Contents

Preface

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

In writing this book, I have made special efforts to

- collect the prerequisites in the first chapter 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 close to the minimum,
- encourage the student 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 a logical and coherent story out of each class taken.

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。我们学习数值逼近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 0

Preliminaries

0.1 Sets and logic

0.1.1 First-order logic

Definition 0.1. A *set* $\mathcal{S}$ is a collection of *distinct* objects x's, often denoted with the following notation

$$\mathcal{S} = \{x \mid \text{ the conditions that } x \text{ satisfies. }\}. \tag{0.1}$$

Notation 1. $\mathbb{R}, \mathbb{Z}, \mathbb{N}, \mathbb{Q}, \mathbb{C}$ 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 0.2. $\mathcal{S}$ 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 0.3 (Statements of first-order logic)**.** A *universal statement* is a logical statement of the form

$$\mathsf{U} = (\forall x \in \mathcal{S},\ \mathsf{A}(x)). \tag{0.2}$$

An *existential statement* has the form

$$\mathsf{E} = (\exists x \in \mathcal{S},\ \text{s.t. } \mathsf{A}(x)), \tag{0.3}$$

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

$$\mathsf{A} \Rightarrow \mathsf{B}. \tag{0.4}$$

Example 0.1. 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 0.4. *Uniqueness quantification* or *unique existential quantification*, written $\exists!$ or $\exists_{=1}$, indicates that exactly one object with a certain property exists.

Exercise 0.2. Express the logical statement $\exists! x,\ \text{s.t. } \mathsf{A}(x)$ with $\exists$, $\forall$, and $\Leftrightarrow$.

Definition 0.5. A *universal-existential statement* is a logical statement of the form

$$\mathsf{U}_E = (\forall x \in \mathcal{S},\ \exists y \in \mathcal{T} \text{ s.t. } \mathsf{A}(x, y)). \tag{0.5}$$

An *existential-universal statement* has the form

$$\mathsf{E}_U = (\exists y \in \mathcal{T},\ \text{s.t. } \forall x \in \mathcal{S},\ \mathsf{A}(x, y)). \tag{0.6}$$

Example 0.3. 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 0.4 (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 0.6. The existential-universal statement implies the corresponding universal-existential statement, but not vice versa.

Example 0.5 (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 0.7 (First-order negation of logical statements)**.** The negations of the statements in Definition 0.3 are

$$\neg\mathsf{U} = (\exists x \in \mathcal{S},\ \text{s.t. } \neg\mathsf{A}(x)). \tag{0.7}$$
$$\neg\mathsf{E} = (\forall x \in \mathcal{S},\ \neg\mathsf{A}(x)). \tag{0.8}$$

Rule 0.8. 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.

One might need to group quantifiers of like type.

Example 0.6 (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 0.7. Negate the logical statement in Definition 0.51.

Axiom 0.9 (Contraposition)**.** A conditional statement is logically equivalent to its contrapositive.

$$(\mathsf{A} \Rightarrow \mathsf{B}) \Leftrightarrow (\neg\mathsf{B} \Rightarrow \neg\mathsf{A}) \tag{0.9}$$

Example 0.8. "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 0.9. Draw an Euler diagram of subsets to illustrate Example 0.8.

Exercise 0.10. 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.

0.1.2 Ordered sets

Definition 0.10. 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 $\mathcal{Y}$:

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) \mid x \in \mathcal{X},\ y \in \mathcal{Y}\}. \tag{0.10}$$

Axiom 0.11 (Fundamental principle of counting)**.** A task 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 then

$$\prod_{i=1}^{k} n_i = n_1 n_2 \cdots n_k. \tag{0.11}$$

Example 0.11. 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 0.12 (Maximum and minimum)**.** Consider $\mathcal{S} \subseteq \mathbb{R}$, $\mathcal{S} \neq \emptyset$. If $\exists s_m \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S}$, $x \le s_m$, then s_m is the *maximum* of $\mathcal{S}$ and denoted by $\max \mathcal{S}$. If $\exists s_m \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S}$, $x \ge s_m$, then s_m is the *minimum* of $\mathcal{S}$ and denoted by $\min \mathcal{S}$.

Definition 0.13 (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 $\forall x \in \mathcal{S}$, $x \le a$; then the set $\mathcal{S}$ is said to be *bounded above.* a is a *lower bound* of $\mathcal{S}$ if $\forall x \in \mathcal{S}$, $x \ge a$; then the set $\mathcal{S}$ is said to be *bounded below.* $\mathcal{S}$ is *bounded* if it is bounded above and bounded below.

Definition 0.14 (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 0.12. If a set $\mathcal{S} \subset \mathbb{R}$ has a maximum, we have $\max \mathcal{S} = \sup \mathcal{S}$.

Example 0.13. $\sup[a, b] = \sup[a, b) = \sup(a, b] = \sup(a, b)$.

Axiom 0.15 (Completeness of $\mathbb{R}$)**.** Every nonempty subset of $\mathbb{R}$ that is bounded above has a least upper bound.

Corollary 0.16. Every nonempty subset of $\mathbb{R}$ that is bounded below has a greatest lower bound.

Definition 0.17. 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 0.18. A binary relation "$\le$" on some set $\mathcal{S}$ is a *total order* or *linear order* on $\mathcal{S}$ iff, $\forall a, b, c \in \mathcal{S}$,

- $a \le b$ and $b \le a$ imply $a = b$ (antisymmetry);
- $a \le b$ and $b \le c$ imply $a \le c$ (transitivity);
- $a \le b$ or $b \le a$ (totality).

A set equipped with a total order is a *chain* or *totally ordered set.*

Example 0.14. The real numbers with less or equal.

Example 0.15. The English letters of the alphabet with dictionary order.

Example 0.16. The Cartesian product of a set of totally ordered sets with the *lexicographical order.*

Example 0.17. Sort your book in lexicographical order and save a lot of time. $\log_{26} N \ll N$!

Definition 0.19. A binary relation "$\le$" on some set $\mathcal{S}$ is a *partial order* on $\mathcal{S}$ iff, $\forall a, b, c \in \mathcal{S}$, antisymmetry, transitivity, and reflexivity ($a \le a$) hold.
A set equipped with a partial order is a called a *poset.*

Example 0.18. The set of subsets of a set $\mathcal{S}$ ordered by inclusion "$\subseteq$."

Example 0.19. The natural numbers equipped with the relation of divisibility.

Example 0.20. 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 0.21. Inheritance ("is-a" relation) is a partial order. $A \to B$ reads "B is a special type of A".

Example 0.22. Composition ("has-a" relation) is also a partial order. $A \rightsquigarrow B$ reads "B *has an* instance/object of A."

Example 0.23. Implication "$\Rightarrow$" is a partial order on the set of logical statements.

Example 0.24. 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.

0.2 Basic analysis

0.2.1 Limits and continuity

Definition 0.20. 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 0.21. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be *injective* or *one-to-one* iff

$$\forall x_1 \in \mathcal{X}, \forall x_2 \in \mathcal{X}, \quad x_1 \neq x_2 \ \Rightarrow\ f(x_1) \neq f(x_2). \tag{0.12}$$

It is *surjective* or *onto* iff

$$\forall y \in \mathcal{Y},\ \exists x \in \mathcal{X},\ \text{s.t. } y = f(x). \tag{0.13}$$

It is *bijective* iff it is both injective and surjective.

Definition 0.22. 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 0.25. Are the integers countable? Are the rationals countable? Are the real numbers countable?

Definition 0.23. A *scalar function* is a function whose range is a subset of $\mathbb{R}$.

Definition 0.24 (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

$$\forall \epsilon > 0, \exists \delta > 0,\ \text{s.t. } \forall x \in I(c, \delta),\ |f(x) - L| < \epsilon. \tag{0.14}$$

Example 0.26. 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 0.24. □

Definition 0.25. $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at c iff

$$\lim_{x\to c} f(x) = f(c). \tag{0.15}$$

f is *continuous on* (a, b), written $f \in \mathcal{C}(a, b)$ if (0.15) holds $\forall x \in (a, b)$.

Definition 0.26. Let $I = (a, b)$. A function $f : I \to \mathbb{R}$ is *uniformly continuous* on I iff

$$\begin{aligned}&\forall \epsilon > 0, \exists \delta > 0,\ \text{s.t.}\\ &\forall x, y \in I,\ |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon.\end{aligned} \tag{0.16}$$

Example 0.27. 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. (0.16), yields $\exists \epsilon > 0$ s.t. $\forall \delta > 0\ \exists x, y > 0$ s.t. $|x - y| < \delta \Rightarrow |f(x) - f(y)| \geq \epsilon$.

We prove a stronger version: $\forall \epsilon > 0$, $\forall \delta > 0\ \exists x, y > 0$ s.t. $|x - y| < \delta \Rightarrow |\frac{1}{x} - \frac{1}{y}| \geq \epsilon$.

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 0.28. On (a, ∞), $f(x) = \frac{1}{x^2}$ is uniformly continuous if $a > 0$ and is not so if $a = 0$.

Theorem 0.27. Uniform continuity implies continuity but the converse is not true.

Proof. exercise. □

Theorem 0.28. $f : \mathbb{R} \to \mathbb{R}$ is *uniformly continuous* on (a, b) iff it can be extended to a continuous function $\tilde{f}$ on $[a, b]$.

Definition 0.29. The *derivative* of a function $f : \mathbb{R} \to \mathbb{R}$ at a is the limit

$$f'(a) = \lim_{h\to 0} \frac{f(a + h) - f(a)}{h}. \tag{0.17}$$

If the limit exists, f is *differentiable* at a.

Example 0.29. 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 0.30. 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 0.31. A scalar function f is bounded on $[a, b]$ if $f \in \mathcal{C}[a, b]$.

Theorem 0.32 (Intermediate value)**.** A scalar function $f \in \mathcal{C}[a, b]$ satisfies

$$\forall y \in [m, M],\ \exists \xi \in [a, b],\ \text{s.t. } y = f(\xi) \tag{0.18}$$

where $m = \inf_{x\in[a,b]} f(x)$ and $M = \sup_{x\in[a,b]} f(x)$.

Theorem 0.33. 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 0.34 (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 0.32, 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 0.33 completes the proof. □

Theorem 0.35 (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 0.34, $\exists \xi \in [a, b]$ such that $g'(\xi) = 0$, which completes the proof. □

0.2.2 Taylor series

Definition 0.36. A *sequence* is a map on $\mathbb{N}^+$ or $\mathbb{N}$.

Example 0.30. Whether the sequence starts from 0 or 1 is a matter of convention and convenience according to the context.

Definition 0.37 (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

$$\forall \epsilon > 0,\ \exists N, \text{ s.t. } \forall n > N,\ |a_n - L| < \epsilon. \tag{0.19}$$

If such a limit L exists, we say that $\{a_n\}$ *converges* to L.

Example 0.31 (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.

$$\pi \approx 3.141592653589793\ldots \tag{0.20}$$

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.

Theorem 0.38 (Bolzano-Weierstrass)**.** Every bounded sequence has a convergent subsequence.

Definition 0.39. A *series* associated with an infinite sequence $\{a_n\}$ is defined as $\sum_{i=0}^{\infty} a_n$, the sum of all terms of the sequence.

Definition 0.40. The *sequence of partial sums* S_n associated to a 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

$$S_n = \sum_{i=0}^{n} a_i. \tag{0.21}$$

Lemma 0.41. A series converges to L iff the associated sequence of partial sums converges to L.

Definition 0.42. A *power series* centered at c is a series of the form

$$p(x) = \sum_{n=0}^{\infty} a_n (x - c)^n, \tag{0.22}$$

where a_n's are the *coefficients*. The *interval of convergence* is the set of values of x for which the series converges:

$$I_c(p) = \{x \mid p(x) \text{ converges}\}. \tag{0.23}$$

Definition 0.43. 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

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x - c)^k \tag{0.24}$$

is called the nth *Taylor polynomial* for $f(x)$ at c.
In particular, the *linear approximation* for $f(x)$ at c is

$$T_1(x) = f(c) + f'(c)(x - c). \tag{0.25}$$

Example 0.32. If $f \in \mathcal{C}^\infty$, then $\forall n \in \mathbb{N}$, we have

$$T_n^{(m)}(x) = \begin{cases} \sum_{k=m}^{n} \frac{f^{(k)}(c)}{(k-m)!}(x - c)^{k-m}, & m \in \mathbb{N}, m \le n; \\ 0, & m \in \mathbb{N}, m > n. \end{cases}$$

This can be proved by induction. In the inductive step, we regroup the summation into a constant term and another shifted summation.

Definition 0.44. The *Taylor series* (or Taylor expansion) for $f(x)$ at c is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!}(x-c)^k. \tag{0.26}$$

Definition 0.45. The *remainder* of the nth *Taylor polynomial* in approximating $f(x)$ is

$$E_n(x) = f(x) - T_n(x). \tag{0.27}$$

Theorem 0.46. Let T_n be the nth Taylor polynomial for $f(x)$ at c.

$$\lim_{n\to\infty} E_n(x) = 0 \;\Leftrightarrow\; \lim_{n\to\infty} T_n(x) = f(x). \tag{0.28}$$

Lemma 0.47. $\forall m = 0, 1, 2, \ldots, n$, $E_n^{(m)}(c) = 0$.

Proof. This follows from Definition 0.43 and Example 0.32. □

Theorem 0.48 (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

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}. \tag{0.29}$$

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

$$g(t) := E_n(t) - \frac{M(t-c)^{n+1}}{(n+1)!}. \tag{0.30}$$

Clearly $g(x) = 0$. By Lemma 0.47, $g^{(k)}(c) = 0$ 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,

$$\exists x_{n+1} \in (c, x_n) \text{ s.t. } g^{(n+1)}(x_{n+1}) = 0. \tag{0.31}$$

Since T_n is a polynomial of degree n, we have $T_n^{(n+1)}(t) = 0$, which, together with (0.31) and (0.30), yields

$$f^{(n+1)}(x_{n+1}) - M = 0.$$

The proof is completed by identifying ξ with x_{n+1}. □

Example 0.33. 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 0.44, the Taylor series of e^x at $c = 0$ is

$$e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}.$$

By Theorem 0.48, 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.

0.2.3 Riemann integral

Definition 0.49. A *partition of an interval* $I = [a, b]$ is a finite ordered subset $T_n \subseteq I$ of the form

$$T_n(a, b) = \{a = x_0 < x_1 < \cdots < x_n = b\}. \tag{0.32}$$

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,

$$h_n = h(T_n) = \max(x_i - x_{i-1}), \qquad i = 1, 2, \ldots, n. \tag{0.33}$$

Definition 0.50. The *Riemann sum* of $f : \mathbb{R} \to \mathbb{R}$ over a partition T_n is

$$S_n(f) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}), \tag{0.34}$$

where $x_i^* \in I_i$ is a *sample point* of the ith subinterval.

Definition 0.51. $f : \mathbb{R} \to \mathbb{R}$ is *integrable* (or more precisely *Riemann integrable*) on $[a, b]$ iff

$$\begin{aligned} &\exists L \in \mathbb{R}, \text{ s.t. } \forall \epsilon > 0, \; \exists \delta > 0 \text{ s.t.} \\ &\forall T_n(a, b) \text{ with } h(T_n) < \delta, \; |S_n(f) - L| < \epsilon. \end{aligned} \tag{0.35}$$

Example 0.34. The following function $f : [a, b] \to \mathbb{R}$ is not Riemann integrable.

$$f(x) = \begin{cases} 1 & x \text{ is rational;} \\ 0 & x \text{ is irrational.} \end{cases}$$

To see this, we first negate the logical statement in (0.35) to get

$$\begin{aligned} &\forall L \in \mathbb{R}, \exists \epsilon > 0, \text{ s.t. } \forall \delta > 0 \\ &\exists T_n(a, b) \text{ with } h(T_n) < \delta, \;\text{ s.t. } |S_n(f) - L| \geq \epsilon. \end{aligned}$$

If $|L| < \frac{b-a}{2}$, we choose all x_i^*'s to be rational so that $f(x_i^*) \equiv 1$; then (0.34) 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 (0.34) yields $S_n(f) = 0$. For $\epsilon = \frac{b-a}{4}$, the formula $|S_n(f) - L| \geq \epsilon$ clearly holds.

