lim_{k\to\infty}\|T^k\| = 0$ and Definition E.22 implies $\lim_{k\to\infty} T^k = \mathbf{0}$. Then Theorem 8.20 yields $\rho(T) < 1$. □

Corollary 8.28. For a square matrix T with $\|T\| < 1$ for some matrix norm induced from a vector norm, the fixed point iteration in (8.2) converges for any initial guess. Furthermore, the error norm of the kth iterate satisfies

$$\|\mathbf{e}^{(k)}\| \le \frac{\|T\|^k}{1 - \|T\|}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|; \tag{8.22}$$

$$\|\mathbf{e}^{(k)}\| \le \frac{\|T\|}{1 - \|T\|}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|, \tag{8.23}$$

where $\mathbf{e}^{(k)} := \mathbf{x} - \mathbf{x}^{(k)}$.

Proof. The first statement follows directly from Theorem 8.27 and Lemma 8.25. (8.23) follows from

$$\begin{aligned}\|\mathbf{x}^{(k)} - \mathbf{x}\| &= \|T(\mathbf{x}^{(k-1)} - \mathbf{x})\| \le \|T\|\|\mathbf{x}^{(k-1)} - \mathbf{x}\|\\ &\le \|T\|\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\| + \|T\|\|\mathbf{x}^{(k)} - \mathbf{x}\|\end{aligned}$$

while (8.22) from

$$\begin{aligned}&\|\mathbf{x}^{(0)} - \mathbf{x}\| \le \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| + \|T\|\|\mathbf{x}^{(0)} - \mathbf{x}\|\\ &\Rightarrow\; \|\mathbf{x}^{(0)} - \mathbf{x}\| \le \tfrac{1}{1-\|T\|}\|\mathbf{x}^{(0)} - \mathbf{x}^{(1)}\|;\\ &\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \|T\|^k\|\mathbf{x}^{(0)} - \mathbf{x}\| \le \tfrac{\|T\|^k}{1-\|T\|}\|\mathbf{x}^{(0)} - \mathbf{x}^{(1)}\|.\end{aligned}$$ □

Corollary 8.29. Convergence of the SOR iteration implies $\omega \in (0, 2)$.

Proof. Lemma 8.11 states that

$$T_{SOR} = (I - \omega D^{-1}L)^{-1}[(1 - \omega)I + \omega D^{-1}U].$$

Since $\det(I - \omega D^{-1}L) = 1$, we have

$$|\det(T_{SOR})| = |1 - \omega|^n = |\lambda_1\lambda_2\cdots\lambda_n|.$$

The convergence and Theorem 8.27 imply $\rho(T_{SOR}) < 1$. Then the above equation yields $\omega \in (0, 2)$. □

8.2.5 Convergence rates

Definition 8.30. The *averaged convergence factor* of a fixed-point iteration in Definition 8.1 during k iterations is the averaged reduction ratio of error norms,

$$\psi_k(T) := \left(\sup_{\mathbf{e}^{(0)} \in \mathbb{R}^n \setminus \{\mathbf{0}\}} \frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|} \right)^{\frac{1}{k}} \tag{8.24}$$

while the *(general) convergence factor* of the fixed-point iteration is the reduction ratio of error norms per iteration in the asymptotic range,

$$\phi(T) := \lim_{k \to +\infty} \psi_k(T). \tag{8.25}$$

The *convergence rate* of the fixed-point iteration is

$$\tau(T) := -\ln \phi(T). \tag{8.26}$$

Theorem 8.31. The general convergence factor is the spectral radius of the iteration matrix, i.e.,

$$\phi(T) = \rho(T). \tag{8.27}$$

Proof. Definition 8.30 yields

$$\begin{aligned} \phi(T) &= \lim_{k \to +\infty} \left(\max_{\mathbf{e}^{(0)} \in \mathbb{R}^n \setminus \{\mathbf{0}\}} \frac{\|T^k \mathbf{e}^{(0)}\|}{\|\mathbf{e}^{(0)}\|} \right)^{\frac{1}{k}} \\ &= \lim_{k \to +\infty} \|T^k\|^{\frac{1}{k}} = \rho(T), \end{aligned}$$

where the second step follows from Definition 7.23 and the last one from the Gelfand formula (Theorem 8.24). □

8.3 Specific convergence analysis

8.3.1 Reducible matrices

Definition 8.32. A *permutation matrix* is a matrix whose columns are a permutation of the columns of the identity matrix.

Lemma 8.33. A permutation matrix P satisfies $P^T P = I$.

Definition 8.34. A matrix $A \in \mathbb{R}^{n \times n}$ is *reducible* if there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_{11} & \mathbf{0} \\ A_{21} & A_{22} \end{bmatrix}, \tag{8.28}$$

where $A_{11} \in \mathbb{R}^{r \times r}$, $A_{21} \in \mathbb{R}^{(n-r) \times r}$, and $A_{22} \in \mathbb{R}^{(n-r) \times (n-r)}$ are block matrices. An *irreducible* matrix is a square matrix that is not reducible.

Definition 8.35. A *partition of a set* S is a set of pairwise disjoint non-empty subsets $S_1, S_2, \ldots, S_p$ whose union equals S.

Definition 8.36. A *decomposition of a set* S is a set of non-empty subsets $S_1, S_2, \ldots, S_p$ whose union equals S.

Lemma 8.37. A matrix $A \in \mathbb{C}^{n \times n}$ is reducible if and only if there exists a partition $\{\mathcal{I}, \mathcal{J}\}$ of $\mathcal{W} = \{1, 2, \ldots, n\}$ such that

$$\forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \quad a_{i,j} = 0. \tag{8.29}$$

Exercise 8.38. Prove Lemma 8.37.

8.3.2 Diagonally dominant matrices

Definition 8.39. A matrix $A \in \mathbb{C}^{n \times n}$ is *(weakly) diagonally dominant* iff

$$\forall i = 1, 2, \ldots, n, \quad |a_{ii}| \geq \sum_{j \neq i, j=1}^{n} |a_{ij}|, \tag{8.30}$$

strictly diagonally dominant iff

$$\forall i = 1, 2, \ldots, n, \quad |a_{ii}| > \sum_{j \neq i, j=1}^{n} |a_{ij}|, \tag{8.31}$$

and *irreducibly diagonally dominant* iff it is irreducible and diagonally dominant with at least one strict inequality in (8.30).

Theorem 8.40 (Gershgorin)**.** Any eigenvalue λ of a square matrix A is located in one of the closed discs of the complex plane centered at a_{ii},

$$\forall \lambda \in \sigma(A), \exists i \in [1, n] \text{ s.t. } |\lambda - a_{ii}| \leq \sum_{j \neq i, j=1}^{n} |a_{ij}|. \tag{8.32}$$

Proof. Choose the eigenvector associated with λ to be $\mathbf{u}$ such that $|u_i| = 1$ and $|u_j| \leq 1$ for each $j \neq i$. Then we have

$$\begin{aligned} & A\mathbf{u} = \lambda \mathbf{u} \ \Rightarrow \ \textstyle\sum_{j=1}^{n} a_{ij} u_j = \lambda u_i \\ \Rightarrow \ & (\lambda - a_{ii}) u_i = \textstyle\sum_{j \neq i, j=1}^{n} a_{ij} u_j \\ \Rightarrow \ & |\lambda - a_{ii}| \leq \textstyle\sum_{j \neq i, j=1}^{n} |a_{ij}||u_j| \leq \sum_{j \neq i, j=1}^{n} |a_{ij}|. \end{aligned}$$ □

Theorem 8.41. A matrix A is nonsingular if it is strictly diagonally dominant or irreducibly diagonally dominant.

Proof. For a strictly diagonally dominant matrix A, the Gershgorin theorem 8.40 implies that 0 is not an eigenvalue of A and thus A is nonsingular. Suppose an irreducibly diagonally dominant matrix A is singular, i.e.,

$$\exists \mathbf{u} \in \mathbb{R}^n \text{ s.t. } \|\mathbf{u}\|_\infty = 1, \quad A\mathbf{u} = \mathbf{0}.$$

Then there exists a partition $\{\mathcal{I}, \mathcal{J}\}$ of $\mathcal{W} = \{1, 2, \ldots, n\}$ such that

$$\mathcal{I} = \{i \in \mathcal{W} : |u_i| = 1\}, \ \mathcal{J} = \{j \in \mathcal{W} : |u_j| < 1\}.$$

In addition, the irreducibly diagonal dominance of A implies $\mathcal{J} \neq \emptyset$. Then for $i \in \mathcal{I}$, the ith equation of $A\mathbf{u} = \mathbf{0}$ yields

$$\begin{aligned} & a_{ii} u_i = -\textstyle\sum_{j \in \mathcal{I}; j \neq i} a_{ij} u_j - \sum_{j \in \mathcal{J}} a_{ij} u_j \\ \Rightarrow \ & |a_{ii}| \leq \textstyle\sum_{j \in \mathcal{I}; j \neq i} |a_{ij}| + \sum_{j \in \mathcal{J}} |a_{ij}||u_j| < \sum_{j \neq i} |a_{ij}|, \end{aligned}$$

which contradicts (8.30). □

Theorem 8.42. For a linear system $A\mathbf{x} = \mathbf{b}$ with A being strictly diagonally dominant or irreducibly diagonally dominant, both Jacobi and Gauss-Seidel iterations converge.

Proof. By the diagonal dominance, the diagonal matrix D_A of A is nonsingular. Then we have

$$\forall |\lambda| \geq 1, \ \det(\lambda I - T_J) = \det(D^{-1}) \det(\lambda D - L - U) \neq 0,$$

because $|\lambda| \geq 1$, the diagonal dominance of $A = D - L - U$, and Theorem 8.41 imply $\det(\lambda D - L - U) \neq 0$. Similarly

$$\forall |\lambda| \geq 1,\ \det(\lambda I - T_{GS}) = \det((D-L)^{-1}) \det(\lambda D - \lambda L - U) \neq 0.$$

Therefore, we have $\rho(T_J) < 1$ and $\rho(T_{GS}) < 1$. □

Theorem 8.43. For a linear system $A\mathbf{x} = \mathbf{b}$ with A being strictly diagonally dominant or irreducibly diagonally dominant, the SOR iteration converges for any $\omega \in (0, 1)$.

Exercise 8.44. Prove Theorem 8.43.

8.3.3 Normal matrices

Notation 5. In contrast to Definition B.187, in this chapter we denote by A^H the adjoint or conjugate transpose of a matrix $A \in \mathbb{C}^{n\times m}$. Unless explicitly stated otherwise, the scalar field of vector spaces in this chapter defaults to $\mathbb{C}$.

Definition 8.45. A *normal matrix* A is a square matrix that commutes with its conjugate transpose,

$$A^H A = A A^H. \tag{8.33}$$

Lemma 8.46. If a normal matrix is triangular, then it is a diagonal matrix.

Exercise 8.47. Prove Lemma 8.46.

Theorem 8.48. A matrix is normal if and only if it is unitarily similar to a diagonal matrix.

Proof. It is straightforward to verify the sufficiency from Definition 8.45. As for the necessity, we express the matrix in the Schur form $A = QRQ^H$. Definition 8.45 yields

$$QR^H RQ^H = QRR^H Q^H,$$

which implies $R^H R = RR^H$, i.e. the triangular matrix R is normal. The proof is then completed by Lemma 8.46. □

Corollary 8.49. Any two eigenvectors of a normal matrix associated with two distinct eigenvalues are orthogonal.

Proof. This follows directly from Theorem 8.48. □

Theorem 8.50. A matrix is normal if and only if each of its eigenvector is also an eigenvector of A^H.

Proof. This is a restatement of Lemma B.202. □

Definition 8.51. Let V be a finite-dimensional vector space with $\mathbb{F}$ as its scalar field. The *Rayleigh quotient of a linear operator* $T \in \mathcal{L}(V)$ is a functional $\mu_T : V \setminus \{\mathbf{0}\} \to \mathbb{F}$ given by

$$\mu_T(\mathbf{x}) = \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}. \tag{8.34}$$

The *numerical range* or *field of values of a linear operator* T is the range of its Rayleigh quotients.

Example 8.52. $\mu_A(\mathbf{x}) = \lambda$ for each eigenpair $(\lambda, \mathbf{x})$ of A.

Theorem 8.53. The field of values of a normal operator A is the convex hull of its spectrum.

Proof. By Theorem 8.48, A is unitarily similar to a diagonal matrix Λ, i.e., $A = Q\Lambda Q^H$ with $Q = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n]$ satisfying $Q^H Q = I$. Then each $(\lambda_i, \mathbf{q}_i)$ is an eigenpair of A. Consequently, any vector $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ can be expressed as $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{q}_i$, for which we have

$$\mu_A(\mathbf{x}) = \frac{\langle \sum_{i=1}^n x_i \lambda_i \mathbf{q}_i,\ \mathbf{x}\rangle}{\langle \mathbf{x},\ \mathbf{x}\rangle} = \frac{\sum_{i=1}^n |x_i|^2 \lambda_i}{\sum_{i=1}^n |x_i|^2} = \sum_{i=1}^n \beta_i \lambda_i,$$

where $\beta_i = \frac{|x_i|^2}{\sum_{j=1}^n |x_j|^2}$. The proof is completed by the facts that $\sum_{i=1}^n \beta_i = 1$ and $\beta_i \geq 0$ for each i. □

Theorem 8.54 (Hausdorff)**.** The field of values of an arbitrary matrix is a convex set that contains the convex hull of its spectrum.

Definition 8.55. The *numerical radius of a square matrix* $A \in \mathbb{C}^{n\times n}$ is the radius of the smallest disk that contains its field of values, i.e.,

$$\nu(A) := \sup_{\mathbf{x}\in\mathbb{C}^n\setminus\{\mathbf{0}\}} |\mu_A(\mathbf{x})|. \tag{8.35}$$

Lemma 8.56. Any square matrix $A \in \mathbb{C}^{n\times n}$ satisfies

$$\rho(A) \leq \nu(A) \leq \|A\|_2, \tag{8.36}$$

where the equalities hold if A is normal.

Exercise 8.57. Prove Lemma 8.56.

Exercise 8.58. Is ν in (8.35) a norm on $\mathbb{C}^{n\times n}$? Prove your conclusion.

8.3.4 Hermitian matrices

Definition 8.59. A matrix $A \in \mathbb{C}^{n\times n}$ is *Hermitian* if it equals its adjoint A^H, i.e. $A = A^H$; A is *skew Hermitian* or *anti-Hermitian* if $A = -A^H$.

Example 8.60. Both Hermitian and skew Hermitian matrices are normal matrices.

Lemma 8.61. Any skew Hermitian matrix S satisfies

$$\forall \mathbf{x} \in \mathbb{C}^n, \quad \text{Re}\,\langle S\mathbf{x}, \mathbf{x}\rangle = 0. \tag{8.37}$$

Exercise 8.62. Prove Lemma 8.61.

Corollary 8.63. Any eigenvalue of a skew Hermitian matrix is either 0 or purely imaginary.

Proof. This follows directly from Lemma 8.61. □

Lemma 8.64. All eigenvalues of a Hermitian matrix A are real, i.e. $\sigma(A) \subset \mathbb{R}$.

Proof. This is a restatement of Lemma B.196. □

Corollary 8.65. A normal matrix with real eigenvalues is Hermitian.

Proof. By Theorem 8.48, we can write $A = QDQ^H$ where the diagonal matrix D has all eigenvalues of A. If D is real, then $D^H = D$ and thus $A^H = A$. □

Corollary 8.66. Any Hermitian matrix is unitarily similar to a real diagonal matrix.

Proof. This follows directly from Theorem 8.48, Example 8.60, and Lemma 8.64. □

Lemma 8.67. The field of values of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is a compact interval in $\mathbb{R}$, where the Rayleigh quotient function μ_A attains its maximum and minimum.

Proof. By Example 8.60, A is a normal operator on $V = \mathbb{C}^n$. Then Theorem 8.53 and Lemma 8.64 imply that the field of values of A is the convex hull of n points in $\mathbb{R}$. It follows from the extreme value theorem D.222 that both the maximum and the minimum of μ_A are attained on $\mu_A(V)$. □

Theorem 8.68 (Rayleigh)**.** The extremum eigenvalues of a Hermitian matrix A can be characterized as

$$\begin{aligned} \lambda_{\max}(A) &= \max_{\mathbf{x}\neq\mathbf{0}} \mu_A(\mathbf{x}), \\ \lambda_{\min}(A) &= \min_{\mathbf{x}\neq\mathbf{0}} \mu_A(\mathbf{x}). \end{aligned} \tag{8.38}$$

Proof. These identities follow directly from Theorem 8.53 and Lemma 8.67. □

Lemma 8.69. Suppose S_1 and S_2 are two subspaces of a vector space V such that $m := \dim S_1 + \dim S_2 - n \geq 1$ where $n = \dim V \in \mathbb{N}^+$. Then the dimension of the subspace $S_1 \cap S_2$ is at least m.

Proof. The subspace intersection theorem states

$$\dim(S_1 \cap S_2) + \dim(S_1 + S_2) = \dim S_1 + \dim S_2,$$

which yields

$$\begin{aligned} \dim(S_1 \cap S_2) = &\quad \dim S_1 + \dim S_2 - \dim(S_1 + S_2) \\ \geq &\quad \dim S_1 + \dim S_2 - n = m. \end{aligned}$$ □

Theorem 8.70 (Courant-Fisher min-max principle)**.** The spectrum of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is characterized by

$$\begin{aligned} \lambda_k = &\min_{\dim(S)=n-k+1} \max_{\mathbf{x}\in S\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}), \\ \lambda_k = &\max_{\dim(S)=k} \min_{\mathbf{x}\in S\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}), \end{aligned} \tag{8.39}$$

where S is a subspace of $\mathbb{C}^n$ and λ_k is the kth eigenvalue of A in non-increasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

Proof. By Theorem 8.48, we have $A = U\Lambda U^H$ where the columns $\mathbf{u}_i$'s of the unitary matrix U are orthonormal eigenvectors and the diagonal entries λ_i's of Λ are the corresponding eigenvalues.

Define $R := \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k)$ and let S be any subspace of $\mathbb{C}^n$ with $\dim S = n - k + 1$. By Lemma 8.69, the dimension of $R \cap S$ is at least 1. Then we have

$$\begin{aligned} \max_{\mathbf{x}\in S\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}) &\geq \max_{\mathbf{x}\in S\cap R\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}) \geq \min_{\mathbf{x}\in S\cap R\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}) \\ &\geq \min_{\mathbf{x}\in R\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}) = \lambda_k, \end{aligned}$$

where the existence of global extrema follows from Lemma 8.67 and the last step follows from Theorem 8.68. The minimum of the above inequality yields

$$\lambda_k \leq \min_{\dim(S)=n-k+1} \max_{\mathbf{x}\in S\setminus\{\mathbf{0}\}} \mu_A(\mathbf{x}),$$

where the "$\leq$" reduces to "$=$" because we can choose

$$S = \operatorname{span}(\mathbf{u}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_n).$$ □

Exercise 8.71. Prove the second equality in (8.39).

8.3.5 Positive definite matrices

Lemma 8.72. The matrix of a positive definite operator $T \in \mathcal{L}(V)$ (as in Definition B.210) has each of its diagonal entries positive.

Proof. This follows from setting $\mathbf{v}$ in Definition B.210 to be the ith column of the identity matrix. □

Definition 8.73. A *real positive definite matrix* $A \in \mathbb{R}^{n\times n}$ and a *real positive semi-definite matrix* $B \in \mathbb{R}^{n\times n}$ are matrices satisfying respectively

$$\forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\},\ \langle A\mathbf{x}, \mathbf{x}\rangle > 0, \tag{8.40}$$

$$\forall \mathbf{x} \in \mathbb{R}^n,\ \langle B\mathbf{x}, \mathbf{x}\rangle \geq 0. \tag{8.41}$$

Definition 8.74. A *symmetric positive definite* (SPD) matrix is a real matrix $A \in \mathbb{R}^{n\times n}$ that is both symmetric and positive definite.

Example 8.75. A real positive definite matrix does not have to be symmetric; e.g., the matrix

$$R = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \tag{8.42}$$

is real positive definite because $[a, b]R[a, b]^T = a^2 + b^2 > 0$, but R is not symmetric. Conversely, a symmetric real matrix needs not to be positive definite. Hence, symmetry and positive-definiteness are orthogonal issues for real matrices, which justifies the coexistence of Definitions 8.73 and 8.74.

Lemma 8.76. Suppose a matrix $A \in \mathbb{C}^{n\times n}$ satisfies

$$\forall \mathbf{x} \in \mathbb{C}^n, \quad \operatorname{Im}\langle A\mathbf{x}, \mathbf{x}\rangle = 0. \tag{8.43}$$

Then A is Hermitian.

Exercise 8.77. Prove Lemma 8.76.

Definition 8.78. A matrix $A \in \mathbb{C}^{n\times n}$ is *Hermitian positive definite* (HPD) iff

$$\forall \mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\},\ \langle A\mathbf{x}, \mathbf{x}\rangle > 0. \tag{8.44}$$

Lemma 8.79. Any eigenvalue of an HPD matrix is positive.

Proof. By Theorem 8.48, A is unitarily similar to a diagonal matrix Λ, i.e., $A = Q\Lambda Q^H$ with $Q = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n]$ satisfying $Q^H Q = I$. Then each $(\lambda_i, \mathbf{q}_i)$ is an eigenpair of A. Consequently, any vector $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ can be expressed as $\mathbf{x} = \sum_{i=1}^n x_i\mathbf{q}_i$, for which we have

$$0 < \langle A\mathbf{x}, \mathbf{x}\rangle = \left\langle \sum_{i=1}^n x_i\lambda_i\mathbf{q}_i,\ \mathbf{x} \right\rangle = \sum_{i=1}^n |x_i|^2\lambda_i.$$

The proof is completed by choosing $\mathbf{x}$ to be the standard basis vectors. □

Lemma 8.80. The real part of any eigenvalue of a real positive definite matrix A must be positive.

Proof. Let $(x+\mathbf{i}y, \mathbf{u}+\mathbf{i}\mathbf{v})$ be an eigenpair of A. Then

$$\begin{aligned} & A(\mathbf{u}+\mathbf{i}\mathbf{v}) = (x+\mathbf{i}y)(\mathbf{u}+\mathbf{i}\mathbf{v}) \\ \Rightarrow & \begin{cases} A\mathbf{u} = x\mathbf{u} - y\mathbf{v} \\ A\mathbf{v} = x\mathbf{v} + y\mathbf{u} \end{cases} \\ \Rightarrow & \begin{cases} \mathbf{u}^T A\mathbf{u} = x\|\mathbf{u}\|_2^2 - y\mathbf{u}^T\mathbf{v} > 0 \\ \mathbf{v}^T A\mathbf{v} = x\|\mathbf{v}\|_2^2 + y\mathbf{v}^T\mathbf{u} > 0 \end{cases} \\ \Rightarrow & \ x > 0, \end{aligned}$$

where the last step follows from $\mathbf{v}^T\mathbf{u} = \mathbf{u}^T\mathbf{v}$. □

Example 8.81. The converse of Lemma 8.80 is false: the eigenvalues of

$$B = \begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix} \tag{8.45}$$

is 1 and 2, but B is not positive definite. In contrast, the converse of Lemma 8.79 is true: a Hermitian matrix with all eigenvalues positive is positive definite.

Definition 8.82. The *Hermitian part* and the *skew Hermitian part* of a square matrix $A \in \mathbb{C}^{n\times n}$ are respectively the matrices H and $\mathbf{i}S$, where

$$\begin{aligned} H &= \tfrac{1}{2}(A + A^H), \\ S &= \tfrac{1}{2\mathbf{i}}(A - A^H), \end{aligned} \tag{8.46}$$

such that $A = H + \mathbf{i}S$.

Lemma 8.83. The Hermitian part H of a matrix $A \in \mathbb{R}^{n\times n}$ satisfies

$$\forall \mathbf{u} \in \mathbb{R}^n,\ \langle H\mathbf{u}, \mathbf{u}\rangle = \langle A\mathbf{u}, \mathbf{u}\rangle . \tag{8.47}$$

Proof. It is readily verified that both H and S are Hermitian. Then the conclusion follows from the decomposition $A = H + \mathbf{i}S$ and Lemma 8.67. □

Theorem 8.84. For any real positive definite matrix A, there exists $\alpha \in \mathbb{R}^+$ such that

$$\forall \mathbf{u} \in \mathbb{R}^n, \quad \langle A\mathbf{u}, \mathbf{u}\rangle \geq \alpha\|\mathbf{u}\|_2^2. \tag{8.48}$$

Proof. By Lemma 8.83, we have $\langle A\mathbf{u}, \mathbf{u}\rangle = \langle H\mathbf{u}, \mathbf{u}\rangle$. Then the proof is completed by Lemma 8.79 and the Rayleigh theorem 8.68 with $\alpha = \lambda_{\min}(H)$. □

Theorem 8.85. Any eigenvalue λ_j of a square matrix A can be bounded as

$$\begin{aligned} \lambda_{\min}(H) \leq \mathrm{Re}\,(\lambda_j) \leq \lambda_{\max}(H), \\ \lambda_{\min}(S) \leq \mathrm{Im}\,(\lambda_j) \leq \lambda_{\max}(S), \end{aligned} \tag{8.49}$$

where H, S are defined in (8.46).

Proof. Any eigenpair $(\lambda_j, \mathbf{u}_j)$ of A with $\|\mathbf{u}_j\| = 1$ satisfies

$$\lambda_j = \langle A\mathbf{u}_j, \mathbf{u}_j\rangle = \langle H\mathbf{u}_j, \mathbf{u}_j\rangle + i\,\langle S\mathbf{u}_j, \mathbf{u}_j\rangle .$$

The rest of the proof follows from Lemma 8.67 and the Rayleigh theorem 8.68. □

Example 8.86. The estimation of the range of eigenvalues of a matrix A from its field of values via Theorem 8.85 may be inaccurate. For example, eigenvalues of the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 10^4 & 1 \end{bmatrix} \tag{8.50}$$

are -99 and 101, but those of H are $1 \pm \frac{1}{2}(10^4+1)$ and those of $\mathbf{i}S$ are $\pm\frac{i}{2}(10^4-1)$.

Lemma 8.87. The numerical radius and the 2-norm of a square matrix A satisfy

$$\frac{1}{2}\|A\|_2 \leq \nu(A) \leq \|A\|_2. \tag{8.51}$$

Exercise 8.88. Prove Lemma 8.87.

Lemma 8.89. If a symmetric matrix $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant or irreducibly diagonally dominant and satisfies $a_{ii} > 0$ for each $i = 1, 2, \ldots, n$, then A is SPD.

Proof. By Theorem 8.41, $A - \lambda I$ is nonsingular for any $\lambda \leq 0$. Hence the eigenvalue of A can not be any value in $(-\infty, 0]$. The proof is completed by the symmetry of A. □

Theorem 8.90. For a linear system $A\mathbf{x} = \mathbf{b}$ with $A \in \mathbb{R}^{n\times n}$ being SPD, the SOR iteration converges for any $\omega \in (0, 2)$.

Proof. For any eigenpair $(\lambda, \mathbf{u})$ of T_{SOR} in (8.12), we have

$$[(1-\omega)D + \omega L^T]\mathbf{u} = \lambda(D - \omega L)\mathbf{u},$$

where we have applied the symmetry of A. Then we have

$$\begin{aligned} & \delta := \mathbf{u}^H D\mathbf{u},\ \alpha + \mathbf{i}\beta := \mathbf{u}^H L\mathbf{u} \\ \Rightarrow & \ (1-\omega)\delta + \omega(\alpha - \mathbf{i}\beta) = \lambda[\delta - \omega(\alpha + \mathbf{i}\beta)] \\ \Rightarrow & \ |\lambda|^2 = \tfrac{[(1-\omega)\delta + \omega\alpha]^2 + \omega^2\beta^2}{(\delta - \omega\alpha)^2 + \omega^2\beta^2}. \end{aligned}$$

The positive definiteness of A, the splitting $A = D - L - U$, and the symmetry $L^T = U$ imply $\delta - 2\alpha > 0$ while Lemma 8.72 yields $\delta > 0$. Then $\omega \in (0, 2)$ implies

$$\begin{aligned} & [(1-\omega)\delta + \omega\alpha]^2 + \omega^2\beta^2 - (\delta - \omega\alpha)^2 - \omega^2\beta^2 \\ = & \ \omega\delta(\delta - 2\alpha)(\omega - 2) < 0, \end{aligned}$$

which yields $|\lambda| < 1$. □

8.3.6 Nonnegative matrices

Definition 8.91. A *nonnegative matrix* is a matrix whose entries are nonnegative. A *positive matrix* is a matrix whose entries are positive.

Notation 6. For two matrices $A, B \in \mathbb{R}^{m\times n}$ with $m, n \in \mathbb{N}^+$, we write $A \leq B$ iff $a_{i,j} \leq b_{i,j}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. In particular, $A \geq \mathbf{0}$ and $A > \mathbf{0}$ mean that A is a nonnegative and positive matrix, respectively.

Lemma 8.92. The relation "$\leq$" for matrices is reflexive ($A \leq A$), antisymmetric ($(A \leq B \wedge B \leq A) \Rightarrow A = B$), and transitive ($(A \leq B \wedge B \leq C) \Rightarrow A \leq C$).

Proof. This follows directly from Notation 6. □

Lemma 8.93. Two nonnegative matrices A and B satisfy

- both AB and $A+B$ are nonnegative;
- A^n is nonnegative for $n \in \mathbb{N}^+$;
- $\mathbf{0} \le A \le B$ implies $\|A\|_1 \le \|B\|_1$ and $\|A\|_\infty \le \|B\|_\infty$.

Proof. This follows directly from Definition 8.91. □

Theorem 8.94 (Perron-Frobenius)**.** The spectral radius of a nonnegative irreducible matrix $A \in \mathbb{R}^{n\times n}$ is a simple eigenvalue of A, with which the associated eigenvector $\mathbf{u}$ can be chosen to be nonnegative.

Lemma 8.95. Nonnegative matrices A, B, C, D satisfy

$$A \le B \;\Rightarrow\; AC \le BC, \quad DA \le DB. \tag{8.52}$$

Lemma 8.96. For nonnegative matrices A and B, we have

$$A \le B \;\Rightarrow\; \forall n \in \mathbb{N}^+,\; A^n \le B^n. \tag{8.53}$$

Exercise 8.97. Prove Lemmas 8.95 and 8.96.

Theorem 8.98. For square matrices A and B, we have

$$\mathbf{0} \le A \le B \;\Rightarrow\; \rho(A) \le \rho(B). \tag{8.54}$$

Proof. This follows directly from Lemma 8.96 and Gelfand's formula (Theorem 8.24) by choosing the norm to be, say, 1-norm, and applying Notation 6. □

Theorem 8.99. A nonnegative square matrix A satisfies $\rho(A) < 1$ if and only if $I - A$ is invertible and $(I-A)^{-1} \ge \mathbf{0}$.

Proof. $\rho(A) < 1$ and Lemma 8.21 imply $I - A$ being invertible. Then Theorem 8.23 and $A \ge \mathbf{0}$ yield $(I-A)^{-1} \ge \mathbf{0}$.

As for the sufficiency, the Perron-Frobenius theorem 8.94 implies the existence of a nonnegative eigenvector $\mathbf{u}$ of A such that $A\mathbf{u} = \rho(A)\mathbf{u}$, i.e.,

$$(I-A)^{-1}\mathbf{u} = \frac{1}{1-\rho(A)}\mathbf{u}.$$

Then $(I-A)^{-1} \ge \mathbf{0}$ implies $1-\rho(A) > 0$, i.e. $\rho(A) < 1$. □

8.3.7 M-matrices and regular splittings

Definition 8.100. An *M-matrix* A is a square matrix of size n that satisfies

(MMT-1) $\forall i = 1, 2, \ldots, n,\; a_{ii} > 0$;

(MMT-2) $\forall i \ne j,\; i, j = 1, 2, \ldots, n,\; a_{ij} < 0$;

(MMT-3) $\det(A) \ne 0$;

(MMT-4) $A^{-1} \ge \mathbf{0}$.

Theorem 8.101. Suppose a square matrix A satisfies (MMT-1,2). Then A is an M-matrix if and only if

$$\rho(I - D_A^{-1}A) < 1, \tag{8.55}$$

where D_A is the diagonal of A.

Proof. The conclusion follows from applying Theorem 8.99 to $B := I - D_A^{-1}A$. □

Example 8.102. For the square matrix A in (7.13), elements of the matrix $B := I - D_A^{-1}A$ are given by

$$b_{i,j} = \begin{cases} \frac{1}{2} & \text{if } i-j = \pm 1; \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\|B\|_1 = 1$ and thus Lemma 8.25 yields $\rho(B) \le 1$.

Suppose $\rho(B) = 1$, then there exists an eigenvector $\mathbf{u}$ with $u_1 = 1$ and $B\mathbf{u} = \mathbf{u}$. The first $m-1$ equations imply $u_i = i$, which contradicts the last equation $u_m = \frac{1}{2}u_{m-1}$. Hence $\rho(B) \ne 1$ and we have $\rho(B) < 1$. By Theorem 8.101, the matrix A in (7.13) is an M-matrix.

Theorem 8.103. If a square matrix A satisfies (MMT-2,3,4), then it satisfies (MMT-1) and (8.55).

Proof. Write $B = A^{-1}$. From $AB = I$ we have

$$\sum_{k=1}^{n} a_{ik}b_{ki} = 1 \;\Rightarrow\; a_{ii}b_{ii} = 1 - \sum_{k\ne i;k=1}^{n} a_{ik}b_{ki}.$$

(MMT-2) states that $a_{ik} \le 0$ and (MMT-4) states that $b_{ki} \ge 0$. Hence the RHS of the last equation is no less than 1 and $b_{ii} \ge 0$ yields $a_{ii} > 0$. Then the proof is completed by Theorem 8.101. □

Theorem 8.104. Suppose a square matrix B satisfies (MMT-2) and another square matrix A satisfies $A \le B$. Then if A is an M-matrix, so is B.

Proof. By $A \le B$ and (MMT-1), we have $D_B \ge D_A \ge \mathbf{0}$ and $D_A - A \ge D_B - B \ge \mathbf{0}$. It follows that

$$I - D_A^{-1}A \ge D_A^{-1}(D_B - B) \ge D_B^{-1}(D_B - B) = I - D_B^{-1}B \ge \mathbf{0},$$

where the second inequality follows from Lemma 8.95 and the fact that both D_A and D_B are nonnegative diagonal matrices. Then Theorems 8.98 and 8.101 yield

$$\rho(I - D_B^{-1}B) \le \rho(I - D_A^{-1}A) < 1$$

and B being an M-matrix follows from Theorem 8.101. □

Definition 8.105. A *regular splitting* of a square matrix A is a pair of matrices M, N satisfying

(RSM-1) $A = M - N$,

(RSM-2) $\det(M) \ne 0$,

(RSM-3) $M^{-1} \ge \mathbf{0}$,

(RSM-4) $N \ge \mathbf{0}$.

Example 8.106. The fixed-point iteration in Definition 8.1 can be viewed as a method for solving the linear system $(I-T)\mathbf{x} = M^{-1}\mathbf{b}$, which is equivalent to

$$M^{-1}A\mathbf{x} = M^{-1}\mathbf{b}. \tag{8.56}$$

This system, having the same solution as the original system (8.1), is often called the *preconditioned system* and M is called the *preconditioning matrix* or the *preconditioner*. Therefore, a relaxation scheme is equivalent to a fixed-point

iteration on a preconditioned system. For the Jacobi, Gauss-Seidel, SOR, and SSOR iterations, the preconditioners are, respectively,

$$M_J = D, \tag{8.57}$$
$$M_{GS} = D - L, \tag{8.58}$$
$$M_{SOR} = \frac{1}{\omega}(D - \omega L), \tag{8.59}$$
$$M_{SSOR} = \frac{1}{\omega(2-\omega)}(D - \omega L)D^{-1}(D - \omega U). \tag{8.60}$$

Theorem 8.107. Let M, N be a regular splitting of A. Then $\rho(M^{-1}N) < 1$ if and only if A is invertible and $A^{-1} \geq \mathbf{0}$.

Proof. Write $G := M^{-1}N$ and we have $A = M(I - G)$. If $\rho(G) < 1$, Lemma 8.21 implies that $I - G$ is nonsingular and (RSM-2) further yields that A is invertible. By Theorem 8.99, we have $(I - G)^{-1} \geq \mathbf{0}$, and (RSM-3) further yields $A^{-1} = (I - G)^{-1}M^{-1} \geq \mathbf{0}$.

As for the sufficiency, $\det(A) \neq 0$ and $\det(M) \neq 0$ imply that $I - G = M^{-1}A$ is invertible. By (RSM-3,4), we have $G \geq \mathbf{0}$. Then the Perron-Frobenious theorem 8.94 guarantees the existence of an eigenvector $\mathbf{x} \geq \mathbf{0}$ associated with the eigenvalue $\rho(G)$ such that $G\mathbf{x} = \rho(G)\mathbf{x}$. Then

$$\begin{aligned} & A^{-1}N = (M(I - M^{-1}N))^{-1}N = (I - G)^{-1}G \\ \Rightarrow \quad & A^{-1}N\mathbf{x} = \tfrac{\rho(G)}{1-\rho(G)}\mathbf{x}. \end{aligned}$$

Furthermore, $A^{-1} \geq \mathbf{0}$ and (RSM-4) yield $\frac{\rho(G)}{1-\rho(G)} \geq 0$, which implies $\rho(G) \in [0, 1]$. $I - G$ being invertible implies $\rho(G) \neq 1$, which completes the proof. □

Example 8.108. For the Jacobi iteration in Definition 8.2, $M_J = D$ and $N = L + U$ form a regular splitting of the matrix A in (7.13). By Example 8.102, A is an M-matrix and thus (by Definition 8.100) is invertible and satisfies $A^{-1} \geq \mathbf{0}$. Then it follows from Lemma 8.3 and Theorem 8.107 that $\rho(T_J) < 1$ and therefore the Jacobi iteration converges.

Exercise 8.109. Show that for the Gauss-Seidel iteration in Lemma 8.6, $M_{GS} = D - L$ and $N = U$ form a regular splitting of the matrix A in (7.13). Therefore the Gauss-Seidel iteration converges by the same arguments in Example 8.108.

Definition 8.110. The *approximate inverse* of a square matrix A is a matrix B that satisfies

$$\|I - BA\| < 1. \tag{8.61}$$

Example 8.111. Referring to (8.57) and (8.58), both M_J^{-1} and M_{GS}^{-1} are approximate inverses of A, c.f. Example 8.108 and Exercise 8.109.

Definition 8.112. The *Richardson iteration* is a fixed point iteration in Definition 8.1 with $T = I - A$ and $\mathbf{c} = \mathbf{b}$.

Chapter 9

Multigrid Methods

9.1 The residual equation

Definition 9.1. In solving a linear system $A\mathbf{x} = \mathbf{b}$, the *error of an approximate solution* $\tilde{\mathbf{x}}$ is

$$\mathbf{e}(\tilde{\mathbf{x}}) := \mathbf{x} - \tilde{\mathbf{x}} \tag{9.1}$$

and the *residual* of $\tilde{\mathbf{x}}$ is

$$\mathbf{r}(\tilde{\mathbf{x}}) := \mathbf{b} - A\tilde{\mathbf{x}}. \tag{9.2}$$

Lemma 9.2. The error and the residual of an approximate solution $\tilde{\mathbf{x}}$ satisfy the *residual equation*

$$A\mathbf{e} = \mathbf{r}. \tag{9.3}$$

Proof. This follows from Definition 9.1 in the same way that Lemma 7.18 follows from Lemma 7.17. □

Definition 9.3. The *condition number of a matrix* A is

$$\mathrm{cond}(A) := \|A\|_2 \|A^{-1}\|_2. \tag{9.4}$$

Theorem 9.4. The relative error of an approximate solution is bounded by its relative residual.

$$\frac{1}{\mathrm{cond}(A)} \frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2} \le \frac{\|\mathbf{e}\|_2}{\|\mathbf{x}\|_2} \le \mathrm{cond}(A) \frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2}. \tag{9.5}$$

Exercise 9.5. Prove Theorem 9.4.

9.2 The model problem

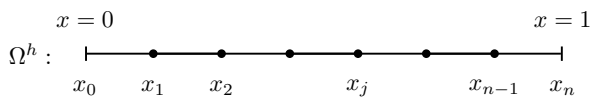
Definition 9.6. The *model problem* for our exposition of multigrid methods is the one-dimensional Poisson equation with homogeneous boundary condition

$$\begin{cases} -\Delta u = f & \text{in } \Omega := (0,1); \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{9.6}$$

Example 9.7. As a special case of Example 7.10 with $\alpha = 0$, $\beta = 0$, and $m = n - 1$, our discretization of (9.6) yields a linear system

$$A\mathbf{u} = \mathbf{f}, \tag{9.7}$$

where the unit interval Ω is discretized by uniform grid size $h = \frac{1}{n}$ into n cells with cell boundaries at $x_j = jh = \frac{j}{n}$ for $j = 0, 1, \ldots, n$, the knowns $f_j = f(x_j)$ and the unknowns u_j are located at the internal nodes x_j with $j = 1, 2, \ldots, n-1$.



The matrix $A \in \mathbb{R}^{(n-1)\times(n-1)}$ is the same as that in (7.13), i.e., a Toeplitz matrix given by

$$a_{ij} = \begin{cases} \frac{2}{h^2} & \text{if } i = j; \\ -\frac{1}{h^2} & \text{if } i - j = \pm 1; \\ 0 & \text{otherwise.} \end{cases} \tag{9.8}$$

By Lemma 7.25, the eigenvalues and eigenvectors of A are

$$\lambda_k(A) = \frac{4}{h^2} \sin^2 \frac{k\pi}{2n} = \frac{4}{h^2} \sin^2 \frac{kh\pi}{2}, \tag{9.9}$$

$$w_{k,j} = \sin \frac{jk\pi}{n} = \sin(x_j k\pi), \tag{9.10}$$

where $j, k = 1, 2, \ldots, n-1$.

Exercise 9.8. What are the values of $\mathrm{cond}(A)$ for A in (7.13) for $n = 8$ and $n = 1024$?

9.3 Algorithmic components

9.3.1 Fourier modes on Ω^h

Notation 7. Ω^h denotes the uniform grid of n intervals that discretizes the problem domain Ω. Occasionally we also abuse the notation to mean the corresponding vector space of grid functions $\{\Omega^h \to \mathbb{R}\}$.

Definition 9.9. The *wavelength of a sinusoidal function* is the distance of one sinusoidal period. The *wavenumber of a sinusoidal function* k is the number of half sinusoidal waves in unit length.

Lemma 9.10. The kth Fourier mode with its jth component as $w_{k,j} = \sin(x_j k\pi)$ has wavelength $L = \frac{2}{k}$.

Proof. By Definition 9.9, $\sin(x_j k\pi) = -\sin(x_j + \frac{L}{2})k\pi$ implies $x_j k\pi = (x_j + \frac{L}{2})k\pi - \pi$. Hence $k = \frac{2}{L}$. □

Exercise 9.11. For $\Omega = (0, 1)$, plot to show that the maximum wavenumber that is representable on Ω^h is $n_{\max} = \frac{1}{h}$. What if we require that the Fourier mode be 0 at the boundary points?

Lemma 9.12 (Aliasing)**.** For $k \in (n, 2n)$ on Ω^h, the Fourier mode $\mathbf{w}_k$ of which the jth component is $w_{k,j} = \sin(x_j k\pi)$ is actually represented as the additive inverse of the mode $\mathbf{w}_{k'}$ where $k' = 2n - k$.

Proof. It is readily verified that

$$\begin{aligned}\sin(x_j k\pi) &= -\sin(2j\pi - x_j k\pi) = -\sin(x_j(2n-k)\pi)\\ &= -\sin(x_j k'\pi) = -w_{k',j}.\end{aligned}$$ □

Example 9.13. According to Lemma 9.12, the mode with $k = \frac{3}{2}n$ is represented by $k = \frac{1}{2}n$.

Exercise 9.14. Plot the case of $n = 6$ for Example 9.13.

Definition 9.15. On Ω^h, the Fourier modes with wavenumbers $k \in [1, \frac{n}{2})$ are called the *low-frequency* (LF) or *smooth* modes, those with $k \in [\frac{n}{2}, n)$ the *high-frequency* (HF) or *oscillatory* modes.

9.3.2 Relaxation

Lemma 9.16. For the linear system (9.7), the weighted Jacobi in Definition 8.9 has the iteration matrix

$$T_\omega = (1-\omega)I + \omega D^{-1}(L+U) = I - \frac{\omega h^2}{2}A, \tag{9.11}$$

whose eigenvectors are the same as those of A, with the corresponding eigenvalues as

$$\lambda_k(T_\omega) = 1 - 2\omega \sin^2 \frac{k\pi}{2n}, \tag{9.12}$$

where $k = 1, 2, \ldots, n-1$.

Exercise 9.17. Prove Lemma 9.16.

Exercise 9.18. Write a program to reproduce Fig. 2.7 in the book by Briggs et al. [2000]. For $n = 64$, $\omega \in [0, 1]$, verify $\rho(T_\omega) \geq 0.9986$ and hence slow convergence.

Definition 9.19. The *smoothing factor* μ is the maximal factor of damping for HF modes. An iterative method is said to have the *smoothing property* if μ is small and independent of the grid size.

Example 9.20. The smoothing factor of the weighted Jacobi is determined by the optimization problem

$$\mu = \min_{\omega\in(0,1]} \max_{k\in[\frac{n}{2},n)} |\lambda_k(T_\omega)|. \tag{9.13}$$

Since $\lambda_k(T_\omega)$ is a monotonically decreasing function, the minimum is obtained by setting

$$\lambda_{\frac{n}{2}}(T_\omega) = -\lambda_n(T_\omega) = -1 + 2\omega,$$

which implies $\omega = \frac{2}{3}$. Consequently we have $|\lambda_k| \leq \mu = \frac{1}{3}$.

Exercise 9.21. Write a program to reproduce Figure 2.8 in the book by Briggs et al. [2000], verifying that regular Jacobi is only good for damping modes $16 \leq k \leq 48$. In contrast, for $\omega = \frac{2}{3}$, the modes $16 \leq k < 64$ are all damped out quickly.

9.3.3 Restriction and prolongation

Lemma 9.22. The kth LF mode on Ω^h becomes the kth mode (LF or HF) on Ω^{2h}:

$$w^h_{k,2j} = w^{2h}_{k,j}. \tag{9.14}$$

LF modes $k \in [\frac{n}{4}, \frac{n}{2})$ of Ω^h will become HF modes on Ω^{2h}.

Proof. It is readily verified that

$$w^h_{k,2j} = \sin\frac{2jk\pi}{n} = \sin\frac{jk\pi}{\frac{n}{2}} = w^{2h}_{k,j}, \tag{9.15}$$

where $k \in [1, \frac{n}{2})$. Because of the smaller range of k on Ω^{2h}, the modes with $k \in [\frac{n}{4}, \frac{n}{2})$ are HF by definition since the highest wavenumber is $\frac{n}{2}$ on Ω^{2h}. □

Definition 9.23. The *restriction* operator

$$I^{2h}_h : \mathbb{R}^{n-1} \to \mathbb{R}^{\frac{n}{2}-1}$$

maps a vector on the fine grid Ω^h to its counterpart on the coarse grid Ω^{2h}:

$$I^{2h}_h \mathbf{v}^h = \mathbf{v}^{2h}. \tag{9.16}$$

Definition 9.24. The *injection* operator is a restriction operator given by

$$v^{2h}_j = v^h_{2j}, \tag{9.17}$$

where $j = 1, 2, \ldots, \frac{n}{2} - 1$.

Definition 9.25. The *full-weighting* operator is a restriction operator given by

$$v^{2h}_j = \frac{1}{4}\left(v^h_{2j-1} + 2v^h_{2j} + v^h_{2j+1}\right), \tag{9.18}$$

where $j = 1, 2, \ldots, \frac{n}{2} - 1$.

Example 9.26. For $n = 8$, the full-weighting operator is

$$I^{2h}_h = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 & & & \\ & & 1 & 2 & 1 & \\ & & & & 1 & 2 & 1 \end{bmatrix}. \tag{9.19}$$

Definition 9.27. The *prolongation or interpolation* operator

$$I^h_{2h} : \mathbb{R}^{\frac{n}{2}-1} \to \mathbb{R}^{n-1}$$

maps a vector on the coarse grid Ω^{2h} to its counterpart on the fine grid Ω^h:

$$I^h_{2h} \mathbf{v}^{2h} = \mathbf{v}^h. \tag{9.20}$$

Definition 9.28. The *linear interpolation* operator is a prolongation operator given by

$$\begin{aligned} v^h_{2j} &= v^{2h}_j,\\ v^h_{2j+1} &= \tfrac{1}{2}(v^{2h}_j + v^{2h}_{j+1}).\end{aligned} \tag{9.21}$$

Example 9.29. For $n = 8$, the linear interpolation operator is

$$I^h_{2h} = \frac{1}{2}\begin{bmatrix} 1 & & \\ 2 & & \\ 1 & 1 & \\ & 2 & \\ & 1 & 1 \\ & & 2 \\ & & 1 \end{bmatrix}. \tag{9.22}$$

9.3.4 Two-grid correction

Definition 9.30. The *two-grid correction scheme*

$$\mathbf{v}^h \leftarrow \texttt{TG}(\mathbf{v}^h, \mathbf{f}^h, \nu_1, \nu_2) \tag{9.23}$$

solves $A\mathbf{u} = \mathbf{f}$ in (9.7) via steps as follows.

(TG-1) Relax $A^h\mathbf{u}^h = \mathbf{f}^h$ for ν_1 times on Ω^h with initial guess $\mathbf{v}^h$: $\mathbf{v}^h \leftarrow T_\omega^{\nu_1}\mathbf{v}^h + \mathbf{c}_1(f)$,

(TG-2) Compute the fine-grid residual $\mathbf{r}^h = \mathbf{f}^h - A^h\mathbf{v}^h$ and restrict it to the coarse grid by $\mathbf{r}^{2h} = I_h^{2h}\mathbf{r}^h$: $\mathbf{r}^{2h} \leftarrow I_h^{2h}(\mathbf{f}^h - A^h\mathbf{v}^h)$,

(TG-3) Solve $A^{2h}\mathbf{e}^{2h} = \mathbf{r}^{2h}$ on Ω^{2h}: $\mathbf{e}^{2h} \leftarrow (A^{2h})^{-1}\mathbf{r}^{2h}$,

(TG-4) Interpolate the coarse-grid error to the fine grid by $\mathbf{e}^h = I_{2h}^h\mathbf{e}^{2h}$ and correct the fine-grid approximation: $\mathbf{v}^h \leftarrow \mathbf{v}^h + I_{2h}^h\mathbf{e}^{2h}$,

(TG-5) Relax $A^h\mathbf{u}^h = f^h$ for ν_2 times on Ω^h with initial guess $\mathbf{v}^h$: $\mathbf{v}^h \leftarrow T_\omega^{\nu_2}\mathbf{v}^h + \mathbf{c}_2(f)$.

Lemma 9.31. Acting on the error vector, the iteration matrix of the two-grid correction scheme (9.23) is

$$TG = T_\omega^{\nu_2}\left[I - I_{2h}^h(A^{2h})^{-1}I_h^{2h}A^h\right]T_\omega^{\nu_1}. \tag{9.24}$$

Proof. By Definition 9.30, the residual on the fine grid is

$$\mathbf{r}^h(\mathbf{v}^h) = \mathbf{f}^h - A^h\left(T_\omega^{\nu_1}\mathbf{v}^h + \mathbf{c}'(f)\right).$$

The two-grid correction scheme with $\nu_2 = 0$ replaces the initial guess with

$$\mathbf{v}^h \leftarrow T_\omega^{\nu_1}\mathbf{v}^h + \mathbf{c}'(f) + I_{2h}^h(A^{2h})^{-1}I_h^{2h}\mathbf{r}^h(\mathbf{v}^h)$$

which also holds for the exact solution $\mathbf{u}^h$

$$\mathbf{u}^h \leftarrow T_\omega^{\nu_1}\mathbf{u}^h + \mathbf{c}'(f) + I_{2h}^h(A^{2h})^{-1}I_h^{2h}\mathbf{r}^h(\mathbf{u}^h).$$

Subtracting the two equations yields

$$\mathbf{e}^h \leftarrow T_\omega^{\nu_1}\mathbf{e}^h - I_{2h}^h(A^{2h})^{-1}I_h^{2h}A^hT_\omega^{\nu_1}\mathbf{e}^h.$$

Similar arguments apply to step (TG-5) yield (9.24). □

9.3.5 Multigrid cycles

Definition 9.32. The *V-cycle scheme*

$$\mathbf{v}^h \leftarrow \texttt{VC}^h(\mathbf{v}^h, \mathbf{f}^h, \nu_1, \nu_2) \tag{9.25}$$

solves $A\mathbf{u} = \mathbf{f}$ in (9.7) via steps as follows.

(VC-1) Relax ν_1 times on $A^h\mathbf{u}^h = \mathbf{f}^h$ with initial guess $\mathbf{v}^h$.

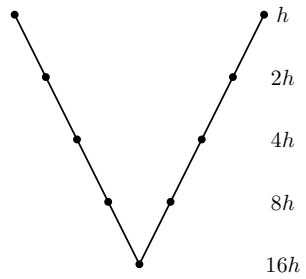
(VC-2) If Ω^h is the coarsest grid, go to (VC-4), otherwise

$$\begin{aligned}\mathbf{f}^{2h} &\leftarrow I_h^{2h}(\mathbf{f}^h - A^h\mathbf{v}^h),\\ \mathbf{v}^{2h} &\leftarrow \mathbf{0},\\ \mathbf{v}^{2h} &\leftarrow \texttt{VC}^{2h}(\mathbf{v}^{2h}, \mathbf{f}^{2h}, \nu_1, \nu_2).\end{aligned}$$

(VC-3) Interpolate error back and correct the solution:

$$\mathbf{v}^h \leftarrow \mathbf{v}^h + I_{2h}^h\mathbf{v}^{2h}.$$

(VC-4) Relax ν_2 times on $A^h\mathbf{u}^h = \mathbf{f}^h$ with initial guess $\mathbf{v}^h$.