Definition 0.52. 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 $[a, b]$:

$$\int_a^b f(x)\mathrm{d}x = \lim_{h_n \to 0} S_n(f). \tag{0.36}$$

Theorem 0.53. A scalar function f is integrable on $[a, b]$ if $f \in \mathcal{C}[a, b]$.

Definition 0.54. 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 0.55. A scalar function is integrable on $[a, b]$ if it is monotonic on $[a, b]$.

Exercise 0.35. True or false: a bijective function is either order-preserving or order-reversing?

Theorem 0.56 (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.

$$\int_a^b w(x)f(x)\mathrm{d}x = f(\xi)\int_a^b w(x)\mathrm{d}x. \tag{0.37}$$

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 $I \ne 0$, hence

$$m \le \frac{1}{I}\int_a^b w(x)f(x)\mathrm{d}x \le M.$$

Applying Theorem 0.32 completes the proof. □

0.2.4 Uniform convergence in metric spaces

Definition 0.57. 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 0.36. 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 $\mathcal{X}$,

$$d(x, y) = \max_{t\in[a,b]} |x(t) - y(t)|. \tag{0.38}$$

Definition 0.58. The *sequence space* ℓ^∞ is a metric space $(\mathcal{X}, d)$, where $\mathcal{X}$ is the set of all bounded sequences of complex numbers,

$$\forall x = (\xi_1, \xi_2, \ldots) \in \mathcal{X}, \exists c_x \in \mathbb{R}, \text{ s.t. } \forall i = 1, 2, \ldots,\ |\xi_i| \le c_x,$$

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 0.37. Let $\mathcal{X}$ be the set of all bounded and unbounded sequences of complex numbers. Show that the following is a metric on $\mathcal{X}$,

$$d(x, y) = \sum_{j=1}^\infty \frac{1}{2^j}\frac{|\xi_j - \eta_j|}{1 + |\xi_j - \eta_j|}, \tag{0.39}$$

where $x = (\xi_j)$ and $y = (\eta_j)$.

Definition 0.59. For a real number $p \ge 1$, the ℓ^p *space* is the metric space $(\mathcal{X}, d)$ with

$$\mathcal{X} = \left\{(\xi_j)_{j=1}^\infty : \xi_j \in \mathbb{C}; \sum_{j=1}^\infty |\xi_j|^p < \infty\right\}; \tag{0.40}$$

$$d(x, y) = \left(\sum_{j=1}^\infty |\xi_j - \eta_j|^p\right)^{1/p}, \tag{0.41}$$

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 0.60. Two positive real numbers p, q are called *conjugate exponents* iff they satisfy

$$p > 1, \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{0.42}$$

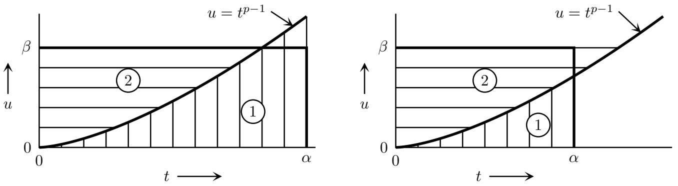
Lemma 0.61. Any two positive real numbers α, β satisfy

$$\alpha\beta \le \frac{\alpha^p}{p} + \frac{\beta^q}{q}, \tag{0.43}$$

where p and q are conjugate exponents.

Proof. By (0.42), 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 $\alpha = 0$ and $\beta = 0$. □

Exercise 0.38. Prove that (0.41) is indeed a metric. In particular, prove that (0.41) satisfies the triangular inequality by showing

(a) Lemma 0.61 implies the *Hölder inequality*, i.e., for conjugate exponents p, q and for any $(\xi_j) \in \ell^p$, $(\eta_j) \in \ell^q$,

$$\sum_{j=1}^\infty |\xi_j\eta_j| \le \left(\sum_{k=1}^\infty |\xi_k|^p\right)^{1/p}\left(\sum_{m=1}^\infty |\eta_m|^q\right)^{1/q}. \tag{0.44}$$

(b) The Hölder inequality implies the *Minkowski inequality*, i.e. for any $p \ge 1$, $(\xi_j) \in \ell^p$, and $(\eta_j) \in \ell^p$,

$$\left(\sum_{j=1}^\infty |\xi_j + \eta_j|^p\right)^{1/p} \le \left(\sum_{k=1}^\infty |\xi_k|^p\right)^{1/p} + \left(\sum_{m=1}^\infty |\eta_m|^p\right)^{1/p}. \tag{0.45}$$

(c) The Minkowski inequality implies that the triangular inequality holds for (0.41).

Definition 0.62. 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

$$B_r(x) := \{y \in \mathcal{X} : d(x,y) < r\}. \tag{0.46}$$

Definition 0.63. 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

$$\forall r > 0,\ E \cap B_r(x_0) \neq \emptyset. \tag{0.47}$$

Definition 0.64 (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

$$\lim_{x \to x_0; x \in E} f(x) = L, \tag{0.48}$$

iff

$$\begin{aligned} &\forall \epsilon > 0, \exists \delta > 0 \text{ s.t. } \forall x \in E, \\ &|x - x_0|_{\mathcal{X}} < \delta \ \Rightarrow\ |f(x) - L|_{\mathcal{Y}} < \epsilon. \end{aligned} \tag{0.49}$$

Notation 2. In Definition 0.64 we used the synonym notation

$$|u - v|_{\mathcal{X}} := d_{\mathcal{X}}(u,v). \tag{0.50}$$

Definition 0.65 (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

$$\forall x \in \mathcal{X}, \qquad \lim_{n \to \infty} f_n(x) = f(x), \tag{0.51}$$

or, equivalently,

$$\forall \epsilon > 0, \forall x \in \mathcal{X},\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > N,\ |f_n(x) - f(x)|_{\mathcal{Y}} < \epsilon. \tag{0.52}$$

Example 0.39. 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 0.40. For the functions in Example 0.39, we have $\lim_{x \to 1; x \in [0,1)} x^n = 1$ for all n and $\lim_{x \to 1; x \in [0,1)} f(x) = 0$; 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 0.41. 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 0.42. Pointwise convergence does not preserve boundedness. For example, the function sequence

$$f_n(x) = \begin{cases} \exp(x) & \text{if } \exp(x) \le n; \\ n & \text{if } \exp(x) > n \end{cases} \tag{0.53}$$

converges pointwise to $f(x) = \exp(x)$. Similarly, the function sequence

$$f_n(x) = \begin{cases} \frac{1}{x} & \text{if } x \ge \frac{1}{n}; \\ 0 & \text{if } x \in (0, \frac{1}{n}) \end{cases} \tag{0.54}$$

converges pointwise to $f(x) = \frac{1}{x}$. As another example, the function sequence

$$f_n(x) = n \sin \frac{x}{n} \tag{0.55}$$

converges pointwise to $f(x) = x$.

Definition 0.66 (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

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall x \in \mathcal{X}, \forall n > N,\ |f_n(x) - f(x)|_{\mathcal{Y}} < \epsilon. \tag{0.56}$$

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 $B_r(x) \cap \mathcal{X}$.

Theorem 0.67. Uniform convergence implies pointwise convergence.

Proof. This follows directly from (0.52), (0.56), and Theorem 0.6. □

Example 0.43 (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 0.43. 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 $+\infty$.

0.3 Linear algebra

0.3.1 Vector spaces and the basis

Definition 0.68. A *field* is a commutative division ring. More commonly, 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 0.69. A *vector space* or *linear space* over a field $\mathbb{F}$ is a set $\mathcal{V}$ together with two binary operations "+" and "×" 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
$\forall \mathbf{u} \in \mathcal{V}$, $\exists \mathbf{0} \in \mathcal{V}$, 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
$\forall \mathbf{u} \in \mathcal{V}$, $\exists 1 \in \mathbb{F}$, s.t. $1\mathbf{u} = \mathbf{u}$;

(VSA-7) distributive laws

$$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V},\ \forall a, b \in \mathbb{F},\ \left\{ \begin{array}{l} (a + b)\mathbf{u} = a\mathbf{u} + b\mathbf{u}, \\ a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}. \end{array} \right.$$

The elements of $\mathcal{V}$ are called *vectors* and the elements of $\mathbb{F}$ are called *scalars*.

Definition 0.70. A vector space with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ is called a *real vector space* or a *complex vector space*, respectively.

Example 0.44. 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.

Definition 0.71. 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 0.72. 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 0.45. 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 0.46. The set of continuous real-valued functions on the interval $[a, b]$ forms a real vector space.

Notation 3. 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 constant function.

Definition 0.73. 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 0.47. $(17, -4, 2)$ is a linear combination of $(2, 1, -3), (1, -2, 4)$ because

$$(17, -4, 2) = 6(2, 1, -3) + 5(1, -2, 4).$$

Example 0.48. $(17, -4, 5)$ 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 0.74. The *span* of a list of vectors $(\mathbf{v}_i)$ is the set of all linear combinations of $(\mathbf{v}_i)$,

$$\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) = \left\{ \sum_{i=1}^{m} a_i \mathbf{v}_i :\ a_i \in \mathbb{F} \right\}. \quad (0.57)$$

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 0.49.

$$\begin{aligned} &(17, -4, 2) \in \operatorname{span}((2, 1, -3), (1, -2, 4)) \\ &(17, -4, 5) \notin \operatorname{span}((2, 1, -3), (1, -2, 4)) \end{aligned}$$

Definition 0.75. A vector space $\mathcal{V}$ is called *finite dimensional* if some list of vectors span $\mathcal{V}$; otherwise it is *infinite dimensional*.

Example 0.50. Let $\mathbb{P}_m(\mathbb{F})$ denote the set of all polynomials with coefficients in $\mathbb{F}$ and degree at most m,

$$\mathbb{P}_m(\mathbb{F}) = \left\{ p : \mathbb{F} \to \mathbb{F};\ p(z) = \sum_{i=0}^{m} a_i z^i, a_i \in \mathbb{F} \right\}. \quad (0.58)$$

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 0.76. A list of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ in $\mathcal{V}$ is called *linearly independent* iff

$$a_1\mathbf{v}_1 + \ldots + a_m\mathbf{v}_m = \mathbf{0} \ \Rightarrow\ a_1 = \cdots = a_m = 0. \quad (0.59)$$

Otherwise the list of vectors is called *linearly dependent*.

Example 0.51. 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 0.52. The list $(1, z, \ldots, z^m)$ is linearly independent in $\mathbb{P}_m(\mathbb{F})$ for each $m \in \mathbb{N}$.

Example 0.53. $(2,3,1)$, $(1,-1,2)$, 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 0.54. Every list of vectors containing the $\mathbf{0}$ vector is linearly dependent.

Lemma 0.77 (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 0.78. 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.

Definition 0.79. 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 0.80. The *standard basis* of $\mathbb{F}^n$ is the list of vectors

$$(1,0,\cdots,0)^T,\ (0,1,0,\cdots,0)^T,\ \ldots,\ (0,\cdots,0,1)^T. \quad (0.60)$$

Example 0.55. $(z^0, z^1, \ldots, z^m)$ is a basis of $\mathbb{P}_m(\mathbb{F})$ in (0.58).

Lemma 0.81. 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

$$\mathbf{u} = \sum_{i=1}^{n} a_i \mathbf{v}_i, \quad (0.61)$$

where $a_i \in \mathbb{F}$.

Lemma 0.82. Every spanning list in a vector space $\mathcal{V}$ can be reduced to a basis of $\mathcal{V}$.

Lemma 0.83. Every linearly independent list of vectors in a finite-dimensional vector space can be extended to a basis of that vector space.

Definition 0.84. 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 0.85. 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 0.86. 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 $\mathcal{V}$.

0.3.2 Inner product spaces

Definition 0.87. Let $\mathbb{F}$ be 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 $\mathcal{V}$.

Exercise 0.56. Deduce *additivity in the second slot* and *conjugate homogeneity in the second slot* from Definition 0.87.

Definition 0.88. The *Euclidean inner product* on $\mathbb{F}^n$ is

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{n} v_i \overline{w_i}. \quad (0.62)$$

Definition 0.89. 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 $\mathcal{V} \to \mathbb{F}$:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \quad (0.63)$$

Definition 0.90. The *Euclidean ℓ_p norm* of a vector $\mathbf{v} \in \mathbb{F}^n$ is

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \quad (0.64)$$

and the *Euclidean ℓ_∞ norm* is

$$\|\mathbf{v}\|_\infty = \max_i |v_i|. \quad (0.65)$$

Theorem 0.91 (Equivalence of norms)**.** Any two norms $\|\cdot\|_N$ and $\|\cdot\|_M$ on a finite dimensional vector space $\mathcal{V} = \mathbb{C}^n$ satisfy

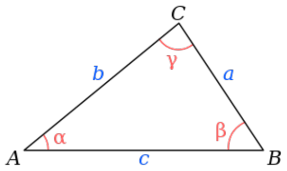
$$\exists c_1, c_2 \in \mathbb{R}^+, \text{ s.t. } \forall \mathbf{x} \in \mathcal{V},\ c_1\|\mathbf{x}\|_M \leq \|\mathbf{x}\|_N \leq c_2\|\mathbf{x}\|_M. \quad (0.66)$$

Definition 0.92. The angle between two vectors $\mathbf{v}, \mathbf{w}$ in an inner product space with $\mathbb{F} = \mathbb{R}$ is the number $\theta \in [0, \pi]$,

$$\theta = \arccos \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\|\|\mathbf{w}\|}. \quad (0.67)$$

Theorem 0.93 (The law of cosines)**.** Any triangle satisfies

$$c^2 = a^2 + b^2 - 2ab\cos\gamma. \quad (0.68)$$



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 (0.67). □

Theorem 0.94 (The law of cosines: abstract version)**.** Any induced norm on a real vector space satisfies

$$\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\langle \mathbf{u}, \mathbf{v}\rangle . \tag{0.69}$$

Proof. Definitions 0.89 and 0.87 and $\mathbb{F} = \mathbb{R}$ yield

$$\begin{aligned}\|\mathbf{u}-\mathbf{v}\|^2 &= \langle \mathbf{u}-\mathbf{v}, \mathbf{u}-\mathbf{v}\rangle\\ &= \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle - \langle \mathbf{v}, \mathbf{u}\rangle\\ &= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\langle \mathbf{u}, \mathbf{v}\rangle .\end{aligned}$$ □

Definition 0.95. 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 0.57. 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 0.96 (Pythagorean)**.** If $\mathbf{u}, \mathbf{v}$ are orthogonal, then $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

Proof. This follows from (0.69) and Definition 0.95. □

Theorem 0.97 (Cauchy-Schwarz inequality)**.**

$$|\langle \mathbf{u}, \mathbf{v}\rangle| \le \|\mathbf{u}\|\|\mathbf{v}\|, \tag{0.70}$$

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

$$\begin{aligned}&\langle \mathbf{u}+\lambda\mathbf{v}, \mathbf{u}+\lambda\mathbf{v}\rangle \ge 0\\ \Rightarrow &\langle \mathbf{u}, \mathbf{u}\rangle + \lambda\langle \mathbf{v}, \mathbf{u}\rangle + \bar{\lambda}\langle \mathbf{u}, \mathbf{v}\rangle + \lambda\bar{\lambda}\langle \mathbf{v}, \mathbf{v}\rangle \ge 0.\end{aligned}$$

If $\mathbf{v} = 0$, (0.70) clearly holds. Otherwise (0.70) follows from substituting $\lambda = -\frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\langle \mathbf{v}, \mathbf{v}\rangle}$ into the above equation. □

Exercise 0.58. To explain the choice of λ in the proof of Theorem 0.97, what is the geometric meaning of (0.70) in the plane? When will the equality hold?

Example 0.59. If $x_i, y_i \in \mathbb{R}$, then for any $n \in \mathbb{N}^+$

$$\left|\sum_{i=1}^n x_i y_i\right|^2 \le \sum_{j=1}^n x_j^2 \sum_{k=1}^n y_k^2.$$

Example 0.60. If $f, g : [a, b] \to \mathbb{R}$ are continuous, then

$$\left|\int_a^b f(x)g(x)\mathrm{d}x\right|^2 \le \left(\int_a^b f^2(x)\mathrm{d}x\right)\left(\int_a^b g^2(x)\mathrm{d}x\right)$$

0.3.3 Normed vector spaces

Definition 0.98. A function $\|\cdot\| : \mathcal{V} \to \mathbb{F}$ is a *norm* for a vector space $\mathcal{V}$ iff it satisfies

(NRM-1) real positivity: $\forall \mathbf{v} \in \mathcal{V}$, $\|\mathbf{v}\| \ge 0$;

(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:
$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|$.

A *normed vector space* or simply a *normed space* is a vector space $\mathcal{V}$ equipped with a norm on $\mathcal{V}$.

Exercise 0.61. Explain how (NRM-1,2,3,4) relate to the geometric meaning of the norm of vectors in $\mathbb{R}^3$.

Lemma 0.99. The norm induced by an inner product is a norm as in Definition 0.98.

Proof. The induced norm as in (0.63) satisfies (NRM-1,2) trivially. For (NRM-3),

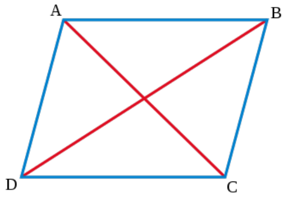
$$\|a\mathbf{v}\|^2 = \langle a\mathbf{v}, a\mathbf{v}\rangle = a\langle \mathbf{v}, a\mathbf{v}\rangle = a\bar{a}\langle \mathbf{v}, \mathbf{v}\rangle = |a|^2\|\mathbf{v}\|^2.$$

To prove (NRM-4), we have

$$\begin{aligned}\|\mathbf{u}+\mathbf{v}\|^2 &= \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}\rangle\\ &= \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle + \langle \mathbf{u}, \mathbf{v}\rangle + \overline{\langle \mathbf{u}, \mathbf{v}\rangle}\\ &\le \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle + 2|\langle \mathbf{u}, \mathbf{v}\rangle|\\ &\le \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\|\\ &= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2 ,\end{aligned}$$

where the second step follows from (IP-5) and the fourth step from Cauchy-Schwarz inequality. □

Theorem 0.100 (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

$$(AB)^2+(BC)^2+(CD)^2+(DA)^2 = (AC)^2+(BD)^2. \tag{0.71}$$

Proof. Apply the law of cosines to the two diagonals, add the two equations, and we obtain (0.71). □

Theorem 0.101 (The parallelogram law: abstract version)**.** Any induced norm (0.63) satisfies

$$2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 = \|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2. \tag{0.72}$$

Proof. Replace $\mathbf{v}$ in (0.69) with $-\mathbf{v}$ and we have

$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\langle \mathbf{u}, \mathbf{v}\rangle .$$

(0.72) follows from adding the above equation to (0.69). □

Exercise 0.62. In the case of Euclidean ℓ_p norms, show that the parallelogram law (0.72) holds if and only if $p = 2$.

Theorem 0.102. The induced norm (0.63) holds for some inner product $\langle \cdot, \cdot \rangle$ if and only if the parallelogram law (0.72) holds for every pair of $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

Exercise 0.63. Prove Theorem 0.102.

Example 0.64. By Theorem 0.102 and Exercise 0.62, the ℓ^1 and ℓ^∞ spaces do not have a corresponding inner product for the ℓ_1 and ℓ_∞ norms.

0.4 Abstract algebra

0.4.1 Binary algebraic structures

Definition 0.103. A *binary operation* on a set $\mathcal{S}$ is a map $\mathcal{S} \times \mathcal{S} \to \mathcal{S}$. A *binary algebraic structure* or a *magma* is an ordered pair $(\mathcal{S}, *)$ where $\mathcal{S}$ is a set and $*$ a binary operation on $\mathcal{S}$.

Definition 0.104. Let $(\mathcal{S}, *)$ and $(\mathcal{S}', *')$ be two binary algebraic structures. A *homomorphism* between $\mathcal{S}$ and $\mathcal{S}'$ is a map $\phi : \mathcal{S} \to \mathcal{S}'$ satisfying

$$\forall a, b \in \mathcal{S}, \qquad \phi(a * b) = \phi(a) *' \phi(b). \tag{0.73}$$

Definition 0.105 (Type of homomorphisms)**.** A *monomorphism* is an injective homomorphism, an *epimorphism* is a surjective homomorphism, An *endomorphism* is a homomorphism $\phi : \mathcal{S} \to \mathcal{S}$. An *isomorphism* is a bijective homomorphism. If such an isomorphism exists between $\mathcal{S}$ and $\mathcal{S}'$, they are said to be *isomorphic*, written $\mathcal{S} \simeq \mathcal{S}'$. An *automorphism* is an isomorphism $\phi : \mathcal{S} \to \mathcal{S}$.

Exercise 0.65. $(\mathbb{R}, +)$ is isomorphic to $(\mathbb{R}^+, \times)$.

Definition 0.106. An element e of a binary structure $(\mathcal{S}, *)$ is an *identity element for* $*$ iff

$$\forall s \in \mathcal{S}, \ e * s = s * e = s. \tag{0.74}$$

Theorem 0.107 (Uniqueness of the identity element)**.** A binary structure has at most one identity element.

Proof. Suppose there are two identity elements e and e'. Then (0.74) implies that they are equal. □

0.4.2 Groups

Definition 0.108. A *group* is an ordered pair $\langle \mathcal{G}, * \rangle$ where $\mathcal{G}$ is a set and '$*$' is a binary operation on $\mathcal{G}$ satisfying the following axioms:

(GRP-1) associativity
$\forall a, b, c \in \mathcal{G}, \quad (a * b) * c = a * (b * c);$

(GRP-2) identity element
$\exists e \in \mathcal{G}, \text{ s.t. } \forall x \in \mathcal{G}, e * x = x * e = x;$

(GRP-3) inverse element
$\forall a \in \mathcal{G}, \exists a' \in \mathcal{G}. \text{ s.t. } a' * a = a * a' = e;$

$\mathcal{G}$ is *abelian* if $*$ is commutative.

Example 0.66. An example of the "is-a" relation is: Abelian group → group → binary algebraic structure.

Exercise 0.67. Find out the definitions of *semigroup*, *monoid*, and *groupoid*; what are their relations to the concepts of magma and group?

Definition 0.109. The *order of a group* G, written $|G|$, is the number of elements in G.

Example 0.68. Each of $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ is an abelian group under addition, and is not under multiplication. With 0 deleted, each resulting set is a group under multiplication.

Example 0.69. $(\mathbb{R}^{m\times n}, +)$ is a group while $(\mathbb{R}^{m\times n}, \times)$ is not. The set of invertible matrices $\mathbb{R}^{n\times n}$ is a group.

Theorem 0.110 (Cancellation laws)**.** For a group $(\mathcal{G}, *)$, the *left and right cancellation laws* hold in $\mathcal{G}$,

$$\forall a, b, c \in \mathcal{G}, \qquad \begin{cases} a * b = a * c \Rightarrow b = c, \\ b * a = c * a \Rightarrow b = c. \end{cases} \tag{0.75}$$

Proof. This follows from multiplying the first equation by a' and applying the associative law and identity element. □

Theorem 0.111. For a group $(\mathcal{G}, *)$, the linear equations $a * x = b$ and $y * a = b$ have unique solutions if $a, b \in \mathcal{G}$.

Proof. The existence follows from multiplying the first equation by a', the associative law, and the identity element. The uniqueness follows from Theorem 0.110. □

Example 0.70. In the case of solving a linear system $A\mathbf{x} = \mathbf{b}$, we know it has a unique solution so long as A belongs to the group of invertible matrices.

Corollary 0.112. The identity element of a group $(\mathcal{G}, *)$ is unique.

Proof. A group *is a* binary algebraic structure and the rest of the proof follows from Theorem 0.107. □

Corollary 0.113. For each element a in a group $\mathcal{G}$, there is only one element $a' \in \mathcal{G}$ such that $a' * a = a * a' = e$.

Proof. This follows from Theorem 0.110. □

Corollary 0.114. Let a' denote the inverse of a in a group $\mathcal{G}$. Then

$$\forall a, b \in \mathcal{G}, \qquad (a * b)' = b' * a'. \tag{0.76}$$

Proof. $(b' * a') * (a * b) = e$. The uniqueness is guaranteed by Corollary 0.113. □

0.4.3 Subgroups and generating sets

Definition 0.115. If a subset $H \subseteq G$ of a group G is closed under the binary operation and H itself forms a group, then H is a *subgroup* of G, written $H \leq G$.

Definition 0.116. The *improper subgroup* of a group G is the subgroup G; all other subgroups of G are *proper subgroups* of G. The subgroup $\{e\}$ is the *trivial subgroup* of G; all other subgroups are *nontrivial subgroups*.

Definition 0.117. Let G be a group and $a \in G$. The *cyclic subgroup of G generated by a* is the subgroup

$$\langle a \rangle = \{a^n \in G : n \in \mathbb{Z}\}. \tag{0.77}$$

Theorem 0.118. The cyclic subgroup of G generated by a is the smallest subgroup of G that contains a.

Definition 0.119. An element a of a group G *generates* G and is a *generator for G* if $\langle a \rangle = G$.

Definition 0.120. The *intersection of the sets S_i* is the set of all elements that are all in the sets S_i,

$$\cap_{i\in I} S_i = \{x : \forall i \in I, \quad x \in S_i\}. \tag{0.78}$$

Theorem 0.121. The intersection of some subgroups H_i of a group G for $i \in I$ is again a subgroup of G.

Definition 0.122. Let G be a group and let $a_i \in G$ for $i \in I$. The smallest subgroup of G containing $\{a_i : i \in I\}$ is the *subgroup generated by* $\{a_i : i \in I\}$. If this subgroup is all of G, then $\{a_i : i \in I\}$ *generates* G and the a_i's are *generators of* G. If there is a finite set $\{a_i : i \in I\}$ that generates G, then G is *finitely generated.*

0.4.4 Permutations and symmetric groups

Definition 0.123. A *permutation of a set A* is a bijective function $\sigma : A \to A$.

Notation 4. A permutation on $\{1, 2, \ldots, n\}$ can be denoted by *Cauchy's two-line notation*,

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ \downarrow & \downarrow & \cdots & \downarrow \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix} \tag{0.79}$$

with arrows often omitted, or *Cauchy's one-line notation*,

$$(\sigma(1)\ \sigma(2)\ \cdots\ \sigma(n)), \tag{0.80}$$

or *cyclic notation*, e.g.,

$$(2, 1, 3)(4, 5) := 2 \mapsto 1 \mapsto 3 \mapsto 2; 4 \mapsto 5 \mapsto 4.$$

The cyclic notation does not display any element that is fixed under the permutation.

Theorem 0.124. Let A be a non-empty set, and let S_A be the collection of all permutations of A. Then $(S_A, \circ, ^{-1}, e)$ is a group where the identity e is the identity function.

Definition 0.125. Let A be the finite set $\{1, 2, \ldots, n\}$. The group of all permutations of A is called the *symmetric group on n letters*, and is denoted by S_n.

Exercise 0.71. Show that S_n has $n!$ elements.

Definition 0.126. The nth *dihedral group D_n* is the group of symmetries of the regular n-gon.

Example 0.72. S_3 is also called the *group D_3 of symmetries of an equilateral triangle.* Label the vertices of an equilateral triangle by $X = \{1, 2, 3\}$. Then the six elements in S_X can be interpreted as three reflections and three rotations around the centroid of $\frac{2i}{3}\pi$ with $i = 1, 2, 3$.

Example 0.73. S_4 is also called the *octic group* or the *group D_4 of symmetries of the square.* Label the vertices of an equilateral triangle by $X = \{1, 2, 3, 4\}$. Then the eight elements in S_X can be interpreted as four rotations, two reflections along the two diagonal directions, and two reflections along the horizontal and vertical axis.

Lemma 0.127. Let G and G' be groups and let $\phi : G \to G'$ be an injective homomorphism. then $\phi(G)$, the image of G, is a subgroup of G' and ϕ is an isomorphism.

Exercise 0.74. Prove Lemma 0.127.

Theorem 0.128 (Cayley)**.** Every group G is isomorphic to a subgroup of S_G.

Proof. For $x \in G$, define $\lambda_x : G \to G$ as $\lambda_x(g) = xg$ for all $g \in G$. λ_x is surjective because

$$\forall c \in G, \exists x^{-1}c \in G, \text{ s.t. } \lambda_x(x^{-1}c) = c.$$

λ_x is injective because

$$\lambda_x(a) = \lambda_x(b) \;\Rightarrow\; xa = xb \;\Rightarrow\; a = b.$$

Therefore λ_x is a permutation of G.

Define $\phi : G \to S_G$ by $\phi(x) = \lambda_x$ for all $x \in G$. ϕ is injective because

$$\phi(x) = \phi(y) \;\Rightarrow\; \lambda_x(e) = \lambda_y(e) \;\Rightarrow\; x = y.$$

ϕ is a homomorphism because

$$\begin{aligned} \forall g \in G, \phi(xy)(g) = \lambda_{xy}(g) = (xy)g &= \lambda_x(\lambda_y(g)) \\ &= (\lambda_x\lambda_y)(g). \end{aligned}$$

The proof is completed by Lemma 0.127. □

0.4.5 Group action on a set

Definition 0.129. An *action of a group G on a set X* is a map $* : G \times X \to X$ such that

(1) $\forall x \in X$, $ex = x$,

(2) $\forall x \in X$, $\forall g_1, g_2 \in G$, $(g_1g_2)(x) = g_1(g_2(x))$.

X is called a *G-set* if G has an action on X.

Example 0.75. The set $X = \{1, 2, \ldots, n\}$ in Examples 0.72 and 0.73 is a S_X-set since the action of S_X on X can be defined as $\cdot(\sigma, x) = \sigma x$.

Theorem 0.130. Let X be a G-set. For each $g \in G$, the function $\sigma_g : X \to X$ defined by $\sigma_g(x) = gx$ is a permutation of X. Also, the map $\phi : G \to S_G$ defined by $\phi(g) = \sigma_g$ is a homomorphism.

Proof. The proof is similar to that of Theorem 0.128. □

Exercise 0.76. Let H be a subgroup of G. Show that G is an H-set under conjugation with

$$\forall g \in G, \forall h \in H, \ \cdot(h, g) = hgh^{-1}.$$

0.4.6 Orbits and alternating groups

Definition 0.131. For a permutation $\sigma : A \to A$, define an equivalence relation by

$$\forall a, b \in A, \quad a \sim b \ \Leftrightarrow \ \exists n \in \mathbb{Z} \text{ s.t. } b = \sigma^n(a). \tag{0.81}$$

The equivalence classes in A determined by (0.81) are called the *orbits of the permutation* σ.

Exercise 0.77. Show that (0.81) is indeed an equivalence relation.

Example 0.78. The orbits of the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 8 & 6 & 7 & 4 & 1 & 5 & 2 \end{pmatrix}$$

are $\{1, 3, 6\}$, $\{2, 8\}$, and $\{4, 5, 7\}$.

Definition 0.132. A permutation $\sigma \in S_n$ is a *cycle* if at most one of its orbits contains more than one element. Two cycles are *disjoint* if any integer is moved by at most one of these cycles.