Lemma 9.33. In a D-dimensional domain with $n = 2^m$ cells ($m \in \mathbb{N}^+$) along each dimension, the storage cost of V-cycles is

$$2n^{\mathrm{D}}\left(1 + 2^{-\mathrm{D}} + 2^{-2\mathrm{D}} + \cdots + 2^{-m\mathrm{D}}\right) < \frac{2n^{\mathrm{D}}}{1 - 2^{-\mathrm{D}}}. \tag{9.26}$$

Let WU denote the computational cost of performing one relaxation sweep on the finest grid. After neglecting the intergrid transfer, the computational cost of a single V-cycle with $\nu_1 = \nu_2 = 1$ is

$$2\mathrm{WU}\left(1 + 2^{-\mathrm{D}} + 2^{-2\mathrm{D}} + \cdots + 2^{-m\mathrm{D}}\right) < \frac{2}{1 - 2^{-\mathrm{D}}}\mathrm{WU}. \tag{9.27}$$

Proof. On each grid, both vectors of errors and residuals must be stored, and this justifies the factor of 2 in (9.26); the rest of (9.26) follows from Definition 9.32. A similar argument yields (9.27). □

Definition 9.34. The *full multigrid V-cycle*

$$\mathbf{v}^h \leftarrow \texttt{FMG}^h(\mathbf{f}^h, \nu_1, \nu_2) \tag{9.28}$$

solves $A\mathbf{u} = \mathbf{f}$ in (9.7) via steps as follows.

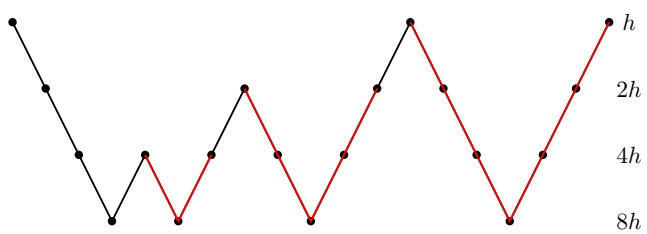
(FMG-1) If Ω^h is the coarsest grid, set $\mathbf{v}^h \leftarrow \mathbf{0}$ and go to (FMG-3), otherwise

$$\begin{aligned}\mathbf{f}^{2h} &\leftarrow I_h^{2h}\mathbf{f}^{\mathbf{h}},\\ \mathbf{v}^{2h} &\leftarrow \texttt{FMG}^{2h}(\mathbf{f}^{2h}, \nu_1, \nu_2).\end{aligned}$$

(FMG-2) Correct $\mathbf{v}^h \leftarrow I_{2h}^h\mathbf{v}^{2h}$.

(FMG-3) Perform a V-cycle with the initial guess as $\mathbf{v}^h$:

$$\mathbf{v}^h \leftarrow \texttt{VC}^h(\mathbf{v}^h, \mathbf{f}^h, \nu_1, \nu_2).$$



Exercise 9.35. Show that, for $\nu_1 = \nu_2 = 1$, the computational cost of an FMG cycle is less than $\frac{2}{(1-2^{-\mathrm{D}})^2}$ WU. Give upper bounds as tight as possible for computational costs of an FMG cycle for $\mathrm{D} = 1, 2, 3$.

9.4 Convergence analysis

9.4.1 The spectral picture

Definition 9.36. A Fourier mode $\mathbf{w}_k^h$ with $k \in [1, \frac{n}{2})$ and the mode $\mathbf{w}_{k'}^h$ with $k' = n - k$ are called *complementary modes* on Ω^h.

Lemma 9.37. A pair of complementary modes satisfy

$$w_{k',j}^h = (-1)^{j+1} w_{k,j}^h. \tag{9.29}$$

Proof. This follows from

$$w_{k',j}^h = \sin\frac{(n-k)j\pi}{n} = \sin\left(j\pi - \frac{kj\pi}{n}\right) = (-1)^{j+1} w_{k,j}^h.$$

□

Lemma 9.38. The action of the full-weighting operator on a pair of complementary modes on Ω^h is

$$I_h^{2h}\mathbf{w}_k^h = c_k \mathbf{w}_k^{2h} := \cos^2\frac{k\pi}{2n}\mathbf{w}_k^{2h}, \tag{9.30a}$$

$$I_h^{2h}\mathbf{w}_{k'}^h = -s_k \mathbf{w}_k^{2h} := -\sin^2\frac{k\pi}{2n}\mathbf{w}_k^{2h}, \tag{9.30b}$$

where $k \in [1, \frac{n}{2})$, $k' = n - k$. In addition, $I_h^{2h}\mathbf{w}_{\frac{n}{2}}^h = \mathbf{0}$.

Proof. For the smooth mode k, we have

$$\begin{aligned}
&\left(I_h^{2h}\mathbf{w}_k^h\right)_j \\
=&\frac{1}{4}\sin\frac{(2j-1)k\pi}{n} + \frac{1}{2}\sin\frac{2jk\pi}{n} + \frac{1}{4}\sin\frac{(2j+1)k\pi}{n} \\
=&\frac{1}{2}\left(1+\cos\frac{k\pi}{n}\right)\sin\frac{2jk\pi}{n} = \cos^2\frac{k\pi}{2n} w_{k,j}^{2h},
\end{aligned}$$

where the last step follows from Lemma 9.22. (9.30b) can be proved by similar steps by replacing k with $n-k$. □

Lemma 9.39. The action of the linear interpolation operator on Ω^{2h} is

$$I_{2h}^h \mathbf{w}_k^{2h} = c_k \mathbf{w}_k^h - s_k \mathbf{w}_{k'}^h, \tag{9.31}$$

where $k' = n - k$.

Proof. Lemma 9.37 and trigonometric identities yield

$$\begin{aligned}
c_k w_{k,j}^h - s_k w_{k',j}^h &= \left(\cos^2\frac{k\pi}{2n} + (-1)^j \sin^2\frac{k\pi}{2n}\right) w_{k,j}^h \\
&= \begin{cases} w_{k,j}^h & \text{if } j \text{ is even;} \\ \cos\frac{k\pi}{n} w_{k,j}^h & \text{if } j \text{ is odd.} \end{cases}
\end{aligned}$$

On the other hand, by Definition 9.28, we have

$$\left(I_{2h}^h \mathbf{w}_k^{2h}\right)_j = \begin{cases} w_{k,j}^h, & \text{if } j \text{ is even,} \\ \frac{1}{2}\sin\frac{k\pi(j-1)}{n} + \frac{1}{2}\sin\frac{k\pi(j+1)}{n} & \text{if } j \text{ is odd,} \end{cases}$$

where last expression simplifies to $\cos\frac{k\pi}{n} w_{k,j}^h$. □

Theorem 9.40. The two-grid correction operator is invariant on the subspace $W_k^h = \operatorname{span}\{\mathbf{w}_k^h, \mathbf{w}_{k'}^h\}$.

$$TG\mathbf{w}_k = \lambda_k^{\nu_1+\nu_2} s_k \mathbf{w}_k + \lambda_k^{\nu_1}\lambda_{k'}^{\nu_2} s_k \mathbf{w}_{k'} \tag{9.32a}$$

$$TG\mathbf{w}_{k'} = \lambda_{k'}^{\nu_1}\lambda_k^{\nu_2} c_k \mathbf{w}_k + \lambda_{k'}^{\nu_1+\nu_2} c_k \mathbf{w}_{k'}, \tag{9.32b}$$

where λ_k is the eigenvalue of T_ω.

Proof. Recall from (9.9) that $\frac{4}{h^2}s_k$ is the eigenvalue of A^h. Consider first the case of $\nu_1 = \nu_2 = 0$.

$$A^h\mathbf{w}_k^h = \frac{4s_k}{h^2}\mathbf{w}_k^h \tag{9.33a}$$

$$\Rightarrow I_h^{2h}A^h\mathbf{w}_k^h = \frac{4c_k s_k}{h^2}\mathbf{w}_k^{2h} \tag{9.33b}$$

$$\Rightarrow (A^{2h})^{-1}I_h^{2h}A^h\mathbf{w}_k^h = \frac{4c_k s_k}{h^2}\frac{(2h)^2}{4\sin^2\frac{k\pi}{n}}\mathbf{w}_k^{2h} = \mathbf{w}_k^{2h} \tag{9.33c}$$

$$\Rightarrow -I_{2h}^h(A^{2h})^{-1}I_h^{2h}A^h\mathbf{w}_k^h = -c_k\mathbf{w}_k^h + s_k\mathbf{w}_{k'}^h \tag{9.33d}$$

$$\Rightarrow \left[I - I_{2h}^h(A^{2h})^{-1}I_h^{2h}A^h\right]\mathbf{w}_k^h = s_k\mathbf{w}_k^h + s_k\mathbf{w}_{k'}^h. \tag{9.33e}$$

Similarly, we have

$$A^h\mathbf{w}_{k'}^h = \frac{4s_{k'}}{h^2}\mathbf{w}_{k'}^h = \frac{4c_k}{h^2}\mathbf{w}_{k'}^h \tag{9.34a}$$

$$\Rightarrow I_h^{2h}A^h\mathbf{w}_{k'}^h = -\frac{4c_k s_k}{h^2}\mathbf{w}_k^{2h} \tag{9.34b}$$

$$\Rightarrow (A^{2h})^{-1}I_h^{2h}A^h\mathbf{w}_{k'}^h = -\frac{4c_k s_k}{h^2}\frac{(2h)^2}{4\sin^2\frac{k\pi}{n}}\mathbf{w}_k^{2h} = -\mathbf{w}_k^{2h} \tag{9.34c}$$

$$\Rightarrow -I_{2h}^h(A^{2h})^{-1}I_h^{2h}A^h\mathbf{w}_{k'}^h = c_k\mathbf{w}_k^h - s_k\mathbf{w}_{k'}^h \tag{9.34d}$$

$$\Rightarrow \left(I - I_{2h}^h(A^{2h})^{-1}I_h^{2h}A^h\right)\mathbf{w}_{k'}^h = c_k\mathbf{w}_k^h + c_k\mathbf{w}_{k'}^h, \tag{9.34e}$$

where $c_k = s_{k'}$ is applied in (9.34a).

Adding pre-smoothing incurs a scaling of $\lambda_k^{\nu_1}$ for (9.33e) and $\lambda_{k'}^{\nu_1}$ for (9.34e). In contrast, adding post-smoothing incurs a scaling of $\lambda_k^{\nu_2}$ for $\mathbf{w}_k^h$ and a scaling of $\lambda_{k'}^{\nu_2}$ for $\mathbf{w}_{k'}^h$ in both (9.33e) and (9.34e). Hence (9.32) holds. □

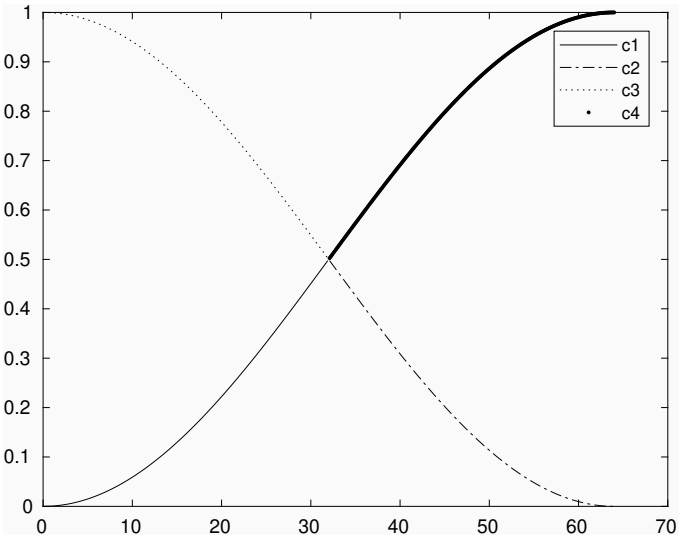
Exercise 9.41. Rewrite (9.32) as

$$TG\begin{bmatrix}\mathbf{w}_k \\ \mathbf{w}_{k'}\end{bmatrix} = \begin{bmatrix}\lambda_k^{\nu_1+\nu_2}s_k & \lambda_k^{\nu_1}\lambda_{k'}^{\nu_2}s_k \\ \lambda_{k'}^{\nu_1}\lambda_k^{\nu_2}c_k & \lambda_{k'}^{\nu_1+\nu_2}c_k\end{bmatrix}\begin{bmatrix}\mathbf{w}_k \\ \mathbf{w}_{k'}\end{bmatrix} = \begin{bmatrix}c_1 & c_2 \\ c_3 & c_4\end{bmatrix}\begin{bmatrix}\mathbf{w}_k \\ \mathbf{w}_{k'}\end{bmatrix}. \tag{9.35}$$

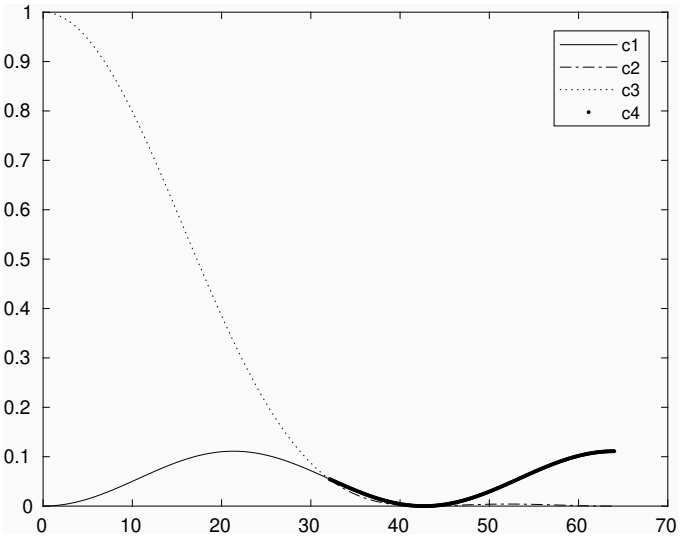
Explain why the magnitude of all four c_i's are small. Deduce the main conclusion $\rho(TG) \approx 0.1$ by reproducing the following plots of the damping coefficients of two-grid correction with weighted Jacobi for $n = 64$ and $\omega = \frac{2}{3}$. The x-axis represents the wavenumber k. Repeat the plots for $n = 128$ to show the independence of $\rho(TG) \approx 0.1$ from the grid size.

Hint: It is tricky to plot the coefficients defined in (9.35). Since c_2, c_4 act on HF modes, one has to ensure that the components in the vectors s_k and c_k indeed correspond to those in $\mathbf{w}_{k'}$. If s_k and c_k are computed from an increasing order of the frequencies, then their components will have to be reversed for plotting. Physical intuition helps in this case: c_1 and c_4 should form one curve while c_2 and c_3 should form another.

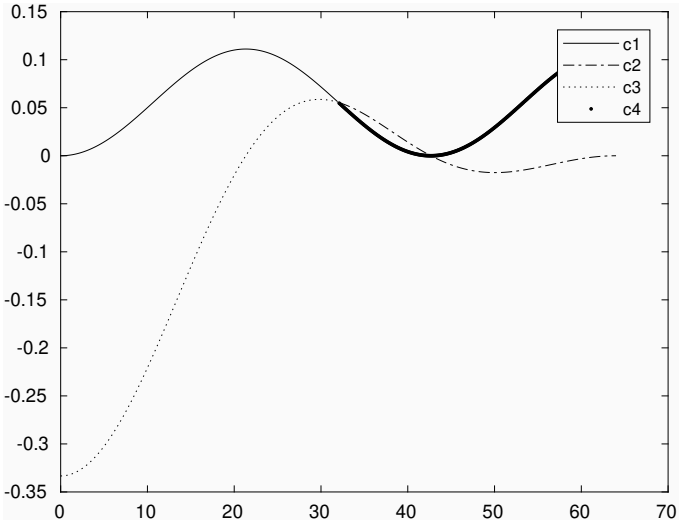
$\nu_1 = 0$, $\nu_2 = 0$.



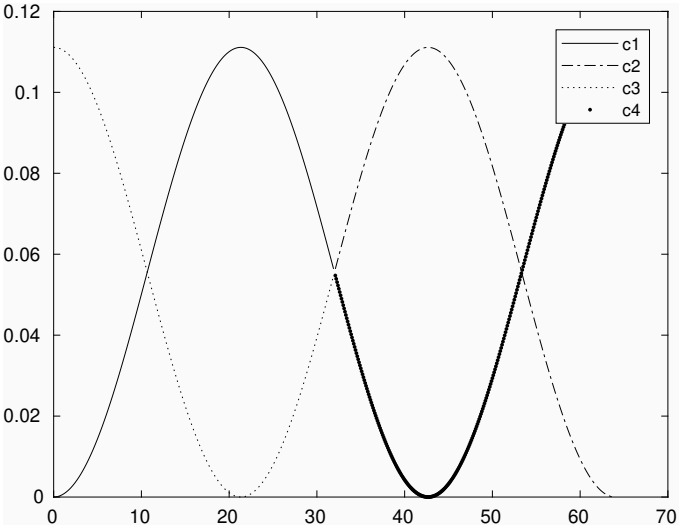
$\nu_1 = 0$, $\nu_2 = 2$.



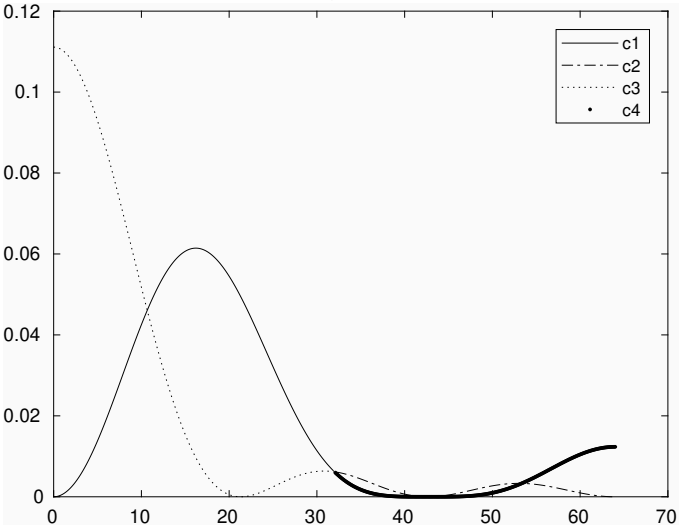
$\nu_1 = 1$, $\nu_2 = 1$.



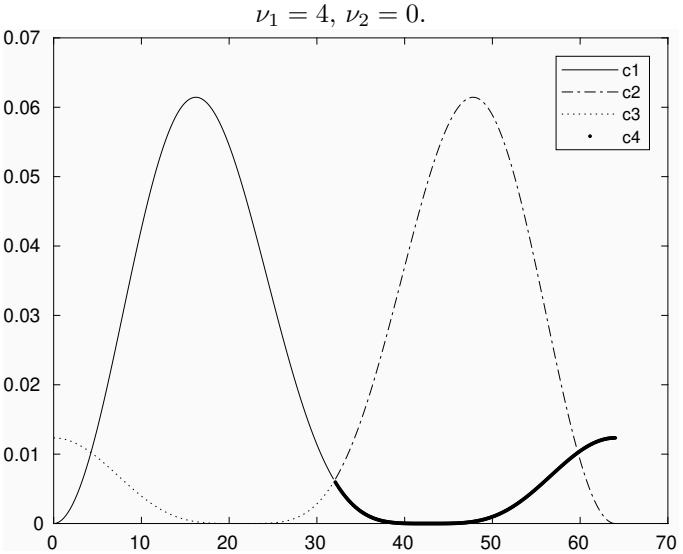
$\nu_1 = 2$, $\nu_2 = 0$.



$\nu_1 = 2$, $\nu_2 = 2$.



$\nu_1 = 4$, $\nu_2 = 0$.



9.4.2 The algebraic picture

Lemma 9.42. The full-weighting operator and the linear interpolation operator satisfy the *variational properties*

$$I_{2h}^{h} = c(I_{h}^{2h})^{T}, \tag{9.36a}$$

$$I_h^{2h} A^h I_{2h}^h = A^{2h}, \tag{9.36b}$$

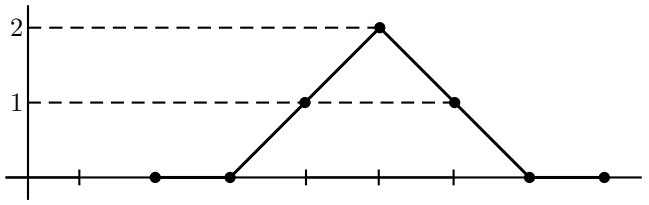
where $c = 2$ and the property (9.36b) is also called the *Galerkin condition*.

Proof. The conclusions follow from (9.8) and Definitions 9.25 and 9.28. □

Lemma 9.43. A basis for the range of the linear interpolation operator $\mathcal{R}(I_{2h}^h)$ is given by its columns, hence the range and the null space of I_{2h}^h satisfy

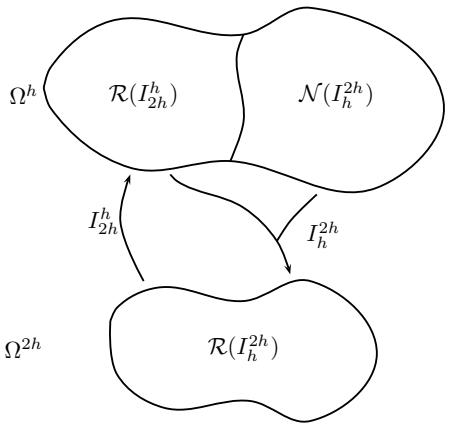
$$\dim \mathcal{R}\left(I_{2h}^h\right) = \frac{n}{2} - 1, \quad \mathcal{N}\left(I_{2h}^h\right) = \{\mathbf{0}\}. \tag{9.37}$$

Proof. $\mathcal{R}(I_{2h}^h) = \{I_{2h}^h \mathbf{v}^{2h} : \mathbf{v}^{2h} \in \Omega^{2h}\}$. The maximum dimension of $\mathcal{R}(I_{2h}^h)$ is thus $\frac{n}{2} - 1$. Any $\mathbf{v}^{2h}$ can be expressed as $\mathbf{v}^{2h} = \sum v_j^{2h} \mathbf{e}_j^{2h}$. It is obvious that the columns of I_{2h}^h are linearly independent. □



Lemma 9.44. The full-weighting operator satisfies

$$\dim \mathcal{R}(I_h^{2h}) = \frac{n}{2} - 1, \qquad \dim \mathcal{N}(I_h^{2h}) = \frac{n}{2}. \tag{9.38}$$



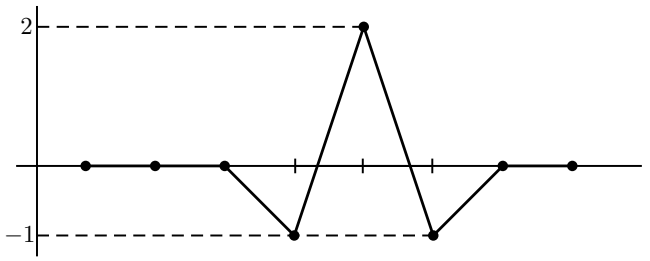
Exercise 9.45. Prove Lemma 9.44.

Lemma 9.46. A basis for the null space of the full-weighting operator is given by

$$\mathcal{N}(I_h^{2h}) = \text{span}\{A^h \mathbf{e}_j^h : j \text{ is odd}\}, \tag{9.39}$$

where $\mathbf{e}_j^h$ is the jth unit vector on Ω^h.

Proof. Consider $I_h^{2h} A^h$. The jth row of I_h^{2h} has $2(j-1)$ leading zeros and the next three nonzero entries are $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. Since the bandwidth of A^h is 3, it suffices to consider only five columns of A^h for potentially non-zero dot-product $\sum_i (I_h^{2h})_{ji} (A^h)_{ik}$. For $2j \pm 1$, these dot products are zero; for $2j$, the dot product is $\frac{1}{2}$; for $2j \pm 2$, the dot product is $-\frac{1}{4}$. Hence for any odd j, we have $I_h^{2h} A^h \mathbf{e}_j^h = \mathbf{0}$. □



Theorem 9.47. The null space of the two-grid correction operator (without relaxation) is the range of linear interpolation:

$$\mathcal{N}(TG) = \mathcal{R}(I_{2h}^h). \tag{9.40}$$

Proof. If $\mathbf{s}^h \in \mathcal{R}(I_{2h}^h)$, then $\mathbf{s}^h = I_{2h}^h \mathbf{q}^{2h}$.

$$TG\mathbf{s}^h = \left[I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right] I_{2h}^h \mathbf{q}^{2h} = \mathbf{0},$$

where the last step comes from (9.36b). Hence we have $\mathcal{R}(I_{2h}^h) \subseteq \mathcal{N}(TG)$. Furthermore, $\mathbf{t}^h \in \mathcal{N}(I_h^{2h} A^h)$ implies

$$TG\mathbf{t}^h = \left[I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right] \mathbf{t}^h = \mathbf{t}^h,$$

i.e., TG is the identity operator when acting on $\mathcal{N}(I_h^{2h} A^h)$. As shown in the plot below Lemma 9.44, the dimension of $\mathcal{N}(TG)$ is no greater than the dimension of $\mathcal{R}(I_h^{2h} A^h)$, which is the same as $\dim \mathcal{R}(I_{2h}^h)$ since A^h is a bijection with full rank on $\mathbb{R}^{n-1}$. This implies that $\dim \mathcal{N}(TG) \leq \dim \mathcal{R}(I_{2h}^h)$, which completes the proof. □

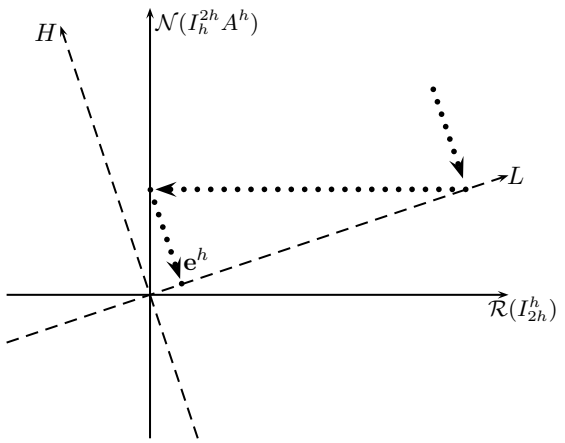
Definition 9.48. Let A be an $n \times n$ symmetric positive definite matrix. The *A-inner product* or *energy inner product* of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle_A := \langle A\mathbf{u}, \mathbf{v} \rangle, \tag{9.41}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product on $\mathbb{R}^n$. Naturally, the *A-norm* or *energy norm* is defined as

$$\|\mathbf{u}\|_A := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle_A}. \tag{9.42}$$

Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are *A-orthogonal* iff $\langle \mathbf{u}, \mathbf{v} \rangle_A = 0$.



9.4.3 The optimal complexity of FMG

Definition 9.49. The error of a computed value v_i^h from the corresponding exact solution $u(x_i)$ is

$$E_i^h := v_i^h - u(x_i) = v_i^h - u_i^h + u_i^h - u(x_i).$$

The *discretization error* is the error $u_i^h - u(x_i)$ incurred by truncating the Taylor series of exact values and the *algebraic error* is the error $v_i^h - u_i^h$ incurred by inexact solution of the linear system.

Lemma 9.50. When linearly interpolating errors from a coarse grid to the fine grid, we have

$$\sqrt{c}\left\|\mathbf{v}^{2h} - \mathbf{u}^{2h}\right\|_{A^{2h}} = \left\|I_{2h}^h \mathbf{v}^{2h} - I_{2h}^h \mathbf{u}^{2h}\right\|_{A^h}. \tag{9.43}$$

where $c \in \mathbb{R}^+$ is the constant in (9.36).

Proof. Definition 9.48 and Lemma 9.42 yield

$$\begin{aligned}
&\|\mathbf{v}^{2h} - \mathbf{u}^{2h}\|_{A^{2h}}^2 \\
=&\left\langle A^{2h}(\mathbf{v}^{2h} - \mathbf{u}^{2h}), \mathbf{v}^{2h} - \mathbf{u}^{2h}\right\rangle \\
=&\left\langle I_h^{2h} A^h I_{2h}^h (\mathbf{v}^{2h} - \mathbf{u}^{2h}), \mathbf{v}^{2h} - \mathbf{u}^{2h}\right\rangle \\
=&\left\langle A^h I_{2h}^h (\mathbf{v}^{2h} - \mathbf{u}^{2h}), \frac{1}{c} I_{2h}^h (\mathbf{v}^{2h} - \mathbf{u}^{2h})\right\rangle \\
=&\frac{1}{c}\|I_{2h}^h \mathbf{v}^{2h} - I_{2h}^h \mathbf{u}^{2h}\|_{A^h}^2,
\end{aligned}$$

where the second step and the third step follow from Lemma 9.42. □

Lemma 9.51. Suppose there exists a constant $K \in \mathbb{R}^+$ independent of the grid size h such that

$$\left\|I_{2h}^h \mathbf{u}^{2h} - \mathbf{u}^h\right\|_{A^h} \le Kh^p. \tag{9.44}$$

Then a single FMG cycle in Definition 9.34 reduces the algebraic error from $O(1)$ to $O(h^p)$, i.e.,

$$\left\|\mathbf{e}^h\right\|_{A^h} \le Kh^p, \tag{9.45}$$

where $p = 2$ is the order of accuracy of the discrete Laplacian (9.8).

Proof. We prove (9.45) by induction. On the coarsest grid, FMG is exact and thus (9.45) holds for the induction basis.

For the induction hypothesis, we assume that the linear system on Ω^{2h} has been solved to the level of discretization error so that

$$(*): \quad \|\mathbf{e}^{2h}\|_{A^{2h}} \le K(2h)^p.$$

We need to show that $(*)$ implies (9.45).

The initial algebraic error on Ω^h is

$$\mathbf{e}_0^h = I_{2h}^h \mathbf{v}^{2h} - \mathbf{u}^h,$$

which yields

$$\begin{aligned}
\left\|\mathbf{e}_0^h\right\|_{A^h} &\le \left\|I_{2h}^h \mathbf{v}^{2h} - I_{2h}^h \mathbf{u}^{2h}\right\|_{A^h} + \left\|I_{2h}^h \mathbf{u}^{2h} - \mathbf{u}^h\right\|_{A^h} \\
&= \sqrt{c}\left\|\mathbf{v}^{2h} - \mathbf{u}^{2h}\right\|_{A^{2h}} + \left\|I_{2h}^h \mathbf{u}^{2h} - \mathbf{u}^h\right\|_{A^h} \\
&\le \sqrt{c}K(2h)^p + Kh^p = (1 + \sqrt{c}2^p)Kh^p,
\end{aligned}$$

where the second step follows from Lemma 9.50 and the third step from (9.44) and the induction hypothesis. Then we have $1 + \sqrt{c}2^p < 7$ from $p = 2$ and $c = 2$. Exercise 9.41 states that $\rho(TG) \approx 0.1$ for a $\mathtt{VC}(2,1)$-cycle and hence one V-cycle is enough to reduce $\left\|\mathbf{e}_0^h\right\|_{A^h}$ to less than Kh^p. □

Theorem 9.52. For the FD discretization (in Example 9.7) of the model problem in Definition 9.6, a single FMG cycle is sufficient to achieve second-order accuracy, with each computated result on Ω^h produced in $O(\frac{1}{h})$ time.

Proof. By Definition 9.49, we have

$$\left\|\mathbf{E}^h\right\| \le \left\|\mathbf{u}^h - \mathbf{u}\right\| + \left\|\mathbf{u}^h - \mathbf{v}^h\right\|.$$

Taylor expansion yields $\left\|\mathbf{u}^h - \mathbf{u}\right\| = O(h^2)$. Then the proof is completed by Exercise 9.35 and Lemma 9.51. □

9.5 Programming assignments

Write a C++ package to implement the one-dimensional multigrid method discussed in this chapter to solve the model problem in Definition 9.6.

I. Your package must give the user the following options:

   (a) boundary conditions: Dirichlet, Neumann, or mixed (partly Dirichlet and partly Neumann).

   (b) restriction operators: full weighting and injection;

   (c) interpolation operators: linear and quadratic;

   (d) cycles: V-cycle and FMG;

   (e) stopping criteria: the number of maximum iterations and the relative accuracy ϵ of the solution;

   (f) the initial guess.

   As for the bottom solver, you can either implement a Gaussian elimination in your own package or use the one in BLAS or LAPACK.

II. For the function in (7.93) derive the corresponding $f(x)$ and the boundary conditions. For $\epsilon = 10^{-8}$ and the zero-vector initial guess, test your multigrid solver for all combinations of (b,c,d) in I on grids with $n = 32, 64, 128, 256$, report the residual and the reduction rate of the residuals for each V-cycle. Report the maximum norm of the error vector and the corresponding convergence rates on the four grids. You should also design at least two of your own test functions and carry out the same process.

III. Gradually reduce ϵ towards 2.2×10^{-16}, under which critical value of ϵ does your program fail to achieve the preset accuracy? Why?

The requirements III-VI in Section 7.7 should also be met in this assignment.

Appendix A

Sets, Logic, and Functions

A.1 First-order logic

Definition A.1. A *set* $\mathcal{S}$ is a collection of *distinct* objects that share a common quality; it is often denoted with the following notation

$$\mathcal{S} = \{x \mid \text{ the conditions that } x \text{ satisfies. } \}. \tag{A.1}$$

Notation 8. $\mathbb{R}, \mathbb{Z}, \mathbb{N}, \mathbb{Q}, \mathbb{C}$ denote the sets of real numbers, integers, natural numbers, rational numbers and complex numbers, respectively. $\mathbb{R}^+, \mathbb{Z}^+, \mathbb{N}^+, \mathbb{Q}^+$ the sets of positive such numbers. In particular, $\mathbb{N}$ contains the number zero while $\mathbb{N}^+$ does not.

Definition A.2. $\mathcal{S}$ is a *subset* of $\mathcal{U}$, written $\mathcal{S} \subseteq \mathcal{U}$, if and only if (iff) $x \in \mathcal{S} \Rightarrow x \in \mathcal{U}$. $\mathcal{S}$ is a *proper subset* of $\mathcal{U}$, written $\mathcal{S} \subset \mathcal{U}$, if $\mathcal{S} \subseteq \mathcal{U}$ and $\exists x \in \mathcal{U}$ s.t. $x \notin \mathcal{S}$.

Definition A.3 (Statements of first-order logic)**.** A *universal statement* is a logical statement of the form

$$\mathsf{U} = (\forall x \in \mathcal{S},\ \mathsf{A}(x)). \tag{A.2}$$

An *existential statement* has the form

$$\mathsf{E} = (\exists x \in \mathcal{S},\ \text{s.t. } \mathsf{A}(x)), \tag{A.3}$$

where $\forall$ ("for each") and $\exists$ ("there exists") are the *quantifiers*, $\mathcal{S}$ is a set, "s.t." means "such that," and $\mathsf{A}(x)$ is the *formula*.
A statement of *implication/conditional* has the form

$$\mathsf{A} \Rightarrow \mathsf{B}. \tag{A.4}$$

Example A.4. Universal and existential statements:
$\forall x \in [2, +\infty),\ x > 1$;
$\forall x \in \mathbb{R}^+,\ x > 1$;
$\exists p, q \in \mathbb{Z},\ \text{s.t. } p/q = \sqrt{2}$;
$\exists p, q \in \mathbb{Z},\ \text{s.t. } \sqrt{p} = \sqrt{q} + 1$.

Definition A.5. *Uniqueness quantification* or *unique existential quantification*, written $\exists!$ or $\exists_{=1}$, indicates that exactly one object with a certain property exists.

Exercise A.6. Express the logical statement $\exists! x,\ \text{s.t. } \mathsf{A}(x)$ with $\exists$, $\forall$, and $\Leftrightarrow$.

Definition A.7. A *universal-existential statement* is a logical statement of the form

$$\mathsf{U}_E = (\forall x \in \mathcal{S},\ \exists y \in \mathcal{T} \text{ s.t. } \mathsf{A}(x, y)). \tag{A.5}$$

An *existential-universal statement* has the form

$$\mathsf{E}_U = (\exists y \in \mathcal{T},\ \text{s.t. } \forall x \in \mathcal{S},\ \mathsf{A}(x, y)). \tag{A.6}$$

Example A.8. True or false:
$\forall x \in [2, +\infty),\ \exists y \in \mathbb{Z}^+ \text{ s.t. } x^y < 10^5$;
$\exists y \in \mathbb{R} \text{ s.t. } \forall x \in [2, +\infty),\ x > y$;
$\exists y \in \mathbb{R} \text{ s.t. } \forall x \in [2, +\infty),\ x < y$.

Example A.9 (Translating an English statement into a logical statement)**.** Goldbach's conjecture states *every even natural number greater than 2 is the sum of two primes.* Let $\mathbb{P} \subset \mathbb{N}^+$ denote the set of prime numbers. Then Goldbach's conjecture is $\forall a \in 2\mathbb{N}^+ + 2, \exists p, q \in \mathbb{P}, \text{ s.t. } a = p + q$.

Theorem A.10. The existential-universal statement implies the corresponding universal-existential statement, but not vice versa.

Example A.11 (Translating a logical statement to an English statement)**.** Let $\mathcal{S}$ be the set of all human beings.
U_E =($\forall p \in \mathcal{S}, \exists q \in \mathcal{S}$ s.t. q is p's mom.)
E_U=($\exists q \in \mathcal{S}$ s.t. $\forall p \in \mathcal{S}$, q is p's mom.)
U_E is probably true, but E_U is certainly false.
If E_U were true, then U_E would be true. Why?

Axiom A.12 (First-order negation of logical statements)**.** The negations of the statements in Definition A.3 are

$$\neg\mathsf{U} = (\exists x \in \mathcal{S},\ \text{s.t. } \neg\mathsf{A}(x)). \tag{A.7}$$
$$\neg\mathsf{E} = (\forall x \in \mathcal{S},\ \neg\mathsf{A}(x)). \tag{A.8}$$

Rule A.13. The negation of a more complicated logical statement abides by the following rules:

- switch the type of each quantifier until you reach the last formula without quantifiers;
- negate the last formula.

In particular, the negation of an implication formula $P \Rightarrow Q$, is $P \wedge \neg Q$.

Example A.14 (The negation of Goldbach's conjecture)**.** $\exists a \in 2\mathbb{N}^+ + 2 \text{ s.t. } \forall p, q \in \mathbb{P},\ a \neq p + q$.

Exercise A.15. Negate the logical statement in Definition C.65.

Axiom A.16 (Contraposition)**.** A conditional statement is logically equivalent to its contrapositive.

$$(\mathsf{A} \Rightarrow \mathsf{B}) \Leftrightarrow (\neg\mathsf{B} \Rightarrow \neg\mathsf{A}) \tag{A.9}$$

Example A.17. "If Jack is a man, then Jack is a human being." is equivalent to "If Jack is not a human being, then Jack is not a man."

Exercise A.18. Draw an Euler diagram of subsets to illustrate Example A.17.

Exercise A.19. Rewrite each of the following statements and its *negation* into *logical statements* using symbols, quantifiers, and formulas.

(a) The only even prime is 2.

(b) Multiplication of integers is associative.

(c) Goldbach's conjecture has at most a finite number of counterexamples.

A.2 Ordered sets

Definition A.20. The *Cartesian product* $\mathcal{X} \times \mathcal{Y}$ between two sets $\mathcal{X}$ and $\mathcal{Y}$ is the set of all possible ordered pairs with first element from $\mathcal{X}$ and second element from $\mathcal{Y}$:

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) \mid x \in \mathcal{X},\ y \in \mathcal{Y}\}. \tag{A.10}$$

Axiom A.21 (Fundamental principle of counting)**.** Consider a task that consists of a sequence of k independent steps. Let n_i denote the number of different choices for the i-th step, the total number of distinct ways to complete the task is

$$\prod_{i=1}^{k} n_i = n_1 n_2 \cdots n_k. \tag{A.11}$$

Example A.22. Let A, E, D be the set of appetizers, main entrees, desserts in a restaurant. $A \times E \times D$ is the set of possible dinner combos. If $\#A = 10$, $\#E = 5$, $\#D = 6$, $\#(A \times E \times D) = 300$.

Definition A.23 (Maximum and minimum)**.** Consider $\mathcal{S} \subseteq \mathbb{R}$, $\mathcal{S} \neq \emptyset$. If $\exists s_m \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S}$, $x \leq s_m$, then s_m is the *maximum* of $\mathcal{S}$ and denoted by $\max \mathcal{S}$. If $\exists s_m \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S}$, $x \geq s_m$, then s_m is the *minimum* of $\mathcal{S}$ and denoted by $\min \mathcal{S}$.

Definition A.24 (Upper and lower bounds)**.** Consider $\mathcal{S} \subseteq \mathbb{R}$, $\mathcal{S} \neq \emptyset$. a is an *upper bound* of $\mathcal{S} \subseteq \mathbb{R}$ if $\forall x \in \mathcal{S}$, $x \leq a$; then the set $\mathcal{S}$ is said to be *bounded above*. a is a *lower bound* of $\mathcal{S}$ if $\forall x \in \mathcal{S}$, $x \geq a$; then the set $\mathcal{S}$ is said to be *bounded below*. $\mathcal{S}$ is *bounded* if it is bounded above and bounded below.

Definition A.25 (Supremum and infimum)**.** Consider a nonempty set $\mathcal{S} \subseteq \mathbb{R}$. If $\mathcal{S}$ is bounded above and $\mathcal{S}$ has a least upper bound then we call it the *supremum* of $\mathcal{S}$ and denote it by $\sup \mathcal{S}$. If $\mathcal{S}$ is bounded below and $\mathcal{S}$ has a greatest lower bound, then we call it the *infimum* of $\mathcal{S}$ and denote it by $\inf \mathcal{S}$.

Example A.26. If a set $\mathcal{S} \subset \mathbb{R}$ has a maximum, we have $\max \mathcal{S} = \sup \mathcal{S}$.

Example A.27. $\sup[a, b] = \sup[a, b) = \sup(a, b] = \sup(a, b)$.

Theorem A.28 (Existence and uniqueness of least upper bound)**.** Every nonempty subset of $\mathbb{R}$ that is bounded above has exactly one least upper bound.

Corollary A.29. Every nonempty subset of $\mathbb{R}$ that is bounded below has a greatest lower bound.

Definition A.30. A *binary relation between two sets* $\mathcal{X}$ and $\mathcal{Y}$ is an ordered triple $(\mathcal{X}, \mathcal{Y}, \mathcal{G})$ where $\mathcal{G} \subseteq \mathcal{X} \times \mathcal{Y}$.
A *binary relation on* $\mathcal{X}$ is the relation between $\mathcal{X}$ and $\mathcal{X}$.
The statement $(x, y) \in R$ is read "x is R-related to y," and denoted by xRy or $R(x, y)$.

Definition A.31. An *equivalence relation* "$\sim$" on $\mathcal{A}$ is a binary relation on $\mathcal{A}$ that satisfies $\forall a, b, c \in \mathcal{A}$,

- $a \sim a$ (reflexivity);
- $a \sim b$ implies $b \sim a$ (symmetry);
- $a \sim b$ and $b \sim c$ imply $a \sim c$ (transitivity).

Definition A.32. A binary relation "$\leq$" on some set $\mathcal{S}$ is a *total order* or *linear order* on $\mathcal{S}$ iff, $\forall a, b, c \in \mathcal{S}$,

- $a \leq b$ and $b \leq a$ imply $a = b$ (antisymmetry);
- $a \leq b$ and $b \leq c$ imply $a \leq c$ (transitivity);
- $a \leq b$ or $b \leq a$ (totality).

A set equipped with a total order is a *chain* or *totally ordered set*.

Example A.33. The real numbers with less or equal.

Example A.34. The English letters of the alphabet with dictionary order.

Example A.35. The Cartesian product of a set of totally ordered sets with the *lexicographical order*.

Example A.36. Sort your book in lexicographical order and save a lot of time. $\log_{26} N \ll N!$

Definition A.37. A binary relation "$\leq$" on some set $\mathcal{S}$ is a *partial order* on $\mathcal{S}$ iff, $\forall a, b, c \in \mathcal{S}$, antisymmetry, transitivity, and reflexivity ($a \leq a$) hold.
A set equipped with a partial order is called a *poset*.

Example A.38. The set of subsets of a set $\mathcal{S}$ ordered by inclusion "$\subseteq$."

Example A.39. The natural numbers equipped with the relation of divisibility.

Example A.40. The set of stuff you will put on your body every morning with the time ordered: undershorts, pants, belt, shirt, tie, jacket, socks, shoes, watch.

Example A.41. Inheritance ("is-a" relation) is a partial order. $A \to B$ reads "B is a special type of A".

Example A.42. Composition ("has-a" relation) is also a partial order. $A \rightsquigarrow B$ reads "B *has an* instance/object of A."

Example A.43. Implication "$\Rightarrow$" is a partial order on the set of logical statements.

Example A.44. The set of definitions, axioms, propositions, theorems, lemmas, etc., is a poset with inheritance, composition, and implication. It is helpful to relate them with these partial orderings.

Definition A.45. An *upper bound* of a subset W of a poset M is an element $u \in M$ such that $x \leq u$ for each $x \in W$. A *maximal element* of a poset M is an $m \in M$ such that

$$\forall x \in M,\ x \geq m \ \Rightarrow\ x = m. \tag{A.12}$$

Axiom A.46 (Zorn's lemma)**.** For a nonempty poset M, if every chain in M has an upper bound, then M has at least one maximal element.

Lemma A.47 (The Union Lemma)**.** Let X be a set and $\mathcal{C}$ be a collection of subsets of X. Assume that for each $x \in X$, there is a set A_x in $\mathcal{C}$ such that $x \in A_x$. Then $\cup_{x\in X} A_x = X$.

A.3 Functions

Definition A.48. A *function/map/mapping* f from $\mathcal{X}$ to $\mathcal{Y}$, written $f : \mathcal{X} \to \mathcal{Y}$ or $\mathcal{X} \mapsto \mathcal{Y}$, is a subset of the Cartesian product $\mathcal{X} \times \mathcal{Y}$ satisfying that $\forall x \in \mathcal{X}$, there is exactly one $y \in \mathcal{Y}$ s.t. $(x, y) \in \mathcal{X} \times \mathcal{Y}$. $\mathcal{X}$ and $\mathcal{Y}$ are the *domain* and *range* of f, respectively.

Definition A.49. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be *injective* or *one-to-one* iff

$$\forall x_1 \in \mathcal{X}, \forall x_2 \in \mathcal{X},\ \ x_1 \neq x_2 \ \Rightarrow\ f(x_1) \neq f(x_2). \tag{A.13}$$

It is *surjective* or *onto* iff

$$\forall y \in \mathcal{Y},\ \exists x \in \mathcal{X},\ \text{s.t. } y = f(x). \tag{A.14}$$

It is *bijective* iff it is both injective and surjective.

Definition A.50. A set $\mathcal{S}$ is *countably infinite* iff there exists a bijective function $f : \mathcal{S} \to \mathbb{N}^+$ that maps $\mathcal{S}$ to $\mathbb{N}^+$. A set is *countable* if it is either finite or countably infinite.

Example A.51. Are the integers countable? Are the rationals countable? Are the real numbers countable?

Definition A.52. A *binary function* or a *binary operation* on a set $\mathcal{S}$ is a map $\mathcal{S} \times \mathcal{S} \to \mathcal{S}$.

Appendix B

Linear Algebra

B.1 Vector spaces

Definition B.1. A *field* $\mathbb{F}$ is a set together with two binary operations, usually called "addition" and "multiplication" and denoted by "$+$" and "$*$", such that $\forall a, b, c \in \mathbb{F}$, the following axioms hold,

- commutativity: $a + b = b + a$, $ab = ba$;
- associativity: $a + (b + c) = (a + b) + c$, $a(bc) = (ab)c$;
- identity: $a + 0 = a$, $a1 = a$;
- invertibility: $a + (-a) = 0$, $aa^{-1} = 1$ $(a \neq 0)$;
- distributivity: $a(b + c) = ab + ac$.

Definition B.2. A *vector space* or *linear space* over a field $\mathbb{F}$ is a set $\mathcal{V}$ together with two binary operations "$+$" and "$\times$" respectively called vector addition and scalar multiplication that satisfy the following axioms:

(VSA-1) commutativity
$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$;

(VSA-2) associativity
$\forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$;

(VSA-3) compatibility
$\forall \mathbf{u} \in \mathcal{V}$, $\forall a, b \in \mathbb{F}$, $(ab)\mathbf{u} = a(b\mathbf{u})$;

(VSA-4) additive identity
$\exists \mathbf{0} \in \mathcal{V}$, $\forall \mathbf{u} \in \mathcal{V}$, s.t. $\mathbf{u} + \mathbf{0} = \mathbf{u}$;

(VSA-5) additive inverse
$\forall \mathbf{u} \in \mathcal{V}$, $\exists \mathbf{v} \in \mathcal{V}$, s.t. $\mathbf{u} + \mathbf{v} = \mathbf{0}$;

(VSA-6) multiplicative identity
$\exists 1 \in \mathbb{F}$, s.t. $\forall \mathbf{u} \in \mathcal{V}$, $1\mathbf{u} = \mathbf{u}$;

(VSA-7) distributive laws

$$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V},\ \forall a, b \in \mathbb{F},\ \begin{cases} (a + b)\mathbf{u} = a\mathbf{u} + b\mathbf{u}, \\ a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}. \end{cases}$$

The elements of $\mathcal{V}$ are called *vectors* and the elements of $\mathbb{F}$ are called *scalars*.

Definition B.3. A *real vector space* or a *complex vector space* is a vector space with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, respectively.

Exercise B.4. Show that a complex vector space always induces another real vector space.

Example B.5. The simplest vector space is $\{\mathbf{0}\}$. Another simple example of a vector space over a field $\mathbb{F}$ is $\mathbb{F}$ itself, equipped with its standard addition and multiplication.

B.1.1 Subspaces

Definition B.6. A subset $\mathcal{U}$ of $\mathcal{V}$ is called a *subspace* of $\mathcal{V}$ if $\mathcal{U}$ is also a vector space when equipped with the same addition and scalar multiplication on V.

Definition B.7. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subsets of $\mathcal{V}$. The *sum* of $\mathcal{U}_1, \ldots, \mathcal{U}_m$ is the set of all possible sums of elements of $\mathcal{U}_1, \ldots, \mathcal{U}_m$:

$$\mathcal{U}_1 + \ldots + \mathcal{U}_m := \left\{ \sum_{j=1}^{m} \mathbf{u}_j : \mathbf{u}_j \in \mathcal{U}_j \right\}. \tag{B.1}$$

Example B.8. For $\mathcal{U} = \{(x, x, y, y) \in \mathbb{F}^4 : x, y \in \mathbb{F}\}$ and $\mathcal{W} = \{(x, x, x, y) \in \mathbb{F}^4 : x, y \in \mathbb{F}\}$, we have

$$\mathcal{U} + \mathcal{W} = \{(x, x, z, y) \in \mathbb{F}^4 : x, y, z \in \mathbb{F}\}.$$

Lemma B.9. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subspaces of $\mathcal{V}$. Then $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ is the smallest subspace of $\mathcal{V}$ that contains $\mathcal{U}_1, \ldots, \mathcal{U}_m$.

Definition B.10. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subspaces of $\mathcal{V}$. The sum $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ is called a *direct sum* if each element in $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ can be written in only one way as a sum $\sum_{j=1}^{m} \mathbf{u}_j$ with $\mathbf{u}_j \in \mathcal{U}_j$ for each $j = 1, \ldots, m$. In this case we write the direct sum as $\mathcal{U}_1 \oplus \ldots \oplus \mathcal{U}_m$.

Exercise B.11. Show that $\mathcal{U}_1 + \mathcal{U}_2 + \mathcal{U}_3$ is not a direct sum:

$$\begin{aligned} \mathcal{U}_1 &= \{(x, y, 0) \in \mathbb{F}^3 : x, y \in \mathbb{F}\}, \\ \mathcal{U}_2 &= \{(0, 0, z) \in \mathbb{F}^3 : z \in \mathbb{F}\}, \\ \mathcal{U}_3 &= \{(0, y, y) \in \mathbb{F}^3 : y \in \mathbb{F}\}. \end{aligned}$$

Lemma B.12. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subspaces of $\mathcal{V}$. Then $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ is a direct sum if and only if the only way to write $\mathbf{0}$ as a sum $\sum_{j=1}^{m} \mathbf{u}_j$, where $\mathbf{u}_j \in \mathcal{U}_j$ for each $j = 1, \ldots, m$, is by taking each $\mathbf{u}_j$ equal to $\mathbf{0}$.

Theorem B.13. Suppose $\mathcal{U}$ and $\mathcal{W}$ are subspaces of $\mathcal{V}$. Then $\mathcal{U} + \mathcal{W}$ is a direct sum if and only if $\mathcal{U} \cap \mathcal{W} = \{\mathbf{0}\}$.

B.1.2 Span and linear independence

Definition B.14. A *list of length* n or *n-tuple* is an ordered collection of n elements (which might be numbers, other lists, or more abstract entities) separated by commas and surrounded by parentheses: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Definition B.15. A vector space composed of all the n-tuples of a field $\mathbb{F}$ is known as a *coordinate space*, denoted by $\mathbb{F}^n$ ($n \in \mathbb{N}^+$).