Example 0.79. Cyclic notations now make perfect sense.

$$(1, 3, 5, 4) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 5 & 1 & 4 \end{pmatrix}$$

Clearly $(1, 3, 5, 4) = (3, 5, 4, 1) = (5, 4, 1, 3) = (4, 1, 3, 5)$.

Theorem 0.133. Every permutation σ of a finite set is a product of disjoint cycles.

Proof. Let $B_1, B_2, \ldots, B_r$ be the orbits of σ, and define corresponding cycles as

$$\mu_i(x) = \begin{cases} \sigma(x) & \text{if } x \in B_i; \\ x & \text{otherwise.} \end{cases} \tag{0.82}$$

Clearly $\sigma = \mu_1 \mu_2 \cdots \mu_r$. Because the orbits are pairwise disjoint, so are the cycles. □

Example 0.80. The multiplication of disjoint cycles is commutative.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 2 & 4 & 3 & 1 \end{pmatrix} = (1, 6)(2, 5, 3) = (2, 5, 3)(1, 6).$$

Example 0.81. If two cycles are not disjoint, their multiplication is not commutative; in fact, their multiplication might not even be a cycle:

$$(1, 4, 5, 6)(2, 1, 5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 4 & 3 & 5 & 2 & 1 \end{pmatrix},$$

$$(2, 1, 5)(1, 4, 5, 6) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 1 & 3 & 2 & 6 & 5 \end{pmatrix}.$$

Definition 0.134. A *transposition* is a cycle of length 2.

Example 0.82. The permutation (2 1 3) in Cauchy's one-line notation is a transposition $(1, 2)$.

Exercise 0.83. Consider the set of standard basis vectors for $\mathbb{R}^n$ as in Definition 0.80,

$$E := \{e_1, e_2, \ldots, e_n\}. \tag{0.83}$$

Show that

(1) Every set X of n elements is isomorphic to E.

(2) Every permutation $\sigma : E \to E$ is a matrix $\mathbb{Z}_2^{n\times n}$ where there is exactly one 1 in each column and each row.

(3) In particular, a transposition $\tau_{i,j}$ is an elementary matrix A of type II, i.e., A is the identity matrix except $a_{i,i} = a_{j,j} = 0$ and $a_{i,j} = a_{j,i} = 1$.

Lemma 0.135. Any permutation of a finite set X of at least two elements is a product of transpositions. In other words, the set of transpositions generates the symmetric group.

Proof. It is easily verified that

$$(a_1, a_2, \ldots, a_n) = (a_1, a_n)(a_1, a_{n-1}) \cdots (a_1, a_2)$$

and the rest of the proof follows from Theorem 0.133. □

Example 0.84. In S_n for $n \geq 2$, the identity permutation is the product $(1, 2)(1, 2)$ of transpositions.

Lemma 0.136. For a permutation $\sigma \in S_n$ and a transposition $\tau = (i, j)$ in S_n, the numbers of orbits of σ and of $\tau\sigma$ differ by 1.

Proof. Suppose i and j are in the same orbit of σ. By Theorem 0.133 we can write σ as a product of disjoint cycles with the first cycle of the form

$$(a, i, c, \ldots, b, j, d, \ldots).$$

Then we have

$$\tau\sigma = (i, j)(a, i, c, \ldots, b, j, d, \ldots) = (a, j, d, \ldots)(b, i, c, \ldots).$$

Suppose i and j are in different orbits of σ. WLOG, we write the first two cycles as

$$(b, j, d, \cdots)(a, i, c, \ldots).$$

Then we have

$$\tau\sigma = (i, j)(b, j, d, \cdots)(a, i, c, \ldots) = (a, j, d, \ldots, b, i, c, \ldots).$$

The statement is proved for other cases similarly. □

Theorem 0.137. No permutations in S_n can be expressed both as a product of an even number of transpositions and as a product of an odd number of transpositions.

Proof. By Lemma 0.135, any $\sigma \in S_n$ can be expressed as

$$\sigma = \tau_1 \tau_2 \cdots \tau_m I,$$

where the identity I has n orbits. Then Lemma 0.136 completes the proof. □

Definition 0.138. The *signature of a permutation* σ, denoted by $\text{sgn}(\sigma)$, is $+1$ or -1 if σ can be expressed as an even or odd number of transpositions, respectively; we also say that the permutation is an *even permutation* or an *odd permutation*, respectively.

Example 0.85. The identity in S_n is an even permutation.

$$\sigma = (1,4,5,6)(2,1,5) = (1,6)(1,5)(1,4)(2,5)(2,1)$$

is an odd permutation.

Theorem 0.139. If $n \geq 2$, then the collection of all even permutations of $\{1,2,\ldots,n\}$ forms a subgroup of order $n!/2$ of the symmetric group S_n.

Exercise 0.86. Prove Theorem 0.139.

Definition 0.140. The *alternating group* A_n *on* n *letters* is the subgroup of S_n consisting of all even permutations of n letters.

0.4.7 Determinants

Definition 0.141. 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., $\forall j = 1, \ldots, n,\ \forall c \in \mathbb{R}$,

$$\begin{aligned} &\delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v} + c\mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n) \\ &= \delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n) \\ &+ c\delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n). \end{aligned} \tag{0.84}$$

Lemma 0.142. Adding a multiple of one vector to another does not change the determinant.

Proof. This follows directly from (SVP-2,3). □

Lemma 0.143. 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 0.144. The signed volume δ is alternating, i.e.,

$$\delta(\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_n) = -\delta(\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n). \tag{0.85}$$

Exercise 0.87. Prove Lemma 0.144 using (SVP-2,3).

Lemma 0.145. Let M_σ denote the matrix of a permutation $\sigma : E \to E$ where E is in (0.83). Then we have $\delta(M_\sigma) = \text{sgn}(\sigma)$.

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 0.144, each transposition yields a multiplication factor -1. Definition 0.138 and (SVP-1) give $\delta(M_\sigma) = \text{sgn}(\sigma)\delta(I) = \text{sgn}(\sigma)$. □

Definition 0.146 (Leibniz formula of determinants)**.** The *determinant of a square matrix* $A \in \mathbb{R}^{n\times n}$ is

$$\det A = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n a_{\sigma(i),i}, \tag{0.86}$$

where the sum is over all permutations σ in the symmetric group and $a_{\sigma(i),i}$ is the element of A at the $\sigma(i)$th row and the ith column.

Exercise 0.88. Show that the determinant formula in (0.86) reduces to

$$\det \begin{bmatrix} a & c \\ b & d \end{bmatrix} = ad - bc \tag{0.87}$$

for $n = 2$. Give a geometric proof that $ad - bc$ is the signed volume of the parallelogram determined by the vectors $(a,b)^T$ and $(c,d)^T$ on the plane.

Theorem 0.147. The signed volume function satisfying (SVP-1,2,3) in Definition 0.141 is unique and is the same as the determinant in (0.86).

Proof. Let the parallelotope be spanned by the column vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. We have

$$\begin{aligned}
&\delta \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{bmatrix} \\
&= \sum_{i_1=1}^{n} v_{i_1 1} \delta \begin{bmatrix} | & v_{12} & \cdots & v_{1n} \\ e_{i_1} & v_{22} & \cdots & v_{2n} \\ | & \vdots & \ddots & \vdots \\ | & v_{n2} & \cdots & v_{nn} \end{bmatrix} \\
&= \sum_{i_1, i_2=1}^{n} v_{i_1 1} v_{i_2 2} \delta \begin{bmatrix} | & | & v_{13} & \cdots & v_{1n} \\ e_{i_1} & e_{i_2} & v_{23} & \cdots & v_{2n} \\ | & | & \vdots & \ddots & \vdots \\ | & | & v_{n2} & \cdots & v_{nn} \end{bmatrix} \\
&= \cdots \\
&= \sum_{i_1, i_2, \ldots, i_n=1}^{n} v_{i_1 1} v_{i_2 2} \cdots v_{i_n n} \delta \begin{bmatrix} | & | & \cdots & | \\ e_{i_1} & e_{i_2} & \cdots & e_{i_n} \\ | & | & \cdots & | \end{bmatrix} \\
&= \sum_{\sigma \in S_n} v_{\sigma(1),1} v_{\sigma(2),2} \cdots v_{\sigma(n),n} \delta \begin{bmatrix} | & | & \cdots & | \\ e_{\sigma(1)} & e_{\sigma(2)} & \cdots & e_{\sigma(n)} \\ | & | & \cdots & | \end{bmatrix} \\
&= \sum_{\sigma \in S_n} v_{\sigma(1),1} v_{\sigma(2),2} \cdots v_{\sigma(n),n} \mathrm{sgn}(\sigma) \\
&= \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} v_{\sigma(i),i},
\end{aligned}$$

where the first four steps follow from (SVP-3), the sixth step from Lemma 0.145, 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 0.89. Use the formula in (0.86) to show that $\det A = \det A^T$.

Definition 0.148. The i, j *cofactor* of $A \in \mathbb{R}^{n\times n}$ is

$$C_{ij} = (-1)^{i+j} M_{ij}, \tag{0.88}$$

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 0.149 (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

$$\det A = \sum_{j'=1}^{n} a_{ij'} C_{ij'} = \sum_{i'=1}^{n} a_{i'j} C_{i'j}. \tag{0.89}$$

Exercise 0.90. Prove Theorem 0.149 by induction.

Chapter 1

Computer Arithmetic

1.1 Floating-point number systems

Definition 1.1. A *bit* is the basic unit of information in computing; it can have only one of two values 0 and 1.

Definition 1.2. 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 1.3. 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 1.1. 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 1.4 (Floating point numbers)**.** A *floating point number* (FPN) is a number of the form

$$x = \pm m \times \beta^e, \tag{1.1}$$

where $e \in [L, U]$ and the *significand* (or *mantissa*) m has the form

$$m = \left(d_0 + \frac{d_1}{\beta} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right), \tag{1.2}$$

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 portion $.d_1 d_2 \cdots d_{p-1}$ is called the *fraction*.

Algorithm 1.5. 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 greater 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 1.2. Convert 156 to binary number:

$$156 = (10011100)_2.$$

Example 1.3. What is the normalized binary form of $\frac{2}{3}$?

$$\begin{aligned} \frac{2}{3} &= (0.a_1 a_2 a_3 \cdots)_2 = (0.1010 \cdots)_2 \\ &= (1.0101010 \cdots)_2 \times 2^{-1}. \end{aligned}$$

Definition 1.6 (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* $[L, U]$.

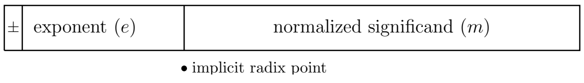
Definition 1.7. An FPN is *normalized* if its mantissa satisfies $1 \le m < \beta$.

Definition 1.8 (IEEE standard 754-2008)**.** The *single precision and double precision* FPNs of current IEEE (Institute of Electrical and Electronics Engineers) standard 754 are normalized FPN systems with the respective characterizations,

$$\beta = 2,\ p = 23 + 1,\ e \in [-126, 127], \tag{1.3a}$$

$$\beta = 2,\ p = 52 + 1,\ e \in [-1022, 1023]. \tag{1.3b}$$

Example 1.4. IEEE 754 has some further details.



(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 in the 2008 standard, $\pm m \times \beta^{-127}$ is reserved for ± 0 and $\pm m \times \beta^{128}$ is reserved for any number with a magnitude too large to be representable by the FPN system.

Definition 1.9. The *machine precision* of a normalized FPN system $\mathcal{F}$ is the distance between 1.0 and the next larger FPN in $\mathcal{F}$,

$$\epsilon_M := \beta^{1-p}. \tag{1.4}$$

Definition 1.10. The underflow limit (UFL) and the overflow limit (OFL) of a normalized FPN system $\mathcal{F}$ are respectively

$$\mathrm{UFL}(\mathcal{F}) := \min |\mathcal{F} \setminus \{0\}| = \beta^L, \tag{1.5}$$

$$\mathrm{OFL}(\mathcal{F}) := \max |\mathcal{F}| = \beta^U (\beta - \beta^{1-p}). \tag{1.6}$$

Example 1.5. 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 $\mathrm{OFL}(\mathcal{F})$

$$\max |\mathcal{F}| = \beta^U (\beta - \beta^{1-p}) \approx 1.80 \times 10^{308}.$$

Corollary 1.11 (Cardinality of $\mathcal{F}$)**.** For a normalized binary FPN system $\mathcal{F}$,

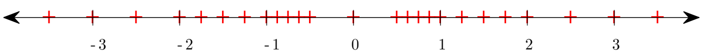
$$\#\mathcal{F} = 2^p (U - L + 1) + 1. \tag{1.7}$$

Proof. The cardinality can be proved by Axiom 0.11. The factor 2^p comes from the sign bit and the mantissa. By Example 1.4, $U - L + 1$ is the number of exponents represented in $\mathcal{F}$. The trailing "+1" in (1.7) accounts for the number 0. □

Definition 1.12. The *range* of a normalized FPN system is a subset of $\mathbb{R}$,

$$\mathcal{R}(\mathcal{F}) := \{x : x \in \mathbb{R}, \mathrm{UFL}(\mathcal{F}) \le |x| \le \mathrm{OFL}(\mathcal{F})\}. \tag{1.8}$$

Example 1.6. 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.

Definition 1.13. Two normalized FPNs a, b are *adjacent* to each other in $\mathcal{F}$ iff

$$\forall c \in \mathcal{F} \setminus \{a, b\}, \quad |a - b| < |a - c| + |c - b|. \tag{1.9}$$

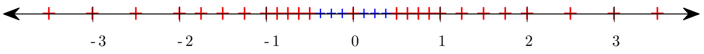
Lemma 1.14. Let a, b be two adjacent normalized FPNs satisfying $|a| < |b|$ and $ab > 0$. Then

$$\beta^{-1} \epsilon_M |a| < |a - b| \le \epsilon_M |a|. \tag{1.10}$$

Proof. Consider $a > 0$, then $\Delta a := b - a > 0$. By Definitions 1.4 and 1.7, $a = m \times \beta^e$ 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. □

Definition 1.15. The *subnormal* or *denormalized* numbers are FPNs of the form (1.1) with $e = L$ and $m \in (0, 1)$. A normalized FPN system can be *extended* by including the subnormal numbers.

Example 1.7. Add subnormal FPNs to the FPN system in Example 1.6 and we have the following plot.



1.2 Rounding error analysis

1.2.1 Rounding a single number

Definition 1.16 (Rounding)**.** *Rounding* is a map $\mathrm{fl} : \mathbb{R} \to \mathcal{F} \cup \{\mathrm{NaN}\}$. 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 1.17. 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 1.18. The *unit roundoff* of $\mathcal{F}$ is the number

$$\epsilon_u := \frac{1}{2}\epsilon_M = \frac{1}{2}\beta^{1-p}. \tag{1.11}$$

Theorem 1.19. For $x \in \mathcal{R}(\mathcal{F})$ as in (1.8), we have

$$\mathrm{fl}(x) = x(1 + \delta), \qquad |\delta| < \epsilon_u. \tag{1.12}$$

Proof. By Definition 0.18, $\mathcal{R}(\mathcal{F})$ 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 (1.12) clearly holds. Otherwise $x_L < x < x_R$. Then Lemma 1.14 and Definitions 1.13 and 1.16 yield

$$|\mathrm{fl}(x)-x| \le \frac{1}{2}|x_R-x_L| \le \epsilon_u \min(|x_L|,|x_R|) < \epsilon_u|x|. \quad (1.13)$$

Hence $-\epsilon_u|x| < \mathrm{fl}(x) - x < \epsilon_u|x|$, which yields (1.12). □

Theorem 1.20. For $x \in \mathcal{R}(\mathcal{F})$, we have

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \qquad |\delta| \le \epsilon_u. \quad (1.14)$$

Proof. The proof is the same as that of Theorem 1.19, except that we replace the last inequality "$< \epsilon_u|x|$" in (1.13) by "$\le \epsilon_u|\mathrm{fl}(x)|$." Consequently, the equality in (1.14) holds when $x = \frac{1}{2}(x_L + x_R)$ and $\mathrm{fl}(x) = x_L$ has $m = 1.0$. □

Example 1.8. 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 1.3, we have

$$\begin{aligned}\frac{2}{3} &= (0.1010\cdots)_2 = (1.0101010\cdots)_2 \times 2^{-1}.\\ x_L &= (1.010\cdots10)_2 \times 2^{-1};\\ x_R &= (1.010\cdots11)_2 \times 2^{-1},\end{aligned}$$

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

$$\begin{aligned}x - x_L &= \frac{2}{3} \times 2^{-24};\\ x_R - x_L &= 2^{-24},\\ x_R - x &= (x_R - x_L) - (x - x_L) = \frac{1}{3} \times 2^{-24}.\end{aligned}$$

Thus Definition 1.16 implies $\mathrm{fl}(x) = x_R$.

1.2.2 Binary floating-point operations

Definition 1.21 (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 $|M_c| \ge \beta$, set $M_c \leftarrow M_c/\beta$ and $e_c \leftarrow e_c + 1$.
- If $|M_c| \in (0, 1)$, repeat $M_c \leftarrow M_c\beta$, $e_c \leftarrow e_c - 1$ until $|M_c| \in [1, \beta)$.

(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 $c \leftarrow M_c \times \beta^{e_c}$.

Example 1.9. 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 1.10. 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 1.11. Repeat Example 1.9 with $b = 8.769 \times 10^4$, $b = -5.678 \times 10^0$, and $b = -5.678 \times 10^3$.

Lemma 1.22. For $a, b \in \mathcal{F}$, $a + b \in \mathcal{R}(\mathcal{F})$ implies

$$\mathrm{fl}(a + b) = (a + b)(1 + \delta), \qquad |\delta| < \epsilon_u. \quad (1.15)$$

Proof. The round-off error in step (v) always dominates that in step (ii), which, because of the $2p$ precision, is nonzero only in the case of $e_a - e_b = p + 1$. Then (1.15) follows from Theorem 1.19. □

Definition 1.23 (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 $p + 2$ 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 with rounding to nearest.

(iii) Normalization:
- If $|M_c| \ge \beta$, set $M_c \leftarrow M_c/\beta$ 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 $c \leftarrow M_c \times \beta^{e_c}$.

Example 1.12. Consider the calculation of $c := \text{fl}(ab)$ 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) $M_c \leftarrow 15.9202$.

(iii) $m_c \leftarrow 1.59202$, $e_c \leftarrow 5$.

(iv) do nothing.

(v) $m_c \leftarrow 1.592$.

(vi) $c = 1.592 \times 10^5$.

Lemma 1.24. For $a, b \in \mathcal{F}$, $|ab| \in \mathcal{R}(\mathcal{F})$ implies

$$\text{fl}(ab) = (ab)(1+\delta), \qquad |\delta| < \epsilon_u. \tag{1.16}$$

Proof. The error only comes from the round-off in steps (ii) and (v). Then (1.16) follows from Theorem 1.19. □

Definition 1.25 (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 $c = \text{fl}\left(\frac{a}{b}\right) \in \mathcal{F}$ 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 1.26. For $a, b \in \mathcal{F}$, $\frac{a}{b} \in \mathcal{R}(\mathcal{F})$ implies

$$\text{fl}\left(\frac{a}{b}\right) = \frac{a}{b}(1+\delta), \qquad |\delta| < \epsilon_u. \tag{1.17}$$

Proof. In the case of $|M_a| = |M_b|$, there is no rounding error in Definition 1.25 and (1.17) clearly holds. Hereafter we denote by M_{c1} and M_{c2} the results of steps (ii) and (v) in Definition 1.25, respectively.

In the case of $|M_a| > |M_b|$, the condition $a, b \in \mathcal{F}$, Definition 1.9, and $|M_a|, |M_b| \in [1, \beta)$ imply

$$\left|\frac{M_a}{M_b}\right| \geq \frac{\beta - \epsilon_M}{\beta - 2\epsilon_M} > 1 + \beta^{-1}\epsilon_M, \tag{1.18}$$

which further implies that the normalization step (iii) in Definition 1.25 is not invoked. By Definitions 1.16, 1.9, and 1.18, 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 1.25 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 (1.18) 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 1.25 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

$$|\delta| = \left|\frac{\beta\delta_1 + \delta_2}{\beta M_a/M_b}\right| < \frac{\beta^{-1}\epsilon_u\epsilon_M + \epsilon_u}{1 + \beta^{-1}\epsilon_M} = \epsilon_u.$$ □

Theorem 1.27 (Model of machine arithmetic)**.** Denote by $\mathcal{F}$ a normalized FPN system with precision p. For each arithmetic operation $\odot = +, -, \times, /$, we have

$$\forall a, b \in \mathcal{F},\ a \odot b \in \mathcal{R}(\mathcal{F}) \ \Rightarrow\ \text{fl}(a \odot b) = (a \odot b)(1+\delta) \tag{1.19}$$

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 1.22, 1.24, and 1.26. □

1.2.3 The propagation of rounding errors

Theorem 1.28. If $\forall i = 0, 1, \cdots, n$, $a_i \in \mathcal{F}$, $a_i > 0$, then

$$\text{fl}\left(\sum_{i=0}^{n} a_i\right) = (1 + \delta_n)\sum_{i=0}^{n} a_i, \tag{1.20}$$

where $|\delta_n| < (1 + \epsilon_u)^n - 1 \approx n\epsilon_u$.

Proof. Define $s_k := \sum_{i=0}^{k} a_i$,

$$\begin{cases} s_0 & := a_0; \\ s_{k+1} & := s_k + a_{k+1}, \end{cases} \qquad \begin{cases} s_0^* & := a_0; \\ s_{k+1}^* & := \mathrm{fl}(s_k^* + a_{k+1}), \end{cases}$$

$$\delta_k := \frac{s_k^* - s_k}{s_k}, \qquad \epsilon_k := \frac{s_{k+1}^* - (s_k^* + a_{k+1})}{s_k^* + a_{k+1}},$$

and 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 1.19 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

$$\begin{aligned} \forall k \in \mathbb{N},\ |\delta_{k+1}| &< \epsilon_u \sum_{i=0}^{k} (1+\epsilon_u)^i \qquad (1.21) \\ &= \epsilon_u \frac{(1+\epsilon_u)^{k+1} - 1}{1 + \epsilon_u - 1} = (1+\epsilon_u)^{k+1} - 1, \end{aligned}$$

where the second step follows from the summation formula of geometric series. The proof is completed by the binomial theorem. □

Exercise 1.13. 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 1.28. Can you give some examples?

Exercise 1.14. 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 1.29. For given $\mu \in \mathbb{R}^+$ and a positive integer $n \le \lfloor \frac{\ln 2}{\mu} \rfloor$, suppose $|\delta_i| \le \mu$ for each $i = 1, 2, \ldots, n$. Then

$$1 - n\mu \le \prod_{i=1}^{n}(1+\delta_i) \le 1 + n\mu + (n\mu)^2, \qquad (1.22)$$

or equivalently, for $I_n := [-\frac{1}{1+n\mu}, 1]$,

$$\exists \theta \in I_n \text{ s.t. } \prod_{i=1}^{n}(1+\delta_i) = 1 + \theta(n\mu + n^2\mu^2). \qquad (1.23)$$

Proof. The condition $|\delta_i| \le \mu$ implies

$$(1-\mu)^n \le \prod_{i=1}^{n}(1+\delta_i) \le (1+\mu)^n.$$

Taylor expansion of $f(\mu) = (1-\mu)^n$ at $\mu = 0$ with Lagrangian remainder yields

$$(1-\mu)^n \ge 1 - n\mu,$$

which implies the first inequality in (1.22). On the other hand, the Taylor series of e^x for $x \in \mathbb{R}^+$ satisfies

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ &= 1 + x + \frac{x^2}{2!}\left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right) \\ &\le 1 + x + \frac{x^2}{2}e^x. \end{aligned}$$

Set $x = n\mu$ in the above inequality, apply the condition $n\mu \le \ln 2$, and we have

$$e^{n\mu} \le 1 + n\mu + (n\mu)^2,$$

which, together with the inequality $(1+\mu)^n \le e^{n\mu}$, yields the second inequality in (1.22).

Finally, (1.22) 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. □

1.3 Accuracy and stability

1.3.1 Avoiding catastrophic cancellation

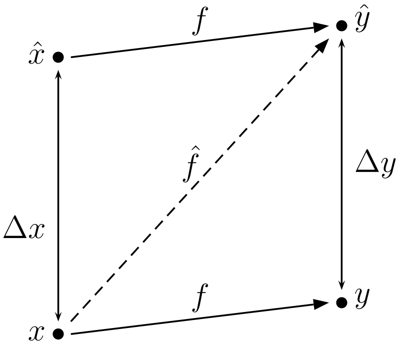
Definition 1.30. Let $\hat{x}$ be an approximation to $x \in \mathbb{R}$. The accuracy of $\hat{x}$ can be measured by its *absolute error*

$$E_{\mathrm{abs}}(\hat{x}) = |\hat{x} - x| \qquad (1.24)$$

and/or its *relative error*

$$E_{\mathrm{rel}}(\hat{x}) = \frac{|\hat{x} - x|}{|x|}. \qquad (1.25)$$

Definition 1.31. 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 1.32 (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. $\forall x \in \mathrm{dom}(f)$, $E_{\mathrm{rel}}(\hat{f}(x)) \le c\epsilon_u$ where c is a small constant.

Example 1.15 (Catastrophic cancellation)**.** For two real numbers $x, y \in \mathcal{R}(\mathcal{F})$, Theorems 1.19 and 1.27 imply

$$\begin{aligned}\mathrm{fl}(\mathrm{fl}(x) \odot \mathrm{fl}(y)) &= (\mathrm{fl}(x) \odot \mathrm{fl}(y))(1+\delta_3)\\ &= (x(1+\delta_1) \odot y(1+\delta_2))(1+\delta_3)\end{aligned}$$

where $|\delta_i| \le \epsilon_u$. From Theorems 1.27 and 1.29, we know that *multiplication is accurate*:

$$\begin{aligned}\mathrm{fl}(\mathrm{fl}(x) \times \mathrm{fl}(y)) &= xy(1+\delta_1)(1+\delta_2)(1+\delta_3)\\ &= xy(1+\theta(3\epsilon_u + 9\epsilon_u^2)),\end{aligned}$$

where $\theta \in [-1, 1]$. Similarly, *division is also accurate*:

$$\begin{aligned}\mathrm{fl}(\mathrm{fl}(x)/\mathrm{fl}(y)) &= \frac{x(1+\delta_1)}{y(1+\delta_2)}(1+\delta_3)\\ =&\frac{x}{y}(1+\delta_1)(1-\delta_2+\delta_2^2-\cdots)(1+\delta_3)\\ \approx&\frac{x}{y}(1+\delta_1)(1-\delta_2)(1+\delta_3).\end{aligned}$$

However, *addition and subtraction might not be accurate*:

$$\begin{aligned}&\mathrm{fl}(\mathrm{fl}(x)+\mathrm{fl}(y)) = (x(1+\delta_1)+y(1+\delta_2))(1+\delta_3)\\ =&(x+y+x\delta_1+y\delta_2)(1+\delta_3)\\ =&(x+y)\left(1+\delta_3+\frac{x\delta_1+y\delta_2}{x+y}+\delta_3\frac{x\delta_1+y\delta_2}{x+y}\right).\end{aligned}$$

In other words, the relative error of addition or subtraction can be arbitrarily large when $x + y \to 0$.

Theorem 1.33 (Loss of most significant digits)**.** Suppose $x, y \in \mathcal{F}$, $x > y > 0$, and

$$\beta^{-t} \le 1 - \frac{y}{x} \le \beta^{-s}. \tag{1.26}$$

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 1.21 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 1.34. Catastrophic cancellation should be avoided whenever possible.

Example 1.16. 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

$$\begin{aligned}x - \sin x &= x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right)\\ &= \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots\end{aligned}$$

1.3.2 Backward stability and numerical stability

Definition 1.35 (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.

$$\begin{aligned}&\forall x \in \mathrm{dom}(f),\ \exists \hat{x} \in \mathrm{dom}(f),\ \text{ s.t.}\\ &\hat{f}(x) = f(\hat{x}) \ \Rightarrow\ E_{\mathrm{rel}}(\hat{x}) \le c\epsilon_u,\end{aligned} \tag{1.27}$$

where c is a small constant.

Definition 1.36. An algorithm $\hat{f}(x_1, x_2)$ for computing $y = f(x_1, x_2)$ is *backward stable* if

$$\begin{aligned}&\forall (x_1, x_2) \in \mathrm{dom}(f),\ \exists (\hat{x}_1, \hat{x}_2) \in \mathrm{dom}(f) \text{ s.t.}\\ &\hat{f}(x_1, x_2) = f(\hat{x}_1, \hat{x}_2) \ \Rightarrow \begin{cases} E_{\mathrm{rel}}(\hat{x}_1) \le c_1\epsilon_u,\\ E_{\mathrm{rel}}(\hat{x}_2) \le c_2\epsilon_u,\end{cases}\end{aligned} \tag{1.28}$$

where c_1, c_2 are two small constants.

Corollary 1.37. For $f(x_1, x_2) = x_1 - x_2$, $x_1, x_2 \in \mathcal{R}(\mathcal{F})$, the algorithm $\hat{f}(x_1, x_2) = \mathrm{fl}(\mathrm{fl}(x_1) - \mathrm{fl}(x_2))$ is backward stable.

Proof. We have $\hat{f}(x_1, x_2) = (\mathrm{fl}(x_1) - \mathrm{fl}(x_2))(1 + \delta_3)$ from Theorem 1.27. Then Theorem 1.19 implies

$$\begin{aligned}\hat{f}(x_1, x_2) &= (x_1(1+\delta_1) - x_2(1+\delta_2))(1+\delta_3)\\ &= x_1(1+\delta_1+\delta_3+\delta_1\delta_3) - x_2(1+\delta_2+\delta_3+\delta_2\delta_3).\end{aligned}$$

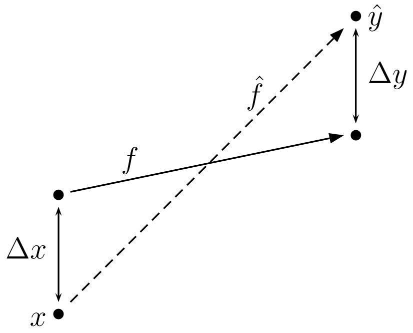
Take $\hat{x}_1$ and $\hat{x}_2$ to be the two terms in the above line and we have

$$\begin{aligned}E_{\mathrm{rel}}(\hat{x}_1) &= |\delta_1 + \delta_3 + \delta_1\delta_3|,\\ E_{\mathrm{rel}}(\hat{x}_2) &= |\delta_2 + \delta_3 + \delta_2\delta_3|.\end{aligned}$$

Then Definition 1.36 completes the proof. □

Example 1.17. For $f(x) = 1 + x$, $x \in (0, \mathrm{OFL})$, show that the algorithm $\hat{f}(x) = \mathrm{fl}(1.0 + \mathrm{fl}(x))$ is not backward stable.

We prove a stronger statement that implies the negation of (1.27). For each $x \in (0, \epsilon_u)$, Definition 1.16 yields $\hat{f}(x) = 1.0$. Then $\hat{f}(x) = f(\hat{x})$ implies $\hat{x} = 0$, which further implies $E_{\mathrm{rel}}(\hat{x}) = 1$.