Example B.16. The properties of forces or velocities in the real world can be captured by a coordinate space $\mathbb{R}^2$ or $\mathbb{R}^3$.

Example B.17. The set of continuous real-valued functions on the interval $[a, b]$ forms a real vector space.

Notation 9. For a set $\mathcal{S}$, define a vector space

$$\mathbb{F}^{\mathcal{S}} := \{f : \mathcal{S} \to \mathbb{F}\}.$$

$\mathbb{F}^n$ is a special case of $\mathbb{F}^{\mathcal{S}}$ because n can be regarded as the set $\{1, 2, \ldots, n\}$ and each element in $\mathbb{F}^n$ can be considered as a function $\{1, 2, \ldots, n\} \mapsto \mathbb{F}$.

Definition B.18. A *linear combination* of a list of vectors $\{\mathbf{v}_i\}$ is a vector of the form $\sum_i a_i \mathbf{v}_i$ where $a_i \in \mathbb{F}$.

Example B.19. $(17, -4, 2)$ is a linear combination of $(2, 1, -3), (1, -2, 4)$ because

$$(17, -4, 2) = 6(2, 1, -3) + 5(1, -2, 4).$$

Example B.20. $(17, -4, 5)$ is not a linear combination of $(2, 1, -3), (1, -2, 4)$ because there do not exist numbers a_1, a_2 such that

$$(17, -4, 5) = a_1(2, 1, -3) + a_2(1, -2, 4).$$

Solving from the first two equations yields $a_1 = 6$, $a_2 = 5$, but $5 \neq -3 \times 6 + 4 \times 5$.

Definition B.21. The *span* of a list of vectors $(\mathbf{v}_i)$ is the set of all linear combinations of $(\mathbf{v}_i)$,

$$\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) = \left\{ \sum_{i=1}^{m} a_i \mathbf{v}_i : \ a_i \in \mathbb{F} \right\}. \tag{B.2}$$

In particular, the span of the empty set is $\{\mathbf{0}\}$. We say that $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ *spans* $\mathcal{V}$ if $\mathcal{V} = \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$.

Example B.22.

$$(17, -4, 2) \in \operatorname{span}((2, 1, -3), (1, -2, 4))$$
$$(17, -4, 5) \notin \operatorname{span}((2, 1, -3), (1, -2, 4))$$

Definition B.23. A vector space $\mathcal{V}$ is called *finite dimensional* if some list of vectors span $\mathcal{V}$; otherwise it is *infinite dimensional*.

Example B.24. Let $\mathbb{P}_m(\mathbb{F})$ denote the set of all polynomials with coefficients in $\mathbb{F}$ and degree at most m,

$$\mathbb{P}_m(\mathbb{F}) = \left\{ p : \mathbb{F} \to \mathbb{F};\ p(z) = \sum_{i=0}^{m} a_i z^i, a_i \in \mathbb{F} \right\}. \tag{B.3}$$

Then $\mathbb{P}_m(\mathbb{F})$ is a finite-dimensional vector space for each non-negative integer m. The set of all polynomials with coefficients in $\mathbb{F}$, denoted by $\mathbb{P}(\mathbb{F}) := \mathbb{P}_{+\infty}(\mathbb{F})$, is infinite-dimensional. Both are subspaces of $\mathbb{F}^{\mathbb{F}}$ for $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Definition B.25. A list of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ in $\mathcal{V}$ is called *linearly independent* iff

$$a_1 \mathbf{v}_1 + \ldots + a_m \mathbf{v}_m = \mathbf{0} \ \Rightarrow\ a_1 = \cdots = a_m = 0. \tag{B.4}$$

Otherwise the list of vectors is called *linearly dependent*.

Example B.26. The empty list is declared to be linearly independent. A list of one vector $(\mathbf{v})$ is linearly independent iff $\mathbf{v} \neq \mathbf{0}$. A list of two vectors is linearly independent iff neither vector is a scalar multiple of the other.

Example B.27. The list $(1, z, \ldots, z^m)$ is linearly independent in $\mathbb{P}_m(\mathbb{F})$ for each $m \in \mathbb{N}$.

Example B.28. $(2, 3, 1)$, $(1, -1, 2)$, and $(7, 3, 8)$ is linearly dependent in $\mathbb{R}^3$ because

$$2(2, 3, 1) + 3(1, -1, 2) + (-1)(7, 3, 8) = (0, 0, 0).$$

Example B.29. Every list of vectors containing the $\mathbf{0}$ vector is linearly dependent.

Lemma B.30 (Linear dependence lemma)**.** Suppose $V = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m)$ is a linearly dependent list in $\mathcal{V}$. Then there exists $j \in \{1, 2, \ldots, m\}$ such that

- $\mathbf{v}_j \in \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1})$;
- if the jth term is removed from V, the span of the remaining list equals $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$.

Lemma B.31. In a finite-dimensional vector space, the length of every linearly independent list of vectors is less than or equal to the length of every spanning list of vectors.

B.1.3 Bases

Definition B.32. A *basis* of a vector space $\mathcal{V}$ is a list of vectors in $\mathcal{V}$ that is linearly independent and spans $\mathcal{V}$.

Definition B.33. The *standard basis* of $\mathbb{F}^n$ is the list of vectors

$$(1, 0, \cdots, 0)^T,\ (0, 1, 0, \cdots, 0)^T,\ \ldots,\ (0, \cdots, 0, 1)^T. \tag{B.5}$$

Example B.34. $(z^0, z^1, \ldots, z^m)$ is a basis of $\mathbb{P}_m(\mathbb{F})$ in (B.3).

Lemma B.35. A list of vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is a basis of $\mathcal{V}$ iff every vector $\mathbf{u} \in \mathcal{V}$ can be written uniquely as

$$\mathbf{u} = \sum_{i=1}^{n} a_i \mathbf{v}_i, \tag{B.6}$$

where $a_i \in \mathbb{F}$.

Lemma B.36. Every spanning list in a vector space $\mathcal{V}$ can be reduced to a basis of $\mathcal{V}$.

Lemma B.37. Every linearly independent list of vectors in a finite-dimensional vector space can be extended to a basis of that vector space.

B.1.4 Dimension

Lemma B.38. Any two bases of a finite-dimensional vector space have the same length.

Proof. Suppose B_1 and B_2 are two bases of V. Then B_1 is linearly independent in V and B_2 spans V. By Lemma B.31, the length of B_1 is no greater than B_2. The proof is completed by switching the roles of B_1 and B_2. □

Definition B.39. The *dimension* of a finite-dimensional vector space $\mathcal{V}$, denoted $\dim \mathcal{V}$, is the length of any basis of the vector space.

Lemma B.40. If $\mathcal{V}$ is finite-dimensional, then every spanning list of vectors in $\mathcal{V}$ with length $\dim \mathcal{V}$ is a basis of $\mathcal{V}$.

Lemma B.41. If $\mathcal{V}$ is finite-dimensional, then every linearly independent list of vectors in $\mathcal{V}$ with length $\dim \mathcal{V}$ is a basis of $\mathcal{V}$.

B.2 Linear maps

Definition B.42. A *linear map* or *linear transformation* between two vector spaces $\mathcal{V}$ and $\mathcal{W}$ is a function $T : \mathcal{V} \to \mathcal{W}$ that satisfies

(LNM-1) additivity
$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $T(\mathbf{u} + \mathbf{v}) = T\mathbf{u} + T\mathbf{v}$;

(LNM-2) homogeneity
$\forall a \in \mathbb{F}$, $\forall \mathbf{v} \in \mathcal{V}$, $T(a\mathbf{v}) = a(T\mathbf{v})$,

where $\mathbb{F}$ is a scalar field. In particular, a linear map $T : \mathcal{V} \to \mathcal{W}$ is called a *(linear) operator* if $\mathcal{W} = \mathcal{V}$.

Notation 10. The set of all linear maps from $\mathcal{V}$ to $\mathcal{W}$ is denoted by $\mathcal{L}(\mathcal{V}, \mathcal{W})$. The set of all linear operators from $\mathcal{V}$ to itself is denoted by $\mathcal{L}(\mathcal{V})$.

Example B.43. The differentiation operator on $\mathbb{R}[x]$ is a linear map $T \in \mathcal{L}(\mathbb{R}[x], \mathbb{R}[x])$

Example B.44. $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m) \simeq \mathbb{F}^{m\times n}$ is a vector space with the zero map $\mathbf{0}$ as the additive identity.

Lemma B.45. The set $\mathcal{L}(\mathcal{V}, \mathcal{W})$, equipped with scalar multiplication $(aT)\mathbf{v} = a(T\mathbf{v})$ and vector addition $(S + T)\mathbf{v} = S\mathbf{v} + T\mathbf{v}$, is a vector space.

Proof. The scalar field $\mathbb{F}$ of $\mathcal{L}(\mathcal{V}, \mathcal{W})$ is the same as that of $\mathcal{V}$ and $\mathcal{W}$. So multiplicative identity is still 1, the same as that of $\mathbb{F}$. However, the additive identity is the zero map $\mathbf{0} \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. □

Definition B.46. The *identity map*, denoted I, is the function on a vector space that assigns to each element the same element:

$$I\mathbf{v} = \mathbf{v}. \tag{B.7}$$

Definition B.47. A *complex linear functional* is a linear map $T : \mathcal{V} \to \mathbb{C}$ with $\mathbb{C}$ being the underlying field of $\mathcal{V}$. A *real linear functional* is a map $T : \mathcal{V} \to \mathbb{R}$ such that (LNM-1) and (LNM-2) in Definition B.42 hold for $\mathbb{F} = \mathbb{R}$.

Lemma B.48. Let V be a complex vector space and f a complex linear functional on V. Then the real part $\operatorname{Re} f(x) = u(x)$ is related to f by

$$\forall x \in V, \quad f(x) = u(x) - \mathbf{i}u(\mathbf{i}x). \tag{B.8}$$

Proof. Any $\alpha, \beta \in \mathbb{R}$ and $z = \alpha + \mathbf{i}\beta \in \mathbb{C}$ satisfy

$$z = \operatorname{Re} z - \mathbf{i}\operatorname{Re}(\mathbf{i}z).$$

Set $z = f(x)$ and we have

$$\begin{aligned} f(x) &= \operatorname{Re} f(x) - \mathbf{i}\operatorname{Re}(\mathbf{i}f(x)) \\ &= u(x) - \mathbf{i}\operatorname{Re}(f(\mathbf{i}x)) = u(x) - \mathbf{i}u(\mathbf{i}x). \end{aligned}$$ □

Lemma B.49. Let V be a complex vector space and $u : V \to \mathbb{R}$ a real linear functional on V. Then the function $f : V \to \mathbb{C}$ defined by (B.8) is a complex linear functional.

Proof. The additivity (LNM-1) of f follows from the additivity of u and (B.8). For any $c \in \mathbb{R}$, we have $f(cx) = cf(x)$ from (B.8). The rest follows from the additivity of f and

$$\begin{aligned} f(\mathbf{i}x) &= u(\mathbf{i}x) - \mathbf{i}u(\mathbf{i}^2x) \\ &= u(\mathbf{i}x) + \mathbf{i}u(x) = \mathbf{i}(u(x) - \mathbf{i}u(\mathbf{i}x)) = \mathbf{i}f(x). \end{aligned}$$ □

B.2.1 Null spaces and ranges

Definition B.50. The *null space* of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ is the subset of $\mathcal{V}$ consisting of those vectors that T maps to the additive identity $\mathbf{0}$:

$$\operatorname{null} T = \{\mathbf{v} \in \mathcal{V} : T\mathbf{v} = \mathbf{0}\}. \tag{B.9}$$

Example B.51. The null space of the differentiation map in Example B.43 is $\mathbb{R}$.

Theorem B.52. A linear map $T \in \mathcal{L}(V, W)$ is injective if and only if $\operatorname{null} T = \{\mathbf{0}\}$.

Definition B.53. The *range* of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ is the subset of $\mathcal{W}$ consisting of those vectors that are of the form $T\mathbf{v}$ for some $\mathbf{v} \in \mathcal{V}$:

$$\operatorname{range} T = \{T\mathbf{v} : \mathbf{v} \in \mathcal{V}\}. \tag{B.10}$$

Example B.54. The range of $A \in \mathbb{C}^{m\times n}$ is the span of its column vectors.

Theorem B.55. The range of a linear map $T \in \mathcal{L}(V, W)$ is a subspace of W.

Theorem B.56 (The counting theorem or the fundamental theorem of linear maps)**.** If $\mathcal{V}$ is a finite-dimensional vector space and $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, then range T is a finite-dimensional subspace of $\mathcal{W}$ and

$$\dim \mathcal{V} = \dim \operatorname{null} T + \dim \operatorname{range} T. \tag{B.11}$$

Theorem B.57. For an operator $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional vector space $\mathcal{V}$, the following are equivalent:

(a) T is invertible;

(b) T is injective;

(c) T is surjective.

B.2.2 The matrix of a linear map

Definition B.58. The *matrix of a linear map* $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ *with respect to the bases* $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ of $\mathcal{V}$ and $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)$ of $\mathcal{W}$, denoted by

$$M_T := M(T, (\mathbf{v}_1, \ldots, \mathbf{v}_n), (\mathbf{w}_1, \ldots, \mathbf{w}_m)), \tag{B.12}$$

is the $m \times n$ matrix $A(T)$ whose entries $a_{i,j} \in \mathbb{F}$ satisfy the linear system

$$\forall j = 1, 2, \ldots, n, \qquad T\mathbf{v}_j = \sum_{i=1}^{m} a_{i,j}\mathbf{w}_i. \tag{B.13}$$

Corollary B.59. The matrix M_T in (B.12) of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ satisfies

$$T[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n] = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m]M_T. \tag{B.14}$$

Proof. This follows directly from (B.13). □

B.2.3 Duality

Dual vector spaces

Definition B.60. The *dual space* of a vector space V is the vector space of all linear functionals on V,

$$V' = \mathcal{L}(V, \mathbb{F}). \tag{B.15}$$

Definition B.61. For a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of a vector space V, its *dual basis* is the list $\varphi_1, \ldots, \varphi_n$ where each $\varphi_j \in V'$ is

$$\varphi_j(\mathbf{v}_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } k \neq j. \end{cases} \tag{B.16}$$

Exercise B.62. Show that the dual basis is a basis of the dual space.

Lemma B.63. A finite-dimensional vector space V satisfies

$$\dim V' = \dim V. \tag{B.17}$$

Proof. This follows from Definition B.60 and the identity $\dim \mathcal{L}(V, W) = \dim(V) \dim(W)$. □

Definition B.64. The *double dual space* of a vector space V, denoted by V'', is the dual space of V'.

Lemma B.65. The function $\Lambda : V \to V''$ defined as

$$\forall v \in V, \forall \varphi \in V', \qquad (\Lambda v)(\varphi) = \varphi(v) \tag{B.18}$$

is a linear bijection.

Proof. It is easily verified that Λ is a linear map. The rest follows from Definitions B.60, B.64, and Lemma B.63. □

Dual linear maps

Definition B.66. The *dual map* of a linear map $T : V \to W$ is the linear map $T' : W' \to V'$ defined as

$$\forall \varphi \in W', \qquad T'(\varphi) = \varphi \circ T. \tag{B.19}$$

Exercise B.67. Denote by D the linear map of differentiation $Dp = p'$ on the vector space $\mathcal{P}(\mathbb{R})$ of polynomials with real coefficients. Under the dual map of D, what is the image of the linear functional $\varphi(p) = \int_0^1 p$ on $\mathcal{P}(\mathbb{R})$?

Theorem B.68. The matrix of T' is the transpose of the matrix of T.

Proof. Let $(\mathbf{v}_1, \cdots, \mathbf{v}_n)$, $(\varphi_1, \cdots, \varphi_n)$, $(\mathbf{w}_1, \cdots, \mathbf{w}_m)$, $(\psi_1, \cdots, \psi_m)$, be bases of V, V', W, W', respectively. Denote by A and C the matrices of $T : V \to W$ and $T' : W' \to V'$, respectively. We have

$$\psi_j \circ T = T'(\psi_j) = \sum_{r=1}^{n} c_{r,j}\varphi_r.$$

By Corollary B.59, applying this equation to $\mathbf{v}_k$ yields

$$(\psi_j \circ T)(\mathbf{v}_k) = \sum_{r=1}^{n} c_{r,j}\varphi_r(\mathbf{v}_k) = c_{k,j}.$$

On the other hand, we have

$$\begin{aligned}(\psi_j \circ T)(\mathbf{v}_k) = \psi_j(T\mathbf{v}_k) &= \psi_j\left(\sum_{r=1}^{m} a_{r,k}w_r\right) \\ &= \sum_{r=1}^{m} a_{r,k}\psi_j(w_r) = a_{j,k}.\end{aligned}$$

□

Definition B.69. The *double dual map* of a linear map $T : V \to W$ is the linear map $T'' : V'' \to W''$ defined as $T'' = (T')'$.

Theorem B.70. For $T \in \mathcal{L}(V)$ and Λ in (B.18), we have

$$T'' \circ \Lambda = \Lambda \circ T. \tag{B.20}$$

Proof. Definition B.69 and equation (B.18) yields

$$\begin{aligned}&\forall v \in V, \ \forall \varphi \in V', \\ &(T'' \circ \Lambda)v\varphi = ((T')'\Lambda v)\varphi = (\Lambda v \circ T')\varphi = \Lambda v(T'\varphi) \\ &\qquad = (T'\varphi)(v) = \varphi(Tv) = \Lambda(Tv)(\varphi) \\ &\qquad = (\Lambda \circ T)v\varphi,\end{aligned}$$

where the third step is natural since T' send V' to V'. □

Corollary B.71. For $T \in \mathcal{L}(V)$ where V is finite-dimensional, the double dual map is

$$T'' = \Lambda \circ T \circ \Lambda^{-1}. \tag{B.21}$$

Proof. This follows directly from Theorem B.70 and Lemma B.65. □

The null space and range of the dual of a linear map

Definition B.72. For $U \subset V$, the *annihilator* of U, denoted U^0, is defined by

$$U^0 := \{\varphi \in V' : \ \forall \mathbf{u} \in U, \ \varphi(\mathbf{u}) = 0\}. \tag{B.22}$$

Exercise B.73. Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5$ denote the standard basis of $V = \mathbb{R}^5$, and $\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5$ its dual basis of V'. Suppose

$$U = \text{span}(\mathbf{e}_1, \mathbf{e}_2) = \{(x_1, x_2, 0, 0, 0) \in \mathbb{R}^5 : x_1, x_2 \in \mathbb{R}\}.$$

Show that $U^0 = \text{span}(\varphi_3, \varphi_4, \varphi_5)$.

Exercise B.74. Let $i : U \hookrightarrow V$ be an inclusion. Show that $\text{null} i' = U^0$.

Lemma B.75. Suppose $U \subset V$. Then U^0 is a subspace of V'.

Exercise B.76. Suppose V is finite-dimensional. Prove every linear map on a subspace of V can be extended to a linear map on V.

Lemma B.77. Suppose V is finite-dimensional and U is a subspace of V. Then

$$\dim U + \dim U^0 = \dim V. \tag{B.23}$$

Proof. Apply Theorem B.56 to the dual of an inclusion $i' : V' \to U'$ and we have

$$\begin{aligned} &\dim \text{range} i' + \dim \text{null} i' = \dim V' \\ \Rightarrow &\dim \text{range} i' + \dim U^0 = \dim V, \end{aligned}$$

where the second line follows from Exercise B.74 and Lemma B.63. For any $\varphi \in U'$, Exercise B.76 states that $\varphi \in U'$ can be extended to $\psi \in V'$ such that $i'(\psi) = \varphi$. Hence i' is surjective and we have $U' = \text{range} i'$. The proof is then completed by Lemma B.63. □

Lemma B.78. Any linear map $T \in \mathcal{L}(V, W)$ satisfies

$$\text{null} T' = (\text{range} T)^0. \tag{B.24}$$

Proof. Definitions B.50, B.53, B.66, and B.72 yield

$$\begin{aligned} \varphi \in \text{null} T' &\Leftrightarrow 0 = T'(\varphi) = \varphi \circ T \\ &\Leftrightarrow \forall v \in V, \varphi(Tv) = 0 \\ &\Leftrightarrow \varphi(\text{range} T) = 0 \\ &\Leftrightarrow \varphi \in (\text{range} T)^0. \end{aligned}$$

□

Lemma B.79. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V, W)$ satisfies

$$\dim \text{null} T' = \dim \text{null} T + \dim W - \dim V. \tag{B.25}$$

Proof. Lemma B.78 and Theorem B.56 yield

$$\begin{aligned} \dim \text{null} T' &= \dim(\text{range} T)^0 = \dim W - \dim(\text{range} T) \\ &= \dim W - \dim V + \dim(\text{null} T) \\ &= \dim \text{null} T + \dim W - \dim V. \end{aligned}$$

□

Corollary B.80. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V, W)$ is surjective if and only if T' is injective.

Proof. T is surjective $\Leftrightarrow W = \text{range} T \Leftrightarrow (\text{range} T)^0 = \{0\}$ $\Leftrightarrow \text{null} T' = \{0\} \Leftrightarrow T'$ is injective. The second step follows from Lemma B.77 applied to W:

$$\dim W = \dim(\text{range} T) + \dim(\text{range} T)^0.$$ □

Lemma B.81. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V, W)$ satisfies

$$\dim \text{range} T' = \dim \text{range} T. \tag{B.26}$$

Proof. Theorem B.56, Lemma B.78, and Lemma B.77 yield

$$\begin{aligned} \dim \text{range} T' &= \dim W - \dim \text{null} T' \\ &= \dim W - \dim(\text{range} T)^0 \\ &= \dim(\text{range} T). \end{aligned}$$

□

Lemma B.82. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V, W)$ satisfies

$$\text{range} T' = (\text{null} T)^0. \tag{B.27}$$

Proof. Theorem B.56, Lemma B.78, and Lemma B.77 yield

$$\begin{aligned} \varphi \in \text{range} T' &\Rightarrow \exists \psi \in W' \text{ s.t } T'(\psi) = \varphi \\ &\Rightarrow \forall v \in \text{null} T, \varphi(v) = \psi(Tv) = 0 \\ &\Rightarrow \varphi \in (\text{null} T)^0. \end{aligned}$$

The proof is completed by

$$\begin{aligned} \dim \text{range} T' &= \dim(\text{range} T) \\ &= \dim V - \dim \text{null} T \\ &= \dim(\text{null} T)^0. \end{aligned}$$

□

Corollary B.83. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V, W)$ is injective if and only if T' is surjective.

Proof. T is injective $\Leftrightarrow \text{null} T = \{0\} \Leftrightarrow (\text{null} T)^0 = V' \Leftrightarrow$ $\text{range} T' = V' \Leftrightarrow T'$ is surjective. The second step follows from Lemmas B.77 and B.63, and the third step follows from Lemma B.82. □

Matrix ranks

Definition B.84. For a matrix $A \in \mathbb{F}^{m \times n} : \mathbb{F}^n \to \mathbb{F}^m$, its *column space* (or range or image) consists of all linear combinations of its columns, its *row space* (or coimage) is the column space of A^T, its *null space* (or kernel) is the null space of A as a linear operator, and the *left null space* (or cokernel) is the null space of A^T.

Definition B.85. The *column rank* and *row rank* of a matrix $A \in \mathbb{F}^{m \times n}$ is the dimension of its column space and row space, respectively.

Lemma B.86. Let A_T denote the matrix of a linear operator $T \in \mathcal{L}(V, W)$. Then the column rank of A_T is the dimension of range$T$.

Proof. For $\mathbf{u} = \sum_i c_i \mathbf{v}_i$, Corollary B.59 yields

$$T\mathbf{u} = \sum_i c_i T\mathbf{v}_i = T[\mathbf{v}_1, \dots, \mathbf{v}_n]\mathbf{c} = [\mathbf{w}_1, \dots, \mathbf{w}_m]A_T\mathbf{c}.$$

Hence we have

$$\{T\mathbf{u} : \mathbf{c} \in \mathbb{F}^n\} = \{[\mathbf{w}_1, \dots, \mathbf{w}_m]A_T\mathbf{c} : \mathbf{c} \in \mathbb{F}^n\}.$$

The LHS is range$T$ while $\{A_T\mathbf{c} : \mathbf{c} \in \mathbb{F}^n\}$ is the column space of A_T. Since $(\mathbf{w}_1, \dots, \mathbf{w}_m)$ is a basis, by Definition B.85 the column rank of the matrix $[\mathbf{w}_1, \dots, \mathbf{w}_m]$ is m. Taking dim to both sides of the above equation yields the conclusion. Note that the RHS is a subspace of $\mathbb{F}^m$ (why?) and the dimension of it does not depend on the special choice of its basis, hence we can choose $(\mathbf{w}_1, \dots, \mathbf{w}_m)$ to be the standard basis and then $[\mathbf{w}_1, \dots, \mathbf{w}_m]$ is simply the identity matrix. □

Theorem B.87. For any $A \in \mathbb{F}^{m\times n}$, its row rank equals its column rank.

Proof. Define a linear map $T : \mathbb{F}^n \to \mathbb{F}^m$ as $T\mathbf{x} = A\mathbf{x}$. Clearly, A is the matrix of T for the standard bases of $\mathbb{F}^n$ and $\mathbb{F}^m$. Then we have,

$$\begin{aligned}\text{column rank of } A &= \dim \text{range} T \\ &= \dim \text{range} T' \\ &= \text{column rank of the matrix of } T' \\ &= \text{column rank of } A^T \\ &= \text{row rank of } A,\end{aligned}$$

where the first step follows from Lemma B.86, the second from Lemma B.81, the third from Lemma B.86, the fourth from Theorem B.68, and the last from the definition of matrix transpose and matrix products. □

Definition B.88. The *rank* of a matrix is its column rank.

Theorem B.89 (Fundamental theorem of linear algebra)**.** For a matrix $A \in \mathbb{F}^{m\times n} : \mathbb{F}^n \to \mathbb{F}^m$, its column space and row space both have dimension $r \le \min(m, n)$; its null space and left null space have dimensions $n - r$ and $m - r$, respectively. In addition, we have

$$\mathbb{F}^m = \text{range} A \oplus \text{null} A^T, \tag{B.28a}$$

$$\mathbb{F}^n = \text{range} A^T \oplus \text{null} A, \tag{B.28b}$$

where range$A \perp$ nullA^T and range$A^T \perp$ null$A$.

Proof. The first sentence is a rephrase of Theorem B.87 and follows from Theorem B.56. For the second sentence, we only prove (B.28b). $\mathbf{x} \in$ null$A$ implies $\mathbf{x} \in \mathbb{F}^n$ and $A\mathbf{x} = \mathbf{0}$. The latter expands to

$$\begin{bmatrix}\mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \cdots \\ \mathbf{a}_m^T\end{bmatrix}\mathbf{x} = \begin{bmatrix}0 \\ 0 \\ \cdots \\ 0\end{bmatrix},$$

which implies that $\forall j = 1, 2, \dots, m$, $\mathbf{a}_j \perp \mathbf{x}$. Hence $\mathbf{x}$ is orthogonal to each basis vector of rangeA^T. The rest of the proof follows from Lemma B.81, Theorem B.68, Theorem B.56. □

B.3 Eigenvalues, eigenvectors, and invariant subspaces

B.3.1 Invariant subspaces

Definition B.90. Under a linear operator $T \in \mathcal{L}(\mathcal{V})$, a subspace $\mathcal{U}$ of $\mathcal{V}$ is *invariant* if $\mathbf{u} \in \mathcal{U}$ implies $T\mathbf{u} \in \mathcal{U}$.

Example B.91. Under $T \in \mathcal{L}(\mathcal{V})$, each of the following subspaces of $\mathcal{V}$ is invariant: $\{\mathbf{0}\}$, $\mathcal{V}$, null$\,T$, and range$\,T$.

Definition B.92. A number $\lambda \in \mathbb{F}$ is called an *eigenvalue of an operator* $T \in \mathcal{L}(\mathcal{V})$ if there exists $\mathbf{v} \in \mathcal{V}$ such that $T\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{v} \neq \mathbf{0}$. Then the vector $\mathbf{v}$ is called an *eigenvector* of T corresponding to λ.

Lemma B.93. Suppose V is finite-dimensional. $\lambda \in V$ is an eigenvalue of $T \in \mathcal{L}(V)$ if and only if $T - \lambda I$ is not injective.

Proof. This follows directly from Definition B.92. □

Example B.94. For each eigenvector $\mathbf{v}$ of $T \in \mathcal{L}(\mathcal{V})$, the subspace span$(\mathbf{v})$ is a one-dimensional invariant subspace of $\mathcal{V}$.

Lemma B.95. Suppose $\lambda_1, \dots, \lambda_m$ are distinct eigenvalues of $T \in \mathcal{L}(\mathcal{V})$ with corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_m$. Then $\mathbf{v}_1, \dots, \mathbf{v}_m$ is linearly independent.

Lemma B.96. Suppose $\mathcal{V}$ is finite-dimensional. Then each operator on $\mathcal{V}$ has at most dim$\,\mathcal{V}$ distinct eigenvalues.

Definition B.97. Suppose $T \in \mathcal{L}(V)$ and U is an invariant subspace of V under T. The *restriction operator* $T|_U \in \mathcal{L}(U)$ is defined by

$$\forall \mathbf{u} \in U, \quad T|_U(\mathbf{u}) = T\mathbf{u}. \tag{B.29}$$

B.3.2 Existence of eigenvalues

Notation 11. Suppose $T \in \mathcal{L}(\mathcal{V})$ and $p \in \mathbb{P}(\mathbb{F})$ is a polynomial given by

$$p(z) = a_0 + a_1 z + \cdots + a_m z^m$$

for $z \in \mathbb{F}$. Then $p(T)$ is the operator given by

$$p(T) = a_0 I + a_1 T + \cdots + a_m T^m,$$

where $I = T^0$ is the identity operator.

Example B.98. Suppose $D \in \mathcal{L}(\mathbb{P}(\mathbb{R}))$ is the differentiation operator defined by $Dq = q'$ and p is the polynomial defined by $p(x) = 7 - 3x + 5x^2$. Then we have

$$p(D) = 7 - 3D + 5D^2, \qquad (p(D))q = 7q - 3q' + 5q''.$$

Definition B.99. The *product polynomial* of two polynomials $p, q \in \mathbb{P}(\mathbb{F})$ is the polynomial defined by

$$\forall z \in \mathbb{F}, \qquad (pq)(z) := p(z)q(z). \tag{B.30}$$

Lemma B.100. Any $T \in \mathcal{L}(\mathcal{V})$ and $p, q \in \mathbb{P}(\mathbb{F})$ satisfy

$$(pq)(T) = p(T)q(T) = q(T)p(T). \tag{B.31}$$

Theorem B.101 (Existence of eigenvalues)**.** Every operator $T \in \mathcal{L}(V)$ on a finite-dimensional, nonzero, complex vector space V has an eigenvalue.

Proof. Write $n := \dim V$. For a nonzero $\mathbf{v} \in V$, the $n+1$ vectors $(\mathbf{v}, T\mathbf{v}, T\mathbf{v}, \cdots, T^n\mathbf{v})$ must be linear dependent, i.e.,

$$\mathbf{0} = a_0\mathbf{v} + a_1T\mathbf{v} + \cdots + a_nT^n\mathbf{v} = (a_0 + a_1T + \cdots + a_nT^n)\mathbf{v}$$

implies that there exists $j \in [1, n]$ such that $a_j \neq 0$. By the fundamental theorem of algebra, the polynomial $\sum_{i=0}^{n} a_iT^i$ has m roots, say, $\lambda_1, \ldots, \lambda_m$, and thus for some $c \in \mathbb{C} \setminus \{0\}$, we have

$$\mathbf{0} = c(T - \lambda_1 I)(T - \lambda_2 I) \cdots (T - \lambda_m I)\mathbf{v}.$$

Hence $T - \lambda_j I$ is not injective for some λ_j and the proof is completed by Lemma B.93. □

B.3.3 Upper-triangular matrices

Definition B.102. The *matrix of a linear operator* $T \in \mathcal{L}(\mathcal{V})$ is the matrix of the linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{V})$, c.f. Definition B.58.

Theorem B.103. Suppose $T \in \mathcal{L}(\mathcal{V})$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathcal{V}$. Then the following are equivalent:

(a) the matrix of T with respect to $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is upper triangular;

(b) $T\mathbf{v}_j \in \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_j)$ for each $j = 1, \ldots, n$;

(c) $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_j)$ is invariant under T for each $j = 1, \ldots, n$.

Theorem B.104. Every linear operator $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional complex vector space $\mathcal{V}$ has an upper-triangular matrix with respect to some basis of $\mathcal{V}$.

Theorem B.105. Suppose $T \in \mathcal{L}(\mathcal{V})$ has an upper-triangular matrix with respect to some basis of $\mathcal{V}$. Then T is invertible if and only if all the entries on the diagonal of that upper-triangular matrix are nonzero.

Theorem B.106. Suppose $T \in \mathcal{L}(\mathcal{V})$ has an upper-triangular matrix with respect to some basis of $\mathcal{V}$. Then the eigenvalues of T are precisely the entries on the diagonal of that upper-triangular matrix.

B.3.4 Eigenspaces and diagonal matrices

Definition B.107. A *diagonal entry* of a matrix is an entry of the matrix of which the row index equals the column index. The *diagonal* of a matrix consists of all diagonal entries of the matrix. A *diagonal matrix* is a square matrix that is zero everywhere except possibly along the diagonal.

Definition B.108. The *eigenspace* of $T \in \mathcal{L}(\mathcal{V})$ corresponding to $\lambda \in \mathbb{F}$ is

$$E(\lambda, T) := \operatorname{null}(T - \lambda I). \tag{B.32}$$

Lemma B.109. Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional space $\mathcal{V}$. Then

$$E(\lambda_1, T) + \cdots + E(\lambda_m, T)$$

is a direct sum and

$$\dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) \leq \dim \mathcal{V}. \tag{B.33}$$

Definition B.110. An operator $T \in \mathcal{L}(\mathcal{V})$ is *diagonalizable* if it has a diagonal matrix with respect to some basis of $\mathcal{V}$.

Theorem B.111 (Conditions of diagonalizability)**.** Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional space $\mathcal{V}$. Then the following are equivalent:

(a) T is diagonalizable;

(b) $\mathcal{V}$ has a basis consisting of eigenvectors of T;

(c) there exist one-dimensional subspaces $U_1, \ldots, U_n$ of $\mathcal{V}$, each invariant under T, such that $\mathcal{V} = U_1 \oplus \cdots \oplus U_n$;

(d) $\mathcal{V} = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T)$;

(e) $\dim \mathcal{V} = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T)$.

Corollary B.112. An operator $T \in \mathcal{L}(\mathcal{V})$ is diagonalizable if T has $\dim \mathcal{V}$ distinct eigenvalues.

B.4 Operators on complex vector spaces

B.4.1 Generalized eigenvectors

Lemma B.113. For a linear operator $T \in \mathcal{L}(V)$, we have

$$\{\mathbf{0}\} = \operatorname{null} T^0 \subseteq \operatorname{null} T^1 \subseteq \cdots \subseteq \operatorname{null} T^k \subseteq \operatorname{null} T^{k+1} \subseteq \cdots. \tag{B.34}$$

Proof. Suppose $\mathbf{v} \in \operatorname{null} T^k$ for $k \in \mathbb{N}$. Thus $T^k\mathbf{v} = 0$. Then $T^{k+1}\mathbf{v} = TT^k\mathbf{v} = T\mathbf{0} = \mathbf{0}$ and therefore $\mathbf{v} \in \operatorname{null} T^{k+1}$. □

Lemma B.114. Suppose a linear operator $T \in \mathcal{L}(V)$ satisfies $\operatorname{null} T^m = \operatorname{null} T^{m+1}$. Then we have

$$\operatorname{null} T^m = \operatorname{null} T^{m+1} = \operatorname{null} T^{m+2} = \cdots. \tag{B.35}$$

Proof. By Lemma B.113, it suffices to show

$$\forall k \in \mathbb{N}^+, \quad \operatorname{null} T^{m+k+1} \subseteq \operatorname{null} T^{m+k},$$

which indeed holds because

$$\begin{aligned} \mathbf{v} \in \operatorname{null} T^{m+k+1} &\Rightarrow \quad T^{m+k+1}\mathbf{v} = \mathbf{0} \;\Rightarrow\; T^{m+1}(T^k\mathbf{v}) = \mathbf{0} \\ &\Rightarrow \quad T^k\mathbf{v} \in \operatorname{null} T^{m+1} = \operatorname{null} T^m \\ &\Rightarrow \quad T^mT^k\mathbf{v} = \mathbf{0} \;\Rightarrow\; T^{m+k}\mathbf{v} = \mathbf{0} \\ &\Rightarrow \quad \mathbf{v} \in \operatorname{null} T^{m+k}. \quad \square \end{aligned}$$

Lemma B.115. A linear operator $T \in \mathcal{L}(V)$ satisfies

$$\operatorname{null} T^n = \operatorname{null} T^{n+1} = \operatorname{null} T^{n+2} = \cdots, \tag{B.36}$$

where $n = \dim V$.

Proof. By Lemma B.114, it suffices to show

$$\operatorname{null} T^n = \operatorname{null} T^{n+1}.$$

Suppose this is not true. Then Lemmas B.113 and B.114 yield

$$\{\mathbf{0}\} = \operatorname{null} T^0 \subset \operatorname{null} T^1 \subset \cdots \subset \operatorname{null} T^n \subset \operatorname{null} T^{n+1},$$

where the symbol"$\subset$" means strict inclusion, c.f. Definition A.2. At each strict inclusion in the above chain, the dimension of the space increases by at least 1, and thus $\dim \operatorname{null} T^{n+1} > n$. But the dimension of any subspace of V cannot exceed that of V. □

Theorem B.116. A linear operator $T \in \mathcal{L}(V)$ satisfies

$$V = \operatorname{null} T^n \oplus \operatorname{range} T^n \tag{B.37}$$

where $n = \dim V$.

Proof. We first show $\operatorname{null} T^n \cap \operatorname{range} T^n = \{\mathbf{0}\}$. Indeed, if $\mathbf{v} \in \operatorname{null} T^n \cap \operatorname{range} T^n$, then $T^n\mathbf{v} = \mathbf{0}$ and there exists $\mathbf{u}$ such that $\mathbf{v} = T^n\mathbf{u}$. Hence $T^{2n}\mathbf{u} = \mathbf{0}$. Lemma B.115 further implies $T^n\mathbf{u} = \mathbf{0}$ and thus $\mathbf{v} = \mathbf{0}$.

By Theorem B.13, $\operatorname{null} T^n + \operatorname{range} T^n$ is a direct sum. Then (B.37) follows from

$$\begin{aligned}\dim(\operatorname{null} T^n \oplus \operatorname{range} T^n) &= \dim \operatorname{null} T^n + \dim \operatorname{range} T^n \\ &= \dim V,\end{aligned}$$

where the second step follows from the fundamental theorem of linear maps (Theorem B.56). □

Example B.117. For the operator $T \in \mathcal{L}(\mathbb{C}^3)$ given by

$$T(z_1, z_2, z_3) = (4z_2, 0, 5z_3), \tag{B.38}$$

$\operatorname{null} T + \operatorname{range} T$ is not a direct sum of $\mathbb{C}^3$ because

$$\begin{aligned}&\operatorname{null} T = \{(z_1, 0, 0) : z_1 \in \mathbb{C}\},\\ &\operatorname{range} T = \{(z_1, 0, z_3) : z_1, z_3 \in \mathbb{C}\}.\end{aligned}$$

In contrast, $T^3(z_1, z_2, z_3) = (0, 0, 125z_3)$ and thus

$$\begin{aligned}&\operatorname{null} T^3 = \{(z_1, z_2, 0) : z_1, z_2 \in \mathbb{C}\},\\ &\operatorname{range} T^3 = \{(0, 0, z_3) : z_3 \in \mathbb{C}\},\\ &\operatorname{null} T^3 \oplus \operatorname{range} T^3 = \mathbb{C}^3.\end{aligned}$$

Definition B.118. A *generalized eigenvector of a linear operator* $T \in \mathcal{L}(V)$ corresponding to the eigenvalue λ of T is a nonzero vector $\mathbf{v} \in V$ satisfying

$$\exists j \in \mathbb{N}^+ \text{ s.t. } (T - \lambda I)^j \mathbf{v} = \mathbf{0}. \tag{B.39}$$

Definition B.119. The *generalized eigenspace of a linear operator* $T \in \mathcal{L}(V)$ corresponding to the eigenvalue λ of T, denoted $G(\lambda, T)$, is the set of all generalized eigenvectors of T corresponding to λ along with the zero vector.

Lemma B.120. A generalized eigenspace $G(\lambda, T)$ satisfies

$$\forall T \in \mathcal{L}(V), \forall \lambda \in \mathbb{F},\ G(\lambda, T) = \operatorname{null}(T - \lambda I)^{\dim V}. \tag{B.40}$$

Proof. Suppose $\mathbf{v} \in \operatorname{null}(T - \lambda I)^{\dim V}$. Then Definitions B.118 and B.119 imply $\mathbf{v} \in G(\lambda, T)$. Conversely, $\mathbf{v} \in G(\lambda, T)$ implies that $(T - \lambda I)^j \mathbf{v} = \mathbf{0}$ for some $j \in \mathbb{N}^+$. Then we have $\mathbf{v} \in \operatorname{null}(T - \lambda I)^{\dim V}$ from Lemmas B.113 and B.115. □

Definition B.121. The *multiplicity or algebraic multiplicity of an eigenvalue* λ of an operator T is the dimension of the corresponding generalized eigenspace,

$$m_a(\lambda) := \dim G(\lambda, T) = \dim \operatorname{null}(T - \lambda I)^{\dim V} \tag{B.41}$$

while the *geometric multiplicity of an eigenvalue* λ of T is the dimension of the corresponding eigenspace,

$$m_g(\lambda) := \dim E(\lambda, T) = \dim \operatorname{null}(T - \lambda I). \tag{B.42}$$

The *index of an eigenvalue* λ of T is the smallest integer k for which

$$\operatorname{null}(T - \lambda I)^k = \operatorname{null}(T - \lambda I)^{k+1}. \tag{B.43}$$

Corollary B.122. Geometric multiplicity and algebraic multiplicity satisfy

$$1 \le m_g(\lambda) \le m_a(\lambda). \tag{B.44}$$

Proof. The case $j = 1$ in Definition B.118 implies that every eigenvector of T is a generalized eigenvector of T. Hence

$$\forall T \in \mathcal{L}(V),\ \forall \lambda \in \mathbb{F},\ E(\lambda, T) \subseteq G(\lambda, T).$$

Then Definition B.121 completes the proof. □

Definition B.123. An eigenvalue λ of A is *defective* iff

$$m_g(\lambda) < m_a(\lambda). \tag{B.45}$$

A is *defective* iff A has one or more defective eigenvalues.

Example B.124. Eigenvalues of the operator $T \in \mathcal{L}(\mathbb{C}^3)$ in (B.38) are 0 and 5, with the corresponding eigenspaces as

$$\begin{aligned}&E(0, T) = \{(z_1, 0, 0) : z_1 \in \mathbb{C}\},\\ &E(5, T) = \{(0, 0, z_3) : z_3 \in \mathbb{C}\}.\end{aligned}$$

The generalized eigenspaces are deduced as follows,

$$\begin{aligned}&T^3(z_1, z_2, z_3) = (0, 0, 125z_3),\\ &G(0, T) = \{(z_1, z_2, 0) : z_1, z_2 \in \mathbb{C}\},\\ &(T - 5I)^3(z_1, z_2, z_3) = (-125z_1 + 300z_2, -125z_2, 0),\\ &G(5, T) = E(5, T) = \{(0, 0, z_3) : z_3 \in \mathbb{C}\}.\end{aligned}$$

Since $m_g(5) = 1 = m_a(5)$ and $m_g(0) = 1 < m_a(0) = 2$, the eigenvalue 5 is not defective but the eigenvalue 0 is. Hence the operator T is defective.

Lemma B.125. Generalized eigenvectors of distinct eigenvalues of an operator $T \in \mathcal{L}(V)$ are linearly independent.

Proof. Let $(\lambda_i, \mathbf{v}_i)$ be a generalized eigenpair of T. Define

$$\mathbf{w} := (T - \lambda_1 I)^k \mathbf{v}_1,$$

where $k \in \mathbb{N}$ is the largest integer such that $\mathbf{w} \neq \mathbf{0}$. By Definition B.118, $(T - \lambda_1 I)\mathbf{w} = \mathbf{0}$ and thus $\mathbf{w}$ is an eigenvector of T. Consequently, we have

$$(*): \quad \forall j \in \mathbb{N}^+,\ \forall \lambda \in \mathbb{C}, \quad (T - \lambda I)^j \mathbf{w} = (\lambda_1 - \lambda)^j \mathbf{w}.$$

Write $n := \dim V$ and define a polynomial of T by

$$p(T) = (T - \lambda_1 I)^k (T - \lambda_2 I)^n \cdots (T - \lambda_m I)^n;$$

By Lemma B.100, the factors of $p(T)$ commute.

Suppose $\mathbf{0} = \sum_{i=1}^m a_i \mathbf{v}_i$ where each $a_i \in \mathbb{C}$. Then the application of $p(T)$ to this equation yields $a_1 = 0$ because of the distinctness of the eigenvalues and

$$\begin{aligned} \mathbf{0} &= a_1 (T - \lambda_1 I)^k (T - \lambda_2 I)^n \cdots (T - \lambda_m I)^n \mathbf{v}_1 \\ &= a_1 (T - \lambda_2 I)^n \cdots (T - \lambda_m I)^n \mathbf{w} \\ &= a_1 (\lambda_1 - \lambda_2)^n \cdots (\lambda_1 - \lambda_m)^n \mathbf{w}, \end{aligned}$$

where the first equality follows from Lemma B.120, the second from the definition of $\mathbf{w}$, and the third from $(*)$.

Similarly, the above arguments applied to the other generalized eigenpairs yield $a_2 = \cdots = a_m = 0$, which implies the linear independence of generalized eigenvectors. □

B.4.2 Nilpotent operators

Definition B.126. An operator $N \in \mathcal{L}(V)$ is *nilpotent* iff $N^k = \mathbf{0}$ for some $k \in \mathbb{N}^+$.

Example B.127. The differentiation operator on the vector space of polynomials of degree at most m is nilpotent. The operator $N \in \mathcal{L}(\mathbb{F}^3)$ given by $N(x, y, z) = (y, z, 0)$ is nilpotent because $N^3 = \mathbf{0}$.

Lemma B.128. A nilpotent operator $N \in \mathcal{L}(V)$ satisfies

$$N^{\dim V} = \mathbf{0}. \tag{B.46}$$

Proof. Definitions B.126 and B.119 imply $G(0, N) = V$. The rest follows from Lemma B.120. □

Lemma B.129. Any nilpotent operator $N \in \mathcal{L}(V)$ has a strictly upper triangular matrix M, i.e., $\forall i \geq j,\ M_{i,j} = 0$.

Proof. Write $n := \dim V$. Choose a basis of $\operatorname{null} N$, extend it to a basis of $\operatorname{null} N^2$, …, $\operatorname{null} N^n$, and we have a matrix U whose columns are the vectors of this basis:

$$U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n].$$

Corollary B.59 states that

$$N[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] M.$$

The fact of $\mathbf{u}_1 \in \operatorname{null} N$ dictates that all entries in the first column of M must be zero. Let $\mathbf{u}_j$ be the first basis vector in $\operatorname{null} N^2$ but not in $\operatorname{null} N$. Then $N\mathbf{u}_j \in \operatorname{null} N$, i.e., $N\mathbf{u}_j$ must be a linear combination of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{j-1}$. Hence the first $j - 1$ entries in the jth column of M could be nonzero while all other entries must be zero. Proceeding in this fashion completes the proof. □

Lemma B.130. If an operator $N \in \mathcal{L}(V)$ is nilpotent, then $I + N$ has a square root, i.e., there exists an operator $M \in \mathcal{L}(V)$ such that $M^2 = I + N$.

Proof. By Definition B.126, $N^m = \mathbf{0}$ for some $m \in \mathbb{N}^+$. For operators of the form

$$M = I + \sum_{i=1}^{m-1} a_i N^i,$$

$M^2 = I + N$ yields $\left(I + \sum_{i=1}^{m-1} a_i N^i\right)^2 = I + N$, which is equivalent to

$$2a_1 = 1,\ 2a_2 + a_1^2 = 0,\ 2a_3 + 2a_1 a_2 = 0,\ \ldots,$$

where $a_1 = \frac{1}{2}$ and each a_j can be uniquely determined from $a_1, a_2, \ldots, a_{j-1}$. □

B.4.3 Operator decomposition

Lemma B.131. Suppose $p(T)$ is a polynomial of an operator $T \in \mathcal{L}(V)$. Then both $\operatorname{null} p(T)$ and $\operatorname{range} p(T)$ are invariant under T.

Proof. $\mathbf{v} \in \operatorname{null} p(T)$ implies $p(T)\mathbf{v} = \mathbf{0}$ and

$$p(T)(T\mathbf{v}) = T(p(T)\mathbf{v}) = T\mathbf{0} = \mathbf{0}.$$

Hence $T\mathbf{v} \in \operatorname{null} p(T)$, and, by Definition B.90, $\operatorname{null} p(T)$ is invariant under T. Similarly,

$$\mathbf{v} \in \operatorname{range} p(T) \ \Rightarrow\ \exists \mathbf{u} \in V \text{ s.t. } p(T)\mathbf{u} = \mathbf{v}$$

and thus

$$T\mathbf{v} = Tp(T)\mathbf{u} = p(T)(T\mathbf{u}).$$

Since $T\mathbf{u} \in V$, we have $T\mathbf{v} \in \operatorname{range} p(T)$ and, by Definition B.90, $\operatorname{range} p(T)$ is invariant under T. □

Theorem B.132 (Decomposing operators on complex vector spaces)**.** Suppose V is a complex vector space, $T \in \mathcal{L}(V)$, and $\lambda_1, \lambda_2, \ldots, \lambda_m$ are all the distinct eigenvalues of T. Then

(DOC-1) $V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T)$;

(DOC-2) each $G(\lambda_j, T)$ is invariant under T;

(DOC-3) each $(T - \lambda_j I)|_{G(\lambda_j, T)}$ is nilpotent.

Proof. Write $n := \dim V$. Lemma B.120 states that for each $j = 1, 2, \ldots, m$ we have $G(\lambda_j, T) = \operatorname{null}(T - \lambda_j I)^n$. Then (DOC-3) follows from Definition B.126 and (DOC-2) follows from Lemma B.131 by choosing $p(z) = (z - \lambda_j)^n$.

(DOC-1) can be proven by an induction on n. The induction basis of $n = 1$ clearly holds. As the induction hypothesis, (DOC-1) holds on all vector spaces of which the dimension is smaller than n.

By Theorem B.101, T has an eigenvalue λ_1. The application of Lemma B.116 to $T - \lambda_1 I$ yields

$$(*): \quad V = G(\lambda_1, T) \oplus U,$$

where $U = \operatorname{range}(T - \lambda_1 I)^n$. If U is empty, (DOC-1) holds trivially; otherwise Lemma B.131 implies that U is invariant

under T, furnishing the restriction operator $T|_U \in \mathcal{L}(U)$ in (B.29). By $(*)$ and Theorem B.101, $T|_U$ has distinct eigenvalues $\lambda_2, \ldots, \lambda_m$, each of which is different from λ_1. Since $\dim G(\lambda_1, T) \geq 1$, we have $\dim U < n$ and thus we can apply the induction hypothesis to U to obtain

$$(**): \quad U = G(\lambda_2, T|_U) \oplus \cdots \oplus G(\lambda_m, T|_U).$$

Combining $(**)$ with $(*)$ completes the proof if we can show

$$\forall j = 2, \ldots, m, \quad G(\lambda_j, T|_U) = G(\lambda_j, T).$$

Indeed, apply $\cap G(\lambda_j, T)$ to both sides of $(*)$ and we have

$$\begin{aligned} G(\lambda_j, T) &= (G(\lambda_1, T) \cap G(\lambda_j, T)) \oplus (U \cap G(\lambda_j, T)) \\ &= G(\lambda_j, T|_U), \end{aligned}$$

where the second step follows from Lemma B.125 and the identity $G(\lambda_j, T|_U) = G(\lambda_j, T) \cap U$. □

Example B.133. For the operator $T \in \mathcal{L}(\mathbb{C}^3)$ in (B.38), T does not have enough eigenvectors to span $\mathbb{C}^3$. Example B.124 shows that

$$\begin{aligned} E(0, T) \oplus E(5, T) &\subset \mathbb{C}^3, \\ G(0, T) \oplus G(5, T) &= \mathbb{C}^3. \end{aligned}$$

Corollary B.134. Suppose V is a complex vector space and $T \in \mathcal{L}(V)$. Then there is a basis of V that consists of generalized eigenvectors of T.

Proof. This follows from (DOC-1) in Theorem B.132. □

Definition B.135. A *block diagonal matrix* A is a square matrix of the form

$$\begin{pmatrix} A_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & A_m \end{pmatrix}, \tag{B.47}$$

where $A_1, \ldots, A_m$ are square matrices along the diagonal and all other entries of A is 0.