Definition 1.38. An algorithm $\hat{f}(x)$ for computing $y = f(x)$ is *stable* or *numerically stable* iff

$$\forall x \in \mathrm{dom}(f),\ \exists \hat{x} \in \mathrm{dom}(f) \text{ s.t. } \begin{cases} \left|\frac{\hat{f}(x)-f(\hat{x})}{f(\hat{x})}\right| \le c_f \epsilon_u, \\ E_{\mathrm{rel}}(\hat{x}) \le c\epsilon_u, \end{cases} \tag{1.29}$$

where c_f, c are two small constants.

Corollary 1.39. If an algorithm is backward stable, then it is numerically stable.

Proof. By Definition 1.35, $f(\hat{x}) = \hat{f}(x)$, hence $c_f = 0$. The other condition also follows trivially. □

Example 1.18. For $f(x) = 1 + x$, $x \in (0, \mathrm{OFL})$, show that the algorithm $\hat{f}(x) = \mathrm{fl}(1.0 + \mathrm{fl}(x))$ 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 1.18 and 1.12) yield $x \in \mathcal{R}(\mathcal{F})$. By Theorem 1.19, $\hat{f}(x) = (1 + x(1 + \delta_1))(1 + \delta_2)$, 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

$$\begin{aligned} &E_{\mathrm{rel}}(\hat{x}) = |\delta_1 + \delta_2 + \delta_1\delta_2| < 3\epsilon_u, \\ \Rightarrow &\left|\frac{\hat{f}(x) - f(\hat{x})}{f(\hat{x})}\right| = \left|\frac{\delta_2}{1 + x(1 + \delta_1 + \delta_2 + \delta_1\delta_2)}\right| \le \epsilon_u, \end{aligned}$$

where the denominator is never close to zero since $x > 0$.

1.3.3 Condition numbers: scalar functions

Definition 1.40. 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,

$$C_f(x) = \left|\frac{xf'(x)}{f(x)}\right|. \tag{1.30}$$

Definition 1.41. 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 1.19. Definition 1.40 yields

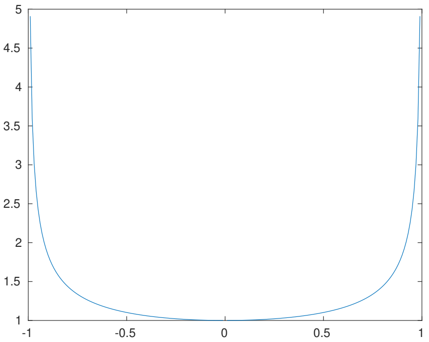
$$E_{\mathrm{rel}}(\hat{y}) \lessapprox C_f E_{\mathrm{rel}}(\hat{x}). \tag{1.31}$$

The approximation mark "≈" refers to the fact that the quadratic term $(\Delta x)^2$ has been ignored. As one way to interpret (1.31) and to understand Definition 1.40, *the computed solution to an ill-conditioned problem may have a large forward error.*

Example 1.20. For the function $f(x) = \arcsin(x)$, its condition number, according to Definition 1.40, 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 $x \to \pm 1$.



Corollary 1.42. 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

$$h \approx -\epsilon\frac{g(r)}{f'(r)}. \tag{1.32}$$

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 1.21 (Wilkinson)**.** Define

$$\begin{aligned} f(x) &:= \prod_{k=1}^{p}(x - k), \\ g(x) &:= x^p. \end{aligned}$$

How is the root $x = p$ affected by perturbing f to $f + \epsilon g$?

By Corollary 1.42, 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.

1.3.4 Condition numbers: vector functions

Definition 1.43. The *condition number of a vector function* $\mathbf{f}:\mathbb{R}^m\to\mathbb{R}^n$ is

$$\text{cond}_{\mathbf{f}}(\mathbf{x}) = \frac{\|\mathbf{x}\|\,\|\nabla\mathbf{f}\|}{\|\mathbf{f}(\mathbf{x})\|}, \tag{1.33}$$

where $\|\cdot\|$ denotes a Euclidean norm such as the 1-, 2-, and ∞-norms.

Example 1.22. 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 1.43 yields

$$\text{cond}_{\mathbf{f}}(\mathbf{x}) = \frac{\|\mathbf{b}\|\,\|A^{-1}\|}{\|\mathbf{u}\|} = \frac{\|A\mathbf{u}\|\,\|A^{-1}\|}{\|\mathbf{u}\|}.$$

In practice the input $\mathbf{b}$ can take any value, hence we have

$$\max\text{cond}_{\mathbf{f}}(\mathbf{x}) = \max\frac{\|A\mathbf{u}\|\,\|A^{-1}\|}{\|\mathbf{u}\|} = \|A\|\,\left\|A^{-1}\right\|,$$

where the last expression is the condition number of A defined in linear algebra and we have used the common definition

$$\|A\| := \max_{\|\mathbf{u}\|\neq 0}\frac{\|A\mathbf{u}\|}{\|\mathbf{u}\|}. \tag{1.34}$$

The above discussion explains why the condition number of a matrix A is usually defined as

$$\text{cond}\,A = \|A\|\|A^{-1}\|. \tag{1.35}$$

Definition 1.44. The *componentwise condition number* of a vector function $\mathbf{f}:\mathbb{R}^m\to\mathbb{R}^n$ is

$$\text{cond}_{\mathbf{f}}(\mathbf{x}) = \|A(\mathbf{x})\|, \tag{1.36}$$

where the matrix $A(\mathbf{x})=[a_{ij}(\mathbf{x})]$ and each component is

$$a_{ij}(\mathbf{x}) = \left|\frac{x_j\frac{\partial f_i}{\partial x_j}}{f_i(\mathbf{x})}\right|. \tag{1.37}$$

Example 1.23. 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

$$\nabla\mathbf{f} = -\frac{1}{x_1^2x_2^2}\begin{bmatrix}x_2^2 & x_1^2\\ x_2^2 & -x_1^2\end{bmatrix}.$$

The condition number based on Definition 1.44 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 1.43 fails to capture the ill-conditioning,

$$C_1 = \frac{\|\mathbf{x}\|_1\|\nabla\mathbf{f}\|_1}{\|\mathbf{f}\|_1} = \frac{|x_1|+|x_2|}{|x_1x_2|}\frac{2\max(x_1^2,x_2^2)}{|x_1+x_2|+|x_1-x_2|},$$

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

$$\forall A\in\mathbb{R}^{n\times n},\qquad \|A\|_1 = \max_j\sum_i|a_{ij}|.$$

Definition 1.45. The *Hilbert matrix* $H_n\in\mathbb{R}^{n\times n}$ is

$$h_{i,j} = \frac{1}{i+j-1}. \tag{1.38}$$

Definition 1.46. The *Vandermonde matrix* $V_n\in\mathbb{R}^{n\times n}$ is

$$v_{i,j} = t_j^{i-1}, \tag{1.39}$$

where $t_1,t_2,\ldots,t_n$ are parameters.

1.3.5 Condition numbers: algorithms

Definition 1.47. 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

$$\forall\mathbf{x}\in\mathcal{F}^m,\ \exists\mathbf{x}_A\in\mathbb{R}^m\text{ s.t. }\mathbf{f}_A(\mathbf{x})=\mathbf{f}(\mathbf{x}_A), \tag{1.40}$$

the *condition number of the algorithm* $\mathbf{f}_A$ is defined as

$$\text{cond}_A(\mathbf{x}) = \frac{1}{\epsilon_u}\inf_{\{\mathbf{x}_A\}}\frac{\|\mathbf{x}_A-\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{1.41}$$

Example 1.24. 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 $|\delta|\leq 5\epsilon_u$. 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

$$\begin{aligned}E_{\text{rel}}(x_A) &= \left|\frac{x^{1+\delta}-x}{x}\right| = |x^\delta-1| = |e^{\delta\ln x}-1|\\ &\approx |\delta\ln x| \leq 5|\ln x|\epsilon_u.\end{aligned}$$

Hence A is well conditioned except when $x\to 0^+$.

Theorem 1.48. 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 $|\delta(x)|\leq\varphi(x)\epsilon_u$. If $\text{cond}_f(x)$ is bounded, then $\forall x\in\mathcal{F}$,

$$\text{cond}_A(x) \leq \frac{\varphi(x)}{\text{cond}_f(x)}. \tag{1.42}$$

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

$$\begin{aligned}f(x)(1+\delta) &= f(x_A) = f(x(1+\epsilon_A)) = f(x+x\epsilon_A)\\ &= f(x)+x\epsilon_Af'(x)+O(\epsilon_A^2).\end{aligned}$$

Neglecting the quadratic term yields

$$\begin{aligned}&x\epsilon_Af'(x) = f(x)\delta\\ \Rightarrow&\left|\frac{x_A-x}{x}\right| = |\epsilon_A| = \left|\frac{f(x)}{xf'(x)}\right||\delta(x)|.\end{aligned}$$

Dividing both sides by ϵ_u yields

$$\frac{1}{\epsilon_u}\left|\frac{x_A-x}{x}\right| = \frac{\delta(x)}{\epsilon_u\text{cond}_f(x)}.$$

Take inf with respect to all x_A's, take sup with respect to x, and we have (1.42). □

Example 1.25. 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 1.48 to analyze the condition of the algorithm

$$f_A = \mathrm{fl}\left[\frac{\mathrm{fl}\big(1-\mathrm{fl}(\cos x)\big)}{\mathrm{fl}(\sin x)}\right] \tag{1.43}$$

that computes $f(x) = \frac{1-\cos x}{\sin x}$ for $x \in (0, \pi/2)$.

By Definition 1.40, it is easy to compute that

$$\mathrm{cond}_f\,(x) = \frac{x}{\sin x}.$$

Furthermore, by Theorem 1.27 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, $\mathrm{cond}_A\,(x) \to +\infty$ as $x \to 0$. On the other hand, $\mathrm{cond}_A\,(x) \to \frac{6}{\pi}$ as $x \to \frac{\pi}{2}$.

Exercise 1.26. Repeat Example 1.25 for $f(x) = \frac{\sin x}{1+\cos x}$ on the same interval.

1.3.6 Overall error of a computer solution

Theorem 1.49. Consider using normalized FPN arithmetics to solve a math problem

$$\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n, \qquad \mathbf{y} = \mathbf{f}(\mathbf{x}). \tag{1.44}$$

Denote the computer input and output as

$$\mathbf{x}^* \approx \mathbf{x}, \qquad \mathbf{y}_A^* = \mathbf{f}_A(\mathbf{x}^*), \tag{1.45}$$

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

$$E_{\mathrm{rel}}(\mathbf{y}_A^*) \lessapprox E_{\mathrm{rel}}(\mathbf{x}^*)\mathrm{cond}_{\mathbf{f}}\,(\mathbf{x}) + \epsilon_u \mathrm{cond}_{\mathbf{f}}\,(\mathbf{x}^*)\,\mathrm{cond}_A\,(\mathbf{x}^*)\,, \tag{1.46}$$

where the relative error is defined in (1.25).

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 (1.31), the first term is

$$\begin{aligned}\frac{\|\mathbf{f}(\mathbf{x}^*)-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|} &\lessapprox \mathrm{cond}_{\mathbf{f}}\,(\mathbf{x})\frac{\|\mathbf{x}^*-\mathbf{x}\|}{\|\mathbf{x}\|}\\ &= E_{\mathrm{rel}}(\mathbf{x}^*)\mathrm{cond}_{\mathbf{f}}\,(\mathbf{x})\,.\end{aligned}$$

By (1.31) and Definition 1.47, the second term is

$$\begin{aligned}\frac{\|\mathbf{f}_A(\mathbf{x}^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x})\|} = \frac{\|\mathbf{f}(\mathbf{x}_A^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x})\|} &\approx \frac{\|\mathbf{f}(\mathbf{x}_A^*)-\mathbf{f}(\mathbf{x}^*)\|}{\|\mathbf{f}(\mathbf{x}^*)\|}\\ &\le \mathrm{cond}_{\mathbf{f}}\,(\mathbf{x}^*)\frac{\|\mathbf{x}_A^*-\mathbf{x}^*\|}{\|\mathbf{x}^*\|}\\ &= \epsilon_u \mathrm{cond}_A\,(\mathbf{x}^*)\,\mathrm{cond}_{\mathbf{f}}\,(\mathbf{x}^*)\,,\end{aligned}$$

where the last step follows from the fact that we only consider the $\mathbf{x}_A^*$ that is the least dangerous. □

1.4 Problems

1.4.1 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 $\mathrm{fl}(x)$ 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 $\mathrm{cond}_f(x) \le 1$ 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 $\mathrm{cond}_A(x)$ for $x \in [0, 1]$.
- Plot $\mathrm{cond}_f(x)$ and $\mathrm{cond}_A(x)$ as a function of x on $[0, 1]$. Discuss your results.

X. The math problem of root finding for a polynomial

$$q(x) = \sum_{i=0}^{n} a_i x^i, \qquad a_n = 1, a_0 \neq 0, a_i \in \mathbb{R} \qquad (1.47)$$

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?

1.4.2 Programming assignments

A. Print values of the functions in (1.48) 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.

$$f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1 \qquad (1.48a)$$

$$g(x) = (((((((x - 8)x + 28)x - 56)x + 70)x - 56)x + 28)x - 8)x + 1 \qquad (1.48b)$$

$$h(x) = (x - 1)^8 \qquad (1.48c)$$

Chapter 2

Solving Nonlinear Equations

2.1 The bisection method

Algorithm 2.1. The *bisection method* finds a root of a continuous function $f : \mathbb{R} \to \mathbb{R}$ by repeatedly reducing the interval to the half interval where the root must lie.

```
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 : $f \in \mathcal{C}[a,b]$,
                $\mathrm{sgn}(f(a)) \neq \mathrm{sgn}(f(b))$
Output: $c, h, k$
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

$u \leftarrow f(a)$
$v \leftarrow f(b)$
for $k = 1 : M$ do
    $h \leftarrow b - a$
    $c \leftarrow a + h/2$
    $w \leftarrow f(c)$
    if $|h| < \delta$ or $|w| < \epsilon$ then
        break
    else if $\mathrm{sgn}(w) \neq \mathrm{sgn}(u)$ then
        $b \leftarrow c$
        $v \leftarrow w$
    else
        $a \leftarrow c$
        $u \leftarrow w$
    end
end
```

2.2 The signature of an algorithm

Definition 2.2. 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 2.3. A *precondition* is a condition that holds for the input prior to the execution of an algorithm.

Definition 2.4. A *postcondition* is a condition that holds for the output after the execution of an algorithm.

Definition 2.5. The *signature of an algorithm* consists of its input, output, preconditions, postconditions, and how input parameters violating preconditions are handled.

2.3 Proof of correctness and simplification of algorithms

Definition 2.6. An *invariant* is a condition that holds during the execution of an algorithm.

Definition 2.7. 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 2.1. What are the invariants in Algorithm 2.1? 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 2.8. 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 : $f \in \mathcal{C}[a,b]$,
                $\mathrm{sgn}(f(a)) \neq \mathrm{sgn}(f(b))$
Output: $c, h, k$
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

$h \leftarrow b - a$
$u \leftarrow f(a)$
for $k = 1 : M$ do
    $h \leftarrow h/2$
    $c \leftarrow a + h$
    $w \leftarrow f(c)$
    if $|h| < \delta$ or $|w| < \epsilon$ then
        break
    else if $\mathrm{sgn}(w) = \mathrm{sgn}(u)$ then
        $a \leftarrow c$
    end
end
```

2.4 Q-order convergence

Definition 2.9 (Q-order convergence)**.** A convergent sequence $\{x_n\}$ is said to *converge* to L with *Q-order* p ($p \geq 1$) if

$$\lim_{n\to\infty} \frac{|x_{n+1} - L|}{|x_n - L|^p} = c > 0; \tag{2.1}$$

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

$$\exists c \in (0,1), \exists d > 0, \text{ s.t. } \forall n \in \mathbb{N},\ |x_n - L| \le c^n d. \tag{2.2}$$

In general, the *order of convergence* of a sequence $\{x_n\}$ converging to L is the maximum $p \in \mathbb{R}^+$ satisfying

$$\exists c > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N,\ |x_{n+1} - L| \le c|x_n - L|^p. \tag{2.3}$$

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

Theorem 2.11 (Monotonic sequence theorem)**.** Every bounded monotonic sequence is convergent.

Theorem 2.12 (Convergence of the bisection method)**.** For a continuous function $f : [a_0, b_0] \to \mathbb{R}$ satisfying $\text{sgn}(f(a_0)) \neq \text{sgn}(f(b_0))$, the sequence of iterates in the bisection method converges linearly with asymptotic factor $\frac{1}{2}$,

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \alpha, \tag{2.4}$$

$$f(\alpha) = 0, \tag{2.5}$$

$$|c_n - \alpha| \le 2^{-(n+1)}(b_0 - a_0), \tag{2.6}$$

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

Proof. It follows from the bisection method that

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

In the rest of this proof, "lim" is a shorthand for "$\lim_{n\to\infty}$." By Theorem 2.11, 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 $f(a_n)f(b_n) \le 0$ always holds. Since f is continuous, $\lim f(a_n)f(b_n) = f(\lim a_n)f(\lim b_n)$, then $f^2(\alpha) \le 0$ implies $f(\alpha) = 0$. (2.6) is another important invariant that can be proven by induction. Comparing (2.6) to (2.2) yields convergence of the bisection method. Also, the convergence is linear with asymptotic factor as $c = \frac{1}{2}$. □

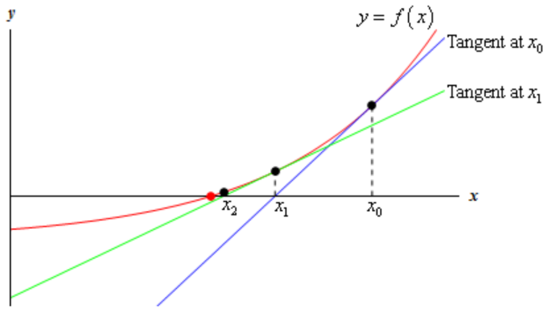
2.5 Newton's method

Algorithm 2.13. *Newton's method* finds the root of $f : \mathbb{R} \to \mathbb{R}$ near an initial guess x_0 by the iteration formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n \in \mathbb{N}. \tag{2.7}$$

```
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 $k = M$

$x \leftarrow x_0$
for $k = 0 : M$ do
    $u \leftarrow f(x)$
    if $|u| < \epsilon$ then
        break
    end
    $x \leftarrow x - u/f'(x)$
end
```



Theorem 2.14 (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 quadratically to the root α, i.e.

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^2} = -\frac{f''(\alpha)}{2f'(\alpha)}. \tag{2.8}$$

Proof. By Taylor's theorem (Theorem 0.48) and the assumption $f \in \mathcal{C}^2$,

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

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 (2.7), 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 (2.3) 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) and completes the proof. □

Theorem 2.15. A continuous function $f:[a,b]\to[c,d]$ is bijective if and only if it is strictly monotonic.

Theorem 2.16. 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 2.15, 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 2.14, we had

$$x_{n+1}-\alpha = (x_n-\alpha)^2\frac{f''(\xi)}{2f'(x_n)}. \tag{2.9}$$

Then $f'>0$ and $f''>0$ further imply that $x_{n+1}>\alpha$ 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 2.11 it converges.

Suppose the sequence $\{x_n\}$ converges to $\alpha+c$ for some fixed $c>0$. Define $\delta=\frac{f(\alpha+c)}{f'(\alpha+c)}$. The Taylor series of $f(\alpha+c)$ expanded at α and $f'(x)>0$ imply $\delta>0$. Because the Newton iteration $\{x_n\}$ converges, we have

$$\forall\epsilon>0, \exists N\in\mathbb{N} \text{ s.t. } \forall n>N, |x_n-x_{n+1}| = \left|\frac{f(x_n)}{f'(x_n)}\right| < \epsilon,$$

which holds in particular for $\epsilon=\frac{1}{2}\delta$. On the other hand,

$$\begin{aligned}\left|x_n-x_{n+1}-\frac{f(\alpha+c)}{f'(\alpha+c)}\right| &\ge \left||x_n-x_{n+1}|-\left|\frac{f(\alpha+c)}{f'(\alpha+c)}\right|\right| \\ &> \delta-\frac{1}{2}\delta=\epsilon.\end{aligned}$$

This contradicts the assumption that the Newton iteration $\{x_n\}$ converges to $\alpha+c$. Together with the first paragraph, this implies that the Newton iteration $\{x_n\}$ converges to α, which is the only root of f.

The quadratic convergence rate can be proved by an induction using (2.9), as in Theorem 2.14. □

Definition 2.17. Let $\mathcal{V}$ be a vector space. A subset $\mathcal{U}\subseteq\mathcal{V}$ is a *convex set* iff

$$\forall x,y\in\mathcal{U}, \forall t\in(0,1), \qquad tx+(1-t)y\in\mathcal{U}. \tag{2.10}$$

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

$$\begin{aligned}&\forall x,y\in\mathcal{U}, \forall t\in(0,1),\\ &f(tx+(1-t)y)\le tf(x)+(1-t)f(y).\end{aligned} \tag{2.11}$$

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

2.6 The secant method

Algorithm 2.18. The *secant method* finds a root of $f:\mathbb{R}\to\mathbb{R}$ near initial guesses x_0, x_1 by the iteration

$$x_{n+1}=x_n-f(x_n)\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}, \qquad n\in\mathbb{N}^+. \tag{2.12}$$