Theorem B.136. Suppose V is a complex vector space, $T \in \mathcal{L}(V)$, and $\lambda_1, \lambda_2, \ldots, \lambda_m$ are all the distinct eigenvalues of T with multiplicities $d_1, d_2, \ldots, d_m$. Then there is a basis of V with respect to which T has a block diagonal form (B.47) where each A_j is a d_j-by-d_j upper triangular matrix of the form

$$\begin{bmatrix} \lambda_j & & * \\ & \ddots & \\ 0 & & \lambda_j \end{bmatrix}. \tag{B.48}$$

Proof. By Theorem B.132, each $(T - \lambda_j I)|_{G(\lambda_j, T)}$ is nilpotent. By Lemma B.129, we can choose a basis of $G(\lambda_j, T)$ such that the matrix of $(T - \lambda_j I)|_{G(\lambda_j, T)}$ is strictly upper triangular. Then the form of (B.48) follows from

$$T|_{G(\lambda_j, T)} = (T - \lambda_j I)|_{G(\lambda_j, T)} + \lambda_j I|_{G(\lambda_j, T)}.$$

The rest follows from (DOC-1) in Theorem B.132. □

Example B.137. For $T \in \mathcal{L}(\mathbb{F}^3)$ given by

$$T(x, y, z) = (6x + 3y + 4z, 6y + 2z, 7z), \tag{B.49}$$

the matrix of T with respect to the standard basis is

$$T_S = \begin{pmatrix} 6 & 3 & 4 \\ 0 & 6 & 2 \\ 0 & 0 & 7 \end{pmatrix}.$$

It is readily verified that the eigenvalues of T are 6 and 7, with the corresponding generalized eigenspaces as

$$\begin{aligned} G(6, T) &= \operatorname{span}\{(1, 0, 0), (0, 1, 0)\}; \\ G(7, T) &= \operatorname{span}\{(10, 2, 1)\}. \end{aligned}$$

The matrix of T with respect to the basis

$$\{\mathbf{v}_1 = (1, 0, 0), \mathbf{v}_2 = (0, 1, 0), \mathbf{v}_3 = (10, 2, 1)\}$$

is of the block diagonal form:

$$T_B = \begin{pmatrix} \begin{bmatrix} 6 & 3 \\ 0 & 6 \end{bmatrix} & 0 \\ 0 \quad 0 & [7] \end{pmatrix}.$$

More precisely, by Corollary B.59, we have

$$T_S[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] T_B.$$

Corollary B.138. Any operator $T \in \mathcal{L}(V)$ on a complex vector space V can be decomposed as $T = \Lambda + N$ where Λ is diagonalizable, N is nilpotent, and $\Lambda N = N\Lambda$.

Proof. By Theorem B.136, T has a block diagonal form where the diagonal matrices are $T_1, T_2, \cdots, T_p$ and each T_j can be decomposed as $T_j = N_j + \Lambda_j$ with $N_j = T_j - \lambda_j I_j$ and $\Lambda_j = \lambda_j I_j$. Clearly N_j is nilpotent and Λ_j is diagonalizable. Also, any matrix commutes with the identity matrix, hence we have $N_j \lambda_j I_j = \lambda_j I_j N_j$. The rest follows from the block diagonal form of T. □

Example B.139. If p is a polynomial of degree k, then $p(a + x)$ can be expressed as a Taylor series,

$$p(a + x) = \sum_{i=0}^{k} \frac{p^{(i)}(a)}{i!} x^i.$$

The formula is an algebraic identity and can be generalized to an operator $T = \Lambda + N$ with the help of $\Lambda N = N\Lambda$,

$$p(\Lambda + N) = \sum_{i=0}^{k} \frac{p^{(i)}(\Lambda)}{i!} N^i = \sum_{i=0}^{m} \frac{p^{(i)}(\Lambda)}{i!} N^i,$$

where the nilpotent operator N satisfies $N^{m+1} = \mathbf{0}$.

The same approach works if p is not a polynomial, but an infinite power series with its radius of convergence as ∞. In particular, if $p(x) = e^x$, we have

$$e^T = \sum_{i=0}^{+\infty} \frac{e^\Lambda}{i!} N^i = e^\Lambda \sum_{i=0}^{m} \frac{1}{i!} N^i, \tag{B.50}$$

which can be adopted as a definition of matrix exponentials.

Theorem B.140. Any invertible operator $T \in \mathcal{L}(V)$ on a complex vector space V has a square root.

Proof. Let $\lambda_1, \lambda_2, \dots, \lambda_m$ be all distinct eigenvalues of T. By (DOC-3) in Theorem B.132, for each $j = 1, 2, \dots, m$ there exists a niplotent operator $N_j \in \mathcal{L}(G(\lambda_j, T))$ such that $T|_{G(\lambda_j,T)} = \lambda_j I + N_j$. Since T is invertible, $\lambda_j \neq 0$ and

$$\forall j = 1, 2, \dots, m, \quad T|_{G(\lambda_j,T)} = \lambda_j \left(I + \frac{N_j}{\lambda_j} \right).$$

By Lemma B.130 and the condition of $\mathbb{F} = \mathbb{C}$, there exists an operator $R_j \in \mathcal{L}(G(\lambda_j, T))$ such that $R_j^2 = T|_{G(\lambda_j,T)}$.

By (DOC-1) in Theorem B.132, any vector $\mathbf{v} \in V$ can be uniquely expressed in the form $\mathbf{v} = \sum_{i=1}^m \mathbf{u}_i$ where $\mathbf{u}_i \in G(\lambda_i, T)$. Then it is straightforward to verify that the operator $R \in \mathcal{L}(V)$ given below satisfies $R^2 = T$:

$$R(\mathbf{v}) = \sum_{j=1}^m R_j \mathbf{u}_j. \qquad \square$$

B.4.4 Jordan basis

Definition B.141. A *Jordan block* of order k has the form

$$J(\lambda, k) = \lambda I_k + S_k, \tag{B.51}$$

where

$$(S_k)_{i,j} = \begin{cases} 1, & i = j - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Example B.142. The Jordan blocks of orders 1, 2, and 3 are

$$J(\lambda, 1) = \lambda, \quad J(\lambda, 2) = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \quad J(\lambda, 3) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}.$$

Example B.143. The nilpotent operator

$$N(x, y, z) = (0, x, y)$$

can be exploited to construct a basis of $\mathbb{F}^3$: $(N^2\mathbf{v}, N\mathbf{v}, \mathbf{v})$ where $\mathbf{v} = (1, 0, 0)$. With respect to this basis, the matrix of N is the Jordan block $J(0, 3)$.

Example B.144. The nilpotent operator

$$N(z_1, z_2, z_3, z_4, z_5, z_6) = (0, z_1, z_2, 0, z_4, 0)$$

can be exploited to construct a basis of $\mathbb{F}^6$:

$$(N^2\mathbf{v}_1, N\mathbf{v}_1, \mathbf{v}_1, N\mathbf{v}_2, \mathbf{v_2}, \mathbf{v}_3)$$

where

$$\begin{aligned} \mathbf{v}_1 &= (1, 0, 0, 0, 0, 0), \\ \mathbf{v}_2 &= (0, 0, 0, 1, 0, 0), \\ \mathbf{v}_3 &= (0, 0, 0, 0, 0, 1). \end{aligned}$$

With respect to this basis, the matrix of N is the block diagonal matrix

$$\begin{pmatrix} J(0,3) & & \\ & J(0,2) & \\ & & J(0,1) \end{pmatrix}.$$

Lemma B.145. For a nilpotent operator $N \in \mathcal{L}(V)$, there exist $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in V$ and $m_1, m_2, \dots, m_n \in \mathbb{N}$ such that

(a) $N^{m_1}\mathbf{v}_1, \cdots, N\mathbf{v}_1, \mathbf{v}_1, \cdots, N^{m_n}\mathbf{v}_n, \cdots, N\mathbf{v}_n, \mathbf{v}_n$ form a basis of V;

(b) $N^{m_1+1}\mathbf{v}_1 = \cdots = N^{m_n+1}\mathbf{v}_n = \mathbf{0}$.

Proof. If $\operatorname{range} N$ is empty, Theorem B.56 implies that $\operatorname{null} N = V$. Then $N = \mathbf{0}$ and both (a) and (b) hold trivially. Hereafter we prove this lemma by an induction on $\dim V$. The induction basis for $\dim V = 1$ clearly holds. Hereafter we assume that $\dim V > 1$ and (a) and (b) hold on all vector spaces of smaller dimensions.

Because N is nilpotent, N is not injective. By Theorem B.57, N is not surjective either. The range of N, a subspace of V (c.f. Theorem B.55), has a smaller dimension than V. Thus we can apply our induction hypothesis to $\operatorname{range} N$ to obtain that, for $N|_{\operatorname{range} N} \in \mathcal{L}(\operatorname{range} N)$, there exist vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in \operatorname{range} N$ and $m_1, m_2, \dots, m_n \in \mathbb{N}$ such that

$$(*): \quad N^{m_1}\mathbf{v}_1, \dots, N\mathbf{v}_1, \mathbf{v}_1, \cdots, N^{m_n}\mathbf{v}_n, \dots, N\mathbf{v}_n, \mathbf{v}_n$$

is a basis of $\operatorname{range} N$ and (b) holds for this basis. Then

$$\forall j = 1, \dots, n, \quad \exists \mathbf{u}_j \in V \text{ s.t. } \mathbf{v}_j = N\mathbf{u}_j.$$

Next, we claim that the following is an independent list of vectors in V:

$$(\square): \quad N^{m_1+1}\mathbf{u}_1, \dots, N\mathbf{u}_1, \mathbf{u}_1, \cdots, N^{m_n+1}\mathbf{u}_n, \dots, N\mathbf{u}_n, \mathbf{u}_n.$$

Indeed, suppose a linear combination of $(\square)$ equals $\mathbf{0}$. Then apply N to it and we get a linear combination of $(*)$ equal to $\mathbf{0}$ and thus all the coefficients in the original linear combination must be 0 except for those of the vectors

$$N^{m_1+1}\mathbf{u}_1 = N^{m_1}\mathbf{v}_1, \ \dots, \ N^{m_n+1}\mathbf{u}_n = N^{m_n}\mathbf{v}_n.$$

Again, the linear independence of the list $(*)$ dictates that all coefficients of the above vectors must be 0.

By Lemma B.37, we can extend $(\square)$ to a basis of V,

$$(\triangle): \quad N^{m_1+1}\mathbf{u}_1, \dots, \mathbf{u}_1, \cdots, N^{m_n+1}\mathbf{u}_n, \dots, \mathbf{u}_n, \mathbf{w}_1, \dots, \mathbf{w}_p.$$

Since $N\mathbf{w}_j \in \operatorname{range} N$, each $N\mathbf{w}_j$ is in the span of $(*)$, thus there exists $\mathbf{x}_j$ in the span of $(\square)$ such that $N\mathbf{w}_j = N\mathbf{x}_j$. Define $\mathbf{u}_{n+j} = \mathbf{w}_j - \mathbf{x}_j$ and we have $N\mathbf{u}_{n+j} = \mathbf{0}$. Therefore the following list of vectors satisfies (a) and (b):

$$N^{m_1+1}\mathbf{u}_1, \dots, \mathbf{u}_1, \cdots, N^{m_n+1}\mathbf{u}_n, \dots, \mathbf{u}_n, \mathbf{u}_{n+1}, \dots, \mathbf{u}_{n+p}.$$

This completes the induction and the proof. $\square$

Definition B.146. A *Jordan basis for a linear operator* $T \in \mathcal{L}(V)$ is a basis of V with respect to which the matrix of T is a block diagonal matrix of the form

$$J = \begin{pmatrix} J(\lambda_1, k_1) & & & \\ & J(\lambda_2, k_2) & & \\ & & \ddots & \\ & & & J(\lambda_p, k_p) \end{pmatrix}, \tag{B.52}$$

where the λ_j's might not be distinct.

Theorem B.147 (Jordan)**.** Any operator $T \in \mathcal{L}(V)$ on a complex vector space has a Jordan basis.

Proof. If T is a nilpotent operator N, we consider the basis given in Lemma B.145(a). For each j, N annihilates $N^{m_j}\mathbf{v}_j$ in the list $N^{m_j}\mathbf{v}_j, \cdots, N\mathbf{v}_j, \mathbf{v}_j$ and sends another vector to its previous. Hence N has a block diagonal matrix where each matrix on the diagonal is the Jordan block $J(0, k_j)$. Thus the statement holds for nilpotent operators.

Otherwise let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be all distinct eigenvalues of T. By Theorem B.132, we have the decomposition

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T),$$

with each $N_j := (T - \lambda_j I)|_{G(\lambda_j, T)}$ being nilpotent. By the previous paragraph, each N_j has a Jordan basis. The combination of all these Jordan bases is a Jordan basis for T. □

B.5 Inner product spaces

B.5.1 Inner products

Definition B.148. Denote by $\mathbb{F}$ the underlying field of a vector space $\mathcal{V}$. The *inner product* $\langle \mathbf{u}, \mathbf{v} \rangle$ on $\mathcal{V}$ is a function $\mathcal{V} \times \mathcal{V} \to \mathbb{F}$ that satisfies

(IP-1) real positivity: $\forall \mathbf{v} \in \mathcal{V}$, $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$;

(IP-2) definiteness: $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ iff $\mathbf{v} = \mathbf{0}$;

(IP-3) additivity in the first slot:
$\forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$;

(IP-4) homogeneity in the first slot:
$\forall a \in \mathbb{F}$, $\forall \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $\langle a\mathbf{v}, \mathbf{w} \rangle = a \langle \mathbf{v}, \mathbf{w} \rangle$;

(IP-5) conjugate symmetry: $\forall \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}$.

An *inner product space* is a vector space $\mathcal{V}$ equipped with an inner product on $\mathcal{V}$.

Corollary B.149. An inner product has additivity in the second slot, i.e. $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$.

Corollary B.150. An inner product has conjugate homogeneity in the second slot, i.e.

$$\forall a \in \mathbb{F},\ \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \qquad \langle \mathbf{v}, a\mathbf{w} \rangle = \bar{a} \langle \mathbf{v}, \mathbf{w} \rangle . \tag{B.53}$$

Exercise B.151. Prove Corollaries B.149 and B.150 from Definition B.148.

Definition B.152. The *Euclidean inner product* on $\mathbb{F}^n$ is

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{n} v_i \overline{w_i}. \tag{B.54}$$

B.5.2 Norms induced from inner products

Definition B.153. Let $\mathbb{F}$ be the underlying field of an inner product space $\mathcal{V}$. The *norm induced by an inner product* on $\mathcal{V}$ is a function $\mathcal{V} \to \mathbb{F}$:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \tag{B.55}$$

Definition B.154. For $p \in [1, \infty)$, the *Euclidean ℓ_p norm* of a vector $\mathbf{v} \in \mathbb{F}^n$ is

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \tag{B.56}$$

and the *Euclidean ℓ_∞ norm* is

$$\|\mathbf{v}\|_\infty = \max_i |v_i|. \tag{B.57}$$

Theorem B.155 (Equivalence of norms)**.** Any two norms $\|\cdot\|_N$ and $\|\cdot\|_M$ on a finite dimensional vector space $\mathcal{V} = \mathbb{C}^n$ satisfy

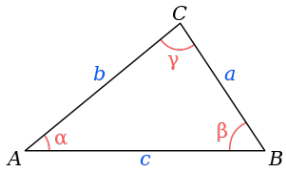
$$\exists c_1, c_2 \in \mathbb{R}^+,\ \text{s.t. } \forall \mathbf{x} \in \mathcal{V},\ c_1 \|\mathbf{x}\|_M \leq \|\mathbf{x}\|_N \leq c_2 \|\mathbf{x}\|_M. \tag{B.58}$$

Definition B.156. The angle between two vectors $\mathbf{v}, \mathbf{w}$ in an inner product space with $\mathbb{F} = \mathbb{R}$ is the number $\theta \in [0, \pi]$,

$$\theta = \arccos \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\| \|\mathbf{w}\|}. \tag{B.59}$$

Theorem B.157 (The law of cosines)**.** Any triangle satisfies

$$c^2 = a^2 + b^2 - 2ab \cos \gamma. \tag{B.60}$$



Proof. The dot product of AB to $AB = CB - CA$ yields

$$c^2 = \langle AB, CB \rangle - \langle AB, CA \rangle .$$

The dot products of CB and CA to $AB = CB - CA$ yield

$$\begin{aligned} \langle CB, AB \rangle &= a^2 - \langle CB, CA \rangle ; \\ - \langle CA, AB \rangle &= - \langle CA, CB \rangle + b^2. \end{aligned}$$

The proof is completed by adding up all three equations and applying (B.59). □

Theorem B.158 (The law of cosines: abstract version)**.** Any induced norm on a real vector space satisfies

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2 \langle \mathbf{u}, \mathbf{v} \rangle . \tag{B.61}$$

Proof. Definitions B.153 and B.148 and $\mathbb{F} = \mathbb{R}$ yield

$$\begin{aligned} \|\mathbf{u} - \mathbf{v}\|^2 &= \langle \mathbf{u} - \mathbf{v}, \mathbf{u} - \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{v}, \mathbf{u} \rangle \\ &= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2 \langle \mathbf{u}, \mathbf{v} \rangle . \end{aligned}$$

□

B.5.3 Norms and induced inner-products

Definition B.159. A function $\|\cdot\| : \mathcal{V} \to \mathbb{F}$ is a *norm* for a vector space $\mathcal{V}$ iff it satisfies

(NRM-1) real positivity: $\forall \mathbf{v} \in \mathcal{V}$, $\|\mathbf{v}\| \geq 0$;

(NRM-2) point separation: $\|\mathbf{v}\| = 0 \Rightarrow \mathbf{v} = \mathbf{0}$.

(NRM-3) absolute homogeneity: $\forall a \in \mathbb{F}$, $\forall \mathbf{v} \in \mathcal{V}$, $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$;

(NRM-4) triangle inequality: $\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

The function $\|\cdot\| : \mathcal{V} \to \mathbb{F}$ is called a *semi-norm* iff it satifies (NRM-1,3,4). A *normed vector space* (or simply a *normed space*) is a vector space $\mathcal{V}$ equipped with a norm on $\mathcal{V}$.

Exercise B.160. Explain how (NRM-1,2,3,4) relate to the geometric meaning of the norm of vectors in $\mathbb{R}^3$.

Lemma B.161. The norm induced by an inner product is a norm as in Definition B.159.

Proof. The induced norm as in (B.55) satisfies (NRM-1,2) trivially. For (NRM-3),

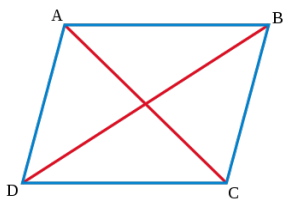
$$\|a\mathbf{v}\|^2 = \langle a\mathbf{v}, a\mathbf{v}\rangle = a\,\langle \mathbf{v}, a\mathbf{v}\rangle = a\bar{a}\,\langle \mathbf{v}, \mathbf{v}\rangle = |a|^2\|\mathbf{v}\|^2.$$

To prove (NRM-4), we have

$$\begin{aligned}\|\mathbf{u}+\mathbf{v}\|^2 &= \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}\rangle \\ &= \langle \mathbf{u},\mathbf{u}\rangle + \langle \mathbf{v},\mathbf{v}\rangle + \langle \mathbf{u},\mathbf{v}\rangle + \overline{\langle \mathbf{u},\mathbf{v}\rangle} \\ &\leq \langle \mathbf{u},\mathbf{u}\rangle + \langle \mathbf{v},\mathbf{v}\rangle + 2|\langle \mathbf{u},\mathbf{v}\rangle| \\ &\leq \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| \\ &= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2,\end{aligned}$$

where the second step follows from (IP-5) and the fourth step from Cauchy-Schwarz inequality. □

Theorem B.162 (The parallelogram law)**.** The sum of squares of the lengths of the four sides of a parallelogram equals the sum of squares of the two diagonals.



More precisely, we have in the above plot

$$(AB)^2+(BC)^2+(CD)^2+(DA)^2 = (AC)^2+(BD)^2. \quad \text{(B.62)}$$

Proof. Apply the law of cosines to the two diagonals, add the two equations, and we obtain (B.62). □

Theorem B.163 (The parallelogram law: abstract version)**.** Any induced norm (B.55) satisfies

$$2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 = \|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2. \quad \text{(B.63)}$$

Proof. Replace $\mathbf{v}$ in (B.61) with $-\mathbf{v}$ and we have

$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\,\langle \mathbf{u},\mathbf{v}\rangle\,.$$

(B.63) follows from adding the above equation to (B.61). □

Exercise B.164. In the case of Euclidean ℓ_p norms, show that the parallelogram law (B.63) holds if and only if $p = 2$.

Theorem B.165. The induced norm (B.55) holds for some inner product $\langle\cdot,\cdot\rangle$ if and only if the parallelogram law (B.63) holds for every pair of $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

Exercise B.166. Prove Theorem B.165.

Example B.167. By Theorem B.165 and Exercise B.164, the ℓ^1 and ℓ^∞ spaces do not have a corresponding inner product for the Euclidean ℓ_1 and ℓ_∞ norms.

B.5.4 Orthonormal bases

Definition B.168. Two vectors $\mathbf{u}, \mathbf{v}$ are called *orthogonal* if $\langle \mathbf{u}, \mathbf{v}\rangle = 0$, i.e., their inner product is the additive identity of the underlying field.

Example B.169. An inner product on the vector space of continuous real-valued functions on the interval $[-1, 1]$ is

$$\langle f, g\rangle = \int_{-1}^{+1} f(x)g(x)\mathrm{d}x.$$

f and g are said to be orthogonal if the integral is zero.

Theorem B.170 (Pythagorean)**.** If $\mathbf{u}, \mathbf{v}$ are orthogonal, then $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

Proof. This follows from (B.61) and Definition B.168. □

Theorem B.171 (Cauchy-Schwarz inequality)**.**

$$|\langle \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\|, \quad \text{(B.64)}$$

where the equality holds iff one of $\mathbf{u}, \mathbf{v}$ is a scalar multiple of the other.

Proof. For any complex number λ, (IP-1) implies

$$\begin{aligned}&\langle \mathbf{u}+\lambda\mathbf{v}, \mathbf{u}+\lambda\mathbf{v}\rangle \geq 0 \\ \Rightarrow\ &\langle \mathbf{u},\mathbf{u}\rangle + \lambda\,\langle \mathbf{v},\mathbf{u}\rangle + \bar{\lambda}\,\langle \mathbf{u},\mathbf{v}\rangle + \lambda\bar{\lambda}\,\langle \mathbf{v},\mathbf{v}\rangle \geq 0.\end{aligned}$$

If $\mathbf{v} = \mathbf{0}$, (B.64) clearly holds. Otherwise (B.64) follows from substituting $\lambda = -\frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{v},\mathbf{v}\rangle}$ into the above equation. □

Exercise B.172. To explain the choice of λ in the proof of Theorem B.171, what is the geometric meaning of (B.64) in the plane? When will the equality hold?

Example B.173. If $x_i, y_i \in \mathbb{R}$, then for any $n \in \mathbb{N}^+$

$$\left|\sum_{i=1}^n x_i y_i\right|^2 \leq \sum_{j=1}^n x_j^2 \sum_{k=1}^n y_k^2.$$

Example B.174. If $f, g : [a, b] \to \mathbb{R}$ are continuous, then

$$\left|\int_a^b f(x)g(x)\mathrm{d}x\right|^2 \leq \left(\int_a^b f^2(x)\mathrm{d}x\right)\left(\int_a^b g^2(x)\mathrm{d}x\right)$$

Definition B.175. A list of vectors $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m)$ is called *orthonormal* if the vectors in it are pairwise orthogonal and each vector has norm 1, i.e.

$$\begin{cases} \forall i = 1, 2, \dots, m, & \|\mathbf{e}_i\| = 1; \\ \forall i \neq j, & \langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0. \end{cases} \tag{B.65}$$

Definition B.176. An *orthonormal basis* of an inner-product space $\mathcal{V}$ is an orthonormal list of vectors in $\mathcal{V}$ that is also a basis of $\mathcal{V}$.

Theorem B.177. If $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ is an orthonormal basis of $\mathcal{V}$, then

$$\forall \mathbf{v} \in \mathcal{V}, \quad \mathbf{v} = \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{e}_i \rangle \, \mathbf{e}_i, \tag{B.66a}$$

$$\|\mathbf{v}\|^2 = \sum_{i=1}^{n} |\langle \mathbf{v}, \mathbf{e}_i \rangle|^2 . \tag{B.66b}$$

Lemma B.178. Every finite-dimensional inner-product space has an orthonormal basis.

Theorem B.179 (Schur)**.** Every linear operator $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional complex vector space $\mathcal{V}$ has an upper-triangular matrix with respect to some orthonormal basis of $\mathcal{V}$.

Proof. This follows from Theorem B.104, Lemma B.178 and the Gram-Schmidt process. □

Definition B.180. A *linear functional* on $\mathcal{V}$ is a linear map from $\mathcal{V}$ to $\mathbb{F}$, or, it is an element of $\mathcal{L}(\mathcal{V}, \mathbb{F})$.

Theorem B.181 (Riesz representation theorem)**.** If $\mathcal{V}$ is a finite-dimensional vector space, then

$$\forall \varphi \in \mathcal{V}', \ \exists! \mathbf{u} \in \mathcal{V} \text{ s.t. } \forall \mathbf{v} \in \mathcal{V}, \quad \varphi(\mathbf{v}) = \langle \mathbf{v}, \mathbf{u} \rangle . \tag{B.67}$$

Proof. Let $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ be an orthonormal basis of $\mathcal{V}$.

$$\begin{aligned} \varphi(\mathbf{v}) &= \varphi\left(\sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{e}_i \rangle \, \mathbf{e}_i \right) = \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{e}_i \rangle \, \varphi(\mathbf{e}_i) \\ &= \sum_{i=1}^{n} \left\langle \mathbf{v}, \overline{\varphi(\mathbf{e}_i)} \mathbf{e}_i \right\rangle = \left\langle \mathbf{v}, \sum_{i=1}^{n} \overline{\varphi(\mathbf{e}_i)} \mathbf{e}_i \right\rangle , \end{aligned}$$

where the last two steps follow from Corollaries B.149 and B.150.

As for the uniqueness, suppose that $\exists \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{V}$ s.t. $\varphi(\mathbf{v}) = \langle \mathbf{v}, \mathbf{u}_1 \rangle = \langle \mathbf{v}, \mathbf{u}_2 \rangle$. Then for each $\mathbf{v} \in \mathcal{V}$,

$$0 = \langle \mathbf{v}, \mathbf{u}_1 \rangle - \langle \mathbf{v}, \mathbf{u}_2 \rangle = \langle \mathbf{v}, \mathbf{u}_1 - \mathbf{u}_2 \rangle .$$

Taking $\mathbf{v} = \mathbf{u}_1 - \mathbf{u}_2$ shows that $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{0}$. □

B.6 Operators on inner-product spaces

B.6.1 Adjoint and self-adjoint operators

Definition B.182. The *adjoint of a linear map* $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ between inner-product spaces is a function $T^* : \mathcal{W} \to \mathcal{V}$ that satisfies

$$\forall \mathbf{v} \in \mathcal{V}, \ \forall \mathbf{w} \in \mathcal{W}, \qquad \langle T\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, T^* \mathbf{w} \rangle . \tag{B.68}$$

Example B.183. Define a linear operator $T : \mathbb{R}^3 \to \mathbb{R}^2$,

$$T(x_1, x_2, x_3) = (x_2 + 3x_3, 2x_1).$$

Then $T^*(y_1, y_2) = (2y_2, y_1, 3y_1)$ because

$$\begin{aligned} \langle (x_1, x_2, x_3), T^*(y_1, y_2) \rangle &= \langle T(x_1, x_2, x_3), (y_1, y_2) \rangle \\ &= \langle (x_2 + 3x_3, 2x_1), (y_1, y_2) \rangle \\ &= x_2 y_1 + 3x_3 y_1 + 2x_1 y_2 \\ &= \langle (x_1, x_2, x_3), (2y_2, y_1, 3y_1) \rangle . \end{aligned}$$

Lemma B.184. If $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, then $T^* \in \mathcal{L}(\mathcal{W}, \mathcal{V})$.

Proof. Use Definition B.42. □

Theorem B.185. The adjoint of a linear map has the following properties.

(ADJ-1) additivity:
$\forall S, T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \qquad (S + T)^* = S^* + T^*;$

(ADJ-2) conjugate homogeneity:
$\forall T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \ \forall a \in \mathbb{F}, \qquad (aT)^* = \bar{a} T^*;$

(ADJ-3) adjoint of adjoint:
$\forall T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \ (T^*)^* = T;$

(ADJ-4) identity: $I^* = I;$

(ADJ-5) products: let $\mathcal{U}$ be an inner-product space,
$\forall T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \ \forall S \in \mathcal{L}(\mathcal{W}, \mathcal{U}), \ (ST)^* = T^* S^*.$

Proof. Use Definitions B.182 and B.148. □

Lemma B.186. $T \in \mathcal{L}(V, W)$ and T^* satisfy

(a) $\operatorname{null} T^* = (\operatorname{range} T)^{\perp}$;

(b) $\operatorname{range} T^* = (\operatorname{null} T)^{\perp}$;

(c) $\operatorname{null} T = (\operatorname{range} T^*)^{\perp}$;

(d) $\operatorname{range} T = (\operatorname{null} T^*)^{\perp}$.

Definition B.187. The *conjugate transpose*, or *Hermitian transpose*, or *Hermitian conjugate*, or *adjoint matrix*, of a matrix $A \in \mathbb{C}^{m \times n}$ is the matrix $A^* \in \mathbb{C}^{n \times m}$ defined by

$$(A^*)_{ij} = \overline{a_{ji}}, \tag{B.69}$$

where $\overline{a_{ji}}$ denotes the complex conjugate of the entry a_{ji}.

Exercise B.188. Show that the conjugate transpose is an adjoint operator in $\mathcal{L}(\mathcal{V}, \mathcal{W})$ with $\mathcal{V} = \mathbb{C}^n$ and $\mathcal{W} = \mathbb{C}^m$.

Definition B.189. A matrix $U \in \mathbb{C}^{n \times n}$ is *unitary* iff $U^* U = I$. A matrix $U \in \mathbb{R}^{n \times n}$ is *orthogonal* iff $U^T U = I$.

Theorem B.190. A matrix $U \in \mathbb{C}^{n \times n}$ is unitary if and only if its columns form an orthonormal basis for $\mathbb{C}^n$.

Proof. This follows from considering the (i, j)th element of $U^* U$ and applying $U^* U = I$ in Definition B.189. □

Corollary B.191. A unitary matrix U preserves norms and inner products. More precisely, we have

$$\forall \mathbf{v}, \mathbf{w} \in \mathbb{C}^n, \qquad \langle U\mathbf{v}, U\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle .$$

Proof. This follows from Definitions B.182 and B.189. □

Theorem B.192. Every unitary matrix $U \in \mathbb{C}^{2\times 2}$ with $\det U = 1$ is of the form

$$U = \begin{bmatrix} a & b \\ -\overline{b} & \overline{a} \end{bmatrix}, \tag{B.70}$$

where $|a|^2 + |b|^2 = 1$.

Proof. Let

$$U = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Then Theorem B.190 and the condition $\det U = 1$ yield

$$\begin{aligned} a\overline{b} + c\overline{d} &= 0, \\ ad - cb &= 1. \end{aligned}$$

In other words, the linear system

$$\begin{bmatrix} \overline{b} & \overline{d} \\ d & -b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

has solution $x = a, y = c$. Furthermore, Theorem B.190 and the form of U yield $|b|^2 + |d|^2 = 1$. Hence the solution $x = a, y = c$ is unique and we have $a = \overline{d}$ and $c = -\overline{b}$, which completes the proof. □

Theorem B.193. Let $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. Suppose $e_1, \dots, e_n$ is an orthonormal basis of $\mathcal{V}$ and $f_1, \dots, f_m$ is an orthonormal basis of $\mathcal{W}$. Then

$$M(T^*, (f_1, \dots, f_m), (e_1, \dots, e_n))$$

is the conjugate transpose of

$$M(T, (e_1, \dots, e_n), (f_1, \dots, f_m)).$$

Proof. By Corollary B.59, we have

$$T[e_1, \dots, e_n] = [f_1, \dots, f_m] M_T.$$

The orthonormality of the two bases and Definition B.152 further imply

$$M_T = \begin{bmatrix} \langle Te_1, f_1\rangle & \langle Te_2, f_1\rangle & \cdots & \langle Te_n, f_1\rangle \\ \langle Te_1, f_2\rangle & \langle Te_2, f_2\rangle & \cdots & \langle Te_n, f_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle Te_1, f_m\rangle & \langle Te_2, f_m\rangle & \cdots & \langle Te_n, f_m\rangle \end{bmatrix}.$$

The proof is completed by repeating the above derivation for T^* and then applying Definitions B.148 and B.182. □

Lemma B.194. Suppose $\mathcal{V}$ is a complex inner product space and $T \in \mathcal{L}(\mathcal{V})$. If

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \langle T\mathbf{v}, \mathbf{v}\rangle = 0, \tag{B.71}$$

then $T = \mathbf{0}$.

Proof. By Definition B.148 and (B.71), we have, $\forall \mathbf{u}, \mathbf{w} \in \mathcal{V}$,

$$\begin{aligned} \langle T\mathbf{u}, \mathbf{w}\rangle =& \frac{\langle T(\mathbf{u}+\mathbf{w}), \mathbf{u}+\mathbf{w}\rangle - \langle T(\mathbf{u}-\mathbf{w}), \mathbf{u}-\mathbf{w}\rangle}{4} \\ &+ \mathbf{i}\frac{\langle T(\mathbf{u}+\mathbf{i}\mathbf{w}), \mathbf{u}+\mathbf{i}\mathbf{w}\rangle - \langle T(\mathbf{u}-\mathbf{i}\mathbf{w}), \mathbf{u}-\mathbf{i}\mathbf{w}\rangle}{4} \\ =& 0. \end{aligned}$$

Setting $\mathbf{w} = T\mathbf{u}$ completes the proof. □

Definition B.195. An operator $T \in \mathcal{L}(\mathcal{V})$ is *self-adjoint* iff $T = T^*$, i.e.

$$\forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \qquad \langle T\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, T\mathbf{w}\rangle. \tag{B.72}$$

Lemma B.196. Every eigenvalue of a self-adjoint operator T is real.

Proof. Let $(\lambda, \mathbf{u})$ be an eigenpair of T. We have

$$\lambda\|\mathbf{u}\|^2 = \langle \lambda\mathbf{u}, \mathbf{u}\rangle = \langle T\mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{u}, T\mathbf{u}\rangle = \langle \mathbf{u}, \lambda\mathbf{u}\rangle = \overline{\lambda}\|\mathbf{u}\|^2,$$

where the third step follows from Definition B.195. Then $\mathbf{u} \neq \mathbf{0}$ implies $\lambda = \overline{\lambda}$. □

Theorem B.197. Suppose $\mathcal{V}$ is a complex inner product space and $T \in \mathcal{L}(\mathcal{V})$. Then T is self-adjoint if and only if

$$\forall \mathbf{v} \in \mathcal{V}, \quad \langle T\mathbf{v}, \mathbf{v}\rangle \in \mathbb{R}. \tag{B.73}$$

Proof. By Definitions B.148, B.182, and B.195, we have

$$\begin{aligned} \langle T\mathbf{v}, \mathbf{v}\rangle - \overline{\langle T\mathbf{v}, \mathbf{v}\rangle} &= \langle T\mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{v}, T\mathbf{v}\rangle \\ &= \langle T\mathbf{v}, \mathbf{v}\rangle - \langle T^*\mathbf{v}, \mathbf{v}\rangle = \langle (T - T^*)\mathbf{v}, \mathbf{v}\rangle. \end{aligned}$$

Then Lemma B.194 completes the proof. □

Lemma B.198. Suppose $\mathcal{V}$ is a real inner product space and $T \in \mathcal{L}(\mathcal{V})$. If T is self-adjoint and satisfies

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \langle T\mathbf{v}, \mathbf{v}\rangle = 0, \tag{B.74}$$

then $T = \mathbf{0}$.

Proof. By the self-adjointness and the underlying field being real, we have

$$\langle T\mathbf{w}, \mathbf{u}\rangle = \langle \mathbf{w}, T\mathbf{u}\rangle = \langle T\mathbf{u}, \mathbf{w}\rangle,$$

which, together with Definition B.148, implies

$$\langle T\mathbf{u}, \mathbf{w}\rangle = \frac{\langle T(\mathbf{u}+\mathbf{w}), \mathbf{u}+\mathbf{w}\rangle - \langle T(\mathbf{u}-\mathbf{w}), \mathbf{u}-\mathbf{w}\rangle}{4}.$$

Setting $\mathbf{w} = T\mathbf{u}$ completes the proof. □

B.6.2 Normal operators

Definition B.199. An operator $T \in \mathcal{L}(\mathcal{V})$ is *normal* iff $TT^* = T^*T$.

Corollary B.200. Every self-adjoint operator is normal.

Lemma B.201. $T \in \mathcal{L}(\mathcal{V})$ is normal if and only if

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \|T\mathbf{v}\| = \|T^*\mathbf{v}\|. \tag{B.75}$$

Proof. By Lemma B.198 and Definition B.182, we have

$$\begin{aligned} T^*T = TT^* &\Leftrightarrow \forall \mathbf{v} \in \mathcal{V},\ \langle (T^*T - TT^*)\mathbf{v}, \mathbf{v}\rangle = 0 \\ &\Leftrightarrow \forall \mathbf{v} \in \mathcal{V},\ \langle T^*T\mathbf{v}, \mathbf{v}\rangle = \langle TT^*\mathbf{v}, \mathbf{v}\rangle \\ &\Leftrightarrow \forall \mathbf{v} \in \mathcal{V},\ \|T\mathbf{v}\|^2 = \|T^*\mathbf{v}\|^2. \end{aligned}$$

The positivity of a norm completes the proof. □

Lemma B.202. $T \in \mathcal{L}(\mathcal{V})$ is normal if and only if each eigenvector of T is also an eigenvector of T^*.

Proof. If T is normal, so is $T - \lambda I$. By Lemma B.201, an eigenpair $(\lambda, \mathbf{u})$ of T satisfies

$$0 = \|(T - \lambda I)\mathbf{u}\| = \|(T - \lambda I)^*\mathbf{u}\| = \|(T^* - \bar{\lambda} I)\mathbf{u}\|,$$

and thus $\mathbf{u}$ is also an eigenvector of T^*.

Conversely, suppose each eigenvector $\mathbf{u}$ of T is also an eigenvector of T^*. Then the above equation implies that the corresponding eigenvalue of T^* is the conjugate of that of T. It suffices to prove that these eigenvectors form a basis of V, because then we have

$$TU = U\Lambda, \quad T^*U = U\Lambda^*,$$

where U is the matrix of these eigenvectors. Thus

$$TT^*U = TU\Lambda^* = U\Lambda\Lambda^* = U\Lambda^*\Lambda = T^*U\Lambda = T^*TU$$

and we have $TT^* = T^*T$ because U is nonsingular. By Theorem B.111, it suffices to show that T is diagonalizable. By Theorem B.147, we only need to show that for any eigenpair $(\lambda, \mathbf{u})$ of T,

$$(A - \lambda I)^2\mathbf{u} = \mathbf{0} \ \Rightarrow\ (A - \lambda I)\mathbf{u} = \mathbf{0},$$

because this condition will annihilate all Jordan blocks of size greater than 1. Define $\mathbf{v} = (A - \lambda I)\mathbf{u}$ and we have

$$\begin{aligned} \langle A\mathbf{u}, \mathbf{v}\rangle &= \langle \mathbf{u}, A^*\mathbf{v}\rangle = \left\langle \mathbf{u}, \bar{\lambda}\mathbf{v}\right\rangle = \langle \lambda\mathbf{u}, \mathbf{v}\rangle, \\ \langle A\mathbf{u}, \mathbf{v}\rangle &= \langle \mathbf{v} + \lambda\mathbf{u}, \mathbf{v}\rangle = \|\mathbf{v}\|^2 + \langle \lambda\mathbf{u}, \mathbf{v}\rangle, \end{aligned}$$

which imply $\mathbf{v} = 0$, i.e., $(A - \lambda I)\mathbf{u} = \mathbf{0}$. □

Theorem B.203. For a linear operator $T \in \mathcal{L}(\mathcal{V})$ on a two-dimensional real inner product space $\mathcal{V}$, the following are equivalent:

(a) T is normal but not self-adjoint.

(b) The matrix of T with respect to every orthonormal basis of $\mathcal{V}$ has the form

$$M(T) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \tag{B.76}$$

where $b \neq 0$.

Proof. (b) ⇒ (a) trivially holds, so we only prove (a) ⇒ (b). Let (e_1, e_2) be an orthonormal basis of $\mathcal{V}$ and set

$$M(T, (e_1, e_2)) = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

By Definition B.199, we have

$$\begin{aligned} \begin{bmatrix} a & c \\ b & d \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} &= \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix} \\ = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} &= \begin{bmatrix} a^2 + b^2 & ac + bd \\ ac + bd & c^2 + d^2 \end{bmatrix}. \end{aligned}$$

$b^2 = c^2$ and the condition of T being not self-adjoint further yields $c = -b \neq 0$, which, together with $ab + cd = ac + bd$, yields $a = d$. □

B.6.3 The spectral theorems

Theorem B.204 (Complex spectral)**.** For a linear operator $T \in \mathcal{L}(\mathcal{V})$ with $\mathbb{F} = \mathbb{C}$, the following are equivalent:

(a) T is normal;

(b) $\mathcal{V}$ has an orthonormal basis consisting of eigenvectors of T;

(c) T has a diagonal matrix with respect to some orthonormal basis of $\mathcal{V}$.

Corollary B.205. A normal operator T whose eigenvalues are real is self-adjoint.

Proof. By Theorem B.204, we write $T = V\Lambda V^{-1}$ and thus $T^* = V\Lambda^* V^{-1}$. Then all entries in Λ being real implies $T = T^*$. □

Theorem B.206 (Real spectral)**.** For a linear operator $T \in \mathcal{L}(\mathcal{V})$ with $\mathbb{F} = \mathbb{R}$, the following are equivalent:

(a) T is self-adjoint;

(b) $\mathcal{V}$ has an orthonormal basis consisting of eigenvectors of T;

(c) T has a diagonal matrix with respect to some orthonormal basis of $\mathcal{V}$.

B.6.4 Isometries

Definition B.207. An operator $S \in \mathcal{L}(\mathcal{V})$ is called a (linear) *isometry* iff

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \|S\mathbf{v}\| = \|\mathbf{v}\|. \tag{B.77}$$

Theorem B.208. An operator $S \in \mathcal{L}(\mathcal{V})$ on a real inner product space is an isometry if and only if there exists an orthonormal basis of $\mathcal{V}$ with respect to which $\mathcal{S}$ has a block diagonal matrix such that each block on the diagonal is a 1-by-1 matrix containing 1 or -1, or, is a 2-by-2 matrix of the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{B.78}$$

where $\theta \in (0, \pi)$.

Corollary B.209. For an operator $\mathcal{S} \in \mathcal{L}(\mathcal{V})$ on a two-dimensional real inner product space, the following are equivalent:

(a) $\mathcal{S}$ is an isometry;

(b) $\mathcal{S}$ is either an identity or a reflection or a rotation.

B.6.5 Singular value decomposition

Definition B.210. An operator $T \in \mathcal{L}(V)$ is *positive semi-definite* iff

$$\forall \mathbf{v} \in V, \quad \langle T\mathbf{v}, \mathbf{v}\rangle \geq 0 \tag{B.79}$$

and is *positive definite* iff

$$\forall \mathbf{v} \in V \setminus \{\mathbf{0}\}, \quad \langle T\mathbf{v}, \mathbf{v}\rangle > 0. \tag{B.80}$$

An operator is *positive* if it is self-adjoint and positive semi-definite.

Corollary B.211. For any linear operator $f \in \mathcal{L}(\mathcal{V})$, both $f^* \circ f$ and $f \circ f^*$ are positive.

Proof. By Definition B.195, $f^* \circ f$ is self-adjoint since

$$\langle (f^* \circ f)\mathbf{u}, \mathbf{v}\rangle = \langle f\mathbf{u}, f\mathbf{v}\rangle = \langle \mathbf{u}, (f^* \circ f)\mathbf{v}\rangle .$$

Suppose $(\lambda, \mathbf{u})$ is an eigenpair of $(f^* \circ f)$. Then we have

$$\begin{aligned} \lambda \langle \mathbf{u}, \mathbf{u}\rangle &= \langle (f^* \circ f)\mathbf{u}, \mathbf{u}\rangle = \langle f\mathbf{u}, f\mathbf{u}\rangle \\ \Rightarrow \lambda &= \frac{\langle f\mathbf{u}, f\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \geq 0. \end{aligned}$$

Similar arguments apply to $f \circ f^*$. □

Definition B.212. The *singular values* of a linear map $f : \mathbb{C}^n \to \mathbb{C}^m$ are the non-negative square roots of the eigenvalues of $f^* \circ f$, usually sorted in non-increasing order as

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0,$$

where r is the rank of f.

Theorem B.213. For any matrix $A \in \mathbb{C}^{m\times n}$ with rank r, there exist orthonormal bases $\mathbb{C}^n = \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ and $\mathbb{C}^m = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ such that

$$\forall j = 1, 2, \ldots, r, \quad \begin{cases} A\mathbf{u}_j = \sigma_j \mathbf{v}_j; \\ A^*\mathbf{v}_j = \sigma_j \mathbf{u}_j, \end{cases} \tag{B.81a}$$

$$\begin{cases} \forall j = r+1, r+2, \ldots, n, & A\mathbf{u}_j = \mathbf{0}; \\ \forall j = r+1, r+2, \ldots, m, & A^*\mathbf{v}_j = \mathbf{0}, \end{cases} \tag{B.81b}$$

where σ_j's are the singular values of A in Definition B.212.

Proof. The matrix $A^*A \in \mathbb{C}^{n\times n}$ is self-adjoint and thus normal. By Theorem B.204, $\mathbb{C}^n$ has an orthonormal basis that are also eigenvectors of A^*A; choose them to be $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$. By Definition B.212 we have

$$A^*A\mathbf{u}_j = \sigma_j^2 \mathbf{u}_j.$$

Then we choose

$$\forall j = 1, 2, \ldots, r, \quad \mathbf{v}_j := \frac{1}{\sigma_j} A\mathbf{u}_j,$$

which implies

$$\langle \mathbf{v}_i, \mathbf{v}_j\rangle = \frac{1}{\sigma_i\sigma_j} \langle A\mathbf{u}_i, A\mathbf{u}_j\rangle = \frac{1}{\sigma_i\sigma_j} \langle \mathbf{u}_i, A^*A\mathbf{u}_j\rangle = \frac{\sigma_j}{\sigma_i} \langle \mathbf{u}_i, \mathbf{u}_j\rangle .$$

Then the orthonormality of $\mathbf{u}_j$'s implies that the first r $\mathbf{v}_j$'s are orthonormal and if $r < m$ we can extend them by the Gram-Schmidt process to arrive at an orthonormal basis of $\mathbb{C}^m$. Therefore (B.81a) holds.

The first line of (B.81b) follows from $\operatorname{null} A = \operatorname{null}(A^*A)$, which is implied by

$$\|A\mathbf{u}\|^2 = \langle A\mathbf{u}, A\mathbf{u}\rangle = \langle \mathbf{u}, A^*A\mathbf{u}\rangle .$$

The rank of A^* is r. If $r = m$, the second line of (B.81b) holds vacuously. Otherwise $r < m$. The fundamental theorem of linear algebra (Theorem B.89) implies that A^* has rank $m - r$, which completes the proof. □

Definition B.214. The *singular value decomposition* (SVD) of a rectangular matrix $A \in \mathbb{C}^{m\times n}$ is the factorization $A = V\Sigma U^*$ where Σ is a diagonal matrix with its diagonal entries as the singular values in Definition B.212 and V and U are unitary matrices whose columns are respectively the vectors $\mathbf{v}_j$'s and $\mathbf{u}_j$'s specified in Theorem B.213, which are also called the *left singular vectors* and the *right singular vectors* of A, respectively. We also refer to the sequence of triples $(\sigma_j, \mathbf{u}_j, \mathbf{v}_j)$ as the *singular system* of A.

Definition B.215. Two matrices $A, B \in \mathbb{R}^{n\times n}$ are called *similar* iff there exists an invertible matrix P such that $B = P^{-1}AP$. The map $A \mapsto P^{-1}AP$ is called a *similarity transformation* or *conjugation of the matrix* A.

B.7 Trace and determinant

Definition B.216. The *trace of a matrix* A, denoted by $\operatorname{Trace} A$, is the sum of the diagonal entries of A.

Lemma B.217. The trace of a matrix is the sum of its eigenvalues, each of which is repeated according to its multiplicity.

Definition B.218. A *permutation of a set* A is a bijective function $\sigma : A \to A$.

Definition B.219. Let σ be a permutation of $A = \{1, 2, \ldots, n\}$ and let s denote the number of pairs of integers (j, k) with $1 \leq j < k \leq n$ such that j appears after k in the list $(m_1, \ldots, m_n)$ given by $m_i = \sigma(i)$. The *sign of the permutation* σ is 1 if s is even and -1 if s is odd.

Definition B.220. The *signed volume of a parallelotope* spanned by n vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ is a function $\delta : \mathbb{R}^{n\times n} \to \mathbb{R}$ that satisfies

(SVP-1) $\delta(I) = 1$;

(SVP-2) $\delta(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = 0$ if $\mathbf{v}_i = \mathbf{v}_j$ for some $i \neq j$;

(SVP-3) δ is linear, i.e., $\forall j = 1, \ldots, n,\ \forall c \in \mathbb{R}$,

$$\begin{aligned} &\delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v} + c\mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n) \\ &= \delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n) \\ &\quad + c\delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n). \end{aligned} \tag{B.82}$$

Exercise B.221. Give a geometric proof that the signed volume of the parallelogram determined by the two vectors $\mathbf{v}_1 = (a, b)^T$ and $\mathbf{v}_2 = (c, d)^T$ is

$$\delta(\mathbf{v}_1, \mathbf{v}_2) = ad - bc = \left\langle \mathbf{v}_1^{\perp}, \mathbf{v}_2 \right\rangle. \tag{B.83}$$

Lemma B.222. Adding a multiple of one vector to another does not change the signed volume.

Proof. This follows directly from (SVP-2,3). □

Lemma B.223. If the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent, then $\delta(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = 0$.

Proof. WLOG, we assume $\mathbf{v}_1 = \sum_{i=2}^n c_i \mathbf{v}_i$. Then the result follows from (SVP-2,3). □

Lemma B.224. The signed volume δ is alternating, i.e.,

$$\delta(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_n) = -\delta(\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n). \tag{B.84}$$

Exercise B.225. Prove Lemma B.224 using (SVP-2,3).

Lemma B.226. Let M_σ denote the matrix of a permutation $\sigma : E \to E$ where E is the set of standard basis vectors in (B.5). Then we have $\delta(M_\sigma) = \operatorname{sgn}(\sigma)$.

Proof. There is a one-to-one correspondence between the vectors in the matrix

$$M_\sigma = [e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}]$$

and the scalars in the one-line notation

$$(\sigma(1)\ \sigma(2)\ \ldots\ \sigma(n)).$$

A sequence of transpositions taking σ to the identity map also takes M_σ to the identity matrix. By Lemma B.224, each transposition yields a multiplication factor -1. Definition B.219 and (SVP-1) give $\delta(M_\sigma) = \operatorname{sgn}(\sigma)\delta(I) = \operatorname{sgn}(\sigma)$. □

Definition B.227 (Leibniz formula of determinants)**.** The *determinant of a square matrix* $A \in \mathbb{R}^{n\times n}$ is

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^n a_{\sigma(i),i}, \tag{B.85}$$

where the sum is over the symmetric group S_n of all permutations and $a_{\sigma(i),i}$ is the element of A at the $\sigma(i)$th row and the ith column.

Lemma B.228. The determinant of a matrix is the product of its eigenvalues, each of which is repeated according to its multiplicity.

Theorem B.229. The signed volume function satisfying (SVP-1,2,3) in Definition B.220 is unique and is the same as the determinant in (B.85).

Proof. Let the parallelotope be spanned by the column vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. We have

$$\begin{aligned}
&\delta \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1n} \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \ldots & v_{nn} \end{bmatrix} \\
&= \sum_{i_1=1}^n v_{i_1 1} \delta \begin{bmatrix} | & v_{12} & \ldots & v_{1n} \\ e_{i_1} & v_{22} & \ldots & v_{2n} \\ | & \vdots & \ddots & \vdots \\ | & v_{n2} & \ldots & v_{nn} \end{bmatrix} \\
&= \sum_{i_1, i_2=1}^n v_{i_1 1} v_{i_2 2} \delta \begin{bmatrix} | & | & v_{13} & \ldots & v_{1n} \\ e_{i_1} & e_{i_2} & v_{23} & \ldots & v_{2n} \\ | & | & \vdots & \ddots & \vdots \\ | & | & v_{n2} & \ldots & v_{nn} \end{bmatrix} \\
&= \cdots \\
&= \sum_{i_1, i_2, \ldots, i_n=1}^n v_{i_1 1} v_{i_2 2} \cdots v_{i_n n} \delta \begin{bmatrix} | & | & \ldots & | \\ e_{i_1} & e_{i_2} & \ldots & e_{i_n} \\ | & | & \ldots & | \end{bmatrix} \\
&= \sum_{\sigma \in S_n} v_{\sigma(1),1} v_{\sigma(2),2} \cdots v_{\sigma(n),n} \delta \begin{bmatrix} | & | & \ldots & | \\ e_{\sigma(1)} & e_{\sigma(2)} & \ldots & e_{\sigma(n)} \\ | & | & \ldots & | \end{bmatrix} \\
&= \sum_{\sigma \in S_n} v_{\sigma(1),1} v_{\sigma(2),2} \cdots v_{\sigma(n),n} \operatorname{sgn}(\sigma) \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^n v_{\sigma(i),i},
\end{aligned}$$

where the first four steps follow from (SVP-3), the sixth step from Lemma B.226, and the fifth step from (SVP-2). In other words, the signed volume $\delta(\cdot)$ is zero for any $i_j = i_k$ and hence the only nonzero terms are those of which $(i_1, i_2, \ldots, i_n)$ is a permutation of $(1, 2, \ldots, n)$. □

Exercise B.230. Use the formula in (B.85) to show that $\det A = \det A^T$.

Definition B.231. The i, j *cofactor* of $A \in \mathbb{R}^{n\times n}$ is

$$C_{ij} = (-1)^{i+j} M_{ij}, \tag{B.86}$$

where M_{ij} is the i, j *minor of a matrix* A, i.e. the determinant of the $(n-1)\times(n-1)$ matrix that results from deleting the i-th row and the j-th column of A.

Theorem B.232 (Laplace formula of determinants)**.** Given fixed indices $i, j \in 1, 2, \ldots, n$, the determinant of an n-by-n matrix $A = [a_{ij}]$ is given by

$$\det A = \sum_{j'=1}^n a_{ij'} C_{ij'} = \sum_{i'=1}^n a_{i'j} C_{i'j}. \tag{B.87}$$

Exercise B.233. Prove Theorem B.232 by induction.

Appendix C

Basic Analysis

C.1 Sequences

Definition C.1. A *sequence* is a function on $\mathbb{N}$.

Definition C.2. The *extended real number system* is the real line $\mathbb{R}$ with two additional elements $-\infty$ and $+\infty$:

$$\mathbb{R}^* := \mathbb{R} \cup \{-\infty, +\infty\}. \tag{C.1}$$

An extended real number $x \in \mathbb{R}^*$ is *finite* if $x \in \mathbb{R}$ and it is *infinite* otherwise.

Definition C.3. The *supremum of a sequence* $(a_n)_{n=m}^{\infty}$ is

$$\sup(a_n)_{n=m}^{\infty} := \sup\{a_n : n \geq m\}, \tag{C.2}$$

and the *infimum of a sequence* $(a_n)_{n=m}^{\infty}$ is

$$\inf(a_n)_{n=m}^{\infty} := \inf\{a_n : n \geq m\}. \tag{C.3}$$

C.1.1 Convergence

Definition C.4 (Limit of a sequence)**.** A sequence $\{a_n\}$ has the *limit* L, written $\lim_{n\to\infty} a_n = L$, or $a_n \to L$ as $n \to \infty$, iff

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N},\ \text{s.t.}\ \ \forall n > N,\ |a_n - L| < \epsilon. \tag{C.4}$$

If such a limit L exists, we say that $\{a_n\}$ *converges* to L.

Example C.5 (A story of π)**.** A famous estimation of π in ancient China is given by Zu, ChongZhi 1500 years ago,

$$\pi \approx \frac{355}{113} \approx 3.14159292.$$

In modern mathematics, we approximate π with a sequence for increasing accuracy, e.g.

$$\pi \approx 3.141592653589793\ldots \tag{C.5}$$

As of March 2019, we human beings have more than 31 trillion digits of π. However, real world applications never use even a small fraction of the 31 trillion digits:

- If you want to build a fence over your backyard swimming pool, several digits of π is probably enough;
- in NASA, calculations involving π use 15 digits for Guidance Navigation and Control;
- if you want to compute the circumference of the entire universe to the accuracy of less than the diameter of a hydrogen atom, you need only 39 decimal places of π.

On one hand, computational mathematics is judged by a metric that is different from that of pure mathematics; this may cause a huge gap between what needs to be done and what has been done. On the other hand, a computational mathematician cannot assume that a fixed accuracy is good enough for all applications. In the approximation a number or a function, she must develop theory and algorithms to provide the user the choice of an ever-increasing amount of accuracy, so long as the user is willing to invest an increasing amount of computational resources. This is one of the main motivations of infinite sequence and series.

Lemma C.6. A convergent sequence has a unique limit.

Definition C.7. A sequence $\{a_n\}$ is *Cauchy* if

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } m, n > N \ \Rightarrow\ |a_n - a_m| < \epsilon. \tag{C.6}$$

Lemma C.8. Every convergent sequence in $\mathbb{R}$ is Cauchy.

Proof. Since (x_n) converges to some $L \in \mathbb{R}$, for any given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n > N$ we have $|x_n - L| < \frac{\epsilon}{2}$. It follows that

$$\begin{aligned} \forall n, m > N,\ |x_n - x_m| &\leq |x_n - L + L - x_m| \\ &\leq |x_n - L| + |x_m - L| \leq \epsilon. \end{aligned}$$ □

Lemma C.9. If a Cauchy sequence contains a convergent subsequence, then the entire sequence converges to the same limit.

Proof. Suppose $\{a_n\}$ is a Cauchy sequence and $\{a_{n_j}\}$ is a subsequence converging to some $a \in \mathbb{R}$. It follows that

$$\begin{aligned} &\forall \epsilon > 0, \exists n_0 \text{ s.t. } \forall m, n \geq n_0, \quad |a_m - a_n| \leq \tfrac{\epsilon}{2}; \\ &\forall \epsilon > 0, \exists j_0 \text{ s.t. } \forall j \geq j_0, \quad |a_{n_j} - a| \leq \tfrac{\epsilon}{2}. \end{aligned}$$

Set $N = \max\{n_0, n_{j_0}\}$ and we have

$$\forall \epsilon > 0, \forall n \geq N, |a_n - a| \leq |a_n - a_N| + |a_N - a| \leq \epsilon,$$

which completes the proof. □

Lemma C.10. Every Cauchy sequence is bounded.

Lemma C.11. Every real sequence has a monotone subsequence.

Theorem C.12. A bounded monotone sequence is convergent.

Theorem C.13 (Bolzano-Weierstrass)**.** Every bounded sequence has a convergent subsequence.

Theorem C.14. Every Cauchy sequence in $\mathbb{R}$ converges to a limit in $\mathbb{R}$.

Proof. By Lemma C.10, the Cauchy sequence (a_n) is bounded. Theorem C.13 implies that $(a_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(a_{n_k})_{k\in\mathbb{N}}$. Then Lemma C.9 completes the proof. □

Theorem C.15 (Completeness of $\mathbb{R}$)**.** A sequence of real numbers is Cauchy if and only if it is convergent.

Proof. This is a summary of Lemma C.8 and Theorem C.14. □

C.1.2 Limit points

Definition C.16. Let $\epsilon > 0$ be a real number. Two real numbers x, y are said to be *ϵ-close* iff $|x - y| \le \epsilon$.

Definition C.17. A real number x is said to be *ϵ-adherent* to a sequence $(a_n)_{n=m}^{\infty}$ of real numbers iff there exists an $n \ge m$ such that a_n is ϵ-close to x. x is *continually ϵ-adherent* to $(a_n)_{n=m}^{\infty}$ iff it is ϵ-adherent to $(a_n)_{n=N}^{\infty}$ for every $N \ge m$.

Definition C.18. A real number x is a *limit point* or *adherent point* of a sequence $(a_n)_{n=m}^{\infty}$ of real numbers if it is continually ϵ-adherent to $(a_n)_{n=m}^{\infty}$ for every $\epsilon \ge 0$.

Definition C.19. The *limit superior* of a sequence $(a_n)_{n=m}^{\infty}$ of real numbers is

$$\limsup_{n\to\infty} a_n := \inf(a_N^+)_{N=m}^{\infty}, \tag{C.7}$$

where $a_N^+ = \sup(a_n)_{n=N}^{\infty}$. The *limit inferior* of $(a_n)_{n=m}^{\infty}$ is

$$\liminf_{n\to\infty} a_n := \sup(a_N^-)_{N=m}^{\infty}, \tag{C.8}$$

where $a_N^- = \inf(a_n)_{n=N}^{\infty}$.

Example C.20. Let $(a_n)_{n=m}^{\infty}$ be the sequence

$$1.1, -1.01, 1.001, -1.0001, 1.00001, \dots$$

Then $(a_n^+)_{n=N}^{\infty}$ is the sequence

$$1.1, 1.001, 1.001, 1.00001, 1.00001, \dots$$

and $(a_n^-)_{n=N}^{\infty}$ is the sequence

$$-1.01, -1.01, -1.0001, -1.0001, -1.000001, -1.000001, \dots$$

Hence we have

$$\limsup_{n\to\infty} a_n = 1, \qquad \liminf_{n\to\infty} a_n = -1.$$

Lemma C.21. Let $(a_n)_{n=m}^{\infty}$ be a sequence of real numbers. For $L^+ = \limsup a_n$ and $L^- = \liminf a_n$, we have

(a) For every $x > L^+$, elements of the sequence are eventually less than x:

$$\forall x > L^+, \exists N \ge m \text{ s.t. } \forall n \ge N, a_n < x.$$

Similarly, for every $x < L^-$, elements of the sequence are eventually greater than x:

$$\forall x < L^-, \exists N \ge m \text{ s.t. } \forall n \ge N, a_n > x.$$

(b) For every $x < L^+$, there are an infinite number of elements in the sequence that are greater than x:

$$\forall x < L^+, \forall N \ge m, \exists n \ge N \text{ s.t. } a_n > x.$$

Similarly, for every $x > L^-$, there are an infinite number of elements in the sequence that are less than x:

$$\forall x > L^-, \forall N \ge m, \exists n \ge N \text{ s.t. } a_n < x.$$

(c) $\inf(a_n)_{n=m}^{\infty} \le L^- \le L^+ \le \sup(a_n)_{n=m}^{\infty}$.

(d) Any limit point c of $(a_n)_{n=m}^{\infty}$ satisfies $L^- \le c \le L^+$.

(e) If L^+ (or L^-) is finite, then it is a limit point of $(a_n)_{n=m}^{\infty}$.

(f) $\lim_{n\to\infty} a_n = c$ if and only if $L^+ = L^- = c$.

Theorem C.22 (Squeeze test or the Sandwich Theorem)**.** Let $(a_n)_{n=m}^{\infty}$, $(b_n)_{n=m}^{\infty}$, and $(c_n)_{n=m}^{\infty}$ be sequences of real numbers that satisfy

$$\exists M \in \mathbb{N} \text{ s.t. } \forall n \ge M, a_n \le b_n \le c_n.$$

Suppose $(a_n)_{n=m}^{\infty}$ and $(c_n)_{n=m}^{\infty}$ both converge to the same limit L. Then $(b_n)_{n=m}^{\infty}$ also converges to L.

Notation 12 (Asymptotic notation)**.** For $g : \mathbb{R}^+ \to \mathbb{R}^+$, $f : \mathbb{R} \to \mathbb{R}$, and $a \in [0, +\infty]$, we write

$$f(x) = O(g(x)) \text{ as } x \to a$$

iff

$$\limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty.$$

In particular, we have

$$f(x) = o(g(x)) \text{ as } x \to a \;\Leftrightarrow\; \lim_{x\to a} \frac{|f(x)|}{g(x)} = 0.$$

We also write

$$f(x) = \Theta(g(x)) \text{ as } x \to a$$

iff

$$0 < \limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty.$$

C.2 Series

Definition C.23 (Finite series)**.** Let m, n be integers and let $(a_i)_{i=m}^n$ be a finite sequence of real numbers. The *finite series* or *finite sum* associated with the sequence $(a_i)_{i=m}^n$ is the number $\sum_{i=m}^n a_i$ given by the recursive formula

$$\sum_{i=m}^{n} a_i := \begin{cases} 0 & \text{if } n < m; \\ a_n + \sum_{i=m}^{n-1} a_i & \text{otherwise.} \end{cases} \tag{C.9}$$

Definition C.24 (Formal infinite series)**.** A (formal) *infinite series* associated with an infinite sequence $\{a_n\}$ is the expression $\sum_{n=0}^{\infty} a_n$.

Definition C.25. The *sequence of partial sums* $(S_n)_{n=0}^{\infty}$ associated with a formal infinite series $\sum_{i=0}^{\infty} a_i$ is defined for each n as the sum of the sequence $\{a_i\}$ from a_0 to a_n

$$S_n = \sum_{i=0}^{n} a_i. \tag{C.10}$$

Definition C.26. A formal infinite series is said to be *convergent* and *converge* to L if its sequence of partial sums converges to some limit L. In this case we write $L = \sum_{n=0}^{\infty} a_n$ and call L the *sum of the infinite series.*

Definition C.27. A formal infinite series is said to be *divergent* if its sequence of partial sums diverges. In this case we do not assign any real number value to this series.

Lemma C.28 (Cauchy criterion)**.** An infinite series $\sum_{n=0}^{\infty} a_n$ of real numbers is convergent if and only if

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ s.t. } \forall p, q \geq N,\ \left| \sum_{n=p}^{q} a_n \right| \leq \epsilon. \tag{C.11}$$

Theorem C.29 (Comparison test)**.** Let $\sum a_n$ be a series where $a_n \geq 0$ for all n.

(i) If $\sum a_n$ converges and $|b_n| \leq a_n$ for all n, then $\sum b_n$ converges.

(ii) If $\sum a_n = +\infty$ and $b_n \geq a_n$ for all n, then $\sum b_n = +\infty$.

Definition C.30. An infinite series $\sum_{n=0}^{\infty} a_n$ is *absolutely convergent* iff the series $\sum_{n=0}^{\infty} |a_n|$ is convergent.

Lemma C.31. An infinite series that is absolutely convergent is convergent.

Theorem C.32 (Ratio test)**.** A series $\sum a_n$ of nonzero terms

(i) converges absolutely if $\limsup_{n\to\infty} |a_{n+1}/a_n| < 1$;

(ii) diverges if $\liminf_{n\to\infty} |a_{n+1}/a_n| > 1$,

Otherwise, this test gives no information about the convergence of $\sum a_n$.

Theorem C.33 (Root test)**.** Let $\sum a_n$ be a series and $\alpha = \limsup_{n\to\infty} |a_n|^{1/n}$. The series $\sum a_n$

(i) converges absolutely if $\alpha < 1$;

(ii) diverges if $\alpha > 1$,

Otherwise, this test gives no information about the convergence of $\sum a_n$.

Proof. Choose $\epsilon > 0$ such that $\alpha + \epsilon \in (0, 1)$. Then

$$\exists N \in \mathbb{N}^+ \text{ s.t. } \forall n \geq N,\ \sup_{n \geq N} \{a_n^{\frac{1}{n}}\} < \alpha + \epsilon.$$

Hence $a_n < (\alpha + \epsilon)^n$ for all $n \geq N$. The comparison test with the geometric series yields (i). (ii) can be shown via proof by contradiction. □

Theorem C.34 (Integral test)**.** Let $f : [0, \infty) \to \mathbb{R}$ be a monotone decreasing function which is non-negative. Then the series $\sum_{n=0}^{\infty} f(n)$ is convergent if and only if $\sup_{N>0} \int_0^N f$ is finite.

C.3 Continuous functions on $\mathbb{R}$

Definition C.35. A *scalar function* is a function whose range is a subset of $\mathbb{R}$.

Definition C.36 (Limit of a scalar function with one variable)**.** Consider a function $f : I \to \mathbb{R}$ with $I(c, r) = (c-r, c) \cup (c, c+r)$. The *limit* of $f(x)$ exists as x approaches c, written $\lim_{x\to c} f(x) = L$, iff

$$\forall \epsilon > 0, \exists \delta > 0,\ \text{s.t. } \forall x \in I(c, \delta),\ |f(x) - L| < \epsilon. \tag{C.12}$$

Example C.37. Show that $\lim_{x\to 2} \frac{1}{x} = \frac{1}{2}$.

Proof. If $\epsilon \geq \frac{1}{2}$, choose $\delta = 1$. Then $x \in (1, 3)$ implies $\left|\frac{1}{x} - \frac{1}{2}\right| < \frac{1}{2}$ since $\frac{1}{x} - \frac{1}{2}$ is a monotonically decreasing function with its supremum at $x = 1$.

If $\epsilon \in (0, \frac{1}{2})$, choose $\delta = \epsilon$. Then $x \in (2-\epsilon, 2+\epsilon) \subset (\frac{3}{2}, \frac{5}{2})$. Hence $\left|\frac{1}{x} - \frac{1}{2}\right| = \frac{|2-x|}{|2x|} < |2-x| < \epsilon$. The proof is completed by Definition C.36. □

Definition C.38. $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at c iff

$$\lim_{x\to c} f(x) = f(c). \tag{C.13}$$

Definition C.39. A scalar function f is *continuous on* (a, b), written $f \in \mathcal{C}(a, b)$, if (C.13) holds $\forall x \in (a, b)$.

Theorem C.40 (Extreme values)**.** A continuous function $f : [a, b] \to \mathbb{R}$ attains its maximum at some point $x_{\max} \in [a, b]$ and its minimum at some point $x_{\min} \in [a, b]$.

Theorem C.41 (Intermediate value)**.** A scalar function $f \in \mathcal{C}[a, b]$ satisfies

$$\forall y \in [m, M],\ \exists \xi \in [a, b],\ \text{s.t. } y = f(\xi) \tag{C.14}$$

where $m = \inf_{x\in[a,b]} f(x)$ and $M = \sup_{x\in[a,b]} f(x)$.

Definition C.42. Let $I = (a, b)$. A function $f : I \to \mathbb{R}$ is *uniformly continuous* on I iff

$$\begin{aligned} &\forall \epsilon > 0, \exists \delta > 0,\ \text{s.t.} \\ &\forall x, y \in I,\ |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon. \end{aligned} \tag{C.15}$$

Example C.43. Show that, on (a, ∞), $f(x) = \frac{1}{x}$ is uniformly continuous if $a > 0$ and is not so if $a = 0$.

Proof. If $a > 0$, then $|f(x) - f(y)| = \frac{|x-y|}{xy} < \frac{|x-y|}{a^2}$.
Hence $\forall \epsilon > 0, \exists \delta = a^2\epsilon$, s.t.
$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \frac{|x-y|}{a^2} < \frac{a^2\epsilon}{a^2} = \epsilon$.

If $a = 0$, negating the condition of uniform continuity, i.e. eq. (C.15), yields $\exists \epsilon > 0$ s.t. $\forall \delta > 0\ \exists x, y > 0$ s.t. $(|x - y| < \delta) \wedge (|\frac{1}{x} - \frac{1}{y}| \geq \epsilon)$.

We prove a stronger version: $\forall \epsilon > 0$, $\forall \delta > 0\ \exists x, y > 0$ s.t. $(|x - y| < \delta) \wedge (|f(x) - f(y)| \geq \epsilon)$.

If $\delta \geq \frac{1}{2\epsilon}$, choose $x = \frac{1}{2\epsilon}$, $y = \frac{1}{4\epsilon}$. This choice satisfies $|x - y| < \delta$ since $x - y = \frac{1}{4\epsilon} < \frac{1}{2\epsilon} \leq \delta$. However, $|f(x) - f(y)| = \frac{|x-y|}{xy} = 2\epsilon > \epsilon$.

If $\delta < \frac{1}{2\epsilon}$, then $2\epsilon\delta < 1$. Choose $x \in (0, \epsilon\delta^2)$ and $y \in (2\epsilon\delta^2, \delta)$. This choice satisfies $|x - y| < \delta$ and $|x - y| > \epsilon\delta^2$. However, $|f(x) - f(y)| = \frac{|x-y|}{xy} > \frac{\epsilon\delta^2}{xy} > \frac{1}{y} > \frac{1}{\delta} > 2\epsilon > \epsilon$. □

Exercise C.44. On (a, ∞), $f(x) = \frac{1}{x^2}$ is uniformly continuous if $a > 0$ and is not so if $a = 0$.

Theorem C.45. Uniform continuity implies continuity but the converse is not true.

Proof. exercise. □

Theorem C.46. $f : \mathbb{R} \to \mathbb{R}$ is uniformly continuous on (a, b) iff it can be extended to a continuous function $\tilde{f}$ on $[a, b]$.

C.4 Differentiation of functions

Definition C.47. The *derivative* of a function $f : \mathbb{R} \to \mathbb{R}$ at a is the limit

$$f'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}. \tag{C.16}$$

If the limit exists, f is *differentiable* at a.

Example C.48. For the power function $f(x) = x^\alpha$, we have $f' = \alpha x^{\alpha-1}$ due to Newton's generalized binomial theorem,

$$(a + h)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} a^{\alpha-n} h^n.$$

Definition C.49. A function $f(x)$ is k times *continuously differentiable* on (a, b) iff $f^{(k)}(x)$ exists on (a, b) and is itself continuous. The set or space of all such functions on (a, b) is denoted by $\mathcal{C}^k(a, b)$. In comparison, $\mathcal{C}^k[a, b]$ is the space of functions f for which $f^{(k)}(x)$ is bounded and uniformly continuous on (a, b).

Theorem C.50. A scalar function f is bounded on $[a, b]$ if $f \in \mathcal{C}[a, b]$.

Theorem C.51. If $f : (a, b) \to \mathbb{R}$ assumes its maximum or minimum at $x_0 \in (a, b)$ and f is differentiable at x_0, then $f'(x_0) = 0$.

Proof. Suppose $f'(x_0) > 0$. Then we have

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} > 0.$$

The definition of a limit implies

$$\exists \delta > 0 \text{ s.t. } a < x_0 - \delta < x_0 + \delta < b,$$

which, together with $|x - x_0| < \delta$, implies $\frac{f(x)-f(x_0)}{x-x_0} > 0$. This is a contradiction to $f(x_0)$ being a maximum when we choose $x \in (x_0, x_0 + \delta)$. □

Theorem C.52 (Rolle's)**.** If a function $f : \mathbb{R} \to \mathbb{R}$ satisfies

(i) $f \in \mathcal{C}[a, b]$ and f' exists on (a, b),

(ii) $f(a) = f(b)$,

then $\exists x \in (a, b)$ s.t. $f'(x) = 0$.

Proof. By Theorem C.41, all values between $\sup f$ and $\inf f$ will be assumed. If $f(a) = f(b) = \sup f = \inf f$, then f is a constant on $[a, b]$ and thus the conclusion holds. Otherwise, Theorem C.51 completes the proof. □

Theorem C.53 (Mean value)**.** If $f \in \mathcal{C}[a, b]$ and if f' exists on (a, b), then $\exists \xi \in (a, b)$ s.t. $f(b) - f(a) = f'(\xi)(b - a)$.

Proof. Construct a linear function $L : [a, b] \to \mathbb{R}$ such that $L(a) = f(a)$, $L(b) = f(b)$, then $\forall x \in (a, b)$, we have $L'(x) = \frac{f(b)-f(a)}{b-a}$. Consider $g(x) = f(x) - L(x)$ on $[a, b]$. $g(a) = 0$, $g(b) = 0$. By Theorem C.52, $\exists \xi \in [a, b]$ such that $g'(\xi) = 0$, which completes the proof. □

C.5 Taylor series

Definition C.54. A *power series* centered at c is a series of the form

$$p(x) = \sum_{n=0}^{\infty} a_n (x - c)^n, \tag{C.17}$$

where a_n's are the *coefficients*. The *interval of convergence* is the set of values of x for which the series converges:

$$I_c(p) = \{x \mid p(x) \text{ converges}\}. \tag{C.18}$$

Definition C.55. If the derivatives $f^{(i)}(x)$ with $i = 1, 2, \ldots, n$ exist for a function $f : \mathbb{R} \to \mathbb{R}$ at $x = c$, then

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x - c)^k \tag{C.19}$$

is called the nth *Taylor polynomial* for $f(x)$ at c.
In particular, the *linear approximation* for $f(x)$ at c is

$$T_1(x) = f(c) + f'(c)(x - c). \tag{C.20}$$

Example C.56. If $f \in \mathcal{C}^\infty$, then $\forall n \in \mathbb{N}$, we have

$$T_n^{(m)}(x) = \begin{cases} \sum_{k=m}^{n} \frac{f^{(k)}(c)}{(k-m)!} (x - c)^{k-m}, & m \in \mathbb{N}, m \leq n; \\ 0, & m \in \mathbb{N}, m > n. \end{cases}$$

This can be proved by induction. In the inductive step, we regroup the summation into a constant term and another shifted summation.

Definition C.57. The *Taylor series* (or Taylor expansion) for $f(x)$ at c is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!} (x - c)^k. \tag{C.21}$$

Definition C.58. The *remainder* of the nth *Taylor polynomial* in approximating $f(x)$ is

$$E_n(x) = f(x) - T_n(x). \tag{C.22}$$

Theorem C.59. Let T_n be the nth Taylor polynomial for $f(x)$ at c.

$$\lim_{n\to\infty} E_n(x) = 0 \;\Leftrightarrow\; \lim_{n\to\infty} T_n(x) = f(x). \tag{C.23}$$

Lemma C.60. $\forall m = 0, 1, 2, \ldots, n$, $E_n^{(m)}(c) = 0$.

Proof. This follows from Definition C.55 and Example C.56. □

Theorem C.61 (Taylor's theorem with Lagrangian form)**.** Consider a function $f : \mathbb{R} \to \mathbb{R}$. If $f \in \mathcal{C}^n[c-d, c+d]$ and $f^{(n+1)}(x)$ exists on $(c-d, c+d)$, then $\forall x \in [c-d, c+d]$, there exists some ξ between c and x such that

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}. \tag{C.24}$$

Proof. Fix $x \neq c$, let M be the unique solution of

$$E_n(x) = f(x) - T_n(x) = \frac{M(x-c)^{n+1}}{(n+1)!}.$$

Consider the function

$$g(t) := E_n(t) - \frac{M(t-c)^{n+1}}{(n+1)!}. \tag{C.25}$$

Clearly $g(x) = 0$. By Lemma C.60, $g^{(k)}(c) = 0$ for each $k = 0, 1, \ldots, n$. Then Rolle's theorem implies that

$$\exists x_1 \in (c, x) \text{ s.t. } g'(x_1) = 0.$$

If $x < c$, change (c, x) above to (x, c). Apply Rolle's theorem to $g'(t)$ on (c, x_1) and we have

$$\exists x_2 \in (c, x_1) \text{ s.t. } g^{(2)}(x_2) = 0.$$

Repeatedly using Rolle's theorem,

$$\exists x_{n+1} \in (c, x_n) \text{ s.t. } g^{(n+1)}(x_{n+1}) = 0. \tag{C.26}$$

Since T_n is a polynomial of degree n, we have $T_n^{(n+1)}(t) = 0$, which, together with (C.26) and (C.25), yields

$$f^{(n+1)}(x_{n+1}) - M = 0.$$

The proof is completed by identifying ξ with x_{n+1}. □

Example C.62. How many terms are needed to compute e^2 correctly to four decimal places?

The requirement of four decimal places means an accuracy of at least $\epsilon = 10^{-5}$. By Definition C.57, the Taylor series of e^x at $c = 0$ is

$$e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}.$$

By Theorem C.61, we have

$$\exists \xi \in [0, 2] \text{ s.t. } E_n(2) = e^{\xi} 2^{n+1}/(n+1)! < e^2 2^{n+1}/(n+1)!$$

Then $e^2 2^{n+1}/(n+1)! \leq \epsilon$ yields $n \geq 12$, i.e., 13 terms.

C.6 Riemann integral

Definition C.63. A *partition of an interval* $I = [a, b]$ is a totally-ordered finite subset $P_n \subseteq I$ of the form

$$P_n(a, b) = \{a = x_0 < x_1 < \cdots < x_n = b\}. \tag{C.27}$$

The interval $I_i = [x_{i-1}, x_i]$ is the ith *subinterval* of the partition. The *norm* of the partition is the length of the longest subinterval,

$$h_n = h(P_n) = \max(x_i - x_{i-1}), \qquad i = 1, 2, \ldots, n. \tag{C.28}$$

Definition C.64. The *Riemann sum* of $f : \mathbb{R} \to \mathbb{R}$ over a partition P_n is

$$S_n(f) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}), \tag{C.29}$$

where $x_i^* \in I_i$ is a *sample point* of the ith subinterval.

Definition C.65. A function $f : \mathbb{R} \to \mathbb{R}$ is *Riemann integrable* on $[a, b]$ iff

$$\begin{aligned} &\exists L \in \mathbb{R}, \text{ s.t. } \forall \epsilon > 0, \ \exists \delta > 0 \text{ s.t.} \\ &\forall P_n(a, b) \text{ with } h(P_n) < \delta, \ |S_n(f) - L| < \epsilon. \end{aligned} \tag{C.30}$$

In this case we write $L = \int_a^b f(x)\mathrm{d}x$ and call it the *Riemann integral* of f on $[a, b]$.

Example C.66. The following function $f : [a, b] \to \mathbb{R}$ is not Riemann integrable.

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational;} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

To see this, we first negate the logical statement in (C.30) to get

$$\begin{aligned} &\forall L \in \mathbb{R}, \exists \epsilon > 0, \text{ s.t. } \forall \delta > 0 \\ &\exists P_n(a, b) \text{ with } h(P_n) < \delta, \ \text{ s.t. } |S_n(f) - L| \geq \epsilon. \end{aligned}$$

If $|L| < \frac{b-a}{2}$, we choose all x_i^*'s to be rational so that $f(x_i^*) \equiv 1$; then (C.29) yields $S_n(f) = b - a$. For $\epsilon = \frac{b-a}{4}$, the formula $|S_n(f) - L| \geq \epsilon$ clearly holds.

If $|L| \geq \frac{b-a}{2}$, we choose all x_i^*'s to be irrational so that $f(x_i^*) \equiv 0$; then (C.29) yields $S_n(f) = 0$. For $\epsilon = \frac{b-a}{4}$, the formula $|S_n(f) - L| \geq \epsilon$ clearly holds.

Definition C.67. If $f : \mathbb{R} \to \mathbb{R}$ is integrable on $[a, b]$, then the limit of the Riemann sum of f is called the *definite integral* of f on $[a, b]$:

$$\int_a^b f(x)\mathrm{d}x = \lim_{h_n \to 0} S_n(f). \tag{C.31}$$

Theorem C.68. A scalar function f is integrable on $[a, b]$ if $f \in \mathcal{C}[a, b]$.

Definition C.69. A *monotonic* function is a function between ordered sets that either preserves or reverses the given order. In particular, $f : \mathbb{R} \to \mathbb{R}$ is *monotonically increasing* if $\forall x, y,\ x \le y \Rightarrow f(x) \le f(y)$; $f : \mathbb{R} \to \mathbb{R}$ is *monotonically decreasing* if $\forall x, y,\ x \le y \Rightarrow f(x) \ge f(y)$.

Theorem C.70. A scalar function is integrable on $[a, b]$ if it is monotonic on $[a, b]$.

Exercise C.71. True or false: a bijective function is either order-preserving or order-reversing?

Theorem C.72 (Integral mean value)**.** Let $w : [a, b] \to \mathbb{R}^+$ be integrable on $[a, b]$. For $f \in \mathcal{C}[a, b]$, $\exists \xi \in [a, b]$ s.t.

$$\int_a^b w(x)f(x)\mathrm{d}x = f(\xi)\int_a^b w(x)\mathrm{d}x. \tag{C.32}$$

Proof. Denote $m = \inf_{x\in[a,b]} f(x)$, $M = \sup_{x\in[a,b]} f(x)$, and $I = \int_a^b w(x)\mathrm{d}x$. Then $mw(x) \le f(x)w(x) \le Mw(x)$ and

$$mI \le \int_a^b w(x)f(x)\mathrm{d}x \le MI.$$

$w > 0$ implies $I \neq 0$, hence

$$m \le \frac{1}{I}\int_a^b w(x)f(x)\mathrm{d}x \le M.$$

Applying Theorem C.41 completes the proof. □

Theorem C.73 (First fundamental theorem of calculus)**.** Let $a < b$ be real numbers. For a continuous function $f : [a, b] \to \mathbb{R}$ that is Riemann integrable, define a function $F : [a, b] \to \mathbb{R}$ by

$$F(x) := \int_a^x f(y)\mathrm{d}y. \tag{C.33}$$

Then F is differentiable and

$$\forall x_0 \in [a, b], \quad F'(x_0) = f(x_0). \tag{C.34}$$

Theorem C.74 (Second fundamental theorem of calculus)**.** Let $a < b$ be real numbers and let $f : [a, b] \to \mathbb{R}$ be a Riemann integrable function. If $F : [a, b] \to \mathbb{R}$ is the antiderivative of f, i.e. $F'(x) = f(x)$, then

$$\int_a^b f = F(b) - F(a). \tag{C.35}$$

C.7 Convergence in metric spaces

Definition C.75. A *metric* is a function $d : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ that satisfies, for all $x, y, z \in \mathcal{X}$,

(1) non-negativity: $d(x, y) \ge 0$;

(2) identity of indiscernibles: $x = y \Leftrightarrow d(x, y) = 0$;

(3) symmetry: $d(x, y) = d(y, x)$;

(4) triangle inequality: $d(x, z) \le d(x, y) + d(y, z)$.

A *metric space* is an ordered pair $(\mathcal{X}, d)$ where $\mathcal{X}$ is a set and d is a metric on $\mathcal{X}$.

Example C.76. Set $\mathcal{X}$ to be $C[a, b]$, the set of continuous functions $[a, b] \to \mathbb{R}$. Then the following is a metric on $\mathcal{X}$,

$$d(x, y) = \max_{t\in[a,b]} |x(t) - y(t)|. \tag{C.36}$$

Definition C.77 (Limiting value of a function)**.** Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be metric spaces. Let E be a subset of $\mathcal{X}$ and $x_0 \in \mathcal{X}$ be an adherent point of E. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to *converge* to $L \in \mathcal{Y}$ as x converges to $x_0 \in E$, written

$$\lim_{x\to x_0; x\in E} f(x) = L, \tag{C.37}$$

iff

$$\begin{aligned}&\forall \epsilon > 0, \exists \delta > 0 \text{ s.t. } \forall x \in E,\\ &\qquad |x - x_0|_\mathcal{X} < \delta \ \Rightarrow\ |f(x) - L|_\mathcal{Y} < \epsilon.\end{aligned} \tag{C.38}$$

Notation 13. In Definition C.77 we used the synonym notation

$$|u - v|_\mathcal{X} := d_\mathcal{X}(u, v). \tag{C.39}$$

Definition C.78 (Pointwise convergence)**.** Let $(f_n)_{n=1}^\infty$ be a sequence of functions from one metric space $(\mathcal{X}, d_\mathcal{X})$ to another $(\mathcal{Y}, d_\mathcal{Y})$, and let $f : \mathcal{X} \to \mathcal{Y}$ be another function. We say that $(f_n)_{n=1}^\infty$ *converges pointwise* to f on $\mathcal{X}$ iff

$$\forall x \in \mathcal{X}, \qquad \lim_{n\to\infty} f_n(x) = f(x), \tag{C.40}$$

or, equivalently,

$$\forall \epsilon > 0, \forall x \in \mathcal{X},\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > N,\ |f_n(x) - f(x)|_\mathcal{Y} < \epsilon. \tag{C.41}$$

Example C.79. Consider $f_n : [0, 1] \to \mathbb{R}$ defined by $f_n(x) := x^n$ and $f : [0, 1] \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} 1 & \text{if } x = 1;\\ 0 & \text{if } x \in [0, 1).\end{cases}$$

The functions f_n are continuous and converge pointwise to f, which is discontinuous. Hence pointwise convergence does not preserve continuity.

Example C.80. For the functions in Example C.79, we have $\lim_{x\to1;x\in[0,1)} x^n = 1$ for all n and $\lim_{x\to1;x\in[0,1)} f(x) = 0$; it follows that

$$\lim_{n\to\infty}\lim_{x\to x_0;x\in\mathcal{X}} f_n(x) \neq \lim_{x\to x_0;x\in\mathcal{X}}\lim_{n\to\infty} f_n(x).$$

Hence pointwise convergence does not preserve limits.

Example C.81. Consider the interval $[a, b] = [0, 1]$, and the function sequence $f_n : [a, b] \to \mathbb{R}$ given by

$$f_n(x) := \begin{cases} 2n & \text{if } x \in \left[\frac{1}{2n}, \frac{1}{n}\right];\\ 0 & \text{otherwise.}\end{cases}$$

Then (f_n) converges pointwise to $f(x) = 0$. However, $\int_a^b f_n = 1$ for every n while $\int_a^b f = 0$. Hence

$$\lim_{n\to\infty}\int_a^b f_n \neq \int_a^b \lim_{n\to\infty} f_n.$$

Hence pointwise convergence does not preserve integral.

Example C.82. Pointwise convergence does not preserve boundedness. For example, the function sequence

$$f_n(x) = \begin{cases} \exp(x) & \text{if } \exp(x) \le n; \\ n & \text{if } \exp(x) > n \end{cases} \tag{C.42}$$

converges pointwise to $f(x) = \exp(x)$. Similarly, the function sequence

$$f_n(x) = \begin{cases} \frac{1}{x} & \text{if } x \ge \frac{1}{n}; \\ 0 & \text{if } x \in (0, \frac{1}{n}) \end{cases} \tag{C.43}$$

converges pointwise to $f(x) = \frac{1}{x}$. As another example, the function sequence

$$f_n(x) = n \sin \frac{x}{n} \tag{C.44}$$

converges pointwise to $f(x) = x$.

Definition C.83 (Uniform convergence)**.** Let $(f_n)_{n=1}^{\infty}$ be a sequence of functions from one metric space $(\mathcal{X}, d_{\mathcal{X}})$ to another $(\mathcal{Y}, d_{\mathcal{Y}})$, and let $f : \mathcal{X} \to \mathcal{Y}$ be another function. We say that $(f_n)_{n=1}^{\infty}$ *converges uniformly* to f on $\mathcal{X}$ iff

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall x \in \mathcal{X}, \forall n > N,\ |f_n(x) - f(x)|_{\mathcal{Y}} < \epsilon. \tag{C.45}$$

The sequence (f_n) is *locally uniformly convergent* to f iff for every point $x \in \mathcal{X}$ there is an $r > 0$ such that $(f_n|_{B_r(x)\cap\mathcal{X}})$ is uniformly convergent to f on $B_r(x) \cap \mathcal{X}$.

Example C.84. Consider $f_n : [0, 1] \to \mathbb{R}$ with $f_n(x) := \frac{x}{n}$. Then (f_n) converges uniformly to $f(x) = 0$.

Theorem C.85. Uniform convergence implies pointwise convergence.

Proof. This follows directly from (C.41), (C.45), and Theorem A.10. □

Example C.86 (Uniform convergence of Taylor series)**.** Consider $f : \mathbb{R} \to \mathbb{R}$ and the sequence of its Taylor polynomial $(T_n)_{n=1}^{\infty}$ in Definition C.55. For any interval $I_r := (a - r, a + r)$, $(T_n)_{n=1}^{\infty}$ converges locally uniformly to $f|_{I_r}$ if r is less or equal to the radius of convergence of f at a. In particular, $(T_n)_{n=1}^{\infty}$ converges locally uniformly to f if the radius of convergence of f is $+\infty$.

Theorem C.87. Consider $b_{ij} \ge 0$ in $\mathbb{R}^* := \mathbb{R} \cup \{+\infty, -\infty\}$. If b_{ij} is monotone increasing in i for each j and is monotone increasing in j for each i, then we have

$$\lim_{i=0}^{\infty} \lim_{j=0}^{\infty} b_{ij} = \lim_{j=0}^{\infty} \lim_{i=0}^{\infty} b_{ij} \tag{C.46}$$

with all the indicated limits existing in $\mathbb{R}^*$.

Theorem C.88. Suppose that $\{f_n : [a, b] \to \mathbb{R}\}$ is a sequence of continuous functions satisfying

- $\{f_n(x_0)\}$ converges for some $x_0 \in [a, b]$,
- each f_n is differentiable on (a, b),
- $\{f_n'\}$ converges uniformly on (a, b).

Then we have

- $\{f_n\}$ converges uniformly on $[a, b]$ to a function f,
- both $f'(x)$ and $\lim_n f_n'(x)$ exist for any $x \in (a, b)$,
- $f'(x) = \lim_n f_n'(x)$ for any $x \in (a, b)$.

C.8 Vector calculus

Lemma C.89. For $E \subset \mathbb{R}$, $f : E \to \mathbb{R}$, $x_0 \in E$, and $L \in \mathbb{R}$, the following two statements are equivalent,

(a) f is differentiable at x_0 and $f'(x_0) = L$;

(b) $\lim_{x\to x_0, x\in E\setminus\{x_0\}} \frac{|f(x)-f(x_0)-L(x-x_0)|}{|x-x_0|} = 0$.

Exercise C.90. Prove Lemma C.89.

Definition C.91 (Total derivative)**.** For $E \subset \mathbb{R}^n$, $f : E \to \mathbb{R}^m$, $x_0 \in E$, f is *differentiable at x_0 with derivative* $L : \mathbb{R}^n \to \mathbb{R}^m$ if

$$\lim_{x\to x_0, x\in E\setminus\{x_0\}} \frac{\|f(x) - f(x_0) - L(x - x_0)\|_2}{\|x - x_0\|_2} = 0. \tag{C.47}$$

We denote the derivative of f with $f'(x_0) = L$ and also call it the *total derivative of f*.

Example C.92. For $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $L : \mathbb{R}^2 \to \mathbb{R}^2$,

$$f(x, y) := (x^2, y^2), \quad L(x, y) := (2x, 4y), \tag{C.48}$$

we claim that f is differentiable at $(1, 2)$ with $f'(x_0) = L$. To show this, we compute

$$\begin{aligned}
&\lim_{\substack{(x,y)\to(1,2)\\(x,y)\neq(1,2)}} \frac{\|f(x,y) - f(1,2) - L((x,y) - (1,2))\|_2}{\|(x,y) - (1,2)\|_2} \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \frac{\|f(1+a, 2+b) - f(1,2) - L(a,b)\|_2}{\|(a,b)\|_2} \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \frac{\|\left((1+a)^2, (2+b)^2\right) - (1,4) - (2a,4b)\|_2}{\|(a,b)\|_2} \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \frac{\|(a^2, b^2)\|_2}{\|(a,b)\|_2} \\
\le&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \left(\frac{\|(a^2,0)\|_2}{\|(a,b)\|_2} + \frac{\|(0,b^2)\|_2}{\|(a,b)\|_2}\right) \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \sqrt{a^2+b^2} \\
=&0.
\end{aligned}$$

Lemma C.93. Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, and $x_0 \in E$ an interior point of E. Suppose f is differentiable at x_0 with derivative L_1 and also differentiable at x_0 with derivative L_2. Then $L_1 = L_2$.

Exercise C.94. Prove Lemma C.93.

Definition C.95 (Directional derivative)**.** Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, $x_0 \in E$ an interior point of E, and $\mathbf{v} \in \mathbb{R}^n$ a vector. If the limit

$$\lim_{t\to 0; t>0, x_0+t\mathbf{v}\in E} \frac{f(x_0+t\mathbf{v})-f(x_0)}{t}$$

exists, we say that f is *differentiable in the direction* $\mathbf{v}$ *at* x_0, and we denote this limit as

$$D_{\mathbf{v}} f(x_0) := \lim_{t\to 0; t>0, x_0+t\mathbf{v}\in E} \frac{f(x_0+t\mathbf{v})-f(x_0)}{t}. \quad \text{(C.49)}$$

Example C.96. For $\mathbf{v} = (3, 4)$ and $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined in (C.48), we have $D_{\mathbf{v}} f(1, 2) = (6, 16)$.

Example C.97. For $f : \mathbb{R} \to \mathbb{R}$, $D_{+1}f(x)$ is the right derivative of f at x (if it exists), and similarly $D_{-1}f(x)$ is the left derivative of f at x (if it exists).

Lemma C.98. Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, $x_0 \in E$ an interior point of E, and $\mathbf{v} \in \mathbb{R}^n$ a vector. If f is differentiable at x_0, then f is also differentiable in the direction $\mathbf{v}$ at x_0, and

$$D_{\mathbf{v}} f(x_0) = f'(x_0)\mathbf{v}. \quad \text{(C.50)}$$

Definition C.99 (Partial derivative)**.** Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, $x_0 \in E$ an interior point of E, and $1 \le j \le n$. The *partial derivative of* f *with respect to the* x_j *variable at* x_0 is defined by

$$\begin{aligned} \frac{\partial f}{\partial x_j}(x_0) &:= \lim_{t\to 0; t>0, x_0+te_j\in E} \frac{f(x_0+te_j)-f(x_0)}{t} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\, f(x_0+te_j)|_{t=0} \end{aligned} \quad \text{(C.51)}$$

provided that the limit exists. Here e_j is the jth standard basis vector of $\mathbb{R}^n$.

Exercise C.100. Show that the existence of partial derivatives at x_0 does not imply that the function is differentiable at x_0 by considering the differentiability of the following function $f : \mathbb{R}^2 \to \mathbb{R}$ at (0,0).

$$f(x, y) = \begin{cases} \frac{x^3}{x^2+y^2} & \text{if } (x, y) \neq (0, 0); \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

Theorem C.101. Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, F a subset of E, and $x_0 \in E$ an interior point of F. If all the partial derivatives $\frac{\partial f}{\partial x_j}$ exist on F and are continuous at x_0, then f is differentiable at x_0, and the linear transformation $f'(x_0) : \mathbb{R}^n \to \mathbb{R}^m$ is defined by

$$f'(x_0)(\mathbf{v}) = \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(x_0). \quad \text{(C.52)}$$

Definition C.102. The *derivative matrix* or *differential matrix* or *Jacobian matrix* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a $m \times n$ matrix,

$$Df := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}. \quad \text{(C.53)}$$

Theorem C.103 (Implicit function theorem)**.** Suppose a $\mathcal{C}^1$ function $\mathbf{g} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ satisfies

(i) $\mathbf{g}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$ where $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^m \times \mathbb{R}^n$;

(ii) the Jacobian matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)$ is invertible.

Then there is a neighborhood U of $\mathbf{x}_0$ and a unique $\mathcal{C}^1$ function $\mathbf{f} : U \to \mathbb{R}^n$ such that

(i) $\mathbf{f}(\mathbf{x}_0) = \mathbf{y}_0$;

(ii) $\mathbf{g}(\mathbf{x}, \mathbf{f}(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in U$.

Furthermore, if $\mathbf{g}$ is analytic or $\mathcal{C}^p$, then $\mathbf{f}$ is analytic or $\mathcal{C}^p$.

Appendix D

Point-set Topology

D.1 Topological spaces

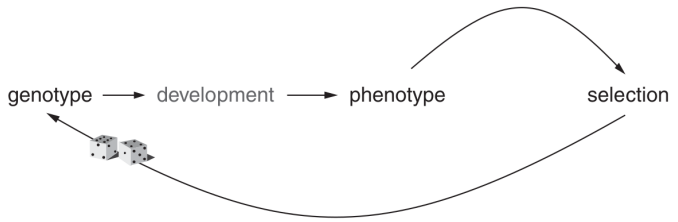
D.1.1 A motivating problem from biology

Definition D.1. *Phenotype* refers to the physical, organizational, and behavioral expression of an organism during its lifetime while *genotype* refers to a heritable repository of information that instructs the production of molecules, whose interactions with the environment generate and maintain the phenotype.

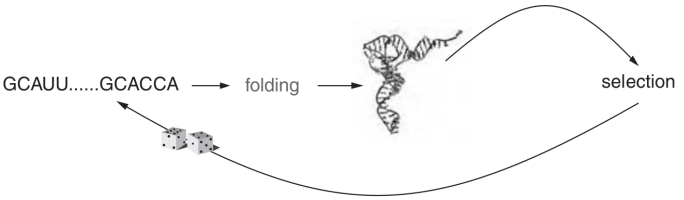
Example D.2. The collection of genes responsible for eye color in a particular individual is a genotype while the observable eye coloration in the individual is the corresponding phenotype.

Phenotype and genotype are two fundamental concepts in the classical framework of evolution.

The genotype-phenotype relationship is of great importance in biology in that evolution is driven by the selection of phenotypes that causes the amplification of their underlying genotypes and the production of novel phenotypes through genetic mutation.



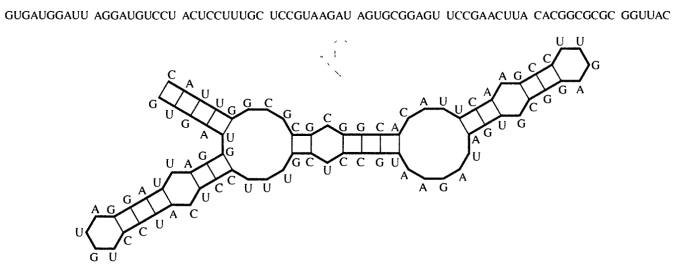
In phenotypic innovation, the heritable modification of a phenotype usually does not involve a direct intervention at the phenotypic level, but proceeds indirectly through changes at the genetic level during a number of processes known as development. While selection is clearly an important driving force of evolution, the dynamics of selection does not tell us much about how evolutionary innovations arise in the first place. A mutation is advantageous if it generates a phenotype favored by selection, but this definition reveals nothing about why or how that mutation could innovate the phenotype. Hence a model of genotype-phenotype relation is needed to illuminate how genetic changes map into phenotypic changes.



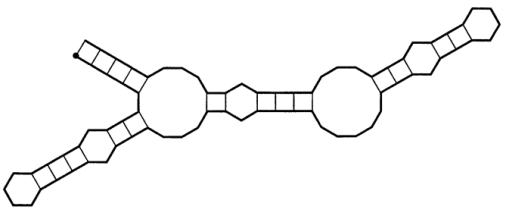
This subsection concerns such a model proposed by Fontana and Schuster [1998a,b] and Stadler et al. [2001] based on the shape of ribonucleic acid (RNA) sequences. Building blocks of strands of RNA are smaller molecules called nucleotides, which have four different types: guanine (G), cytosine (C), adenine (A) and uracil (U). This sequence of an RNA molecule functions as a genotype, since it can be directly replicated by suitable enzymes. Meanwhile, nonadjacent nucleotide pairs undergo additional (weaker) bonding, contorting the sequence into a more complicated three-dimensional structure. It is by this process that an RNA sequence always acquires a physical shape and this shape functions as the phenotype.

Definition D.3. The *primary structure* of an RNA molecule is its unfolded nucleotide chain, often represented by a *genotype sequence* over the alphabet set $\{C, G, A, U\}$. The *secondary structure* or *bonding diagram* or *RNA shape* of an RNA molecule is an unlabeled diagram depicting the bonding that occurs in the resulting RNA molecule.

Example D.4. In the plot below, a genotype sequence gets folded into a three-dimensional structure represented by the planar graph.



The following plot shows the bonding diagram, with the dot as the location of the first nucleotide in the sequence.



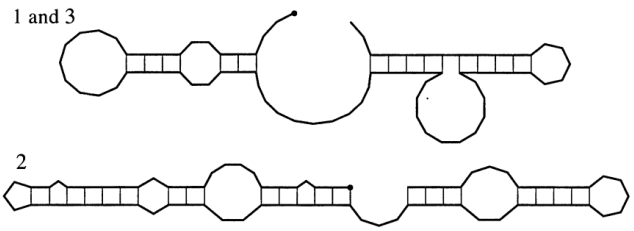
This RNA shape conveys biochemical behavior to an RNA molecule and is therefore subject to selection.

Exercise D.5. How phenotype changes with genotype?

Example D.6. A unique RNA shape can be assigned to each genotype sequence. This function is not injective since multiple genotype sequences may result in the same RNA shape. In the meantime, a single-entry change in the genotype sequence may completely alter the RNA shape. Consider the following genotype sequences.

1. `GGGCAGUCUC CCGGCGUUUA AGGGAUCCUG AACUUCGUCG CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG`
2. `GGGCAGUCUC CCGGCGUUUA AGGAAUCCUG AACUUCGUCG CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG`
3. `GGGCAGUCUC CCGGCCUUUA AGGGAUCCUG AACUUCGUCG CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG`

The corresponding bonding diagrams are as follow.

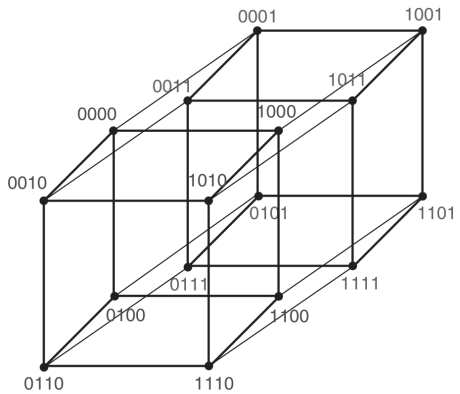


The bonding diagram are the same for sequences 1 and 3, which are identical except for the 16th entry. On the other hand, sequence 2 differs from sequence 1 in only the 24th entry, but their diagrams are very different.

Definition D.7. A *point mutation* is a mutation from one genotype sequence to another by changing a single entry in the sequence. Two sequences are called *neighbors* if they can be converted to each other by a point mutation.

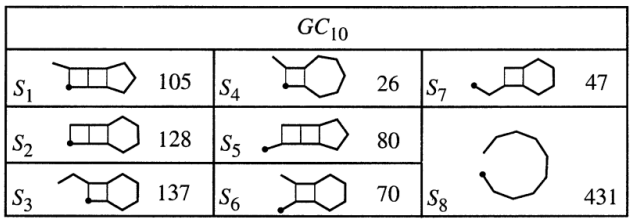
Definition D.8. The *sequence space* of length n (proposed by Eigen [1971]) is a metric space of all genotype sequences of length n with the metric being *the distance between two sequences*, i.e., the smallest number of point mutations required to convert one sequence to the other.

Example D.9. We show below a sequence space of length 4 over the binary alphabet $\{0, 1\}$.



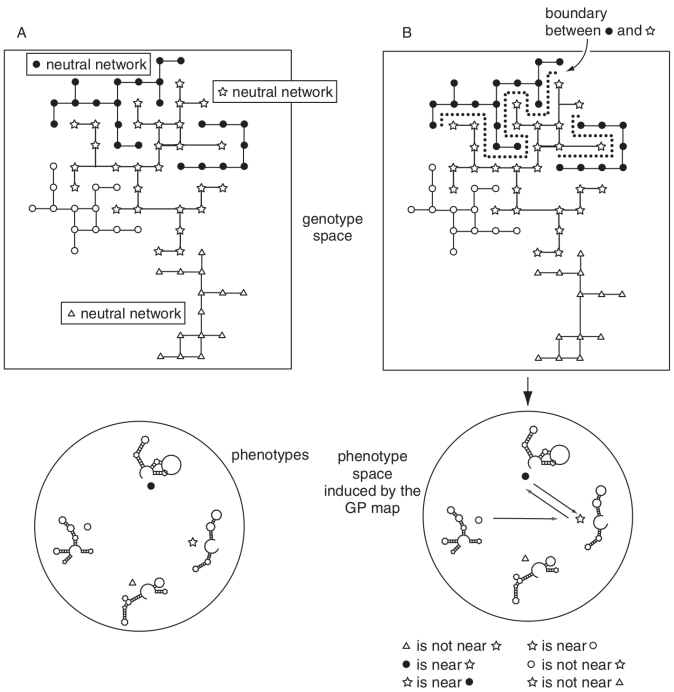
Definition D.10. The *neutral network* of an RNA shape s, denoted by $N(s)$, is the set of all genotype sequences that result in s after folding and bonding.

Example D.11. Consider the sequence space GC_{10} of length 10 over the alphabet $\{G, C\}$. There are 1024 possible sequences, and, after folding and bonding, they result in eight different RNA shapes, as shown below.



The number to the right of a sequence S_i is $\#N(S_i)$.

Example D.12. The possibility of changing the genotype while preserving the phenotype is a manifestation of a certain degree of phenotypic robustness toward genetic mutations. Meanwhile it is a key factor underlying the capacity of a system to evolve.



In the above plots, imagine a population with phenotype 'star' in an evolutionary situation where phenotype 'triangle' would be advantageous or desirable. But phenotype 'triangle' may not be accessible to phenotype 'star' in the vicinity of the population's current location. However, due to the neutral network of 'star,' the population is not stuck, but can drift on that network into far away regions, vastly improving its chances of encountering the neutral network of 'triangle.' Therefore, neutral networks enable phenotypic innovation by permitting the accumulation of neutral mutations.

Exercise D.13. How do we capture and quantify the accessibility of one (favorable) phenotype from another (less favorable) by means of mutations in the sequence space? For any two phenotypes, is there always a directed path from one to the other?

Definition D.14. A *phenotype space* is a set of RNA shapes on which a topology is defined to quantify proximity of RNA shapes.

Definition D.15. The *mutation probability* of an RNA shape r to another RNA shape s is defined as

$$p_{r,s} := \frac{m_{r,s}}{m_{r,*}}, \tag{D.1}$$

where $m_{r,s}$ is the number of point mutations that change a sequence in $N(r)$ to a neighboring sequence in $N(s)$ and $m_{r,*}$ is the number of point mutations that change a sequence in $N(r)$ to a neighboring sequence in any other network including $N(s)$.

Exercise D.16. Show that the mutation probability cannot be a metric on the phenotype space.

Example D.17 (Bubble sort)**.** To sort the sequence `51428`, the first pass of the algorithm goes as follows.

```
( 5 1 4 2 8 ) --> ( 1 5 4 2 8 )
( 1 5 4 2 8 ) --> ( 1 4 5 2 8 )
( 1 4 5 2 8 ) --> ( 1 4 2 5 8 )
( 1 4 2 5 8 ) --> ( 1 4 2 5 8 )
```

The second pass goes as follows.

```
( 1 4 2 5 8 ) --> ( 1 4 2 5 8 )
( 1 4 2 5 8 ) --> ( 1 2 4 5 8 )
( 1 2 4 5 8 ) --> ( 1 2 4 5 8 )
```

Now, the array is already sorted, but the algorithm does not know if it is completed. The algorithm needs one whole pass without any swap to know it is sorted. The third pass goes as follows.

```
( 1 2 4 5 8 ) --> ( 1 2 4 5 8 )
( 1 2 4 5 8 ) --> ( 1 2 4 5 8 )
```

This algorithm is expressed in `C` as follows.

```
void bubble_sort(int* a, int n){
    for (int j=0; j<n-1; j++)
        for (int i = 0; i<n-1-j; i++)
            if(a[i] > a[i+1])
                swap(a[i], a[i+1]);
}

void swap(int& b, int& c){
    int temp = b;
    b = c;
    c = temp;
}
```

As a limit of the above implementation, the program does not apply to the data type `char`, nor any other data type without an implicit conversion, even if the "less than" binary relation for such a data type is natural. You have to manually repeat the above program for each data type. An elegant solution is to use function template in `C++` as follows.

```
template<typename T>
void bubble_sort(T* a, int n){
    for (int i=0; i<n-1; i++)
        for (int j=0; j<n-1-i; j++)
        if (a[j] > a[j+1])
            swap<T>(a[j], a[j+1]);
}

template<typename T>
void swap(T& b, T& c){
    T temp = b;
    b = c;
    c = temp;
}
```

D.1.2 Generalizing continuous maps

Definition D.18. A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous at* a iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \text{ s.t. } |x - a| < \delta \ \Rightarrow \ |f(x) - f(a)| < \epsilon. \tag{D.2}$$

Definition D.19. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *continuous at* $\mathbf{x} = a$ iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \text{ s.t. } f(B(a, \delta)) \subset B(f(a), \epsilon), \tag{D.3}$$

where the n-dimensional open ball $B(p, r)$ is

$$B(p, r) = \{x \in \mathbb{R}^n : \|x - p\|_2 < r\}. \tag{D.4}$$

Definition D.20. A function $f : X \to Y$ with $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ is *continuous at* $\mathbf{x} = a$ iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \text{ s.t. } f(V_a) \subset U_a, \tag{D.5}$$

where the two sets associated with a are

$$V_a := B(a, \delta) \cap X, \qquad U_a := B(f(a), \epsilon) \cap Y. \tag{D.6}$$

Definition D.21. A function $f : X \to Y$ is *continuous* if it is continuous at every point $a \in X$.

Example D.22. Is the function $x \mapsto \frac{1}{x}$ continuous? It depends on whether its domain includes the origin. But it is indeed continuous on domains such as $(0, 1]$, $\mathbb{R} \setminus \{0\}$, and $[1, 2]$. Note that definitions of the one-sided continuity in calculus are nicely incorporated in Definition D.20.

Definition D.23. A function $f : X \to Y$ with $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ is *continuous* iff

$$\forall U_a \in \gamma_Y, \exists V_a \in \gamma_X \text{ s.t. } f(V_a) \subset U_a, \tag{D.7}$$

where γ_X and γ_Y are sets of intersections of the open balls to X and Y, respectively,

$$\gamma_X := \{B(a,\delta) \cap X : a \in X, \delta \in \mathbb{R}^+\};$$
$$\gamma_Y := \{B(f(a),\epsilon) \cap Y : f(a) \in Y, \epsilon \in \mathbb{R}^+\}.$$

Definition D.24. A *basis of neighborhoods* (or a *basis*) on a set X is a collection $\mathcal{B}$ of subsets of X such that

- covering: $\cup\mathcal{B} = X$, and
- refining:

$$\forall U, V \in \mathcal{B}, \forall x \in U \cap V,\ \exists B \in \mathcal{B} \text{ s.t. } x \in B \subset (U \cap V).$$

Definition D.25. For two sets X, Y with bases of neighborhoods $\mathcal{B}_X$, $\mathcal{B}_Y$, a surjective function $f : X \to Y$ is *continuous* iff

$$\forall U \in \mathcal{B}_Y\ \exists V \in \mathcal{B}_X \text{ s.t. } f(V) \subset U. \tag{D.8}$$

Lemma D.26. If a surjective function $f : X \to Y$ is continuous in the sense of Definitions D.20 and D.21, then it is continuous in the sense of Definition D.25.

Proof. By Definition D.24, the following collections are bases of $X \subseteq \mathbb{R}^m$ and $Y = f(X) \subseteq \mathbb{R}^n$, respectively,

$$\mathcal{B}_X = \{B(a,\delta) \cap X : a \in X, \delta > 0\};$$
$$\mathcal{B}_Y = \{B(b,\epsilon) \cap Y : b \in Y, \epsilon > 0\}.$$

The rest follows from Definitions D.25 and D.20. □

Example D.27. The *right rays*

$$\mathcal{B}_{RR} = \{\{x : x > s\} : s \in \mathbb{R}\} \tag{D.9}$$

form a basis of $\mathbb{R}$.

Exercise D.28. Prove that the set of all right half-intervals in $\mathbb{R}$ is a basis of neighborhoods:

$$\mathcal{B} = \{[a,b) : a < b\}. \tag{D.10}$$

Example D.29. A basis on $\mathbb{R}^2$ is the set of all quadrants

$$\mathcal{B}_q = \{Q(r,s) : r, s \in \mathbb{R}\}, \tag{D.11}$$
$$Q(r,s) = \{(x,y) \in \mathbb{R}^2 : x > r, y > s\}. \tag{D.12}$$

Exercise D.30. Denote an open square in $\mathbb{R}^2$ as

$$S((a,b),d) = \{(x,y) : \max(|x-a|, |y-b|) < d\}.$$

Prove that

- for $(m,n) \in \mathbb{R}^2$ and $d > 0$,

$$\forall (a,b) \in S((m,n),d),\ \exists r > 0 \text{ s.t. } S((a,b),r) \subset S((m,n),d).$$

- the set of all open squares in $\mathbb{R}^2$ is a basis of $\mathbb{R}^2$,

$$\mathcal{B}_s = \{S((a,b),d) : (a,b) \in \mathbb{R}^2, d > 0\}.$$

Exercise D.31. Show that the *closed balls* ($r > 0$)

$$\bar{B}(p,r) = \{x \in \mathbb{R}^n : \|x - p\|_2 \le r\} \tag{D.13}$$

do not form a basis of $\mathbb{R}^n$. However, the following collection is indeed a basis:

$$\mathcal{B}_p = \{\bar{B}(a,r) : a \in \mathbb{R}^n, r \ge 0\}, \tag{D.14}$$

which is the union of all closed balls and all singleton sets.

D.1.3 Open sets: from bases to topologies

Definition D.32. A subset U of X is *open* (with respect to a given basis of neighborhoods $\mathcal{B}$ of X) iff

$$\forall x \in U\ \exists B \in \mathcal{B} \text{ s.t. } x \in B \subset U. \tag{D.15}$$

Lemma D.33. Each neighborhood in the basis $\mathcal{B}$ is open.

Proof. This follows from $B \subset B \in \mathcal{B}$ and Definition D.32. □

Exercise D.34. What are the open subsets of $\mathbb{R}$ with respect to the right rays in (D.9)?

Lemma D.35. The intersection of two open sets is open.

Proof. Let U_1 and U_2 be two open sets and fix a point $x \in U_1 \cap U_2$. By Definition D.32, there exists $B_1, B_2 \in \mathcal{B}$ such that $x \in B_1 \subset U_1$ and $x \in B_2 \subset U_2$. Then Definition D.24 implies that there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2 \subset U_1 \cap U_2$. Then the proof is completed by Definition D.32 and x being arbitrary. □

Lemma D.36. The union of two open sets is open.

Lemma D.37. The union of any collection of open sets is open.

Definition D.38. *The topology of X generated by a basis* $\mathcal{B}$ is the collection $\mathcal{T}$ of all open subsets of X in the sense of Definition D.32.

Definition D.39. The *standard topology* is the topology generated by the *standard Euclidean basis*, which is the collection of all open balls in $X = \mathbb{R}^n$.

Theorem D.40. The topology of X generated by a basis satisfies

- $\emptyset, X \in \mathcal{T}$;
- $\alpha \subset \mathcal{T} \Rightarrow \cup_{U\in\alpha} U \in \mathcal{T}$;
- $U, V \in \mathcal{T} \Rightarrow U \cap V \in \mathcal{T}$.

Proof. The first item follows from Definition D.32. The other follows from Lemmas D.35 and D.37. □

Example D.41. The largest basis on a set X is the set of all subsets of X,

$$\mathcal{B}_d(X) = \{A \subset X\} = 2^X, \tag{D.16}$$

and the topology it generates is called *the discrete topology*, which coincides with the basis. This topology is more economically generated by the basis of all singletons,

$$\mathcal{B}_s(X) = \big\{\{x\} : x \in X\big\}. \tag{D.17}$$

The smallest basis on X is simply $\{X\}$ and the topology it generates is called the *trivial/anti-discrete/indiscrete topology* $\mathcal{T}_a = \{\emptyset, X\}$.

Exercise D.42. Show that if U is open with respect to a basis $\mathcal{B}$, then $\mathcal{B} \cup \{U\}$ is also a basis.

D.1.4 Topological spaces: from topologies to bases

Definition D.43. For an arbitrary set X, a collection $\mathcal{T}$ of subsets of X is called a *topology on* X iff it satisfies the following conditions,

(TPO-1) $\emptyset, X \in \mathcal{T}$;

(TPO-2) $\alpha \subset \mathcal{T} \ \Rightarrow\ \cup_{U\in\alpha} U \in \mathcal{T}$;

(TPO-3) $U, V \in \mathcal{T} \ \Rightarrow\ U \cap V \in \mathcal{T}$.

The pair $(X, \mathcal{T})$ is called a *topological space*. The elements of $\mathcal{T}$ are called *open sets*.

Corollary D.44. The topology of X generated by a basis $\mathcal{B}$ as in Definition D.38 is indeed a topology in the sense of Definition D.43.

Proof. This follows directly from Theorem D.40. □

Example D.45. For each $n \in \mathbb{Z}$, define

$$B(n) = \begin{cases} \{n\} & \text{if } n \text{ is odd;} \\ \{n-1, n, n+1\} & \text{if } n \text{ is even.} \end{cases} \tag{D.18}$$

The topology generated by the basis $\mathcal{B} = \{B(n) : n \in \mathbb{Z}\}$ is called the *digital line topology* and we refer to $\mathbb{Z}$ with this topology as the *digital line*.

Theorem D.46. A topology generated by a basis $\mathcal{B}$ equals the collection of all unions of elements of $\mathcal{B}$. (In particular, the empty set is the union of "empty collections" of elements of $\mathcal{B}$.)

Proof. Given a collection of elements of $\mathcal{B}$, Lemma D.33 states that each of them belongs to $\mathcal{T}$. Since $\mathcal{T}$ is a topology, (TPO-2) implies that all unions of these elements are also in $\mathcal{T}$. Conversely, given an open set $U \in \mathcal{T}$, we can choose for each $x \in U$ an element $B_x \in \mathcal{B}$ such that $x \in B_x \subset U$. Hence $U = \cup_{x\in U} B_x$ and this completes the proof. □

Corollary D.47. Let $\mathcal{T}$ be a topology on X generated by the basis $\mathcal{B}$. Then every open set $U \in \mathcal{T}$ is a union of some basis neighborhoods in $\mathcal{B}$. (In particular, the empty set is the union of "empty collections" of elements of $\mathcal{B}$.)

Lemma D.48. Let $(X, \mathcal{T})$ be a topological space. Suppose a collection of open sets $\mathcal{C} \subset \mathcal{T}$ satisfies

$$\forall U \in \mathcal{T}, \forall x \in U, \exists C \in \mathcal{C} \text{ s.t. } x \in C \subset U. \tag{D.19}$$

Then $\mathcal{C}$ is a basis for $\mathcal{T}$.

Proof. We first show that $\mathcal{C}$ is a basis. The covering relation holds trivially by setting $U = X$ in (D.19). As for the refining condition, let $x \in C_1 \cap C_2$ where $C_1, C_2 \in \mathcal{C}$. Since $C_1 \cap C_2$ is open, (D.19) implies that there exists $C_3 \in \mathcal{C}$ such that $x \in C_3 \subset C_1 \cap C_2$. Hence $\mathcal{C}$ is a basis by Definition D.24.

Then we show the topology $\mathcal{T}'$ generated by $\mathcal{C}$ equals $\mathcal{T}$. On one hand, for any $U \in \mathcal{T}$ and any $x \in U$, by (D.19) there exists $C \in \mathcal{C}$ such that $x \in C \subset U$. By Definitions D.32 and D.38, we have $U \in \mathcal{T}'$. On the other hand, it follows from Corollary D.47 that any $W \in \mathcal{T}'$ is a union of elements of $\mathcal{C}$. Since each element of $\mathcal{C}$ is in $\mathcal{T}$, we have $W \in \mathcal{T}$. □

Example D.49. The following countable collection

$$\mathcal{B} = \{(a, b) : a < b, a \text{ and } b \text{ are rational}\} \tag{D.20}$$

is a basis that generates the standard topology on $\mathbb{R}$.

Lemma D.50. A collection of subsets of X is a topology on X if and only if it generates itself.

Proof. The necessity holds trivially since (TPO-1) implies the covering condition and (TPO-3) implies the refining condition. As for the sufficiency, suppose $U, V \in \mathcal{T}$. By Definition D.32, $U \cup V$ is also open, hence $U \cup V \in \mathcal{T}$. This argument holds for the union of an arbitrary number of open sets. □

D.1.5 Generalized continuous maps

Definition D.51. The *preimage of a set* $U \subset Y$ (or the *fiber over* U) under $f : X \to Y$ is

$$f^{-1}(U) := \{x \in X : f(x) \in U\}. \tag{D.21}$$

Exercise D.52. Show that the operation f^{-1} preserves inclusions, unions, intersections, and differences of sets:

$$\begin{cases} B_0 \subseteq B_1 \Rightarrow f^{-1}(B_0) \subseteq f^{-1}(B_1), \\ f^{-1}(B_0 \cup B_1) = f^{-1}(B_0) \cup f^{-1}(B_1), \\ f^{-1}(B_0 \cap B_1) = f^{-1}(B_0) \cap f^{-1}(B_1), \\ f^{-1}(B_0 \setminus B_1) = f^{-1}(B_0) \setminus f^{-1}(B_1). \end{cases} \tag{D.22}$$

In comparison, f only preserves inclusions and unions:

$$\begin{cases} A_0 \subseteq A_1 \Rightarrow f(A_0) \subseteq f(A_1), \\ f(A_0 \cup A_1) = f(A_0) \cup f(A_1), \\ f(A_0 \cap A_1) \subseteq f(A_0) \cap f(A_1), \\ f(A_0 \setminus A_1) \supseteq f(A_0) \setminus f(A_1), \end{cases} \tag{D.23}$$

where the equalities in the last two equations holds if f is injective.

Lemma D.53. For a map $f : X \to Y$, $A \subseteq X$, and $B \subseteq Y$, we have

$$A \subseteq f^{-1}(f(A)),\ f(f^{-1}(B)) \subseteq B, \tag{D.24}$$

where the first inclusion is an equality if f is injective and the second is an equaility if f is surjective or $B \subseteq f(X)$.

Proof. By (D.21), $a \in A$ implies $a \in f^{-1}(f(A))$. Conversely, $a \in f^{-1}(f(A))$ implies $f(a) \in f(A)$. f being injective dictates $a \in A$.

By (D.21), $b \in f(f^{-1}(B))$ implies $b \in B$. Furthermore, if f is surjective or $B \subseteq f(X)$, then for any $b \in B$ we have $f^{-1}(\{b\}) \neq \emptyset$ and thus

$$b \in f(f^{-1}(\{b\})) \subseteq f(f^{-1}(B)). \qquad \square$$

Definition D.54 (Continuous maps between topological spaces)**.** A function $f : X \to Y$ is *continuous* iff the preimage of each open set $U \subset Y$ is open in X.

Lemma D.55. Let $f : X \to Y$ be a continuous function in the sense of Definition D.54. Then the preimage of an open subset $U \subset Y$ satisfies

$$f^{-1}(U) = \cup_{x \in X} V_x, \tag{D.25}$$

where the set V_x is a basis element of X containing x such that $f(V_x) \subset U$.

Proof. Since U is open, Definition D.54 implies that $f^{-1}(U)$ is open. Then by Definition D.32, $x \in f^{-1}(U)$ implies the existence of $V_x \in \mathcal{B}_X$ such that $x \in V_x \subset f^{-1}(U)$. Hence $f^{-1}(U) \subset \cup_{x \in X} V_x$.

Conversely, the condition $f(V_x) \subset U$ and (D.22) yield $f^{-1}f(V_x) \subset f^{-1}(U)$, which, together with Lemma D.53, implies $V_x \subset f^{-1}(U)$. Hence $\cup_{x \in X} V_x \subset f^{-1}(U)$. $\square$

Theorem D.56. If a surjective function is continuous in the sense of Definition D.54, it is continuous in the sense of Definition D.25.

Proof. Consider $U \in \mathcal{B}_Y$. By Lemma D.33, U is open and then Definition D.54 implies that $f^{-1}(U)$ is open in X. The surjectivity of f implies that $f^{-1}(U)$ is not empty. Then by Definition D.32 we have

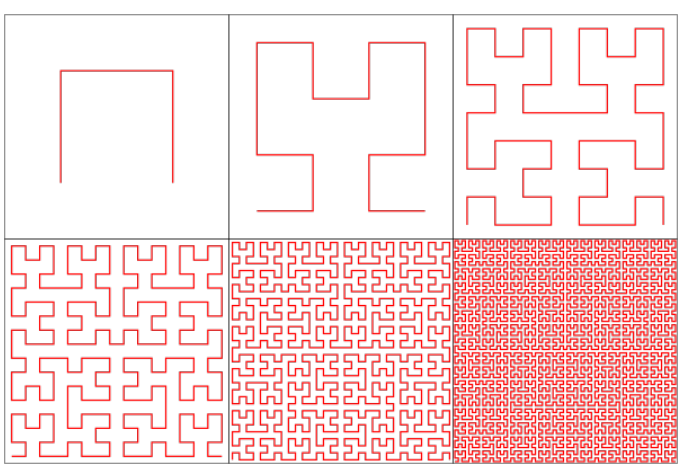
$$\forall x \in f^{-1}(U), \exists V \in \mathcal{B}_X \text{ s.t. } x \in V \subset f^{-1}(U),$$

hence $f(V) \subset f(f^{-1}(U)) = U$, cf. Lemma D.53. $\square$

Exercise D.57. Show that Lemma D.55 holds in the sense of a strengthened continuity of Definition D.25 as follows.

$$\forall U \in \mathcal{B}_Y, \ \forall x \in X \text{ satisfying } f(x) \in U,$$
$$\exists V \in \mathcal{B}_X \text{ satisfying } x \in V \text{ s.t. } f(V) \subset U.$$

Example D.58. A continuous function is not necessarily "well behaved," as exemplified by the following space-filling *Hilbert curve.*



D.1.6 The subbasis topology

Definition D.59. A *subbasis* $\mathcal{S}$ on X is a collection of subsets of X such that the covering condition in Definition D.24 holds.

Example D.60. The set of all open balls with their radii no less than a given $h > 0$, written $\mathcal{B}_h$, is a subbasis but not a basis.

Definition D.61. *The topology of X generated by a subbasis $\mathcal{S}$* is the collection $\mathcal{T}_\mathcal{S}$ of all unions of finite intersections of elements of $\mathcal{S}$.

Exercise D.62. Show that the topology generated by a subbasis $\mathcal{S}$ as in Definition D.61 is indeed a topology in the sense of Definition D.43.

Exercise D.63. Show $\mathcal{S} \subset \mathcal{T}_\mathcal{S}$. In other words, for the topology generated by a subbasis $\mathcal{S}$, every set in $\mathcal{S}$ is an open set in X.

Exercise D.64. Show that if $\mathcal{T}$ is a topology on X containing $\mathcal{S}$, then $\mathcal{T}_\mathcal{S} \subset \mathcal{T}$.

Exercise D.65. Assume that each $x \in X$ is contained in at most finitely many sets in $\mathcal{S}$ and let B_x be the intersection of all sets in $\mathcal{S}$ that contain x. Show that

- the collection $\mathcal{B}_\mathcal{S} := \{B_x : x \in X\}$ is a basis for $\mathcal{T}_\mathcal{S}$;
- if $\mathcal{B}$ is a basis for $\mathcal{T}_\mathcal{S}$, then $\mathcal{B}_\mathcal{S} \subset \mathcal{B}$.

D.1.7 The topology of phenotype spaces

Example D.66. Consider the sequence space GC_{10} in Example D.11. Suppose for GC_{10} the value of $p_{i,j}$, i.e. the mutation probability from s_i to s_j as in Definition D.15, is the number in the ith row and the jth column in the following table.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
s_1	—	0.13	0.15	0.08	0.07	0.09	0.04	0.44
s_2	0.11	—	0.15	0	0.11	0.18	0.05	0.40
s_3	0.12	0.15	—	0.03	0.09	0.06	0.05	0.50
s_4	0.29	0	0.14	—	0.07	0.09	0.06	0.35
s_5	0.08	0.15	0.12	0.02	—	0.08	0.08	0.47
s_6	0.12	0.29	0.09	0.03	0.09	—	0.03	0.35
s_7	0.08	0.12	0.13	0.03	0.13	0.05	—	0.46
s_8	0.18	0.19	0.24	0.04	0.15	0.11	0.09	—

The proximity of RNA shapes is based on the likelihood of a point mutation from one RNA shape to another. By Example D.11, there are eight RNA shapes and hence it is reasonable to assume that $p_{i,j} > 1/7$ implies that the s_j phenotype is accessible to s_i. Hence we define

$$\forall i = 1, 2, \ldots, 8, \quad R_i := \{s_i\} \cup \left\{ s_j : p_{i,j} > \frac{1}{7} \right\}. \tag{D.26}$$

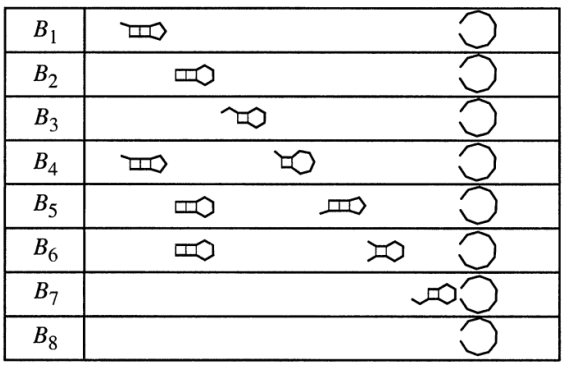
Each R_i is a row in the following table.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
s_1	✓		✓					✓
s_2		✓	✓			✓		✓
s_3		✓	✓					✓
s_4	✓			✓				✓
s_5		✓			✓			✓
s_6		✓				✓		✓
s_7							✓	✓
s_8	✓	✓	✓		✓			✓

The topology $\mathcal{T}_{1/7}$ on GC_{10} is defined as the topology generated by the subbasis

$$\mathcal{R}_{1/7} := \{R_i : i = 1, 2, \ldots, 8\}. \tag{D.27}$$

It can be shown that a topology on a finite set has a unique minimal basis that generates the topology. In this case, the basis is illustrated as follows.



Exercise D.67. Change the threshold value in Example from 1/7 to 1/10 and repeat the entire process. What is the minimum value of q such that $\mathcal{T}_q$ on GC_{10} becomes the discrete topology?

D.1.8 Closed sets

Definition D.68. A subset of X is called *closed* if its complement is open.

Example D.69. The set

$$K = \left\{ \frac{1}{n} : n \in \mathbb{Z}^+ \right\} \tag{D.28}$$

is neither open nor closed. In comparison, $K \cup \{0\}$ is closed.

Theorem D.70. The set σ of all closed subsets of X satisfies the following conditions:

(TPC-1) $\emptyset, X \in \sigma$;

(TPC-2) $\alpha \subset \sigma \ \Rightarrow \ \cap\alpha \in \sigma$;

(TPC-3) $U, V \in \sigma \ \Rightarrow \ U \cup V \in \sigma$.

Example D.71. The following example shows that infinite intersections of open sets might not be open and infinite unions of closed sets might not be closed.

$$\bigcap \left\{ \left(-\frac{1}{n}, \frac{1}{n} \right) : n = 1, 2, \ldots \right\} = \{0\};$$

$$\bigcup \left\{ \left[-1 + \frac{1}{n}, 1 - \frac{1}{n} \right] : n = 1, 2, \ldots \right\} = (-1, 1).$$

Lemma D.72. A function $f : X \to Y$ is continuous if and only if the preimage of any closed set is closed.

Proof. By Definition D.51, we have

$$f^{-1}(U) = f^{-1}(Y \setminus (Y \setminus U)) = X \setminus f^{-1}(Y \setminus U).$$

The rest follows from Definitions D.54 and D.68. □

Definition D.73. The *graph of a function* $f : X \to Y$ is the set $\{(x, y) \in X \times Y : y = f(x)\}$.

Lemma D.74. The graph of a continuous function $f : [a, b] \to \mathbb{R}$ is closed in the space $[a, b] \times \mathbb{R}$.

Exercise D.75. Give an example of the graph of a discontinuous function $f : \mathbb{R} \to \mathbb{R}$ being not closed in $\mathbb{R}^2$. Give another example of the graph of a discontinuous function $f : \mathbb{R} \to \mathbb{R}$ being closed in $\mathbb{R}^2$.

Exercise D.76. Let X be a topological space.

(a) For a continuous function $f : X \to \mathbb{R}$, show that the set $\{x \in X : f(x) = r\}$, i.e. the solution set of any equation with respect to f for some $r \in \mathbb{R}$, is closed.

(b) Show that this fails for a general continuous function $f : X \to Y$ where Y is an arbitrary topological space.

(c) What condition on Y would guarantee that the conclusion holds?

D.1.9 Interior–Frontier–Exterior

Definition D.77. A point $x \in X$ is an *interior point* of A if there is a neighborhood W of x that lies entirely in A. The set of interior points of a set U is called its *interior* and denoted by $\mathrm{Int}(U)$.

Lemma D.78. $\mathrm{Int}(A)$ is open for any A.

Proof. Exercise. □

Example D.79. The interior of a closed ball is the corresponding open ball.

Definition D.80. A point $x \in X$ is an *exterior point* of A if there is a neighborhood W of x that lies entirely in $X \setminus A$. The set of exterior points of a set U is called its *exterior* and denoted by $\mathrm{Ext}(U)$.

Example D.81. The exterior of the set K in (D.28) is $\mathbb{R} \setminus K \setminus \{0\}$. Why not 0?

Definition D.82. A point x is a *closure point* of A if each neighborhood of x contains some point in A.

Example D.83. Any point in the set K in (D.28) is a closure point of K, so is 0.

Definition D.84. A point x is an *accumulation point* (or a *limit point*) of A if each neighborhood of x contains some point $p \in A$ with $p \neq x$.

Example D.85. The only accumulation point of the set K in (D.28) is 0.

Example D.86. Each point in $\mathbb{R}$ is an accumulation point of $\mathbb{Q}$.

Definition D.87. A point x in a set A is *isolated* if there exists a neighborhood of x such that x is the only point of A in this neighborhood.

Example D.88. Every point of the set K in (D.28) is isolated.

Definition D.89. A point x is a *frontier point* of a set A iff it is a closure point for both A and its complement. The set of all frontier points is called *the frontier* $\mathrm{Fr}(A)$ of A.

Theorem D.90. For any set A in X, its interior, its frontier, and its exterior form a partition of X.

Proof. Consider an arbitrary point $a \in X$. If there exists a neighborhood $\mathcal{N}_a$ of a such that $\mathcal{N}_a \subset A$, then Definition D.77 implies $a \in \mathrm{Int}(A)$. If $\mathcal{N}_a \subset X \setminus A$, then Definition D.80 implies $a \in \mathrm{Ext}(A)$. Otherwise, for all neighborhoods of a we have $\mathcal{N}_a \not\subset A$ and $\mathcal{N}_a \not\subset X \setminus A$, which implies that any $\mathcal{N}_a$ contains points both from A and $X \setminus A$. The rest follows from Definition D.89. □

Definition D.91. The closure of A, written $\mathrm{Cl}(A)$ or $\overline{A}$, is the set of all closure points of A.

Lemma D.92. $\mathrm{Int}(A) \subset A \subset \mathrm{Cl}(A)$.

Lemma D.93. $\mathrm{Cl}(A) = \mathrm{Int}(A) \cup \mathrm{Fr}(A)$.

Theorem D.94. The closure of a set A is the smallest closed set containing A:

$$\mathrm{Cl}(A) = \cap\{G : A \subset G, G \text{ is closed in } X\}. \tag{D.29}$$

Proof. Write $\alpha := \{G : A \subset G, G \text{ is closed in } X\}$ and $A^- := \cap\alpha$ and we need to show

- $A^- \subset \mathrm{Cl}(A)$;
- $A^- \supset \mathrm{Cl}(A)$.

We only prove the first part and leave the other as an exercise. Consider $x \notin \mathrm{Cl}(A)$. Then by Definitions D.82 and D.91 there exists an open neighborhood $\mathcal{N}_x$ of x such that $\mathcal{N}_x \cap A = \emptyset$. Hence the set $P := X \setminus \mathcal{N}_x$ contains A. P is also closed because $\mathcal{N}_x$ is open. Therefore $P \in \alpha$ and $x \notin A^-$. □

Exercise D.95. Prove $\mathrm{Cl}(A \cap B) \subset \mathrm{Cl}(A) \cap \mathrm{Cl}(B)$. What if we have infinitely many sets?

Theorem D.96. The interior of a set A is the largest open set contained in A,

$$\mathrm{Int}(A) = \cup\{U : U \subset A,\ U \text{ is open in } X\}. \tag{D.30}$$

Theorem D.97. Let A' be the set of accumulation points of A. Then $\mathrm{Cl}(A) = A \cup A'$.

Proof. Suppose $x \in \mathrm{Cl}(A)$, If $x \in A$, then $x \in A \cup A'$ trivially holds. Otherwise $x \notin A$, Definition D.91 dictates that its neighborhood must contain at least one point in A. Hence Definition D.77 yields $x \in A'$. In both cases we have $x \in A \cup A'$.

Conversely, suppose $x \in A \cup A'$. If $x \in A$, Lemma D.92 implies $x \in \mathrm{Cl}(A)$. If $x \in A'$, x is an accumulation point of A and is thus a closure point A. □

Corollary D.98. A subset of a topological space is closed if and only if it contains all of its accumulation points.

Proof. If A is a superset of A', the set of all accumulation points of A. We have $A \cup A' = A = \mathrm{Cl}(A)$ from Theorem D.97. Definition D.91 implies that A is closed.

Suppose A is closed, but there is an accumulation point x of A such that $x \notin A$. By Definition D.84, in any neighborhood of x there exists a point $p \in A$ such that $p \neq x$; this contradicts the complement of A being open. □

D.1.10 Hausdorff spaces

Definition D.99. Suppose X is a set with a basis of neighborhoods γ. Let $\{x_n : n = 1, 2, \ldots\}$ be a sequence of elements of X and $a \in X$. Then we say the sequence *converges* to a, written

$$\lim_{n\to\infty} x_n = a, \text{ or } x_n \to a \text{ as } n \to \infty,$$

iff

$$\forall U \in \gamma \text{ with } a \in U,\ \exists N \in \mathbb{N}^+ \text{ s.t. } n > N \ \Rightarrow\ x_n \in U. \tag{D.31}$$

Exercise D.100. Prove that the definition remains equivalent if we replace "basis γ" with "topology $\mathcal{T}$."

Exercise D.101. Show that if a sequence converges with respect to a basis γ, it also converges with respect to any basis equivalent to γ.

Theorem D.102. Continuous functions preserve convergence, i.e., for a continuous $f : X \to Y$, $\lim_{n\to\infty} x_n = a$ implies $\lim_{n\to\infty} f(x_n) = f(a)$.

Proof. This follows from Definitions D.99 and D.25. □

Exercise D.103. A sequence $\alpha = \{x_n : n = 1, 2, \ldots\}$ in a topological space X can be viewed as a subset of X, $A = \{x_n : n \in \mathbb{N}^+\}$. Compare the meanings of the closure points of A and the accumulation points of A. What about the limit of α?

Exercise D.104. For metric topology, show that a function $f : X \to Y$ is continuous if and only if the function commutes with limits for any convergent sequence in X.

Example D.105. When do we have $x_n \to a$ for discrete topology?

Example D.106. When do we have $x_n \to a$ for anti-discrete topology?

Definition D.107. A topological space $(X, \mathcal{T})$ is called a *Hausdorff space* iff

$$\forall a, b \in X,\ a \neq b, \exists U, V \in \mathcal{T} \text{ s.t. } a \in U, b \in V, U \cap V = \emptyset. \tag{D.32}$$

Lemma D.108. Every subset of finite points in a Hausdorff space is closed.

Proof. By (TPC-3) in Theorem D.70, it suffices to show that every singleton set is closed. Consider $X \setminus \{x_0\}$. For any $x \neq x_0$, Definition states that there exists $U \supset x$, $V \supset x_0$ such that $U \cap V = \emptyset$, hence $x_0 \notin U$ and $U \in X \setminus \{x_0\}$. Therefore $X \setminus \{x_0\}$ is open. □

Exercise D.109. Does there exist a topological space X that is not Hausdorff but in which every finite point set is closed?

Definition D.110. A topological space is called a *T1 space* iff every finite subset is closed in it.

Theorem D.111. Let X be a T1 space and A a subset of X. A point x is an accumulation point of A if and only if every neighborhood of x intersects with infinitely many points of A.

Proof. The sufficiency follows directly from Definition D.84. As for the necessity, suppose there exists a neighborhood U of x such that $(A \setminus \{x\}) \cap U = \{x_1, x_2, \ldots, x_m\}$. Then by Definition D.110 we know

$$U \cap (X \setminus \{x_1, x_2, \ldots, x_m\}) = U \cap (X \setminus (A \setminus \{x\}))$$

is an open set containing x, yet it does not contain any points in A other than x. This contradicts the condition of x being an accumulation point of A. □

Theorem D.112. A sequence of points in a Hausdorff space X converges to at most one point in X.

Proof. By Definition D.99, a convergence to two points in X would be a contradiction to Definition D.107. □

D.2 Continuous maps

D.2.1 The subspace/relative topology

Lemma D.113. Consider a subset A of a topological space X. Suppose γ_X is a basis of neighborhoods of X. Then

$$\gamma_A := \{W \cap A : W \in \gamma_X\} \tag{D.33}$$

is a basis of neighborhoods of A.

Proof. The covering condition for A holds because the covering condition of X holds. As for the refining condition, for any $U, V \in \gamma_A$ and any $x \in U \cap V$, there exists $U', V' \in \gamma_X$ such that $U = U' \cap A$, $V = V' \cap A$, and $W' \subset U' \cap V'$ for some $W' \in \gamma_X$. Setting $W := W' \cap A$ and we have

$$x \in W \subset (U' \cap V') \cap A = (U' \cap A) \cap (V' \cap A) = U \cap V,$$

which completes the proof. □

Definition D.114. The topology generated by γ_A in (D.33) is called the *relative topology* or *subspace topology* on A generated by the basis γ_X of X.

Lemma D.115. Consider a subset A of a topological space X. Suppose $\mathcal{T}_X$ is a topology on X. Then

$$\mathcal{T}_A := \{W \cap A : W \in \mathcal{T}_X\} \tag{D.34}$$

is a topology on A.

Proof. For (TPO-1), we choose $W = \emptyset, A$. For (TPO-2),

$$\bigcup_{W \in \alpha} (W \cap A) = (\bigcup_{W \in \alpha} W) \cap A,$$

where $\bigcup_{W \in \alpha} W$ is a subset of X. For (TPO-3),

$$(U \cap A) \cap (V \cap A) = (U \cap V) \cap A,$$

where $U \cap V$ is a subset of X. □

Definition D.116 (Subspace and subspace topology)**.** Given a topological space $(X, \mathcal{T})$ and a subset $A \subset X$, the topological space $(A, \mathcal{T}_A)$ is called a *subspace* of X and the topology $\mathcal{T}_A$ in (D.34) is called the *subspace topology or relative topology induced by* X.

Theorem D.117. Let γ_X be a basis that generates the topology $\mathcal{T}_X$ on a topological space X. Then the subspace topology on A induced by $\mathcal{T}_X$ is equivalent to the subspace topology generated by γ_X. In other words, $\mathcal{T}_A$ is generated by γ_A.

$$\begin{array}{ccc} \gamma_X & \xrightarrow{\text{open}} & \mathcal{T}_X \\ \downarrow \cap A & & \downarrow \cap A \\ \gamma_A & \xrightarrow{\text{open}} & \mathcal{T}_A \end{array}$$

Proof. We first show that U is open with respect to (w.r.t.) γ_A for any given $U \in \mathcal{T}_A$. By Lemma D.115, there exists $U' \in \mathcal{T}_X$ such that $U = U' \cap A$. The condition of γ_X being a basis of X yields

$$\forall y \in U',\ \exists B' \in \gamma_X \text{ s.t. } y \in B' \subset U',$$

which implies

$$\forall x \in U \subset U',\ \exists B := (B' \cap A) \in \gamma_A \text{ s.t. } x \in B \subset U.$$

It remains to show that any set U that is open w.r.t. γ_A is in $\mathcal{T}_A$, i.e., we need to find $U' \in \mathcal{T}_X$ such that $U = U' \cap A$. Since U is open w.r.t. γ_A, Definition D.32 yields

$$\forall x \in U,\ \exists N_x \in \gamma_A \text{ s.t. } x \in N_x \subset U,$$

where $N_x = N_x' \cap A$ for some $N_x' \in X$. We then choose

$$U' := \bigcup_{x \in U} N_x'.$$

Theorem D.46 implies that U' is open and $U = U' \cap A$. □

Lemma D.118. Let A be a subspace of X. If U is open in A and A is open in X, then U is open in X.

Proof. Since U is open in A, Definition D.116 yields

$$\exists U' \in X \text{ open s.t. } U = U' \cap A,$$

the rest of the proof follows from A being open in X. □

Lemma D.119 (Closedness in a subspace)**.** Let A be a subspace of X. Then a set $V \subset A$ is closed in A if and only if it equals the intersection of A with a closed subset of X.

Proof. Suppose V is closed in A. Then

$$\exists V' \subset A \text{ s.t. } V \cup V' = A, V' \in \mathcal{T}_A.$$

Since A is a subspace of X, we have from Definition D.116

$$\exists U' \subset X, \text{ s.t. } V' = U' \cap A, U' \in \mathcal{T}_X.$$

Hence the set $U := X \setminus U'$ is closed in X and

$$\begin{aligned} A \cap U &= A \cap (X \setminus U') = A \setminus (X \setminus (X \setminus U')) \\ &= A \setminus U' = A \setminus (U' \cap A) = A \setminus V' = V. \end{aligned}$$

Conversely, suppose

$$\exists U \in X \text{ s.t. } (X \setminus U) \in \mathcal{T}_X, V = U \cap A.$$

Define $V' := (X \setminus U) \cap A$ and we know from Definition D.116 that V' is open in A. The proof is then completed by

$$V \cup V' = (U \cap A) \cup ((X \setminus U) \cap A) = A,$$

where the last step follows from the condition $A \subset X$. □

Corollary D.120 (Transitivity of relative closedness)**.** Let A be a subspace of X. If V is closed in A and A is closed in X, then V is closed in X.

Proof. This follows directly from Lemma D.119 by using $V = V \cap A$. □

D.2.2 New maps from old ones

Theorem D.121. The composition of continuous functions is continuous.

Proof. Suppose we have continuous functions $f = X \to Y$ and $g = Y \to Z$. Let $h = gf : X \to Z$ be their composition. Then for any open set $U \in Z$,

$$h^{-1}(U) = (gf)^{-1}(U) = f^{-1}(g^{-1}(U))$$

is open due to the continuity of g and f and Definition D.54. □

Theorem D.122. Suppose X is a topological space and $f, g : X \to \mathbb{R}$ are continuous functions. Then $f + g$, $f - g$, and $f \cdot g$ are continuous; f/g is also continuous if $g(x) \neq 0$ for all x.

Proof. By Theorem D.184, the function $h : X \to \mathbb{R}^2$ given by $h(x) = (f(x), g(x))$ is continuous. We also know that the function $+ : \mathbb{R}^2 \to \mathbb{R}$ is continuous. Hence the function $f + g = + \circ h$ is continuous. □

Definition D.123. Let X be a topological space and A a subset of X. The *inclusion* $i_A : A \hookrightarrow X$ is given by

$$\forall x \in A, \qquad i_A(x) = x. \tag{D.35}$$

Definition D.124. Let X and Y be topological spaces and A a subset of X. The *restriction of a function* $f : X \to Y$ to A is a function given as

$$\forall x \in A, \qquad f|_A(x) := f(x). \tag{D.36}$$

Theorem D.125 (Restricting the domain)**.** Any restriction of a continuous function is continuous.

Proof. For any open set U in Y, we have $i_A^{-1}(U) = U \cap A$. The rest follows from the relative topology. □

Exercise D.126. Let $i_A : A \hookrightarrow X$ be an inclusion. Suppose the set A is given a topology such that, for every topological space Y and every function $f : Y \to A$, f is continuous if and only if the composition $(i_A \circ f) : Y \to X$ is continuous. Prove that this topology of A is the same as the relative topology of A in X.

Lemma D.127 (Restricting the range)**.** If $f : X \to Y$ is a continuous function, so is $g_f : X \to f(X)$ given by $g_f(x) := f(x)$ for all $x \in X$.

Proof. Of course the topology of $f(X)$ is understood as the subspace topology of Y. The rest follows from Definition D.116. □

Lemma D.128 (Expanding the range)**.** Let $f : X \to Y$ be a continuous function and Y a subspace of Z. Then the function $g : X \to Z$ given by $g(x) := f(x)$ for all $x \in X$ is continuous.

Proof. Write $g = i_Y \circ f$. □

Lemma D.129 (Pasting lemma)**.** Let A, B be two closed subsets of a topological space X such that $X = A \cup B$. Suppose $f_A : A \to Y$ and $f_B : B \to Y$ are continuous functions

$$\forall x \in A \cap B, \qquad f_A = f_B. \tag{D.37}$$

Then the following function $f : X \to Y$ is continuous,

$$f(x) := \begin{cases} f_A(x) & \text{if } x \in A, \\ f_B(x) & \text{if } x \in B. \end{cases} \tag{D.38}$$

Proof. Define $W := f_A(A) \cup f_B(B)$. Then for any $V \subset Y$, (D.38) and the condition (D.37) yields

$$V = (V \cap W) \cup (V \setminus W) = (V \cap f_A(A)) \cup (V \cap f_B(B)) \cup (V \setminus W).$$

If V is closed in Y, then its preimage is

$$\begin{aligned} f^{-1}(V) &= f^{-1}(V \cap f_A(A)) \cup f^{-1}(V \cap f_B(B)) \\ &= f_A^{-1}(V \cap f_A(A)) \cup f_B^{-1}(V \cap f_B(B)) \\ &= g_A^{-1}(V \cap f_A(A)) \cup g_B^{-1}(V \cap f_B(B)), \end{aligned}$$

where g_A and g_B are defined in Lemma D.127. We claim that $f^{-1}(V)$ is also closed in X with arguments as follows.

(i) Since V is closed, Lemma D.119 implies that $V \cap f_A(A)$ and $V \cap f_B(B)$ are closed in $f_A(A)$ and $f_B(B)$, respectively.

(ii) By Lemma D.127, both g_A and g_B are continuous. Hence the two sets to be unioned in the last line of the above equation are closed in A and B, respectively.

(iii) By Corollary D.120, both sets in the last step are closed in X.

The rest of the proof follows from Lemma D.72. □

Exercise D.130. Show that Lemma D.129 fails if A and B are not closed.

Exercise D.131. Formulate the pasting lemma in terms of open sets and prove it.

Exercise D.132. What is the counterpart of the pasting lemma in complex analysis?

Definition D.133 (Expanding the domain)**.** For $A \subset X$ and a given function $f : A \to Y$, a function $F : X \to Y$ is called an *extension* of f if $F|_A = f$.

D.2.3 Homeomorphisms

Definition D.134. A function $f : X \to Y$ between topological spaces X and Y is called a *homeomorphism* iff f is bijective and both f and f^{-1} are continuous. Then X and Y are said to be *homeomorphic* or *topologically equivalent*, written $X \approx Y$.

Lemma D.135. If two spaces X and Y are homeomorphic, then

$$\forall a \in X,\ \exists b \in Y \text{ s.t. } X \setminus \{a\} \approx Y \setminus \{b\}. \tag{D.39}$$

Exercise D.136. Show that the function $f : \{A, B\} \to \{C\}$ given by $f(A) = f(B) = C$ is continuous, but not a homeomorphism. Hence a necessary condition for homeomorphism is the number of connected components.

Example D.137. Consider X the letter "T" and Y a line segment. They are not homeomorphic because removing the junction point in T would result in three pieces while removing any point in the line segment yields at most two connected components.

Exercise D.138. Classify the following symbols of the standard computer keyboard by considering them as 1-dimensional topological spaces.

```
‘ 1 2 3 4 5 6 7 8 9 0 - =
q w e r t y u i o p [ ] \
a s d f g h j k l ; ’
z x c v b n m , . /
~ ! @ # $ % ^ & * ( ) _ +
Q W E R T Y U I O P { } |
A S D F G H J K L : "
Z X C V B N M < > ?
```

Exercise D.139. Consider the identity function $f = \mathbf{I}_X : (X, \mathcal{T}) \to (X, \kappa)$ where κ is the anti-discrete topology and $\mathcal{T}$ is not. Show that f^{-1} is not continuous and hence f is not a homeomorphism.

Exercise D.140. Give an example of a continuous bijection $f : X \to Y$ that isn't a homeomorphism; this time both X and Y are subspaces of $\mathbb{R}^2$.

Exercise D.141. For a continuous function $f : \mathbb{R} \to \mathbb{R}$, define $g : \mathbb{R} \to \mathbb{R}^2$ by $g(x) = (x, f(x))$. Prove that g is continuous and its image, the graph of f, is homeomorphic to $\mathbb{R}$.

Lemma D.142. All closed intervals of a non-zero, finite length are homeomorphic.

Lemma D.143. All open intervals, including infinite ones, are homeomorphic.

Proof. The tangent function gives you a homeomorphism between $(-\frac{\pi}{2}, \frac{\pi}{2})$ and $(-\infty, +\infty)$. □

Lemma D.144. An open interval is not homeomorphic to a closed interval (nor half-open).

Definition D.145. The *n-sphere* is a subset in $\mathbb{R}^{n+1}$,

$$\mathbb{S}^n := \{\mathbf{x} \in \mathbb{R}^{n+1} : \|\mathbf{x}\| = 1\}. \tag{D.40}$$

Its north pole is denoted by $N = (0, 0, \cdots, 0, 1)$.

Definition D.146. The *stereographic projection*

$$P : \mathbb{S}^n \setminus N \to \mathbb{R}^n$$

is given by

$$P(\mathbf{x}) := \left(\frac{x_1}{1 - x_{n+1}}, \frac{x_2}{1 - x_{n+1}}, \ldots, \frac{x_n}{1 - x_{n+1}} \right). \tag{D.41}$$

Lemma D.147. The stereographic projection is a homeomorphism with its inverse as

$$P^{-1}(\mathbf{y}) = \frac{1}{1 + \|\mathbf{y}\|^2} \left(2y_1, 2y_2, \ldots, 2y_n, \|\mathbf{y}\|^2 - 1 \right). \tag{D.42}$$

Exercise D.148. Show that the 2-sphere and the hollow cube are homeomorphic by using the *radial projection* f,

$$f(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}. \tag{D.43}$$

Theorem D.149. Homeomorphisms form an equivalence relation on the set of all topological spaces.

Proof. For a homeomorphism $f : (X, \mathcal{T}_X) \to (Y, \mathcal{T}_Y)$, we can define a function $f_{\mathcal{T}} : \mathcal{T}_X \to \mathcal{T}_Y$ by setting $f_{\mathcal{T}}(V) := f(V)$. It is easy to show that $f_{\mathcal{T}}$ is also a bijection from Definition D.134. □

Definition D.150. An *embedding of X in Y* is a function $f : X \to Y$ that maps X homeomorphically to the subspace $f(X)$ in Y.

Example D.151. For an embedding $f : [0, 1] \to X$, its image is called an *arc* in X. For an embedding $f : \mathbb{S}^1 \to X$, its image is called a *simple closed curve* in X.

D.3 A zoo of topologies

D.3.1 Hierarchy of topologies

Definition D.152. Suppose that $\mathcal{T}$ and $\mathcal{T}'$ are two topologies on a given set X. If $\mathcal{T}' \supset \mathcal{T}$, we say that $\mathcal{T}'$ is *finer/larger* than $\mathcal{T}$; if $\mathcal{T}'$ properly contains $\mathcal{T}$, we say that $\mathcal{T}'$ is *strictly finer/strictly larger* than $\mathcal{T}$. We also say that $\mathcal{T}$ is *coarser/smaller*, or *strictly coarser/strictly smaller*, in these two respective situations. We say $\mathcal{T}$ and $\mathcal{T}'$ are *comparable* if either $\mathcal{T}' \supset \mathcal{T}$ or $\mathcal{T}' \subset \mathcal{T}$.

Lemma D.153. Let $\mathcal{B}$ and $\mathcal{B}'$ be bases for the topologies $\mathcal{T}$ and $\mathcal{T}'$, respectively, on X. $\mathcal{T}'$ is finer than $\mathcal{T}$ if and only if

$$\forall x \in X, \forall B \in \mathcal{B} \text{ with } x \in B, \exists B' \in \mathcal{B}' \text{ s.t. } x \in B' \subset B. \tag{D.44}$$

Proof. The sufficiency $U \in \mathcal{T} \Rightarrow U \in \mathcal{T}'$ follows directly from (D.44) and Definition D.152.

As for the necessity, we start with given $x \in X$ and $B \in \mathcal{B}$ with $x \in B$. By Lemma D.33, B is open, i.e. $B \in \mathcal{T}$. Then by hypothesis $B \in \mathcal{T}'$. Definition D.32 implies that there exists $B' \in \mathcal{B}'$ such that $x \in B' \subset B$, which completes the proof. □

Exercise D.154. The bounded complements of all non-degenerate Jordan curves form a basis of neighborhoods. Is the topology generated by this basis finer than that generated by the open balls?

Definition D.155. The *finite complement topology* on X is

$$\mathcal{T} = \{U \subset X : U = \emptyset \text{ or } X \setminus U \text{ is finite}\}. \tag{D.45}$$

The *countable complement topology* on X is

$$\mathcal{T} = \{U \subset X : U = \emptyset \text{ or } X \setminus U \text{ is countable}\}. \tag{D.46}$$

The *particular point topology* on X is

$$\mathcal{T} = \{U \subset X : U = \emptyset \text{ or } p \in U\}. \tag{D.47}$$

The *excluded point topology* on X is

$$\mathcal{T} = \{U \subset X : U = X \text{ or } p \notin U\}. \tag{D.48}$$

Exercise D.156. Show that each of the topologies in Definition D.155 is indeed a topology.

Exercise D.157. For a three-element set $X = \{a, b, c\}$, enumerate all possible topologies up to the permutation isomorphism.

Exercise D.158. Which topology in the answer of Exercise D.157 has a basis other than itself?

Exercise D.159. Define a directed graph $G = (V, E)$ where the vertex set V contains the topologies in Exercise D.157 and E contains an edge $\mathcal{T}_1 \to \mathcal{T}_2$ iff $\mathcal{T}_2$ is strictly finer than $\mathcal{T}_1$. Plot the graph G.

Definition D.160. The *lower limit topology* $\mathcal{T}_\ell$ on $\mathbb{R}$ is the topology generated by all half-open intervals of the form $[a, b)$ with $a < b$. The space $\mathbb{R}$ endowed with $\mathcal{T}_\ell$ is denoted by $\mathbb{R}_\ell$.

Definition D.161. The *K-topology* $\mathcal{T}_K$ on $\mathbb{R}$ is the topology generated by all open intervals (a, b) and all sets of the form $(a, b) \setminus K$ where K is set in (D.28). The space $\mathbb{R}$ endowed with $\mathcal{T}_K$ is denoted by $\mathbb{R}_K$.

Lemma D.162. The topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are strictly finer than the standard topology on $\mathbb{R}$, but are not comparable with one another.

Proof. For any $x \in (a, b)$, we can always find $[x, b) \in \mathcal{T}_\ell$ and $(a, b) \in \mathcal{T}_K$ such that $x \in [x, b) \subset (a, b)$ and $x \in (a, b) \subset (a, b)$. On the other hand, for any $x \in \mathbb{R}$ and any neighborhood $[x, b) \in \mathbb{R}_\ell$, no open interval in the standard topology simultaneously contains x and is a subset of $[x, b)$. Similarly, for $0 \in \mathbb{R}$ and $B_K := (-1, 1) \setminus K \supset \{0\}$, no open interval simultaneously contains 0 and is a subset of B_K. Hence $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are strictly finer than the standard topology on $\mathbb{R}$.

To show that $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are not comparable, it suffices to give two examples. For any $x \in K \subset \mathbb{R}$ and any neighborhood $[x, b) \in \mathcal{T}_\ell$, no open sets in $\mathcal{T}_K$ simultaneously contains x and is a subset of $[x, b)$. Conversely, for $0 \in \mathbb{R}$ and the above B_K, no interval $[a, b) \in \mathcal{T}_\ell$ simultaneously contains 0 and is a subset of B_K. □

Exercise D.163. The topologies on $\mathbb{R}^2$ generated by the open balls and the open squares are the same topology.

Exercise D.164. Show that the collection

$$\mathcal{C} = \{[a, b) : a < b, a \text{ and } b \text{ are rational}\} \tag{D.49}$$

is a basis that generates a topology $\mathcal{T}_Q$ different from the lower limit topology $\mathcal{T}_\ell$ on $\mathbb{R}$. Compare this to Example D.49.

D.3.2 The order topology

Definition D.165. Let X be a totally ordered set with more than one element. Let $\mathcal{B}$ be the collection of all sets of the following types:

(1) All open intervals (a, b) in X;

(2) All half-open intervals of the form $[a_0, b)$ where a_0 is the smallest element (if any) of X;

(3) All half-open intervals of the form $(a, b_0]$ where b_0 is the largest element (if any) of X.

The *order topology* on X is the topology generated by the basis $\mathcal{B}$.

Exercise D.166. Show that $\mathcal{B}$ is indeed a basis of X in Definition D.165.

Example D.167. The standard topology on $\mathbb{R}$ as in Definition D.39 is the same as the order topology derived from the usual order on $\mathbb{R}$. This is due to the fact that there exists in $\mathbb{R}$ neither the smallest element nor the largest element.

Definition D.168. The *dictionary order* or *lexicopgraphical order* on $\mathbb{R} \times \mathbb{R}$ is a total order defined as

$$(a, b) < (c, d) \Leftrightarrow a < c \text{ or } a = c, b < d. \tag{D.50}$$

Example D.169. A basis for the order topology on $\mathbb{R} \times \mathbb{R}$ with the dictionary order is the collection of all open intervals of the form (1) in Definition D.165.

Example D.170. The order topology of positive integers $\mathbb{Z}^+$ is the same as the discrete topology. For $n > 1$, take the basis interval $(n - 1, n + 1)$; for $n = 1$, take the interval $[1, 2)$.

Exercise D.171. Show that the order topology derived from the dictionary order on the set $X = \{1, 2\} \times \mathbb{Z}^+$ is not the discrete topology.

Exercise D.172. Show that Definition D.165 does not generalize to posets.

Definition D.173. Let X be an ordered set and $a \in X$. The *rays* determined by a are the four subsets of X:

$$(a, +\infty) := \{x : x > a\}; \tag{D.51a}$$
$$(-\infty, a) := \{x : x < a\}; \tag{D.51b}$$
$$[a, +\infty) := \{x : x \geq a\}; \tag{D.51c}$$
$$(-\infty, a] := \{x : x \leq a\}. \tag{D.51d}$$

The first two are *open rays* while the last two *closed rays*.

Exercise D.174. Show that the open rays form a subbasis for the order topology on X.

Definition D.175. The set $[0, 1] \times [0, 1]$ in the dictionary order topology is called the *ordered square*, denoted by I_o^2.

D.3.3 The product topology

Definition D.176. Let X and Y be topological spaces. The *product topology* on $X \times Y$ is the topology generated by the basis

$$\bar{\gamma}_{X\times Y} := \{B_1 \times B_2 : B_1 \in \mathcal{T}_X, B_2 \in \mathcal{T}_Y\}, \tag{D.52}$$

where $\mathcal{T}_X$ and $\mathcal{T}_Y$ are topologies on X and Y, respectively.

Exercise D.177. Check that $\bar{\gamma}_{X\times Y}$ in (D.52) is indeed a basis.

Exercise D.178. Give an example that $\bar{\gamma}_{X\times Y}$ is not a topology.

Exercise D.179. The product of two Hausdorff spaces X and Y is Hausdorff.

Theorem D.180. Let X and Y be topological spaces with bases γ_X and γ_Y, respectively. Then the set

$$\gamma_{X\times Y} := \{B_1 \times B_2 : B_1 \in \gamma_X, B_2 \in \gamma_Y\}, \tag{D.53}$$

is a basis for the topology of $X \times Y$.

Proof. Both the covering and refining conditions hold trivially. □

Definition D.181. For topological spaces X and Y, the functions $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ given by

$$\pi_1(x, y) = x, \qquad \pi_2(x, y) = y \tag{D.54}$$

are called the *projections* of $X \times Y$ onto its first and second factors, respectively.

Lemma D.182. The product topology on $X \times Y$ is the same as the topology generated by the subbasis

$$\mathcal{S} := \left\{\pi_1^{-1}(U) : U \in \mathcal{T}_X\right\} \cup \left\{\pi_2^{-1}(V) : V \in \mathcal{T}_Y\right\}. \tag{D.55}$$

Proof. Let $\mathcal{T}$ denote the product topology in Definition D.176 and $\mathcal{T}'$ denote the topology generated by the subbasis (D.55). Every element in $\mathcal{S}$ belongs to $\mathcal{T}$, so do any unions of finite intersections of elements of $\mathcal{S}$. Hence Definition D.59 yields $\mathcal{T}' \subset \mathcal{T}$.

Conversely, each element in the basis of $\mathcal{T}$ is an intersection of elements in $\mathcal{S}$,

$$B_1 \times B_2 = \pi_1^{-1}(B_1) \cap \pi_2^{-1}(B_2),$$

hence $B_1 \times B_2 \in \mathcal{T}'$ and thus $\mathcal{T} \subset \mathcal{T}'$. □

Corollary D.183. The projections in Definition D.181 are continuous (with respect to the product topology).

Proof. Consider $\pi_1 : X \times Y \to X$. For each open set $U \in \mathcal{T}_X$, Lemma D.182 and Definition D.61 imply that its preimage under π_1 is open in the product topology. □

Theorem D.184 (Product of maps)**.** Given $f_1 : A \to X$ and $f_2 : A \to Y$, the map $f : A \to X \times Y$ with

$$f(a) := (f_1(a), f_2(a)) \tag{D.56}$$

is continuous if and only if both f_1 and f_2 are continuous.

Proof. Write $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$. The necessity follows from Corollary D.183 and Theorem D.121.

As for the sufficiency, we need to show that the preimage $f^{-1}(U \times V)$ of any basis element $U \times V$ is open. By Definition D.176, $U \times V \in \mathcal{B}_{X\times Y}$ implies that $U \in \mathcal{T}_X$ and $V \in \mathcal{T}_Y$. By Definition D.51, any point $a \in f^{-1}(U \times V)$ if and only if $f(a) \in U \times V$, which, by (D.56), is equivalent to $f_1(a) \in U$ and $f_2(a) \in V$. Hence, we have

$$f^{-1}(U \times V) = f_1^{-1}(U) \cap f_2^{-1}(V).$$

The rest of the proof follows from the conditions of both f_1 and f_2 being continuous. □

Example D.185. A parametrized curve $\gamma(t) = (x(t), y(t))$ is continuous if and only if both x and y are continuous.

D.3.4 The metric topology

Definition D.186. For a metric d on X in Definition C.75, the number $d(x, y)$ is called the *distance* between x and y.

Definition D.187. In a metric space $(\mathcal{X}, d)$, an *open ball* $B_r(x)$ centered at $x \in \mathcal{X}$ with radius r is the subset

$$B_r(x) := \{y \in \mathcal{X} : d(x, y) < r\}. \tag{D.57}$$

Lemma D.188. If d is a metric on X, then the collection of all open balls is a basis on X.

Definition D.189. The topology on X generated by the basis of all open balls in Definition D.187 is called the *metric topology* induced by the metric d.

Lemma D.190. A set U is open in the metric topology induced by d if and only if

$$\forall x \in U, \exists r > 0 \text{ s.t. } B_r(x) \subset U.$$

Definition D.191. A topological space X is said to be *metrizable* if there exists a metric d on X that induces the topology of X. A *metric space* is a metrizable topological space together with a specific metric d that gives the topology of X.

Definition D.192. A point x in a normed space X is an *interior point* of A if there is an open ball $B_r(x)$ that lies entirely in A. The set of interior points of a set U is called its *interior* and denoted by $\text{Int}(U)$.

Definition D.193. A point x in a normed space X is an *exterior point* of A if there is an open ball $B_r(x)$ that lies entirely in $X \setminus A$. The set of exterior points of a set U is called its *exterior* and denoted by $\text{Ext}(U)$.

Definition D.194. For metric spaces (X, d_1) and (Y, d_2), a function $f : X \to Y$ is *continuous* iff

$$\forall \epsilon > 0 \ \forall x \in X \ \exists \delta > 0 \text{ s.t. } \forall y \in X \quad d_1(x, y) < \delta \ \Rightarrow \ d_2(f(x), f(y)) < \epsilon \tag{D.58}$$

Definition D.195. For metric spaces (X, d_1) and (Y, d_2), a function $f : X \to Y$ is *uniformly continuous* iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \text{ s.t. } \forall x, y \in X \quad d_1(x, y) < \delta \ \Rightarrow \ d_2(f(x), f(y)) < \epsilon. \tag{D.59}$$

D.4 Connectedness

Definition D.196. Let X be a topological space. A *separation* of X is a pair U, V of disjoint nonempty open subsets of X whose union is X. A topological space is *connected* if there does not exist a separation of X.

Exercise D.197. Why do we define the separation as a pair of disjoint *open* sets? Can we define separation using closed sets?

Example D.198. A space X with indiscrete topology is connected, since there exists no separation of X.

Lemma D.199. For a subspace Y of X, a separation of Y is a pair of disjoint nonempty sets A and B such that $A \cup B = Y$ and neither of them contains a limit point of the other. The space Y is connected if there exists no separation of Y.

Proof. Suppose first that A and B form a separation of Y. By Definition D.196, A and B are both open in Y. Furthermore, A is also closed since its complement B is open in Y. Thus $\overline{A} = A$ and $B \cap \overline{A} = \emptyset$.

Conversely, suppose $A \cup B = Y$, $A \cap \overline{B} = \emptyset$, and $B \cap \overline{A} = \emptyset$. Then we have

$$\overline{A} \cap Y = \overline{A} \cap (A \cup B) = (\overline{A} \cap A) \cup (\overline{A} \cap B) = A.$$

Similarly, $\overline{B} \cap Y = B$. Both A and B are closed in Y, hence they are both open in Y and form a separation of Y. □

Example D.200. Let $Y = [-1, 1]$ be a subspace of $X = \mathbb{R}$. The sets $[-1, 0]$ and $(0, 1]$ are disjoint and nonempty, but they do not form a separation of Y because $[-1, 0]$ is not open in Y. Alternatively, one can use Lemma D.199 to say that $[-1, 0]$ contains a limit point 0 of the other set $(0, 1]$.

Example D.201. Let $Y = [-1, 0) \cup (0, 1]$. Each of the sets $[-1, 0)$ and $(0, 1]$ is nonempty and open in Y; therefore, they form a separation of Y. Again, an alternative argument utilizes Lemma D.199.

Example D.202. The set $\mathbb{Q}$ of rationals is not connected and the only connected subspaces are the one-point spaces. Indeed, for any subset $Y \subset \mathbb{Q}$ that contains more than one point, we can pick distinct $p, q \in Y$ with $p < q$. Then any irrational number $a \in (p, q)$ satisfies

$$Y = (Y \cap (-\infty, a)) \cup (Y \cap (a, +\infty)).$$

By Definition D.196, this separation implies that Y is not connected.

Theorem D.203. Connectedness is preserved by continuous functions; i.e., the image of a connected space under a continuous map is connected.

Proof. Let X be a connected space and $f : X \to Y$ a continuous function. We show that the image space $Z := f(X)$ is connected. Suppose Z is not connected. Then there exists disjoint nonempty open sets U, V such that $Z = U \cup V$. By Definition D.54, $f^{-1}(U)$ and $f^{-1}(V)$ are disjoint open sets and $X = f^{-1}(U) \cup f^{-1}(V)$, which contradicts the condition of X being connected. □

Theorem D.204 (Intermediate value theorem (generalized))**.** Let $f : X \to Y$ be a continuous function where X is a connected space and Y is an ordered set in the order topology. If a and b are two points of X and if r is a point of Y lying between $f(a)$ and $f(b)$, then there exists a point c of X such that $f(c) = r$.

Definition D.205. A *path* in a topological space X is a continuous map $f : I \to X$, where $I := [0, 1]$, and $x_0 := f(0)$ and $x_1 := f(1)$ are called its *initial point* and *final point*, respectively.

Definition D.206. A space X is *path-connected* if, for every $x_0, x_1 \in X$, there exists a path from x_0 to x_1.

Exercise D.207. Prove that if $[a, b]$ is path-connected, so are (a, b) and $[a, b)$.

Theorem D.208. Path-connectedness is preserved by continuous functions; i.e., the image of a path-connected space under a continuous function is path-connected.

Proof. Let X be a connected space and $f : X \to Y$ a continuous function. We show that the image space $Z := f(X)$ is connected. Any $C, D \in Z$ have their preimages $A = f^{-1}(C) \in X$ and $B = f^{-1}(D) \in X$. The path-connectedness of X implies that there exists a continuous function $q : [0, 1] \to X$ such that $q(0) = A$ and $q(1) = B$. By Theorem D.121, the composition $p = fq$ is continuous, $p(0) = f(q(0)) = f(A) = C$, and $p(1) = f(q(1)) = f(B) = D$. Hence Z is path-connected by Definition D.206. □

Lemma D.209. Every path-connected space is connected.

Proof. Suppose a topological space X is not connected but path-connected. Then there exists a separation U, V of X such that $X = U \cup V$. Consider an arbitrary path $f : [0, 1] \to X$. Since $f([0, 1])$ is a continuous image of a connected set, we know from Theorem D.203 that $f([0, 1])$ is connected, hence it must lie entirely in either U or V. Consequently, there is no path in X joining a point of A to a point of B, contradicting the condition of X being path-connected. □

Exercise D.210. A connected space is not necessarily path-connected, c.f. the *topologist's sine curve*. The space

$$S = \left\{ \left(x, \sin \frac{1}{x} \right) : x \in (0, 1] \right\}. \tag{D.60}$$

is connected because it is the image of the connected space $(0, 1]$ under a continuous map. Hence the closure of S

$$\overline{S} = S \cup \{(0, y) : y \in [-1, 1]\}. \tag{D.61}$$

is also connected in $\mathbb{R}^2$. But $\overline{S}$ is not path-connected. Can you prove it?

Exercise D.211. Deduce Theorem C.41 from Theorem D.208.

Theorem D.212 (Fixed points in one dimension)**.** Every continuous function $f : [-1, 1] \to [-1, 1]$ has a fixed point.

Proof. If $f(-1) = -1$ or $f(1) = 1$, we are done; otherwise we have $f(-1) = a > -1$ and $f(1) = b < 1$. Hence none of the following two disjoint sets is empty,

$$A := \{(x, f(x)) : f(x) > x\},\ B := \{(x, f(x)) : f(x) < x\}.$$

By Theorems D.184 and D.203, the graph of f,

$$G := \{(x, f(x)) : x \in [-1, 1]\},$$

is path-connected.

Suppose no x^* satisfies $f(x^*) = x^*$, then $G = A \cup B$. In the topological space G, both A and B are open with respect to a subspace topology of the standard topology. By Definition D.196 and Lemma D.209, this is a contradiction to G being path connected. □

Exercise D.213. Prove Theorem D.212 via connectedness.

Definition D.214. The equivalence classes resulting from connectedness and path-connectedness are called *components* and *path components*, respectively.

Example D.215. The topologist's sine curve $\overline{S}$ in Exercise D.210 has only one component, but has two path components S and $V := \overline{S} \setminus S$. Note that S is open in $\overline{S}$ but not closed, while V is closed in $\overline{S}$ but not open.

If one forms a space from $\overline{S}$ by deleting all points of V having rational second coordinate, one obtains a space that has only one component but uncountably many path components.

Definition D.216. A space X is called *locally connected at* x iff for every neighborhood U of x, there exists a connected neighborhood V of x contained in U. X is *locally connected* iff it is locally connected at each of its points.

Example D.217. $\mathbb{Q}$ is neither connected nor locally connected; the subspace $[-1, 0) \cup (0, +1]$ is not connected but locally connected; the topologist's sine curve is connected but not locally connected; each interval and each ray in the real line is both connected and locally connected.

Definition D.218. A space X is called *locally path-connected at* x iff for every neighborhood U of x, there exists a path-connected neighborhood V of x contained in U. X is *locally path-connected* iff it is locally path-connected at each of its points.

Theorem D.219. A space X is locally connected if and only if for every open set U of X, each component of U is open in X.

Theorem D.220. A space X is locally path-connected if and only if for every open set U of X, each path component of U is open in X.

Theorem D.221. Each path component of a topological space X lies in a component of X. If X is locally path connected, then the components and the path components of X are the same.

Proof. The first statement follows from Lemma D.209. Let C be a component of X, P be a path-component of X. If there is a point $x \in P$ and $x \in C$, we have $P \subset C$. Suppose $P \neq C$. Let Q be the union of all other path components of X, each of which intersects C and thus lies in C. Hence we have $C = P \cup Q$. By Theorem D.220 and the local path-connectedness of X, each path component of X must be open in X. Thus P and Q constitute a separation of X, contradicting the connectedness of C. □

D.5 Compactness

Theorem D.222 (Extreme values)**.** A continuous function attains its extreme values on closed bounded intervals. In other words, if f is continuous on $[a, b]$, there exist $c, d \in [a, b]$ such that

$$f(c) = \max_{x \in [a,b]} f(x), \qquad f(d) = \min_{x \in [a,b]} f(x). \tag{D.62}$$

Definition D.223. A collection α of subsets of a topological space X is said to *cover* X, or to be a covering of X, if the union of all elements of α equals X; it is an *open covering* of X if each element of α is an open subset of X.

Definition D.224. An *(open) cover of a subset X in a topological space Y* is a collection α of (open) subsets in Y such that $X \subset \cup\alpha$. A subcover of X is a subcollection of a cover that also covers X.

Example D.225. Consider K in (D.28) and $X = K \cup \{0\}$. An open cover of K in $\mathbb{R}$ is $\{U_n : n \in \mathbb{N}^+\}$ where

$$U_n = \left(\frac{1}{n} - \epsilon_n, \frac{1}{n} + \epsilon_n \right), \qquad \epsilon_n := \frac{1}{n(n+1)};$$

elements of this open cover are pairwise disjoint for all $n > 1$. An open cover of X in $\mathbb{R}$ is $\{U_n : n \in \mathbb{N}^+\} \cup (-\epsilon, \epsilon)$ with $\epsilon := \frac{1}{N}$ for some $N \in \mathbb{N}^+$.

Example D.226. Consider K in Example D.225 as a space with relative topology induced from $\mathbb{R}$. Each singleton set

$$s_n := \left\{ \frac{1}{n} \right\}$$

is open in K since $s_n = U_n \cap K$ and U_n is open in $\mathbb{R}$. Hence $\{s_n : n \in \mathbb{N}^+\}$ is an infinite open cover of K.

Exercise D.227. Consider X in Example D.225 as a space with relative topology induced from $\mathbb{R}$. Is the collection

$$\{\{0\}\} \cup \{s_n : n \in \mathbb{N}^+\}$$

an open cover of X? If not, can you find an infinite open cover of X whose elements are pairwise disjoint for sufficiently large n? If not, can you give a finite open cover of X?

Exercise D.228. What is the crucial difference between K and X in the space $\mathbb{R}$ in terms of covers and subcovers?

Proof. For any open cover U of X, there exists an element of U containing all but finite many of the points $1/n$. Hence, we have a finite subcover in U for X. This is not true for K. □

Definition D.229. A *compact topological space* is a topological space X where every open cover of X has a finite subcover.

Lemma D.230. A subspace Y of a topological space X is *compact* if and only if every open cover of Y contains a finite subcover of Y.

Lemma D.231. If X is a compact subset of a space Y, then X is compact in relative topology.

Theorem D.232 (Bolzano-Weierstrass)**.** In a compact space, every infinite subset has an accumulation point.

Proof. Suppose there exists an infinite subset A that does not have an accumulation point. Then we can construct an open cover of X

$$\alpha = \{U_x : x \in X\}$$

such that there is at most one element of A in an element of α. By compactness, α contains a finite subcover α' that covers X. However, since each element in the finite set α' only has one element of A and α' covers A, A must be finite, which contradicts the condition of A being infinite. □

Corollary D.233. In a compact space, every sequence has a convergent subsequence.

Definition D.234. A topological space is said to be *locally compact at x* iff there is some compact subspace C of X that contains a neighborhood of x; it is *locally compact* iff it is locally compact at each of its points.

Example D.235. The real line $\mathbb{R}$ is not compact, but locally compact. The subspace $\mathbb{Q}$ is not locally compact.

Theorem D.236. A topological space X is locally compact Hausdorff if and only if there exists a compact Hausdorff space Y such that X is a subspace of Y and $Y \setminus X$ consists of a single point.

Proof. See [Munkres, 2017, page 183]. □

Definition D.237. If Y is a compact Hausdorff space and X is a proper subspace of Y such that $\overline{X} = Y$, then Y is said to be a *compactification* of X. In particular, if $Y \setminus X$ is a singleton set, then Y is called the *one-point compactification* of X.

Example D.238. In Example D.225, X is the one-point compactification of K.

Example D.239. The one-point compactification of the real line $\mathbb{R}$ is homeomorphic with the circle. Similarly, the one-point compactification of the complex plane is homeomorphic with the sphere $\mathbb{S}^2$. The *Riemann sphere* is the space $\mathbb{C} \cup \{\infty\}$.

Theorem D.240. Let X be a Hausdorff space. Then X is locally compact if and only if, given $x \in X$ and a neighborhood U of x, there is a neighborhood V of x such that $\overline{V}$ is compact and $\overline{V} \subset U$.

Corollary D.241. Let X be locally compact Hausdorff and let A be a subspace of X. If A is closed in X or open in X, then A is locally compact.

Corollary D.242. A space X is homeomorphic to an open subspace of a compact Hausdorff space if and only if X is locally compact Hausdorff.

Appendix E

Functional Analysis

Example E.1. A copper mining company mines in a mountain that has an estimated total amount of Q tonnes of copper. Let $x(t)$ denote the amount of copper removed during the period $[0,t]$, with $x(0)=0$ and $x(T)=Q$. Assume x is a continuous function $[0,T]\to\mathbb{R}$ and the cost of extracting copper per unit tonne at time t is

$$c(t)=ax(t)+bx'(t), \tag{E.1}$$

where $a,b\in\mathbb{R}^+$. What is the optimal mining operation $x(t)$ that minimizes the cost function

$$f(x)=\int_0^T (ax(t)+bx'(t))x'(t)\mathrm{d}t?$$

In math terms, we would like to minimize $f:\mathcal{C}^1_Q[0,T]\to\mathbb{R}^+$ where $\mathcal{C}^1_Q[0,T]$ is the set of continuously differentiable functions $x:[0,T]\to\mathbb{R}$ satisfying $x(0)=0$ and $x(T)=Q$.

In calculus, the minimizer x_* of a function $f\in C^2$ is usually found by the condition $f'(x_*)=0$ and $f''(x_*)>0$. However, the above problem does not fit into the usual framework of calculus, since x is not a number but a function that belongs to an infinite-dimensional function space. Solving this problem requires a number of techniques in functional analysis.

E.1 Normed and Banach spaces

E.1.1 Metric spaces

Definition E.2. The ℓ^∞ *sequence space* is a metric space (ℓ^∞,d), where ℓ^∞ is the set of all bounded sequences of complex numbers,

$$\ell^\infty:=\left\{(\xi_1,\xi_2,\ldots):\exists c_x\in\mathbb{R},\ \text{s.t.}\ \sup_{i\in\mathbb{N}^+}|\xi_i|\le c_x.\right\} \tag{E.2}$$

and the metric is given by

$$d(x,y)=\sup_{i\in\mathbb{N}^+}|\xi_i-\eta_i|$$

where $y=(\eta_1,\eta_2,\ldots)\in\mathcal{X}$.

Exercise E.3. Let $\mathcal{X}$ be the set of all bounded and unbounded sequences of complex numbers. Show that the following is a metric on $\mathcal{X}$,

$$d(x,y)=\sum_{j=1}^{\infty}\frac{1}{2^j}\frac{|\xi_j-\eta_j|}{1+|\xi_j-\eta_j|}, \tag{E.3}$$

where $x=(\xi_j)$ and $y=(\eta_j)$.

Definition E.4. For a real number $p\ge 1$, the ℓ^p *sequence space* is the metric space (ℓ^p,d) with

$$\ell^p:=\left\{(\xi_j)_{j=1}^\infty:\xi_j\in\mathbb{C};\sum_{j=1}^{\infty}|\xi_j|^p<\infty\right\}; \tag{E.4}$$

$$d(x,y)=\left(\sum_{j=1}^{\infty}|\xi_j-\eta_j|^p\right)^{1/p}, \tag{E.5}$$

where $x=(\xi_j)$ and $y=(\eta_j)$ are both in $\mathcal{X}$. In particular, the *Hilbert sequence space* ℓ^2 is the ℓ^p space with $p=2$.

Definition E.5. A pair of *conjugate exponents* are two real numbers $p,q\in[1,\infty]$ satisfying

$$p+q=pq,\ \text{i.e.,}\ \frac{1}{p}+\frac{1}{q}=1. \tag{E.6}$$

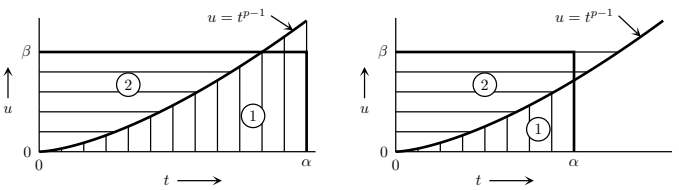
Lemma E.6. Any two positive real numbers α,β satisfy

$$\alpha\beta\le\frac{\alpha^p}{p}+\frac{\beta^q}{q}, \tag{E.7}$$

where p and q are conjugate exponents and the equality holds if $\beta=\alpha^{p-1}$.

Proof. By (E.6), we have

$$u=t^{p-1}\ \Rightarrow\ t=u^{q-1}.$$



It follows that

$$\alpha\beta \le \int_0^\alpha t^{p-1}\mathrm{d}t + \int_0^\beta u^{q-1}\mathrm{d}u = \frac{\alpha^p}{p} + \frac{\beta^q}{q},$$

where the equality holds if $\beta = \alpha^{p-1}$ since $p = q(p-1)$. □

Corollary E.7. A pair of conjugate exponents p, q satisfy

$$\forall a, b \in [0, +\infty), \quad a^{\frac{1}{p}} b^{\frac{1}{q}} \le \frac{a}{p} + \frac{b}{q}. \tag{E.8}$$

Proof. This follows directly from Lemma E.6. □

Theorem E.8 (Hölder's inequality)**.** For $n \in \mathbb{N}^+ \cup \{+\infty\}$, $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ and conjugate exponents $p, q \in [1, \infty]$, we have

$$\sum_{j=1}^n |x_j y_j| \le \|\mathbf{x}\|_p \|\mathbf{y}\|_q, \tag{E.9}$$

where $\|\cdot\|_p$ is the Euclidean norm. For $p, q \in (1, \infty)$, the equality in (E.9) holds if

$$\exists c \in \mathbb{R} \text{ s.t. } \forall j = 1, \dots, n, \ |x_j|^p = c|y_j|^q. \tag{E.10}$$

Proof. If $\sum_{j=1}^n |x_j|^p = 0$ or $\sum_{j=1}^n |y_j|^p = 0$ or $p = \infty$ or $q = \infty$, then (E.9) holds trivially. Otherwise we define

$$a_i := \frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}, \quad b_i := \frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}.$$

It follows from (E.8) that

$$\frac{|x_i y_i|}{\left(\sum_{j=1}^n |x_j|^p\right)^{\frac{1}{p}} \left(\sum_{j=1}^n |y_j|^q\right)^{\frac{1}{q}}} \le \frac{|x_i|^p}{p\sum_{j=1}^n |x_j|^p} + \frac{|y_i|^q}{q\sum_{j=1}^n |y_j|^q}.$$

Sum up all equations for $i = 1, \dots, n$ and we have

$$\frac{\sum_{j=1}^n |x_i y_i|}{\|\mathbf{x}\|_p \|\mathbf{y}\|_q} \le \frac{1}{p} + \frac{1}{q} = 1,$$

which yields (E.9). Substitute (E.10) into (E.9) and we have the equality. □

Example E.9. Cauchy-Schwarz inequality in Theorem B.171 is a special case of the Hölder inequality (E.9) for $p = q = 2$.

Exercise E.10. Prove that (E.5) satisfies the triangular inequality and is indeed a metric.

(a) The Hölder inequality implies the *Minkowski inequality*, i.e. for any $p \ge 1$, $(\xi_j) \in \ell^p$, and $(\eta_j) \in \ell^p$,

$$\left(\sum_{j=1}^\infty |\xi_j + \eta_j|^p\right)^{1/p} \le \left(\sum_{k=1}^\infty |\xi_k|^p\right)^{1/p} + \left(\sum_{m=1}^\infty |\eta_m|^p\right)^{1/p}. \tag{E.11}$$

(b) The Minkowski inequality implies that the triangular inequality holds for (E.5).

E.1.2 Normed spaces

Example E.11. $(\mathbb{R}^n, \|\cdot\|_p)$ is a normed space, where $\|\cdot\|_p$ is the Euclidean norm in Definition B.154 with $p \in [1, \infty)$:

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^n |x_j|^p\right)^{\frac{1}{p}}.$$

Exercise E.12. Prove the backward triangle inequality of a norm $\|\cdot\|$, i.e.,

$$\forall u, v \in V, \quad |\|u\| - \|v\|| \le \|u - v\|. \tag{E.12}$$

Exercise E.13. Use Holder's inequality to verify the triangle inequality for the Euclidean norm in Example E.11.

Definition E.14. In a normed space $(\mathcal{X}, \|\cdot\|)$, an *open ball* $B_r(x)$ centered at $x \in \mathcal{X}$ with radius $r > 0$ is the subset

$$B_r(x) := \{y \in \mathcal{X} : \|x - y\| < r\}. \tag{E.13}$$

Lemma E.15. Any open ball in a normed space is a convex set as in Definition 1.18.

Proof. For $\alpha \in [0, 1]$ and $\mathbf{x}, \mathbf{y} \in B_r(\mathbf{0})$, we have

$$\begin{aligned}\|\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\| &\le \|\alpha\mathbf{x}\| + \|(1-\alpha)\mathbf{y}\| \\ &\le \alpha\|\mathbf{x}\| + (1-\alpha)\|\mathbf{y}\| < \alpha r + (1-\alpha)r = r,\end{aligned}$$

where we have applied properties of norms. Hence $\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in B_r(\mathbf{0})$. □

Exercise E.16. Show that the Euclidean norm $\|\cdot\|_p$ in Example E.11 satisfies a monotonicity property:

$$1 \le p \le q \le \infty \ \Rightarrow \ \forall \mathbf{x} \in \mathbb{R}^n \ \|\mathbf{x}\|_q \le \|\mathbf{x}\|_p.$$

Example E.17. $(\mathbb{R}^n, \|\cdot\|_\infty)$ is a normed space, where $\|\cdot\|_\infty$ is the Euclidean norm in Definition B.154:

$$\|\mathbf{x}\|_\infty = \max_j |x_j|.$$

Definition E.18. Let $\Omega \subset \mathbb{R}^n$ be a bounded open set. The *p-norm of a continuous scalar function* in the linear space $\mathcal{C}(\overline{\Omega})$ is

$$\forall v \in \mathcal{C}(\overline{\Omega}), \quad \|v\|_p := \left[\int_\Omega |v(\mathbf{x})|^p \mathrm{d}\mathbf{x}\right]^{\frac{1}{p}} \tag{E.14}$$

and the ∞*-norm* or *maximum norm* is given by

$$\forall v \in \mathcal{C}(\overline{\Omega}), \quad \|v\|_\infty := \max_{\mathbf{x}\in\overline{\Omega}} |v(\mathbf{x})|. \tag{E.15}$$

Example E.19. $(\mathcal{C}(\overline{\Omega}), \|\cdot\|_\infty)$ in Definition E.18 is a normed space, so is $(\mathcal{C}(\overline{\Omega}), \|\cdot\|_p)$ for any $p \in [1, \infty)$.

Example E.20. For the ℓ^∞ sequence space in (E.2),

$$\ell^\infty := \left\{(a_n)_{n\in\mathbb{N}} : \sup_{n\in\mathbb{N}} |a_n| < \infty.\right\},$$

define $\|\cdot\|_\infty : \ell^\infty \to \mathbb{R}^+ \cup \{0\}$ as

$$\|(a_n)_{n\in\mathbb{N}}\|_\infty = \sup_{n\in\mathbb{N}} |a_n|. \tag{E.16}$$

Then $(\ell^\infty, \|\cdot\|_\infty)$ is a normed space.

Example E.21. For the ℓ^p space in (E.4) with $p \in [1, \infty)$,

$$\ell^p := \left\{ (a_n)_{n\in\mathbb{N}} : a_n \in \mathbb{C}; \sum_{n\in\mathbb{N}} |a_n|^p < \infty \right\},$$

we have

$$\begin{cases} (a_n)_{n\in\mathbb{N}} \in \ell^p,\ (b_n)_{n\in\mathbb{N}} \in \ell^p \\ |a+b|^p \le (|a|+|b|)^p \le 2^p(\max(|a|,|b|))^p \le 2^p(|a|^p+|b|^p) \end{cases}$$
$$\Rightarrow (a_n)_{n\in\mathbb{N}} + (b_n)_{n\in\mathbb{N}} \in \ell^p,$$

where the comparison test is applied.

Then $(\ell^p, \|\cdot\|_p)$ is a normed space where

$$\|(a_n)_{n\in\mathbb{N}^+}\|_p := \left(\sum_{n=1}^{\infty} |a_n|^p \right)^{\frac{1}{p}}. \tag{E.17}$$

Notation 14. Let c_{00} denote the space of all sequences that are eventually 0, c_0 the space of all sequences that converge to 0, and c the space of all sequences that converge.

Definition E.22 (Convergence of sequences)**.** A sequence $\{u_n\}$ in a normed space $(V, \|\cdot\|)$ is *convergent* to $u \in V$ iff

$$\lim_{n\to\infty} \|u_n - u\| = 0. \tag{E.18}$$

E.1.3 The topology of normed spaces

Example E.23. The topology of a normed space is the metric topology in Definition D.189 because a normed space is always a metric space.

Definition E.24. Let $(\mathcal{X}, d)$ be a metric space. A point $x_0 \in \mathcal{X}$ is an *adherent point* or a *closure point* of $E \subset \mathcal{X}$ or a *point of closure* or a *contact point* iff

$$\forall r > 0,\ E \cap B_r(x_0) \neq \emptyset. \tag{E.19}$$

Example E.25. Any point in the set K in (D.28) is a closure point of K, so is 0.

Definition E.26. A point x is an *accumulation point* (or a *limit point*) of A iff

$$\forall r > 0,\ (B_r(x) \setminus \{x\}) \cap A \neq \emptyset. \tag{E.20}$$

Example E.27. The only accumulation point of the set K in (D.28) is 0.

Example E.28. Each number in $\mathbb{R}$ is an accumulation point of $\mathbb{Q}$.

Definition E.29. Let $V_1 \subset V_2$ be two subsets in a normed space V. The set V_1 is *dense* in V_2 iff

$$\forall u \in V_2, \forall \epsilon > 0, \exists v \in V_1 \text{ s.t. } \|v - u\| < \epsilon. \tag{E.21}$$

Theorem E.30. $\mathbb{Q}$ is dense in $\mathbb{R}$.

Exercise E.31. Show that c_{00} is dense in ℓ^2.

Example E.32. The set of polynomials is dense in $(\mathcal{C}[a,b], \|\cdot\|_\infty)$, c.f. Theorem 2.53.

Definition E.33. A normed space is *separable* if it has a countable dense set.

Example E.34. By Definitions E.29 and E.33, $L^p(\Omega)$ is separable since the set of all polynomials with rational coefficients is countable and is dense in $L^p(\Omega)$.

Lemma E.35. ℓ^∞ is not separable.

Proof. Suppose ℓ^∞ is separable. Then there exists in ℓ^∞ a dense subset $D = \{x_1, x_2, x_3, \dots, \}$. For the set A of sequences with each term being either 0 or 1, we have

$$\forall a, b \in A, \quad a \neq b \ \Leftrightarrow\ \|a - b\|_\infty = 1.$$

It follows from D being dense in ℓ^∞ that

$$\forall a \in A, \exists x_{n(a)} \in D \text{ s.t. } x_{n(a)} \in B_{\frac{1}{2}}(a).$$

Because the open balls are pairwise distinct, the map $f : A \to \mathbb{N}$ given by $f(a) = n(a)$ is injective. However, the construction of A implies that A is uncountable because A has a one-to-one correspondence with all real numbers in $[0, 1)$ via binary expansion. Thus $f : A \to \mathbb{N}$ cannot be injective and this completes the proof. □

Exercise E.36. Prove that ℓ^p is separable for all $p \in [1, \infty)$.

E.1.4 Bases of infinite-dimensional spaces

Definition E.37. An infinite dimensional normed space V has a *countably-infinite basis* iff

$$\exists \{v_i\}_{i\ge1} \subset V \text{ s.t. } \forall v \in V, \exists \{\alpha_{n,i}\}_{i=1}^n \text{ where } n \in \mathbb{N}^+, \alpha_{n,i} \in \mathbb{R}$$
$$\text{s.t. } \lim_{n\to\infty} \left\| v - \sum_{i=1}^{n} \alpha_{n,i} v_i \right\| = 0. \tag{E.22}$$

The sequence $\{v_i\}_{i\ge1}$ is a *basis* if any finite subset of it is linearly independent.

Definition E.38. A *Schauder basis* of an infinite-dimensional normed linear space V is a sequence $\{v_n\}_{n\ge1}$ of elements in V such that

$$\forall v \in V, \exists! \{\alpha_n\}_{n\ge1} \text{ where } \alpha_n \in \mathbb{R} \text{ s.t. } v = \sum_{n=1}^{\infty} \alpha_n v_n. \tag{E.23}$$

Example E.39. The sequence space ℓ^2 in Definition E.4 has a Schauder basis

$$\left\{ e_j = (0, \dots, 0, 1, 0, 0, \dots) \right\}_{j=1}^{\infty}$$

since any $\xi = (\xi_1, \xi_2, \dots) \in \ell^2$ can be uniquely written as $\xi = \sum_{j=1}^{\infty} \xi_j e_j$.

Example E.40. It can be proved that the set $\{1, \cos nx, \sin nx\}_{n=1}^{\infty}$ is a Schauder basis of $L^p(-\pi, \pi)$ for $p \in (1, \infty)$.

E.1.5 Sequential compactness

Definition E.41. A subset K of a normed space $(X, \|\cdot\|)$ is *sequentially compact* if every sequence in K has a convergent subsequence that converges in K,

$$\forall (x_n)_{n\in\mathbb{N}} \subset K,\ \exists n_k : \mathbb{N} \to \mathbb{N},\ \exists L \in K \text{ s.t. } \lim_{n\to+\infty} x_{n_k} = L. \tag{E.24}$$

Example E.42. Any interval $[a, b]$ is sequentially compact in $\mathbb{R}$. Indeed, any sequence in $[a, b]$ is bounded, and by the Bolzano-Weierstrass theorem (Theorem C.13) it has a convergent subsequence, of which the limit must be in $[a, b]$, thanks to the completeness of $\mathbb{R}$ (Theorem C.15),

Example E.43. (a, b) is not sequentially compact since the sequence $\left(a + \frac{b-a}{2n}\right)_{n\in\mathbb{N}^+}$ is contained in (a, b), but its limit a is not contained in (a, b).

Example E.44. $\mathbb{R}$ is not sequentially compact because the sequence $(n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ cannot have a convergent subsequence: the distance bewteen any two terms on any subsequence is at least 1.

Lemma E.45. Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.

Proof. We prove this statement by an induction on n. The induction basis is the Bolzano-Weierstrass theorem (Theorem C.13). Suppose the statement holds for $n \geq 1$. For a bounded sequence $(\mathbf{x}_m)_{m\in\mathbb{N}} \subset \mathbb{R}^{n+1}$, we split each $\mathbf{x}_m$ as $\mathbf{x}_m = (\alpha_n, \beta_n)$, where $\alpha_n \in \mathbb{R}^n$ and $\beta_n \in \mathbb{R}$. Since $\mathbf{x}_m$ is bounded and $\|\alpha_n\|_2 \leq \|\mathbf{x}_m\|_2$, α_m is also bounded. By the induction hypothesis, $(\alpha_m)_{m\in\mathbb{N}}$ has a convergent subsequence, say $(\alpha_{m_k})_{k\in\mathbb{N}}$, that converges to $\alpha \in \mathbb{R}^k$. Then $(\beta_{m_k})_{k\in\mathbb{N}}$ is bounded and by Theorem C.13 it has a convergent subsequence $(\beta_{m_{k_p}})_{p\in\mathbb{N}}$ that converges to $\beta \in \mathbb{R}$. Therefore we have

$$\lim_{p\to\infty} \mathbf{x}_{m_{k_p}} = \lim_{p\to\infty} \left(\alpha_{m_{k_p}}, \beta_{m_{k_p}}\right) = (\alpha, \beta) \in \mathbb{R}^{n+1},$$

which completes the proof. □

Theorem E.46. In a metric space, sequential compactness is equivalent to compactness.

Lemma E.47. A sequentially compact subset K of a normed space X must be closed and bounded.

Proof. Supppose K is compact but not bounded. Then

$$\forall n \in \mathbb{N}, \exists \mathbf{x}_n \in K \text{ s.t. } \|\mathbf{x}_n\| \geq n.$$

Hence no subsequence of $(\mathbf{x}_n)_{n\in\mathbb{N}} \subset K$ converges and this contradicts the compactness of K.

For any convergent sequence $(\mathbf{x}_n)_{n\in\mathbb{N}} \subset K$, Definition E.41 implies that it has a convergent subsequence that converges in K. The uniqueness of limit (Lemma C.6) dictates that the two sequences converge to the same limit in K. Now that any convergent sequence converges to some limit point in K, Corollary D.98 implies that K is closed. □

Theorem E.48. A subset K of $\mathbb{R}^n$ is sequentially compact if and only if K is closed and bounded.

Proof. The necessity follows from Lemma E.47, we only prove the sufficiency. Any sequence $(\mathbf{x}_n)_{n\in\mathbb{N}} \subset K$ is bounded because K is bounded. Then Lemma E.45 dictates that $(\mathbf{x}_n)_{n\in\mathbb{N}}$ has a convergent subsequence. Because each term $\mathbf{x}_n \in K$ and K is closed, Corollary D.98 implies that the limit point of $(\mathbf{x}_n)_{n\in\mathbb{N}}$ is also in K. The proof is then completed by Definition E.41. □

Example E.49. The intervals $(a, b]$, $[a, b)$, $(-\infty, b]$, and $[a, +\infty)$ are not sequentially compact in $\mathbb{R}$.

Definition E.50. The *Cantor set* is a subset of $\mathbb{R}$ give by $C := \cap_{n=1}^{+\infty} F_n$ where $F_1 = [0, 1]$ and each F_{n+1} is obtained by deleting from F_n the open middle third of each closed interval.

Example E.51. The Cantor set is an intersection of closed set and thus it is closed. It is also bounded and thus it is sequentially compact.

Corollary E.52. A subset K of a finite-dimensional normed space X is sequentially compact if and only if K is closed and bounded.

Example E.53. The closed unit ball in $(\mathcal{C}[0, 1], \|\cdot\|_\infty)$

$$K := \{f \in \mathcal{C}[0, 1] : \|f\|_\infty \leq 1\} \tag{E.25}$$

is closed and bounded, but K is not sequentially compact. Consider the hat function

$$B_n(x) = \begin{cases} \frac{x-a_n}{b_n-a_n} & x \in [a_n, b_n], \\ \frac{x-c_n}{b_n-c_n} & x \in [b_n, c_n], \\ 0 & \text{otherwise}, \end{cases} \tag{E.26}$$

where $a_n = 1 - \frac{1}{2^n}$, $c_n = a_{n+1}$, and $b_n = \frac{a_n+c_n}{2}$. Then the sequence $(B_n)_{n\in\mathbb{N}}$ has no convergent subsequence.

Example E.54. The closed unit ball in ℓ^2,

$$K := \left\{\mathbf{x} \in \ell^2 : \|\mathbf{x}\|_2 \leq 1\right\}, \tag{E.27}$$

is closed and bounded, but is not sequentially compact. For

$$\mathbf{e}_n = (0, \cdots, 0, 1, 0, \cdots, 0) \in K \subset \ell^2$$

where all terms are zero except than the nth term is 1, the sequence $(\mathbf{e}_n)_{n\in\mathbb{N}^+}$ has no convergent subsequence.

Example E.55. The *Hilbert cube* in the normed space ℓ^2,

$$C := \left\{(x_n)_{n\in\mathbb{N}^+} :\ x_n \in \left[0, \frac{1}{n}\right]\right\}, \tag{E.28}$$

can be shown to be a sequentially compact subset.

Definition E.56. An *open cover of a topological space* X is collection of open subsets of X such that any element of X belongs to some open subset in the collection.

Definition E.57. A subset K in a topological space is *compact* if and only if every open cover of K has a finite subcover.

E.1.6 Continuous maps of normed spaces

Definition E.58. Let X and Y be normed spaces. A function $f : X \to Y$ is *continuous at* $x_0 \in X$ iff

$$\forall \epsilon > 0,\ \exists \delta > 0 \text{ s.t.}$$
$$\forall x \in X,\ \|x - x_0\|_X < \delta \ \Rightarrow\ \|f(x) - f(x_0)\|_Y < \epsilon. \quad \text{(E.29)}$$

The function $f : X \to Y$ is *continuous* iff it is continuous at every $x_0 \in X$.

Lemma E.59. Let X and Y be normed spaces. A function $f : X \to Y$ is *continuous at* $x \in X$ iff, for any sequence with $\lim_{n\to\infty} x_n = x$, we have $\lim_{n\to\infty} f(x_n) = f(x)$.

Exercise E.60. Prove Lemma E.59.

Lemma E.61. The norm function $\|\cdot\|$ is continuous.

Proof. By Definition E.58, we have $\lim_{n\to\infty} \|u_n - u\| = 0$ from $\lim_{n\to\infty} u_n = u$. The rest of the proof follows from the backward triangle inequality (E.12). □

Exercise E.62. For $V = \mathcal{C}[0,1]$ and $x_0 \in [0,1]$, define a function $\ell_{x_0} : V \to \mathbb{R}$ as

$$\ell_{x_0}(v) = v(x_0).$$

Show that ℓ_{x_0} is continuous on $\mathcal{C}[0,1]$.

Example E.63. The function $S : (\mathcal{C}[0,1], \|\cdot\|_\infty) \to (\mathbb{R}, |\cdot|)$,

$$S(f) = \int_0^1 f^2(x)\mathrm{d}x, \quad \text{(E.30)}$$

is continuous. Indeed, for any $g \in \mathcal{C}[0,1]$, we have

$$\begin{aligned} |S(f) - S(g)| &= \left| \int_0^1 f^2(x)\mathrm{d}x - \int_0^1 g^2(x)\mathrm{d}x \right| \\ &\le \int_0^1 |f(x) - g(x)|\,|f(x) - g(x) + 2g(x)|\,\mathrm{d}x \\ &\le \int_0^1 \|f - g\|_\infty \left(\|f - g\|_\infty + 2\|g\|_\infty\right)\mathrm{d}x, \end{aligned}$$

which implies

$$\forall \epsilon > 0, \exists \delta = \min\left(1, \frac{\epsilon}{1 + 2\|g\|_\infty}\right) \text{ s.t. } \|f - g\|_\infty < \delta \ \Rightarrow$$
$$\|f - g\|_\infty \left(\|f - g\|_\infty + 2\|g\|_\infty\right) < \|f - g\|_\infty (1 + 2\|g\|_\infty) < \epsilon$$
$$\Rightarrow\ |S(f) - S(g)| < \epsilon.$$

Example E.64. The differentiation map

$$\frac{\mathrm{d}}{\mathrm{d}t} : (\mathcal{C}^1[a,b], \|\cdot\|_\infty) \to (\mathcal{C}[a,b], \|\cdot\|_\infty)$$

is not continuous, but can be made continuous if we change the norm on $\mathcal{C}^1[a,b]$ to

$$\|f\|_{1,\infty} := \|f\|_\infty + \|f'\|_\infty. \quad \text{(E.31)}$$

Indeed, for $f_n(t) = \frac{1}{\sqrt{n}}\cos(2\pi n t)$, we have

$$\forall n \in \mathbb{N}^+, \quad \|f_n' - \mathbf{0}'\|_\infty = 2\pi\sqrt{n} > 1,$$

yet $\|f_n - \mathbf{0}\|$ can be made arbitrarily small as $n \to \infty$. In contrast, $D : (\mathcal{C}^1[a,b], \|\cdot\|_{1,\infty}) \to (\mathcal{C}[a,b], \|\cdot\|_\infty)$ is continuous because

$$\forall \epsilon > 0,\ \exists \delta = \epsilon,\ \text{s.t.}\ \ \forall f, g \in \mathcal{C}^1[0,1],\ \|f - g\|_{1,\infty} < \delta \ \Rightarrow$$
$$\|Df - Dg\|_\infty = \|f' - g'\|_\infty \le \|f - g\|_{1,\infty} < \delta = \epsilon.$$

Exercise E.65. Show that the arc length function $L : \mathcal{C}^1[0,1] \to \mathbb{R}$,

$$L(f) := \int_0^1 \sqrt{1 + (f'(t))^2}\mathrm{d}t, \quad \text{(E.32)}$$

is not continuous if the norm of $\mathcal{C}^1[0,1]$ is $\|\cdot\|_\infty$, whereas it is continuous if we equip $\mathcal{C}^1[0,1]$ with (E.31).

Exercise E.66. Is the function $S : (c_{00}, \|\cdot\|_\infty) \to (\mathbb{R}, |\cdot|)$,

$$S\left((a_n)_{n\in\mathbb{N}}\right) = \sum_{n=1}^{\infty} a_n^2 \quad \text{(E.33)}$$

continuous?

Theorem E.67. A map $f : X \to Y$ between normed spaces is continuous if and only if the preimage $f^{-1}(V)$ of each open set V in Y is open in X.

Corollary E.68. A map $f : X \to Y$ between normed spaces is continuous if and only if the preimage $f^{-1}(V)$ of each closed set V in Y is closed in X.

Lemma E.69. If $f : X \to Y$ and $g : Y \to Z$ are continuous functions between normed spaces, then the composition map $g \circ f : X \to Z$ is continuous.

Lemma E.70. Let X, Y be normed spaces and let K be a compact subset of X. If $f : X \to Y$ is continuous at each $x \in K$, then $f(K)$ is a compact subset of Y.

Proof. For a sequence $(y_n)_{n\in\mathbb{N}} \subset f(K)$, there exists for each $n \in \mathbb{N}$ an $x_n \in K$ such that $f(x_n) = y_n$. This defines a sequence $(x_n)_{n\in\mathbb{N}} \subset K$. Because K is compact, Definition E.41 implies the existence of a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ that converges to $L \in K$. Since f is continuous, Lemma E.59 implies that $(y_n)_{n\in\mathbb{N}}$ converges to $f(L) \in f(K)$. □

Theorem E.71 (Weierstrass)**.** Suppose K is a nonempty compact subset of a normed space X and the function $f : X \to \mathbb{R}$ is continuous at each $x \in K$. Then

$$\exists a, b \in K \text{ s.t. } \begin{cases} f(a) = \max\{f(x) : x \in K\}, \\ f(b) = \min\{f(x) : x \in K\}. \end{cases}$$

Proof. It suffices to only prove the first clause. By Lemma E.70, $f(K)$ is compact, and thus by Lemma E.47 $f(K)$ is bounded. $f(K)$ is also nonempty because K is nonempty. Then Theorem A.28 implies that $f(K) \subset \mathbb{R}$ must have a unique supremum

$$M := \sup\{f(x) : x \in K\} \in \mathbb{R},$$

and hence there exists a sequence $(x_n)_{n\in\mathbb{N}} \subset K$ satisfying $\lim_{n\to\infty} f(x_n) = M$. By Definition E.41, $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ that converges to some $c \in K$. The continuity of f, Lemma E.59 and Lemma C.9 yield

$$\lim_{k\to\infty} f(x_{n_k}) = f(c) = M. \qquad \square$$

Example E.72. Since the set $K = \{\mathbf{x} \in \mathbf{R}^3 : \|\mathbf{x}\|_2 = 1\}$ is compact in $\mathbb{R}^3$ and the function $\mathbf{x} \mapsto \sum_{j=1}^3 x_j$ is continuous, the optimization problem

$$\begin{cases} \text{minimize } \sum_{j=1}^3 x_j, \\ \text{subject to } \|\mathbf{x}\|_2 = 1, \end{cases}$$

has a minimizer.

E.1.7 Norm equivalence

Example E.73. The optimal mining problem in Example E.1 concerns $\mathcal{C}^1[a,b]$. Since $\mathcal{C}^1[a,b]$ is a subspace of $\mathcal{C}[a,b]$, we could use the norm $\|\cdot\|_\infty$ for $\mathcal{C}[a,b]$ as a norm for $\mathcal{C}^1[a,b]$. But by Example E.64, the differentiation map would not be continuous; instead, if we equip $\mathcal{C}^1[a,b]$ with (E.31), then the differentiation map is continuous. Also, it might be more appropriate to regard two functions in $\mathcal{C}^1[a,b]$ as being close to each other if both their function values and their function derivatives are close.

Definition E.74. Two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ on V are *equivalent*, written $\|\cdot\|_A \sim \|\cdot\|_B$, iff

$$\exists c_1, c_2 \in \mathbb{R}^+ \text{ s.t. } \forall v \in V, \quad c_1\|v\|_A \le \|v\|_B \le c_2\|v\|_A. \tag{E.34}$$

Example E.75. The Euclidean norms in Definition B.154 satisfy

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_p \le \sqrt[p]{n}\|\mathbf{x}\|_\infty. \tag{E.35}$$

Therefore all the Euclidean ℓ_p norms are equivalent.

Exercise E.76. Show that $\sim$ in Definition E.74 defines an equivalence relation on the set of all norms on V.

Exercise E.77. Show that two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ on a linear space V are equivalent if and only if each sequence converging with respect to $\|\cdot\|_A$ also converges with respect to $\|\cdot\|_B$.

Theorem E.78. All norms are equivalent on $\mathbb{R}^n$ or $\mathbb{C}^n$.

Proof. Since $\sim$ is an equivalence relation on the set of all norms, it suffices to prove that any norm is equivalent to the 2-norm. Let $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$ be a basis of $\mathbb{R}^n$. Then any vector $\mathbf{x} \in \mathbb{R}^n$ can be expressed as $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$. Thus

$$\|\mathbf{x}\| = \left\| \sum_{i=1}^n x_i \mathbf{e}_i \right\| \le \sum_{i=1}^n |x_i| \|\mathbf{e}_i\| \le M\|\mathbf{x}\|_2,$$

where $M = \sqrt{\sum_{i=1}^n \|\mathbf{e}_i\|}$ and the last inequality follows from the Cauchy-Schwarz inequality (Theorem B.171). Set

$$K := \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\|_2 = 1\}.$$

Since K is a compact set and the norm $\|\cdot\| : K \to \mathbb{R}$ is a continuous function (Lemma E.61), $\|\cdot\|$ must attain its minimum value m on K. Furthermore, $m > 0$ since $\mathbf{0} \notin K$. For $\mathbf{x} \neq 0$, we have $\mathbf{y} := \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \in K$ since $\|\mathbf{y}\|_2 = 1$. Then $\|y\| \ge m$ implies $m\|\mathbf{x}\|_2 \le \|\mathbf{x}\|$. □

Corollary E.79. Over a finite dimensional space, any two norms are equivalent.

Proof. This follows from Theorem E.78 and the isomorphism of linear spaces. □

Example E.80. In the normed space $V := \mathcal{C}[0,1]$, consider a sequence of functions $\{u_n\}$ given by

$$u_n(x) := \begin{cases} 1 - nx, & x \in [0, \frac{1}{n}]; \\ 0, & x \in (\frac{1}{n}, 1]. \end{cases}$$

For the p-norm in (E.14), we have

$$\|u_n\|_p = [n(p+1)]^{-\frac{1}{p}}$$

and thus the sequence $\{u_n\}$ converges to $u = 0$. However, for the ∞-norm in (E.15), we have

$$\|u_n\|_\infty = 1$$

and thus the sequence $\{u_n\}$ does not converge to $u = 0$.

E.1.8 Banach spaces

Definition E.81. A *Cauchy sequence in a normed space* V is a sequence $\{u_n\} \subset V$ satisfying

$$\lim_{m,n \to +\infty} \|u_m - u_n\| = 0. \tag{E.36}$$

Definition E.82 (Banach spaces)**.** A *Banach space* (or a *complete* normed space) is a normed space V such that every Cauchy sequence in V converges to an element in V.

Example E.83 ($\mathbb{Q}$ is not complete)**.** The sequence

$$x_1 = \frac{3}{2}; \qquad \forall n > 1,\ x_n = \frac{4 + 3x_{n-1}}{3 + 2x_{n-1}} \tag{E.37}$$

is bounded below by $\sqrt{2}$ and is monotonically decreasing. By Theorem C.12, (x_n) is convergent in $\mathbb{R}$. However, although (x_n) is Cauchy in $\mathbb{Q}$, it is not convergent in $\mathbb{Q}$ because

$$L = \frac{4 + 3L}{3 + 2L} \ \Rightarrow\ L = \sqrt{2}.$$

Example E.84. The sequence $(\sum_{k=1}^n \frac{1}{k^k})_{n \in \mathbb{N}}$ is Cauchy, but we do not know yet whether the limit is rational or irrational.

Theorem E.85. $(\mathcal{C}[a,b], \|\cdot\|_\infty)$ is a Banach space.

Proof. It is straightforward to show that $\|\cdot\|_\infty$ is a norm. We only show the completeness in three steps.

First, at any fixed $t \in [a,b]$, we can reduce a Cauchy sequence $\{f_n\}_{n \ge 1} \subset \mathcal{C}[a,b]$ to a sequence $\{f_n(t)\} \subset \mathbb{R}$. The completeness of $\mathbb{R}$ (Theorem C.15) yields $\lim_{n\to\infty} f_n(t) \in \mathbb{R}$. For any Cauchy sequence $\{f_n\} \subset \mathcal{C}[a,b]$, this process furnishes a function $f : [a,b] \to \mathbb{R}$ given by

$$f(t) = \lim_{n \to \infty} f_n(t).$$

Second, we show $f \in \mathcal{C}[a,b]$, i.e., f is continuous. The sequence $\{f_n\} \subset \mathcal{C}[a,b]$ being Cauchy implies

$$\forall \epsilon > 0, \exists N-1 \in \mathbb{N} \text{ s.t. } \forall m, n > N-1, \qquad \|f_m - f_n\|_\infty < \frac{\epsilon}{3}.$$

In particular, set $m = N$, let $n \to \infty$, and we have

$$\forall t \in [a,b], \quad |f_N(t) - f(t)| \le \|f_N - f_n\|_\infty < \frac{\epsilon}{3}.$$

The condition of $f_N \in \mathcal{C}[a,b]$ implies

$$\begin{aligned}&\forall t \in [a,b], \forall \epsilon > 0, \exists \delta > 0 \text{ s.t.}\\ &\qquad |t-\tau| < \delta \;\Rightarrow\; |f_N(t) - f_N(\tau)| < \frac{\epsilon}{3}.\end{aligned}$$

The above two equations yield

$$\begin{aligned}&\forall t \in [a,b], \forall \epsilon > 0, \exists \delta > 0 \text{ s.t. } |t-\tau| < \delta \;\Rightarrow\\ &\quad |f(t) - f(\tau)| \le |f(t) - f_N(t)| + |f_N(t) - f_N(\tau)|\\ &\qquad + |f_N(\tau) - f(\tau)|\\ &\quad \le \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,\end{aligned}$$

which shows that f is continuous at every $t \in [a,b]$.

Finally, we show that $\{f_n\}_{n\ge 1}$ indeed converges to f. The sequence $\{f_n\} \subset \mathcal{C}[a,b]$ being Cauchy implies

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall m, n > N, \qquad \|f_m - f_n\|_\infty < \epsilon.$$

For a fixed $n > N$, we have

$$\forall m > N, \forall t \in [a,b], \quad |f_n(t) - f_m(t)| \le \|f_n - f_m\|_\infty < \epsilon,$$

which implies

$$\forall t \in [a,b], \quad |f_n(t) - f(t)| = \left|f(t) - \lim_{m\to\infty} f_m(t)\right| < \epsilon.$$

It follows that

$$\|f_n - f\|_\infty = \max_{t\in[a,b]} |f_n(t) - f(t)| < \epsilon.$$

In the above process, we could have fixed any $n > N$ at the outset to obtain the same result. Therefore we have

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, \ \|f_n - f\|_\infty < \epsilon,$$

which implies $\lim_{n\to\infty} f_n = f$. □

Exercise E.86. Define $\mathcal{C}_b[0,\infty)$ as the set of all functions f that are continuous on $[0,\infty)$ and satisfy

$$\|f\|_\infty := \sup_{x\ge 0} |f(x)| < \infty.$$

Show $\mathcal{C}_b[0,\infty)$ with this norm is complete.

Exercise E.87. Define $\mathcal{C}^\alpha[a,b]$ as the set of all functions $f \in \mathcal{C}[a,b]$ satisfying

$$M_\alpha(f) := \sup_{x,y\in[a,b];x\ne y} \frac{|f(x) - f(y)|}{|x-y|^\alpha} < \infty.$$

Define $\|f\|_\alpha = \|f\|_\infty + M_\alpha(f)$. Show that $(\mathcal{C}^\alpha[a,b], \|\cdot\|_\alpha)$ is a Banach space.

Example E.88. For $p \in [1,\infty)$, $\left(\mathcal{C}(\overline{\Omega}), \|\cdot\|_p\right)$ is not a Banach space. Consider $(u_n)_{n\in\mathbb{N}} \subset \mathcal{C}[0,1]$ given by

$$u_n(x) = \begin{cases} 0, & x \in [0, \frac{1}{2} - \frac{1}{2n}];\\ nx - \frac{n-1}{2}, & x \in [\frac{1}{2} - \frac{1}{2n}, \frac{1}{2} + \frac{1}{2n}];\\ 1 & x \in [\frac{1}{2} + \frac{1}{2n}, 1]. \end{cases} \tag{E.38}$$

$(u_n)_{n\in\mathbb{N}}$ is clearly Cauchy and we have

$$\lim_{n\to\infty} u_n = u(x) = \begin{cases} 0, & x \in [0, \frac{1}{2});\\ 1, & x \in (\frac{1}{2}, 1]. \end{cases}$$

But $u(x)$ cannot be in $\mathcal{C}(\overline{\Omega})$ no matter how we define $u(\frac{1}{2})$.

Exercise E.89. Show that the sequence space $(\ell^p, \|\cdot\|_p)$ is complete for $p \in [1, +\infty]$.

Theorem E.90. In a Banach space, absolutely convergent series converge. More precisely, if $(x_n)_{n\in\mathbb{N}}$ is a sequence in a Banach space $(X, \|\cdot\|)$ such that $\sum_{n=1}^\infty \|x_n\|$ converges, then $\sum_{n=1}^\infty x_n$ converges in X. Furthermore,

$$\left\|\sum_{n=1}^\infty x_n\right\| \le \sum_{n=1}^\infty \|x_n\|. \tag{E.39}$$

Proof. Since X is Banach, it suffices to prove that the sequence $\left(s_n = \sum_{i=1}^n x_i\right)_{n\in\mathbb{N}}$ is Cauchy. Since the real sequence $\left(\sigma_n = \sum_{i=1}^n \|x_i\|\right)_{n\in\mathbb{N}}$ is Cauchy, we have

$$\forall \epsilon > 0, \ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > m > N, \ \sum_{i=m+1}^n \|x_i\| < \epsilon,$$

which implies that $\left(s_n = \sum_{i=1}^n x_i\right)_{n\in\mathbb{N}}$ is Cauchy:

$$\left\|\sum_{i=m+1}^n x_i\right\| \le \sum_{i=m+1}^n \|x_i\| < \epsilon.$$

Set $L := \sum_{i=m+1}^n x_i$ and we have

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

which implies

$$\|L\| \le \|s_n - L\| + \|s_n\| < \epsilon + \sigma_n < \epsilon + \sum_{n=1}^\infty \|x_n\|,$$

where the second inequality follows from the triangle inequality and the third from n being a finite number. Then (E.39) holds because ϵ can be made arbitrarily small. □

Example E.91. The series $\sum_{n=1}^\infty \frac{1}{n^2}\sin(nx)$ converges in $(\mathcal{C}[0,2\pi], \|\cdot\|_\infty)$ since $\sum_{n=1}^\infty \frac{1}{n^2}$ converges in $\mathbb{R}$. Hence $x \mapsto \sum_{n=1}^\infty \frac{1}{n^2}\sin(nx)$ defines a continuous function.

Exercise E.92. Prove the converse of Theorem E.90, i.e., a normed space X is complete if every absolutely convergent series converges in X.

Theorem E.93. For each normed space V, there exists another normed space W and a dense subspace $\widehat{V} \subset W$ such that one can find an *isometric isomorphism* between V and $\widehat{V}$, i.e., a bijective linear function $\mathcal{I} : V \to \widehat{V}$ satisfying

$$\forall v \in V, \quad \|\mathcal{I}v\|_W = \|v\|_V. \tag{E.40}$$

Furthermore, the complete normed space W is unique up to the isometric isomorphism.

Definition E.94. The normed space W in Theorem E.93 is called the *completion of the normed space* V.

Example E.95. If V is the normed space $\mathbb{Q}$ of rational numbers, then $W = \mathbb{R}$ is a completion of $\mathbb{Q}$, where each element is an equivalence class of Cauchy sequences of rational numbers.

E.2 Continuous linear maps

E.2.1 The space $\mathcal{CL}(X, Y)$

Notation 15. $\mathcal{CL}(X, Y)$ denotes the *set of all continuous linear transformations or bounded linear transformations* from the normed space X to the normed space Y,

$$\mathcal{CL}(X, Y) := \mathcal{C}(X, Y) \cap \mathcal{L}(X, Y). \tag{E.41}$$

For $Y = X$, we write $\mathcal{CL}(X)$.

Theorem E.96. For any map $T \in \mathcal{L}(X, Y)$, the following statements are equivalent:

(1) T is continuous,

(2) T is continuous at $\mathbf{0}$,

(3) $\exists M \in \mathbb{R}^+$ s.t. $\forall x \in X$, $\|Tx\|_Y \le M\|x\|_X$.

Proof. (1)⇒(2) follows from Definition E.58. For (2)⇒(3), the continuity of T at $\mathbf{0}$ implies

$$\text{for } \epsilon = 1, \exists \delta > 0 \text{ s.t. } \|x\| < \delta \;\Rightarrow\; \|Tx\| < 1.$$

Replacing x with $y = \frac{\delta}{2}\frac{x}{\|x\|}$ in the above inequalities yields $\|Tx\| \le M\|x\|$ with $M = \frac{2}{\delta}$. Finally, (3)⇒(1) follows from

$$\begin{aligned}&\forall \epsilon > 0, \exists \delta = \frac{\epsilon}{M} \text{ s.t. } \|x - y\| < \delta \;\Rightarrow\\ &\|Tx - Ty\| = \|T(x - y)\| \le M\|x - y\| = \frac{\epsilon}{\delta}\|x - y\| < \epsilon. \;\square\end{aligned}$$

Example E.97. The left shift operator $L : \ell^2 \to \ell^2$ and right shift operator $R : \ell^2 \to \ell^2$,

$$L(a_1, a_2, a_3, \ldots) = (a_2, a_3, \ldots), \tag{E.42}$$
$$R(a_1, a_2, a_3, \ldots) = (0, a_1, a_2, \ldots), \tag{E.43}$$

are linear operators. Furthermore, $L, R \in \mathcal{CL}(\ell^2)$ because they are bounded:

$$\|L(a_n)_{n\in\mathbb{N}}\| \le \|(a_n)_{n\in\mathbb{N}}\|, \quad \|R(a_n)_{n\in\mathbb{N}}\| = \|(a_n)_{n\in\mathbb{N}}\|.$$

Example E.98. The linear map $T : (\mathcal{C}[a, b], \|\cdot\|_\infty) \to \mathbb{R}$ given by $T(f) = \int_a^b f(t)\mathrm{d}t$ is continuous because

$$|T(f)| = \left|\int_a^b f(t)\mathrm{d}t\right| \le \int_a^b \|f\|_\infty \mathrm{d}t = (b - a)\|f\|_\infty.$$

By Lemma E.59, T preserves convergent sequences:

$$\lim_{n\to\infty} f_n = f \;\Rightarrow\; \lim_{n\to\infty} \int_a^b f_n = \int_a^b f.$$

In other words, the continuity of T under $\|\cdot\|_\infty$ guarantees that T and $\lim_{n\to\infty}$ are commutative; see Section C.7.

Theorem E.99 (Existence and uniqueness of ODEs)**.** The IVP

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = f(x(t), t) \tag{E.44}$$

with initial condition $x(0) = x_0 \in \mathbb{R}$ has a unique solution $x \in \mathcal{C}^1[0, T]$ for some $T > 0$, if $f : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ is Lipschitz continuous in space and continuous in time.

Proof. For existence, we define $y_0(t) = x_0$ and

$$(*): \quad y_{n+1}(t) = x_0 + \int_0^t f(y_n(\tau), \tau)\mathrm{d}\tau.$$

For any $t \in [0, \frac{1}{2L}]$ where L is the Lipschitz constant,

$$\begin{aligned}|y_{n+1}(t) - y_n(t)| &= \left|\textstyle\int_0^t f(y_n(\tau), \tau) - f(y_{n-1}(\tau), \tau)\mathrm{d}\tau\right|\\ &\le \textstyle\int_0^t |f(y_n(\tau), \tau) - f(y_{n-1}(\tau), \tau)|\,\mathrm{d}\tau\\ &\le \textstyle\int_0^t L|y_n(\tau) - y_{n-1}(\tau)|\mathrm{d}\tau\\ &\le \textstyle\int_0^t L\|y_n - y_{n-1}\|_\infty \mathrm{d}\tau\\ &\le \tfrac{1}{2}\|y_n - y_{n-1}\|_\infty.\end{aligned}$$

Hence we have

$$\|y_{n+1} - y_n\|_\infty \le \frac{1}{2}\|y_n - y_{n-1}\|_\infty \le \frac{1}{2^n}\|y_1 - y_0\|_\infty.$$

It follows that $(y_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and there exists $y \in \mathcal{C}^1[0, T]$ such that $\lim_{n\to\infty} y_n = y$. Similarly, $(f(y_n, t))_{n\in\mathbb{N}}$ is a Cauchy sequence and there exists $f(y, t)$ such that $\lim_{n\to\infty} f(y_n, t) = f(y, t)$. Take $\lim_{n\to\infty}(*)$, apply Example E.98, and we have

$$(**): \quad y(t) = x_0 + \int_0^t f(y(\tau), \tau)\mathrm{d}\tau.$$

It is trivial to check that the above $y(t)$ solves (E.44).

For uniqueness, suppose for two solutions x and y of (E.44) there exists $t^* \in (0, T)$ satisfying

$$t^* := \max\{t \in [0, T] : \forall \tau \le t,\ y(\tau) = x(\tau)\}.$$

We choose

$$\begin{aligned}N &:= \max\left\{2, \tfrac{1}{L(T - t^*)}\right\},\\ M &:= \max_{t\in[t_*, t_* + \frac{1}{LN}]} |x(t) - y(t)|\end{aligned}$$

to obtain $t_* + \frac{1}{LN} \le T$. Then $(**)$ implies

$$\begin{aligned}&\forall t \in [t^*, t_* + \tfrac{1}{LN}],\\ &|x(t) - y(t)| = \left|\int_{t^*}^t [f(x(\tau),\tau) - f(y(\tau),\tau)]\,\mathrm{d}\tau\right|\\ &\le \int_{t^*}^t |f(x(\tau),\tau) - f(y(\tau),\tau)|\,\mathrm{d}\tau \le \int_{t^*}^t L\,|x(\tau) - y(\tau)|\,\mathrm{d}\tau\\ &\le LM(t - t^*) \le \tfrac{M}{N},\end{aligned}$$

which yields $M \le \frac{M}{N}$ and contradicts $N \ge 2$. Hence the uniqueness is proved by the non-existence of such a t^*. □

Example E.100. The continuity of differentiation maps in Example E.64 can be determined by Theorem E.96. For $\|\cdot\|_{1,\infty}$, we have $\|D\mathbf{x}\|_\infty \le \|\mathbf{x}\|_{1,\infty}$, and thus the operator $D : (\mathcal{C}[0,1], \|\cdot\|_{1,\infty}) \to (\mathcal{C}[0,1], \|\cdot\|_\infty)$ is continuous.

In comparison, $D : (\mathcal{C}[0,1], \|\cdot\|_\infty) \to (\mathcal{C}[0,1], \|\cdot\|_\infty)$ is not continuous: for $\mathbf{x}_n = t^n$, we have $\|x_n\|_\infty = 1$ yet $\lim_{n\to\infty} \|\mathbf{x}_n'\| = \infty$.

Corollary E.101. For finite-dimensional normed spaces X and Y, we have $\mathcal{L}(X,Y) = \mathcal{CL}(X,Y)$.

Proof. Each linear transformation $T_A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ has a matrix $A \in \mathbb{R}^{m\times n}$ such that

$$\begin{aligned}\|T_A\mathbf{x}\|_2^2 = \|A\mathbf{x}\|_2^2 &= \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij}x_j\right)^2\\ &\le \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij}\right)^2 \left(\sum_{j=1}^n x_j\right)^2 = \|\mathbf{x}\|_2^2 \sum_{i=1}^m\sum_{j=1}^n a_{ij}^2,\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality. The proof is completed by Theorem E.96, Theorem E.78, and the isomorphism of linear spaces. □

Exercise E.102. For an infinite-dimensional matrix A satisfying $\sum_{i=1}^\infty \sum_{j=1}^\infty a_{ij}^2 < \infty$, define $T_A : \ell^2 \to \ell^2$ by

$$\forall \mathbf{x} = (x_j)_{j\in\mathbb{N}} \in \ell^2, \quad T_A\mathbf{x} = A\mathbf{x} = \left(\sum_{j=1}^\infty a_{ij}x_j\right)_{i\in\mathbb{N}^+}.$$

Prove $T_A \in \mathcal{CL}(\ell^2)$.

Exercise E.103. For $A, B \in \mathcal{C}[a,b]$ and

$$S := \{f \in \mathcal{C}^1[a,b] : f(a) = f(b) = 0\}, \tag{E.45}$$

show that the map $L : (S, \|\cdot\|_{1,\infty}) \to \mathbb{R}$ given by

$$L(f) = \int_a^b \big[A(t)f(t) + B(t)f'(t)\big]\,\mathrm{d}t$$

is a bounded linear transformation.

Exercise E.104. For $A \in \mathbb{R}^{m\times n}$, show that the subspace $\ker A$ is closed in $\mathbb{R}^n$.

Theorem E.105. Every subspace of $\mathbb{R}^n$ is closed.

Exercise E.106. Prove Theorem E.105.

Lemma E.107. $\mathcal{CL}(X,Y)$ is a subspace of $\mathcal{L}(X,Y)$.

Exercise E.108. Prove Lemma E.107.

Lemma E.109. The *operator norm* $\|\cdot\| : \mathcal{CL}(X,Y) \to \mathbb{R}$,

$$\forall T \in \mathcal{CL}(X,Y),\ \|T\| := \sup\big\{\|Tx\| : x \in X, \|x\| \le 1\big\}, \tag{E.46}$$

is well defined, i.e., $\|T\|$ is a unique bounded real number.

Proof. By Theorem A.28, it suffices to show that

$$S := \big\{\|Tx\| : x \in X, \|x\| \le 1\big\} \tag{E.47}$$

is a nonempty bounded subset of $\mathbb{R}$. S is nonempty because $\mathbf{0} \in X$ and $T\mathbf{0} = \mathbf{0}_Y$, imply $0 \in S$. The boundedness of S follows from Theorem E.96(3) and $\|x\|_X \le 1$. □

Lemma E.110. For any $T \in \mathcal{CL}(X,Y)$, we have

$$(\forall x \in X, \|Tx\| \le M\|x\|) \ \Rightarrow\ \|T\| \le M. \tag{E.48}$$

Proof. M is an upper bound of the set S in (E.47) while $\|T\|$ is the least upper bound of S. □

Lemma E.111. $\forall T \in \mathcal{CL}(X,Y)$, $\forall x \in X$, $\|Tx\| \le \|T\|\|x\|$.

Proof. The statement holds trivially for $x = \mathbf{0}$. Otherwise for $y = \frac{x}{\|x\|}$ we have $\|Ty\| \in S$ where S is in (E.47). Hence

$$\|Ty\| \le \|T\| \ \Rightarrow\ \|Tx\| \le \|T\|\|x\|. \qquad \square$$

Lemma E.112. $\forall S \in \mathcal{CL}(X,Y)$, $\forall T \in \mathcal{CL}(Y,Z)$, we have $\|ST\| \le \|S\|\|T\|$.

Proof. This follows from Lemmas E.110 and E.111. □

Theorem E.113. $\big(\mathcal{CL}(X,Y), \|\cdot\|\big)$ is a normed space.

Exercise E.114. Prove Theorem E.113.

Lemma E.115. For a normed space X, $\big(\mathcal{CL}(X,Y), \|\cdot\|\big)$ is a Banach space if Y is a Banach space.

Proof. Let $(T_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{CL}(X,Y)$. For any $x \in X$, $(T_nx)_{n\in\mathbb{N}} \subset Y$ is Cauchy as Lemma E.111 yields

$$\|T_nx - T_mx\| \le \|T_n - T_m\|\|x\|.$$

Since Y is complete, $(T_nx)_{n\in\mathbb{N}}$ converges to some $L(x) \in Y$. This defines a map $T(x) = L(x)$.

The second step is to show $T \in \mathcal{CL}(X,Y)$.

The third step is to show $\lim_{n\to\infty} T_n = T$. □

Exercise E.116. Supplement the proof of Lemma E.115 with all details.

Corollary E.117. If X is a normed space over $\mathbb{R}$, then the dual space of X, $X' = \mathcal{CL}(X,\mathbb{R})$, is a Banach space with the operator norm.

Proof. This follows directly from Lemma E.115. □

Corollary E.118. If X is a Banach space, then $\mathcal{CL}(X)$ is a Banach space with the operator norm.

Proof. This follows directly from Lemma E.115. □

Definition E.119. An *algebra* is a vector space V with an associative and distributive multiplication $V \times V \to V$,

$$\forall u, v, w \in V,\ \forall \alpha \in \mathbb{F},\quad \begin{cases} u(vw) = (uv)w, \\ (u+v)w = uw + vw, \quad u(v+w) = uv + uw, \\ \alpha(uv) = u(\alpha v) = (\alpha u)v. \end{cases} \tag{E.49}$$

The *multiplicative identity* is the element $e \in V$ such that $\forall v \in V$, $ev = v = ve$.

Definition E.120. A *normed algebra* is an algebra V with a norm $\|\cdot\|$ satisfying

$$\forall u, v \in V, \quad \|uv\| \le \|u\|\|v\|. \tag{E.50}$$

A *Banach algebra* is a normed algebra that is complete.

E.2.2 The topology of $\mathcal{CL}(X, Y)$

Notation 16. For a vector space X and its subsets A, A_1, A_2, we write

$$\begin{aligned} \forall \alpha \in \mathbb{R}, \quad \alpha A &:= \{\alpha a, a \in A\}; \\ \forall w \in X, \quad A + w &:= \{a + w, a \in A\}. \\ \forall A_1, A_2 \subset X, \quad A_1 + A_2 &:= \{a_1 + a_2 : a_1 \in A_1, a_2 \in A_2\}. \end{aligned} \tag{E.51}$$

Definition E.121. A linear map $T : X \to Y$ between normed spaces X and Y is *open* if its image of any open set is open.

Lemma E.122. Let X and Y be normed spaces. A bounded linear map $T \in \mathcal{CL}(X, Y)$ is open if and only if the image of the unit open ball in X under T contains some open ball centered at $\mathbf{0}_Y$ in Y, i.e.,

$$\exists \delta > 0 \text{ s.t. } B(\mathbf{0}_Y, \delta) \subset T\left(B(\mathbf{0}_X, 1)\right). \tag{E.52}$$

Proof. For necessity, T being an open map implies that the image $T\left(B(\mathbf{0}_X, 1)\right)$ is open. The linearity of T implies $\mathbf{0}_Y \in T\left(B(\mathbf{0}_X, 1)\right)$. Then Lemma D.190 yields (E.52).

For sufficiency, let $U \subset X$ be open, we need to show

$$(*): \quad \forall y_0 \in T(U),\ \exists r_Y > 0 \text{ s.t. } B(y_0, r_Y) \subset T(U).$$

$y_0 \in T(U)$ implies there exists $x_0 \in U$ such that $Tx_0 = y_0$. Since U is open, we have

$$(**): \quad \exists r_X > 0 \text{ s.t. } B(x_0, r_X) \subset U.$$

Choose $r_Y = \delta r_X$ and we have

$$\begin{aligned} B(Tx_0, r_Y) &= Tx_0 + B(\mathbf{0}_Y, \delta r_X) \subset Tx_0 + TB(\mathbf{0}_X, r_X) \\ &= TB(x_0, \delta r_X) \subset T(U), \end{aligned}$$

where the second step follows from (E.52). □

Lemma E.123. If a closed set F in a normed space X does not contain any open set, then $X \setminus F$ is dense in X.

Proof. We need to show

$$\forall x \in X, \exists r > 0 \text{ s.t. } B(x, r) \cap (X \setminus F) \neq \emptyset.$$

If $x \in (X \setminus F)$, then we are done. Otherwise $x \in F$ implies that $B(x, r)$ is not contained in F for any $r > 0$. Therefore,

$$\forall r > 0, \forall x \in X,\ \exists y \in B(x, r) \subset (X \setminus F) \text{ s.t. } \|y - x\| < r.$$

Then the proof is completed by Lemma D.190. □

Theorem E.124 (Baire)**.** Suppose $(F_n)_{n\in\mathbb{N}}$ is a sequence of closed sets in a Banach space X such that $X = \cup_{n\in\mathbb{N}} F_n$. Then there exists an $n \in \mathbb{N}$ and a nonempty open set U such that $U \subset F_n$.

Proof. Suppose that no F_n contains any nonempty open set. Then Lemma E.123 implies that $X \setminus F_n$ is dense in X for each $n \in \mathbb{N}$. Therefore we have

$$\exists x_1 \in (X \setminus F_1), \exists r_1 > 0 \text{ s.t. } \overline{B(x_1, r_1)} \subset (X \setminus F_1).$$

Both $B(x_1, r_1)$ and $(X \setminus F_1)$ are open and thus their intersection $D_2 := B(x_1, r_1) \cap (X \setminus F_1)$ is also open. Hence,

$$\exists x_2 \in D_2, \exists r_2 \in \left(0, \frac{r_1}{2}\right) \text{ s.t. } \overline{B(x_2, r_2)} \subset D_2;$$

Proceed inductively and we have

$$\exists x_n \in D_n, \exists r_n \in \left(0, \frac{r_{n-1}}{2}\right) \text{ s.t. } \overline{B(x_n, r_n)} \subset D_n,$$

where $D_n := B(x_{n-1}, r_{n-1}) \cap (X \setminus F_{n-1})$. By construction, $n > m$ implies $B(x_n, r_n) \subset B(x_m, r_m)$ and

$$\|x_n - x_m\| < r_m < \frac{r_1}{2^{m-1}}.$$

Hence $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and converges to x in the Banach space X. For any $m \in \mathbb{N}$, we have

$$x \in \overline{B(x_m, r_m)} \subset \left(X \setminus \cup_{i=1}^m F_m\right),$$

which contradicts $X = \cup_{n\in\mathbb{N}} F_n$ as $m \to \infty$. □

Lemma E.125 (Unit open ball)**.** Suppose X and Y are Banach spaces and $T \in \mathcal{CL}(X, Y)$ is surjective. Then the image $T(B_0)$ of the open ball $B_0 := B(\mathbf{0}_X, 1)$ contains an open ball about $\mathbf{0}_Y$.

Proof. Define

$$B_n := B\left(\mathbf{0}_X, \frac{1}{2^n}\right)$$

and we show $\overline{T(B_1)}$ contains an open ball. Indeed

$$\forall x \in X,\ \exists k > 2\|x\| \text{ s.t. } x \in kB_1$$

and thus $X = \cup_{k\in\mathbb{N}} kB_1$. T being surjective implies

$$Y = T(X) = T\left(\bigcup_{k\in\mathbb{N}^+} kB_1\right) = \bigcup_{k\in\mathbb{N}^+} kT(B_1) = \bigcup_{k\in\mathbb{N}^+} \overline{kT(B_1)},$$

where the last step follows from the condition of Y being a Banach space. By Theorem E.124, there exists some $\overline{kT(B_1)}$ that contains a nonempty open ball, which implies that $\overline{T(B_1)}$ also contains an open ball, say,

$$B(y_0, \epsilon) \subset \overline{T(B_1)},$$

which implies

$$B(\mathbf{0}_Y,\epsilon) = B(y_0,\epsilon) - y_0 \subset \overline{T(B_1)} + \overline{T(B_1)} \subset \overline{T(B_1)+T(B_1)} = \overline{T(B_0)}.$$

To sum up the above arguments, we have

$$(*): \quad B(\mathbf{0}_Y,\epsilon) \subset \overline{T(B_0)}.$$

Define $V_n := B(\mathbf{0}_Y, \frac{\epsilon}{2^n})$. To complete the proof, we show

$$V_1 = B\left(\mathbf{0}_Y, \frac{\epsilon}{2}\right) \subset T(B_0).$$

The linearity of T and $(*)$ imply

$$(\triangle): \quad \forall n \in \mathbb{N},\ V_n \subset \overline{T(B_n)}.$$

For $y \in V_1$, we have $y \in \overline{T(B_1)}$. Since both T and $\|\cdot\|$ are continuous, the map $x \mapsto \|y - Tx\|$ is also continuous, and therefore

$$\exists x_1 \in B_1 \text{ s.t. } \|y - Tx_1\| < \frac{\epsilon}{4}.$$

By definition of V_n and $(\triangle)$, $y - Tx_1 \in V_2 \subset \overline{T(B_2)}$. Thus

$$\exists x_2 \in B_2 \text{ s.t. } \|(y - Tx_1) - Tx_2\| < \frac{\epsilon}{8}.$$

Proceed inductively and we have

$$\forall k = 1,2,\ldots,n,\ \exists x_k \in B_k \text{ s.t. } \left\| y - T\sum_{k=1}^n x_k \right\| < \frac{\epsilon}{2^{n+1}}.$$

Take limit of the above and we have

$$(\square): \quad y = T\lim_{n\to\infty}\sum_{k=1}^n x_k.$$

For each k, $x_k \in B_k$ implies $\|x_k\| < \frac{1}{2^k}$. Hence

$$\sum_{k=1}^\infty \|x_k\| < \sum_{k=1}^\infty \frac{1}{2^k} = 1.$$

The completeness of X and Theorem E.90 yield

$$\exists x \in X \text{ s.t. } x = \sum_{k=1}^\infty x_k, \quad \|x\| < 1.$$

Then $(\square)$ yields $y = Tx \in T(B_0)$. □

Theorem E.126 (Open mapping)**.** For Banach spaces X and Y, any surjective map $T \in \mathcal{CL}(X,Y)$ is open.

Proof. This follows from Lemmas E.122 and E.125. □

Example E.127. The following function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \begin{cases} x+1 & \text{if } x \in (-\infty,-1]; \\ 0 & \text{if } x \in (-1,+1); \\ x-1 & \text{if } x \in [+1,+\infty), \end{cases}$$

is surjective and continuous; but since $f(-1,1) = \{0\}$ is closed, f is not open. By the open mapping theorem, if a map between two Banach spaces is not open but surjective and continuous, then it must be nonlinear.

E.2.3 Invertible operators

Lemma E.128. In a finite-dimensional vector space X, if two operators $T, S \in \mathcal{L}(X)$ satisfy $TS = I$, then $ST = I$.

Proof. $TS = I$ implies $\ker S = \{\mathbf{0}\}$ because

$$Sx = \mathbf{0} \ \Rightarrow\ TSx = \mathbf{0} \ \Rightarrow\ x = \mathbf{0}.$$

Thus for any basis $(v_i)_{i=1}^n$ of X, $(Sv_i)_{i=1}^n$ is also a basis.

$$\forall x \in X,\ \exists (\beta_i)_{i=1}^n \text{ s.t. } x = \sum_{i=1}^n \beta_i S v_i = S\sum_{i=1}^n \beta_i v_i.$$

It follows that

$$\forall x \in X,\ STx = STS\sum_{i=1}^n \beta_i v_i = S\sum_{i=1}^n \beta_i v_i = x,$$

which implies $ST = I$. □

Example E.129. For the shift operators on ℓ^2 in Example E.97, we have $LR = I$ but $RL \neq I$,

$$RL(1,0,0,\ldots) = (0,0,0,\ldots).$$

Definition E.130. For vector spaces X and Y, a map $A \in \mathcal{L}(X,Y)$ is *invertible* if there exists $B \in \mathcal{L}(Y,X)$ such that $AB = I \in \mathcal{L}(Y)$ and $BA = I \in \mathcal{L}(X)$. Then B is called the *inverse* of A.

Exercise E.131. Prove that the inverse of $A \in \mathcal{L}(X,Y)$ is unique if A is invertible.

Lemma E.132. For any vector spaces X and Y, if a linear map $A \in \mathcal{L}(X,Y)$ is invertible, then A is bijective.

Proof. A is injective because

$$Ax = Ay \ \Rightarrow\ A^{-1}Ax = A^{-1}Ay \ \Rightarrow\ x = y.$$

A is surjective because $\forall y \in Y$, $A^{-1}y \in X$ implies $y = Ax$ for some $x \in X$. □

Lemma E.133. For any vectors space X and Y, if a map $A \in \mathcal{L}(X,Y)$ is invertible, then its inverse A^{-1} is linear.

Proof. For any $x, y \in X$, set $z = A^{-1}(x+y)$ and we have

$$x + y = Az \ \Rightarrow\ A^{-1}x + A^{-1}y = z = A^{-1}(x+y).$$

Similarly, for any $\alpha \in \mathbb{F}$, set $z = A^{-1}(\alpha x)$ and we have

$$Az = \alpha x \ \Rightarrow\ A\frac{z}{\alpha} = x \ \Rightarrow\ \frac{z}{\alpha} = A^{-1}x \Rightarrow \alpha A^{-1}x = z = A^{-1}(\alpha x).$$ □

Lemma E.134. For finite-dimensional vector space X and Y, if a map $A \in \mathcal{L}(X,Y)$ is bijective, then A is invertible.

Proof. For the bijective map A, define a map $B : Y \to X$,

$$\forall v \in X, \quad A(Bv) = v.$$

The existence and uniqueness of Bv are guaranteed by the surjectivity and injectivity of A. Therefore, $AB = I$. Furthermore, $BA = I$ follows from the injectivity of A and

$$\forall v \in X, \quad A(BAv) = (AB)Av = Av.$$

Finally, Lemma E.133 implies that B is a linear map. □

Theorem E.135. Suppose X and Y are finite-dimensional normed spaces. Then a map $A \in \mathcal{CL}(X, Y)$ is invertible with $A^{-1} \in \mathcal{CL}(Y, X)$ if and only if A is bijective.

Proof. This follows from Lemmas E.132, E.133, E.134, and Corollary E.101. □

Example E.136. The map $A : c_{00} \to c_{00}$ given by

$$\forall (x_n)_{n\in\mathbb{N}} \in c_{00}, \quad A(x_1, x_2, x_3, \ldots) = (x_1, \frac{x_2}{2}, \frac{x_3}{3}, \ldots)$$

is linear, bijective, and continuous (since $\|A\mathbf{x}\|_\infty \le \|\mathbf{x}\|_\infty$). However, it is not invertible in $\mathcal{CL}(c_{00})$. Suppose it is and $B \in \mathcal{CL}(c_{00})$ is the inverse of A. Then for the sequences $\mathbf{e}_m := (0, \ldots, 0, 1, 0, \ldots)$ where all terms are 0 except that the mth term is 1, we have

$$1 = \|\mathbf{e}_m\|_\infty = \|BA\mathbf{e}_m\|_\infty \le \|B\|\|A\mathbf{e}_m\|_\infty = \frac{\|B\|}{m}.$$

Hence $\forall m \in \mathbb{N}$, $\|B\| \ge m$ and this contradicts Lemma E.109.

Theorem E.137 (Banach)**.** For Banach spaces X and Y, a map $T \in \mathcal{CL}(X, Y)$ is invertible with $T^{-1} \in \mathcal{CL}(Y, X)$ if and only if T is bijective.

Proof. The necessity follows from Lemma E.132. For sufficiency, the bijective map T induces a map $T^{-1} : Y \to X$,

$$\forall y = Tx \in Y, \quad T^{-1}(y) = x.$$

Since the bijectiveness of T guarantees that T^{-1} is well defined, T^{-1} is indeed an inverse of T. By Lemma E.133, T^{-1} is linear. It remains to show that T^{-1} is continuous. By the surjectivity of T and Theorem E.126, T is open. Hence $T(U)$ is open whenever U is open. Meanwhile we have

$$\begin{aligned}\left(T^{-1}\right)^{-1}(U) &= \left\{y \in Y : T^{-1}y \in U\right\} \\ &= \{y \in Y : y \in T(U)\} = T(U).\end{aligned}$$

Thus T^{-1} is continuous by Theorem E.67. □

Definition E.138. A pair of *isomorphic normed spaces* are Banach spaces X and Y for which there exists a bijective map $T \in \mathcal{CL}(X, Y)$. Then T is called an *isomorphism of normed spaces* and we write $X \simeq Y$.

Theorem E.139 (Closed graph)**.** For two Banach spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$, a map $T \in \mathcal{L}(X, Y)$ is continuous if and only if its graph $\mathcal{G}(T) := \{(x, Tx) : x \in X\}$ is closed in $(X \times Y, \|\cdot\|_\infty)$ where

$$\forall (x, y) \in X \times Y, \quad \|(x, y)\|_\infty := \max(\|x\|_X, \|y\|_Y). \tag{E.53}$$

Exercise E.140. Prove Theorem E.139.

E.2.4 Series of operators

Theorem E.141 (Neumann series)**.** Suppose X is a Banach space and $A \in \mathcal{CL}(X)$ has $\|A\| < 1$. Then we have

(NST-1) $I - A$ is invertible in $\mathcal{CL}(X)$,

(NST-2) $(I - A)^{-1} = I + A + \cdots + A^n + \cdots = \sum_{n=0}^\infty A^n$,

(NST-3) $\left\|(I - A)^{-1}\right\| \le \frac{1}{1-\|A\|}$.

Proof. Since X is a Banach space, Corollary E.118 states that $\mathcal{CL}(X)$ is also a Banach space. By Theorem E.90, the convergence of $\sum_{n=0}^\infty \|A\|^n$ implies that the sequence $(S_n)_{n\in\mathbb{N}}$ with

$$S_n = \sum_{k=0}^n A^k$$

converges to some $S \in \mathcal{CL}(X)$. It follows that

$$\begin{aligned}& S_nA = AS_n = \textstyle\sum_{k=1}^{n+1} A^k = S_{n+1} - I \\ \Rightarrow\ & \begin{cases} \|AS_n - AS\| \le \|A\|\|S_n - S\|, \\ \|S_nA - SA\| \le \|A\|\|s - S_n\|. \end{cases} \\ \Rightarrow\ & SA = AS = S - I \\ \Rightarrow\ & (I - A)S = I = S(I - A) \\ \Rightarrow\ & (I - A)^{-1} = S = \textstyle\sum_{n=0}^\infty A^n,\end{aligned}$$

where the last step follows from Definition E.130. Finally, (NST-3) follows from

$$\left\|(I - A)^{-1}\right\| = \left\|\sum_{n=0}^\infty A^n\right\| \le \sum_{n=0}^\infty \|A^n\| \le \sum_{n=0}^\infty \|A\|^n = \frac{1}{1 - \|A\|}$$

where the first inequality follows from Theorem E.90 and the second inequality from Lemma E.112. □

Theorem E.142. Suppose X is a Banach space. Then the exponential of $A \in \mathcal{CL}(X)$, defined as

$$e^A := \sum_{n=0}^\infty \frac{1}{n!} A^n, \tag{E.54}$$

converges in $\mathcal{CL}(X)$.

Proof. By Lemma E.112, we have

$$\sum_{n=0}^\infty \left\|\frac{1}{n!}A^n\right\| \le \sum_{n=0}^\infty \frac{1}{n!}\|A\|^n = e^{\|A\|}.$$

By the comparison test, $\sum_{n=0}^\infty \frac{1}{n!}A^n$ converges absolutely. The rest of the proof follows from Theorem E.90. □

Lemma E.143. For a Banach space X, $A \in \mathcal{CL}(X)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{tA} := Ae^{tA} = e^{tA}A. \tag{E.55}$$

Lemma E.144. For a Banach space X, if $A, B \in \mathcal{CL}(X)$ commute, i.e. $AB = BA$, then

$$e^{A+B} := e^A e^B. \tag{E.56}$$

Corollary E.145. For a Banach space X and $A \in \mathcal{CL}(X)$, e^A is always invertible with its inverse as e^{-A}.

Theorem E.146 (Existence and uniqueness of ODEs)**.** For a Banach space X and $A \in \mathcal{CL}(X)$, the IVP

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t) \tag{E.57}$$

with initial condition $x(0) = x_0 \in X$ has a unique solution $x(t) = e^{tA}x_0$ for $t \in \mathbb{R}$.

Proof. If $x(t)$ solves (E.57), then

$$\frac{\mathrm{d}}{\mathrm{d}t}(e^{-tA}x(t)) = e^{-tA}(-A)x(t) + e^{-tA}\frac{\mathrm{d}}{\mathrm{d}t}(x(t)) = 0,$$

which implies $e^{-tA}x(t) = x_0$ and thus $x(t) = e^{tA}x_0$. □

E.2.5 Uniform boundedness

Lemma E.147. Suppose X is a normed space and a subset $A \subset X$ satisfies

- A is symmetric, i.e., $-A = A$;
- A is mid-point convex, i.e., $\forall x, y \in A$, $\frac{x+y}{2} \in A$,
- there exists a nonempty open set $U \subset A$.

Then there exists $\delta > 0$ such that $B(\mathbf{0}_X, \delta) \subset A$.

Proof. For $\alpha \neq 0$ and $a \in X$, the maps $x \mapsto x + a$ and $x \mapsto \alpha a$ are both continuous with continuous inverses. By Theorem E.67, U being open in X implies that its preimage $U + \{-a\}$ under $x \mapsto x + a$ is also open in X. Adopting notations in (E.51), we find that the set

$$U + (-A) := \cup_{a\in A}(U + \{-a\})$$

is open since it is a union of open sets. For $a \in U$, we have

$$\mathbf{0}_X = \frac{a-a}{2} \in \frac{U+(-A)}{2} \subset \frac{A+(-A)}{2} = \frac{A+A}{2} = A,$$

where the last two equalities follows from A being symmetric and mid-point convex, respectively. The proof is completed by Lemma D.190 and $\frac{U+(-A)}{2}$ being open. □

Theorem E.148 (Uniform boundedness principle)**.** Suppose X is a Banach space and Y is a normed linear space. For a family of maps $T_i \in \mathcal{CL}(X, Y)$, $i \in I$, "pointwise boundedness" implies "uniform boundedness,"

$$\forall x \in X, \quad \sup_{i\in I} \|T_i x\| < +\infty \;\Rightarrow\; \sup_{i\in I} \|T_i\| < +\infty.$$

Proof. For any given $n \in \mathbb{N}$, we define

$$F_n := \cap_{i\in I}\{x \in X : \|T_i x\| \le n\} = \{x \in X : \sup_{i\in I} \|T_i x\| \le n\}.$$

As intersection of closed sets, each F_n is closed. By pointwise boundedness, we have $X = \cup_{n\in\mathbb{N}} F_n$. The Baire theorem E.124 implies that there exists some F_n that contains a nonempty open subset. Since F_n is also symmetric and mid-point convex, Lemma E.147 implies that F_n contains an open ball $B(\mathbf{0}_X, \delta)$. Consequently, $x \in B(\mathbf{0}_X, \delta) \subset F_n$ implies

$$\|x\| < \delta \;\Rightarrow\; \forall i \in I, \|T_i x\| \le n.$$

Thus for any $x \in X$, there exists $y = \frac{\delta}{2}\frac{x}{\|x\|}$ such that

$$\forall i \in I,\ \|T_i y\| \le n \;\Rightarrow\; \|T_i x\| \le \frac{2n}{\delta}\|x\|,$$

and the proof is completed by Lemma E.110. □

Example E.149. Many PDEs can be written in the form

$$Tx = y,$$

where y is a known vector incorporating initial and boundary conditions, x is the unknown, and T is a continuous linear operator. If the PDE is well-posed, we can often assume that T is a bijection, hence by Theorem E.137 the inverse of T is a bounded linear operator and we write $x = T^{-1}y$. In numerically solving the PDE, we usually approximate y by a grid function y_n and approximate T^{-1} by a discrete operator T_n^{-1}. The convergence usually means

$$\forall y \in \mathcal{C}^r(\overline{\Omega}), \lim_{n\to\infty} y_n = y,\ \lim_{n\to\infty} T_n^{-1} y_n = x,$$

i.e., $\sup_{n\to\infty} \|T_n^{-1} y_n\| < \infty$. Theorem E.148 then implies $\sup_{n\in\mathbb{N}} \|T_n^{-1}\| < \infty$, which usually implies some form of numerical stability.

Theorem E.150 (Banach-Steinhaus)**.** Suppose X and Y are Banach spaces. If a sequence $(T_n)_{n\in\mathbb{N}} \in \mathcal{CL}(X, Y)$ has

$$\forall x \in X,\ \lim_{n\to\infty} \|T_n x\| < \infty,$$

then the map $x \mapsto \lim_{n\to\infty} T_n x$ belongs to $\mathcal{CL}(X, Y)$.

Proof. Clearly the map $T(x) = \lim_{n\to\infty} T_n x$ is linear, it remains to show that it is continuous. Because the limit $\lim_{n\to\infty} T_n x$ exists, we have $\sup_{n\in\mathbb{N}} \|T_n x\| < \infty$ for all $x \in X$. Then Theorem E.148 implies $\sup_{n\in\mathbb{N}} \|T_n\| < \infty$. Hence

$$\exists M \in \mathbb{R},\ \forall x \in X, \forall n \in \mathbb{N},\ \|T_n x\| \le M\|x\|;$$

the limit of the above yields $\forall x \in X,\ \|Tx\| \le M\|x\|$. The proof is completed by Theorem E.96. □

Bibliography

W. L. Briggs, V. E. Henson, and S. F. McCormick. *A multigrid tutorial*. SIAM, 2nd edition, 2000. ISBN: 0-89871-462-1.

M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

W. Fontana and P. Schuster. Continuity in evolution: on the nature of transitions. *Science*, 280:1451–5, 1998a.

W. Fontana and P. Schuster. Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, 194:491–515, 1998b.

E. Isaacson and H. B. Keller. *Analysis of Numerical Methods*. Dover, 1966. ISBN: 0-486-68029-0.

J. Munkres. *Topology*. Pearson, second edition, 2017.

B. M. R. Stadler, G. Wagner, and W. Fontana. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *J. Theor. Biol.*, 213:241–274, 2001.

M. E. Taylor. *Partial Differential Equations I*. Number 115 in Applied Mathematical Sciences. Springer, 2nd edition, 2011.