```
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 $k=M$
$x_n \leftarrow x_1$
$x_{n-1} \leftarrow x_0$
$u \leftarrow f(x_n)$
$v \leftarrow f(x_{n-1})$
for $k=2:M$ do
    if $|u|>|v|$ then
        $x_n \leftrightarrow x_{n-1}$
        $u \leftrightarrow v$
    end
    $s \leftarrow \frac{x_n-x_{n-1}}{u-v}$
    $x_{n-1} \leftarrow x_n$
    $v \leftarrow u$
    $x_n \leftarrow x_n - u\times s$
    $u \leftarrow f(x_n)$
    if $|x_n-x_{n-1}|<\delta$ or $|u|<\epsilon$ then
        break
    end
end
```

Definition 2.19. The sequence $\{F_n\}$ of *Fibonacci numbers* is defined as

$$F_0=0,\ F_1=1, \qquad F_{n+1}=F_n+F_{n-1}. \tag{2.13}$$

Theorem 2.20 (Binet's formula)**.** Denote the golden ratio by $r_0=\frac{1+\sqrt{5}}{2}\approx1.618$ and let $r_1=1-r_0=\frac{1-\sqrt{5}}{2}$, then

$$F_n=\frac{r_0^n-r_1^n}{\sqrt{5}}. \tag{2.14}$$

Corollary 2.21. The ratios r_0, r_1 in Theorem 2.20 satisfy

$$F_{n+1}=r_0F_n+r_1^n. \tag{2.15}$$

Proof. This follows from (2.14) and values of r_0 and r_1. □

Lemma 2.22 (Error relation of the secant method)**.** For the secant method (2.12), 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

$$x_{n+1} - \alpha = (x_n - \alpha)(x_{n-1} - \alpha)\frac{f''(\zeta_n)}{2f'(\xi_n)}. \tag{2.16}$$

Proof. Define a divided difference as

$$f[a, b] = \frac{f(a) - f(b)}{a - b}. \tag{2.17}$$

Then it takes some algebra to show that the formula (2.12) is equivalent to

$$x_{n+1} - \alpha = (x_n - \alpha)(x_{n-1} - \alpha)\frac{\frac{f[x_{n-1},x_n]-f[x_n,\alpha]}{x_{n-1}-\alpha}}{f[x_{n-1}, x_n]}. \tag{2.18}$$

By (2.17) and the mean value theorem (Theorem 0.35), there exists ξ_n between x_{n-1} and x_n such that

$$f[x_{n-1}, x_n] = f'(\xi_n). \tag{2.19}$$

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

$$\frac{f[x_{n-1}, x_n] - f[x_n, \alpha]}{x_{n-1} - \alpha} = g'(\beta) \tag{2.20}$$

for some β between x_{n-1} and α. Compute the derivative of $g'(\beta)$ from (2.17), use the Lagrangian remainder Theorem 0.48, and we have

$$\frac{f[x_{n-1}, x_n] - f[x_n, \alpha]}{x_{n-1} - \alpha} = \frac{f''(\zeta_n)}{2} \tag{2.21}$$

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 (2.19) and (2.21) into (2.18). □

Theorem 2.23 (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 2.22

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

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 $ME_0 < \eta$, $ME_1 < \eta$, $ME_2 < ME_1ME_0 < \eta^2$, $ME_3 < ME_2ME_1 < \eta^3$, $\cdots$, $ME_{n+1} < ME_nME_{n-1} < \eta^{q_n+q_{n-1}} = \eta^{q_{n+1}}$, i.e.

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

$\{q_n\}$ is a Fibonacci sequence starting from $q_0 = 1, q_1 = 1$. By Theorem 2.20, 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:

$$\frac{B_{n+1}}{B_n^{r_0}} = \frac{\frac{1}{M}\eta^{q_{n+1}}}{\left(\frac{1}{M}\right)^{r_0}\eta^{r_0 q_n}} = M^{r_0-1}\eta^{q_{n+1}-r_0q_n} \le M^{r_0-1}\eta^{-1}$$

where $q_{n+1} - r_0q_n = r_1^{n+1} > -1$.

To prove convergence rates, we define

$$m_n := \left|\frac{f''(\zeta_n)}{2f'(\xi_n)}\right|, \qquad m_\alpha := \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|, \tag{2.22}$$

where ζ_n and ξ_n are the same as those in Lemma 2.22. 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 2.19. Then

$$\begin{aligned}\frac{E_{n+1}}{E_n^{r_0}} =& E_1^{F_{n+1}-r_0F_n} E_0^{F_n-r_0F_{n-1}} m_1^{F_n-r_0F_{n-1}} m_2^{F_{n-1}-r_0F_{n-2}} \\ & \cdots m_{n-2}^{F_3-r_0F_2} m_{n-1}^{F_2-r_0F_1} m_n^{F_1} \\ =& E_1^{r_1^n} E_0^{r_1^{n-1}} m_1^{r_1^{n-1}} m_2^{r_1^{n-2}} \cdots m_{n-1}^{r_1^1} m_n^1,\end{aligned} \tag{2.23}$$

where the second step follows from Corollary 2.21. (2.22) and the convergence we just proved yield

$$\lim_{n\to+\infty} m_n = m_\alpha, \tag{2.24}$$

which means

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, m_n \in (m_\alpha - \epsilon, m_\alpha + \epsilon)\,. \tag{2.25}$$

We define

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

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 (2.25)

$$B \le (m_\alpha + \epsilon)^{1+r_1^1+r_1^2+\cdots+r_1^{n-N}},$$

and then

$$\begin{aligned}\lim_{n\to\infty} \frac{E_{n+1}}{E_n^{r_0}} &= \lim_{n\to\infty} A \lim_{n\to\infty} B \\ &= \lim_{n\to\infty} B \le (m_\alpha)^{\frac{1}{1-r_1}} = (m_\alpha)^{\frac{1}{r_0}}.\end{aligned}$$

The proof is then completed by Definition 2.9. □

Corollary 2.24. 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

$$T_N = (1+s)m\lceil \log_2 K\rceil, \tag{2.26}$$
$$T_S = m\lceil \log_{r_0} K\rceil, \tag{2.27}$$

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

$$K = \frac{\log c\epsilon}{\log c|x_0-\alpha|}, \tag{2.28}$$

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 2.14. 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 (2.26) holds.

For the secant method, assume $ME_0 \ge ME_1$. By the proof of Theorem 2.23, we have

$$ME_n \le (ME_0)^{r_0^{n+1}/\sqrt{5}}.$$

Let $j \in \mathbb{N}^+$ denote 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!). Finally, 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. □

2.7 Fixed-point iterations

Definition 2.25. A *fixed point* of a function g is an independent parameter of g satisfying $g(\alpha) = \alpha$.

Example 2.2. A fixed point of $f(x) = x^2 - 3x + 4$ is $x = 2$.

Lemma 2.26. 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 0.32). □

Exercise 2.3. 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 2.27 (Brouwer's fixed point)**.** Any 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.

Exercise 2.4. 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 2.27 to prove that there always exists some point in the crumpled ball that sits above the same point it sat above prior to crumpling.

Example 2.5. 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.

Definition 2.28. A *fixed-point iteration* is a method for finding a fixed point of g with a formula of the form

$$x_{n+1} = g(x_n), \qquad n \in \mathbb{N}. \tag{2.29}$$

Example 2.6. Newton's method is a fixed-point iteration.

Exercise 2.7. 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 2.29. A function $f : [a, b] \to [a, b]$ is a *contraction* or *contractive mapping* on $[a, b]$ if

$$\exists\lambda \in [0, 1) \text{ s.t. } \forall x, y \in [a, b],\ |f(x)-f(y)| \le \lambda|x-y|. \tag{2.30}$$

Example 2.8. Any linear function $f(x) = \lambda x + c$ with $0 \le \lambda < 1$ is a contraction.

Theorem 2.30 (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 (2.29) converges to α for any choice $x_0 \in [a, b]$ and

$$|x_n - \alpha| \le \frac{\lambda^n}{1-\lambda}|x_1 - x_0|. \tag{2.31}$$

Proof. By Lemma 2.26, 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 2.29, $x_{n+1} = g(x_n)$ 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,

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

From the first and last right-hand sides (RHSs), we have $|x_0 - \alpha| \le \frac{1}{1-\lambda}|x_1 - x_0|$, which yields (2.31). □

Theorem 2.31. 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 (2.29) converges to α for any choice $x_0 \in [a, b]$, the error bound (2.31) holds, and

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = g'(\alpha). \tag{2.32}$$

Proof. The mean value theorem (Theorem 0.35) implies that, for all $x, y \in [a, b]$, $|g(x) - g(y)| \le \lambda|x - y|$. Theorem 2.30 yields all the results except (2.32), 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 2.32. 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 2.30 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 2.31 completes the proof. □

Corollary 2.33. Consider $g : [a, b] \to [a, b]$ with a fixed point $g(\alpha) = \alpha \in [a, b]$. The fixed-point iteration (2.29) converges to α with pth-order accuracy ($p > 1$, $p \in \mathbb{N}$) for any choice $x_0 \in [a, b]$ if

$$\begin{cases} g \in \mathcal{C}^p[a, b], \\ \forall k = 1, 2, \ldots, p - 1,\ g^{(k)}(\alpha) = 0, \\ g^{(p)}(\alpha) \ne 0. \end{cases} \tag{2.33}$$

Proof. By Corollary 2.32, the fixed-point iteration converges uniquely to α because $g'(\alpha) = 0$. 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 0.31 implies that $g^{(p)}$ is bounded on $[a, b]$. Hence there exists a constant M such that $E_{\text{abs}}(x_{n+1}) < M E^p_{\text{abs}}(x_n)$. □

Example 2.9. 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	$\frac{-48R(9x^4-18Rx^2+R^2)}{(3x^2+R)^4}$	$\frac{-48R(-8R^2)}{(4R)^4} = \frac{3}{2R} \ne 0$

The rest follows from Corollary 2.33.

2.8 Problems

2.8.1 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 maximum possible distance between the root r and the midpoint of the interval?

II. In using the bisection algorithm with its initial interval as $[a_0, b_0]$, we want to determine the root with its *relative* error no greater than ϵ. Assume $a_0 > 0$. Prove that the number of steps n must satisfy

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

III. 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?

IV. 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.

V. 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

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

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

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

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

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

VIII. 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?

IX. $(*)$ 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:

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

2.8.2 Programming assignments

A. Implement the bisection method and test your program on these 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 $[0, 4]$.

B. Implement Newton's method to solve the equation $x = \tan x$. Find the roots near 4.5 and 7.7.

C. Implement the secant method and test your program on 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.

Chapter 3

Polynomial Interpolation

Definition 3.1. *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 3.1. The interpolating function may be piecewise constant, piecewise linear, polynomial, spline, or other non-polynomial functions.

3.1 The Vandermonde determinant

Definition 3.2. For $n+1$ given points $x_0, x_1, \ldots, x_n \in \mathbb{R}$, the associated *Vandermonde matrix* $V \in \mathbb{R}^{(n+1)\times(n+1)}$ is

$$V(x_0, x_1, \ldots, x_n) = \begin{bmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{bmatrix}. \tag{3.1}$$

Lemma 3.3. The determinant of a Vandermonde matrix can be expressed as

$$\det V(x_0, x_1, \ldots, x_n) = \prod_{i>j}(x_i - x_j). \tag{3.2}$$

Proof. Consider the function

$$\begin{aligned} U(x) &= \det V(x_0, x_1, \ldots, x_{n-1}, x) \\ &= \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix}. \end{aligned} \tag{3.3}$$

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 rows in the determinant. It follows that

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

where A depends only on $x_0, x_1, \ldots, x_{n-1}$. Meanwhile, the expansion of $U(x)$ in (3.3) by minors of its last row implies that the coefficient of x^n is $U(x_0, x_1, \ldots, x_{n-1})$. Hence we have

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

and consequently the recursion

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

An induction based on $U(x_0, x_1) = x_1 - x_0$ yields (3.2). □

Theorem 3.4 (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

$$\forall i = 0, 1, \ldots, n, \qquad p_n(x_i) = f_i. \tag{3.4}$$

Proof. Set up a polynomial $\sum_{i=0}^n a_i x^i$ with $n+1$ undetermined coefficients a_i. The condition (3.4) 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 3.3, 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. □

3.2 The Cauchy remainder

Theorem 3.5 (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 0.34) 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 3.6 (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

$$R_n(f;x) := f(x) - p_n(f;x) \tag{3.5}$$

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

$$R_n(f;x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x - x_i) \tag{3.6}$$

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 $x \ne x_0, x_1, \ldots, x_n$ 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 3.5, $W^{(n+1)}(\xi) = 0$ 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 (3.6) holds. □

Corollary 3.7. Suppose $f(x) \in \mathcal{C}^{n+1}[a,b]$. Then

$$|R_n(f;x)| \le \frac{M_{n+1}}{(n+1)!}\prod_{i=0}^{n}|x - x_i| < \frac{M_{n+1}}{(n+1)!}(b-a)^{n+1}, \tag{3.7}$$

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

Example 3.2. 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 3.7 we have $|R_1| \le 4.42 \times 10^{-7}$. The true error is about 1.10×10^{-7}.

3.3 The Lagrange formula

Definition 3.8. 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

$$p_n(x) = \sum_{k=0}^{n} f_k \ell_k(x), \tag{3.8}$$

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

$$\ell_k(x) = \prod_{i\ne k;i=0}^{n}\frac{x - x_i}{x_k - x_i}. \tag{3.9}$$

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

Example 3.3. 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 3.9. Define a symmetric polynomial

$$\pi_n(x) = \begin{cases} 1, & n = 0; \\ \prod_{i=0}^{n-1}(x - x_i), & n > 0. \end{cases} \tag{3.10}$$

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

$$\forall x \ne x_k, \qquad \ell_k(x) = \frac{\pi_{n+1}(x)}{(x - x_k)\pi'_{n+1}(x_k)}. \tag{3.11}$$

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

$$\sum_{k=0}^{n}\ell_k(x) \equiv 1 \tag{3.12}$$

$$\forall j = 1, \ldots, n, \qquad \sum_{k=0}^{n}(x_k - x)^j \ell_k(x) \equiv 0 \tag{3.13}$$

Proof. By Theorems 3.4 and 3.6, 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 (3.12). Similarly, (3.13) can be proved by interpolating the polynomial $q(u) = (u - x)^j$ for each $j = 1, \ldots, n$ with the Lagrange formula. □

3.4 The Newton formula

Definition 3.11 (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

$$p_n(x) = \sum_{k=0}^{n} a_k \pi_k(x), \tag{3.14}$$

where π_k is defined in (3.10) 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 3.12. Suppose $(i_0, i_1, i_2, \ldots, i_k)$ is a permutation of $(0, 1, 2, \ldots, k)$. Then

$$f[x_0, x_1, \ldots, x_k] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_k}]. \tag{3.15}$$

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 3.4. □

Corollary 3.13. The kth divided difference can be expressed as

$$f[x_0, x_1, \ldots, x_k] = \sum_{i=0}^{k} \frac{f_i}{\prod_{j\neq i; j=0}^{k}(x_i - x_j)} = \sum_{i=0}^{k} \frac{f_i}{\pi'_{k+1}(x_i)}, \tag{3.16}$$

where $\pi_{k+1}(x)$ is defined in (3.10).

Proof. The uniqueness of interpolating polynomials in Theorem 3.4 implies that the two polynomials in (3.8) and (3.14) are the same. Then the first equality follows from (3.9) and Definition 3.11, while the second equality follows from Lemma 3.9. □

Theorem 3.14. Divided differences satisfy the recursion

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, x_2, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0}. \tag{3.17}$$

Proof. By Definition 3.11, $f[x_1, x_2, \ldots, x_k]$ 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 $f[x_0, x_1, \ldots, x_{k-1}]$. 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 $i = 1, 2, \ldots, k-1$. Hence $P(x_i) = P_1(x_i)$ for $i = 1, 2, \ldots, k-1$. Lastly, $P(x_k) = P_2(x_k)$. Therefore, $P(x)$ as above is the interpolating polynomial for 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 the definitions of and the kth divided difference. □

Definition 3.15. The kth divided difference ($k \in \mathbb{N}^+$) on the *table of divided differences*

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

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 3.4. 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 3.15, we can construct the following table of divided difference,

$$\begin{array}{c|llll}
0 & 6 & & & \\
1 & -3 & -9 & & \\
2 & -6 & -3 & 3 & \\
3 & 9 & 15 & 9 & 2
\end{array} \tag{3.18}$$

By Definition 3.11, the interpolating polynomial is generated from the main diagonal and the first column of the above table as follows.

$$p_3 = 6 - 9x + 3x(x-1) + 2x(x-1)(x-2). \tag{3.19}$$

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

Exercise 3.5. Redo Example 3.3 with the Newton formula.

Theorem 3.16. For distinct points $x_0, x_1, \ldots, x_n$ and an arbitrary x, we have

$$\begin{aligned} f(x) = {} & f[x_0] + f[x_0, x_1](x - x_0) + \cdots \\ & + f[x_0, x_1, \cdots, x_n]\prod_{i=0}^{n-1}(x - x_i) \\ & + f[x_0, x_1, \cdots, x_n, x]\prod_{i=0}^{n}(x - x_i). \end{aligned} \tag{3.20}$$

Proof. Take another point $z \neq x_i$. The Newton formula applied to $x_0, x_1, \ldots, x_n, z$ 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 (3.20).

The above argument assumes $x \neq x_i$. Now consider the case of $x = x_j$ for some fixed j. Rewrite (3.20) as $f(x) = p_n(f; x) + R(x)$ where $R(x)$ is clearly the last term in (3.20). We need to show

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

which clearly holds because $R(x_j) = 0$ and the interpolation condition at x_j dictates $p_n(f; x_j) = f(x_j)$. □

Corollary 3.17. 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

$$f[x_0, x_1, \cdots, x_n, x] = \frac{1}{(n+1)!}f^{(n+1)}(\xi(x)) \tag{3.21}$$

where ξ depends on x and $\xi(x) \in (a, b)$.

Proof. This follows from Theorems 3.16 and 3.6. □

Corollary 3.18. If $x_0 < x_1 < \cdots < x_n$ and $f \in \mathcal{C}^n[x_0, x_n]$, we have

$$\lim_{x_n \to x_0} f[x_0, x_1, \cdots, x_n] = \frac{1}{n!} f^{(n)}(x_0). \tag{3.22}$$

Proof. Set $x = x_{n+1}$ in Corollary 3.17, replace $n+1$ by n, and we have $\xi \to x_0$ as $x_n \to x_0$ since each $x_i \to x_0$. □

Definition 3.19. For $n \in \mathbb{N}^+$, the nth *forward difference* associated with a sequence of values $\{f_0, f_1, \ldots\}$ is

$$\begin{aligned} \Delta f_i &= f_{i+1} - f_i, \\ \Delta^{n+1} f_i &= \Delta \Delta^n f_i = \Delta^n f_{i+1} - \Delta^n f_i, \end{aligned} \tag{3.23}$$

and the nth *backward difference* is

$$\begin{aligned} \nabla f_i &= f_i - f_{i-1}, \\ \nabla^{n+1} f_i &= \nabla \nabla^n f_i = \nabla^n f_i - \nabla^n f_{i-1}. \end{aligned} \tag{3.24}$$

Theorem 3.20. The forward difference and backward difference are related as

$$\forall n \in \mathbb{N}^+, \qquad \Delta^n f_i = \nabla^n f_{i+n}. \tag{3.25}$$

Proof. An easy induction. □

Theorem 3.21. The forward difference can be expressed explicitly as

$$\Delta^n f_i = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k}. \tag{3.26}$$

Proof. For $n = 1$, (3.26) reduces to $\Delta f_i = f_{i+1} - f_i$. The rest of the proof is an induction utilizing the identity

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}. \tag{3.27}$$

Suppose (3.26) 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 (3.23), 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 (3.27) and the fact that $(-1)^{n+1} f_i$ and f_{i+n+1} contribute to the first and last terms, respectively. □

Theorem 3.22. On a grid $x_i = x_0 + ih$ with uniform spacing h, the sequence of values $f_i = f(x_i)$ satisfies

$$\forall n \in \mathbb{N}^+, \qquad f[x_0, x_1, \ldots, x_n] = \frac{\Delta^n f_0}{n! h^n}. \tag{3.28}$$

Proof. Of course (3.28) can be proven by induction. Here we provide a more informative proof. For $\pi_{n+1}(x)$ defined in (3.10), we have $\pi'(x_k) = \prod_{i=0, i \neq k}^{n} (x_k - x_i)$. It follows from $x_k - x_i = (k-i)h$ that

$$\pi'(x_k) = \prod_{i=0, i\neq k}^{n} (k-i)h = h^n k!(n-k)!(-1)^{n-k}. \tag{3.29}$$

Then we have

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

where the first step follows from Corollary 3.13, the second from (3.29), and the last from Theorem 3.21. □

Theorem 3.23 (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

$$\forall s \in \mathbb{R}, \qquad p_n(f; x_0 + sh) = \sum_{k=0}^{n} \binom{s}{k} \Delta^k f_0, \tag{3.30}$$

where $\Delta^0 f_0 = f_0$ and

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!}. \tag{3.31}$$

Proof. Set $f(x) = p_n(f; x)$ in Theorem 3.16, apply Theorem 3.22, 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 (3.31). □

3.5 The Neville-Aitken algorithm

Theorem 3.24. 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

$$p_{k+1}^{[i]}(x) = \frac{(x - x_i) p_k^{[i+1]}(x) - (x - x_{i+k+1}) p_k^{[i]}(x)}{x_{i+k+1} - x_i}. \tag{3.32}$$

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 we have

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

which, together with (3.32), implies

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

In addition, (3.32) and the induction hypothesis yield

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

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

Example 3.6. To estimate $f(x)$ for $x = \frac{3}{2}$ directly from the table in Example 3.4, we construct a table by repeating (3.32) with $x_i = i$ for $i = 0, 1, 2, 3$.

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			

(3.33)

The result is the same as that in Example 3.4. In contrast, the calculation and layout of the two tables are distinct.

3.6 Hermite interpolation

Definition 3.25. 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 to find a polynomial p of the lowest degree such that

$$\forall i = 0, 1, \ldots, k, \ \forall \mu = 0, 1, \cdots, m_i, \ \ p^{(\mu)}(x_i) = f_i^{(\mu)}, \tag{3.34}$$

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

Definition 3.26. The nth divided difference at $n+1$ "confluent" (i.e. identical) points is defined as

$$f[x_0, x_0, \cdots, x_0] = \frac{1}{n!} f^{(n)}(x_0), \tag{3.35}$$

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

Theorem 3.27. For the Hermite interpolation problem in Definition 3.25, denote $N = k + \sum_i m_i$. Denote by $p_N(f; x)$ the unique element of $\mathbb{P}_N$ for which (3.34) holds. Suppose $f^{(N+1)}(x)$ exists in (a, b). Then

$$f(x) - p_N(f; x) = \frac{f^{(N+1)}(\xi)}{(N+1)!} \prod_{i=0}^{k} (x - x_i)^{m_i+1}. \tag{3.36}$$

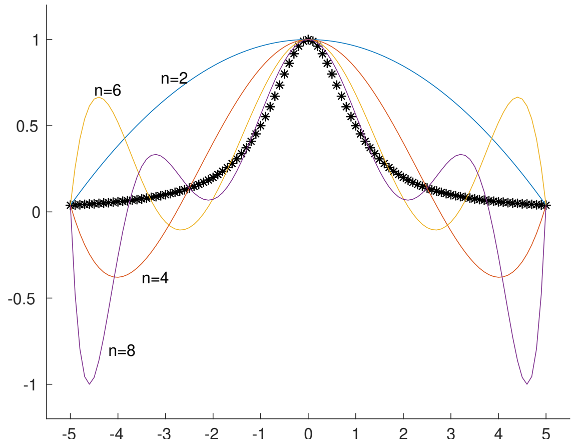
Proof. The proof is similar to that of Theorem 3.6. Pay attention to the difference caused by the multiple roots of the polynomial $\prod_{i=0}^k (x - x_i)^{m_i+1}$. □

3.7 The Chebyshev polynomials

Example 3.7 (Runge phenomenon)**.** The points $x_0, x_1, \ldots, x_n$ in Theorem 3.4 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

$$\forall f \in \mathcal{C}[x_0, x_n], \forall x \in [x_0, x_n], \quad \lim_{n \to +\infty} p_n(f; x) = f(x). \tag{3.37}$$

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

$$f(x) = \frac{1}{1 + x^2} \tag{3.38}$$

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

Definition 3.28. The *Chebyshev polynomial* of degree n of the first kind is a polynomial $T_n : [-1, 1] \to \mathbb{R}$,

$$T_n(x) = \cos(n \arccos x). \tag{3.39}$$

Theorem 3.29.

$$\forall n \in \mathbb{N}^+, \qquad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{3.40}$$

Proof. By trigonometric identities, we have

$$\begin{aligned} \cos(n+1)\theta &= \cos n\theta \cos\theta - \sin n\theta \sin\theta, \\ \cos(n-1)\theta &= \cos n\theta \cos\theta + \sin n\theta \sin\theta. \end{aligned}$$

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

Corollary 3.30. The coefficient of x^n in T_n is 2^{n-1} for each $n > 0$.

Proof. Use (3.40) and $T_1 = x$ in an induction. □

Theorem 3.31. $T_n(x)$ has simple zeros at the n points

$$x_k = \cos\frac{2k-1}{2n}\pi, \tag{3.41}$$

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

$$x'_k = \cos\frac{k}{n}\pi, \qquad k = 0, 1, \ldots, n, \tag{3.42}$$

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

Proof. (3.39) and (3.41) 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 (3.39) 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, $\forall k = 1, 2, \ldots, n-1$,

$$T'_n(x'_k) = n\left(1-\cos^2\frac{k\pi}{n}\right)^{-\frac{1}{2}}\sin(k\pi) = 0;$$

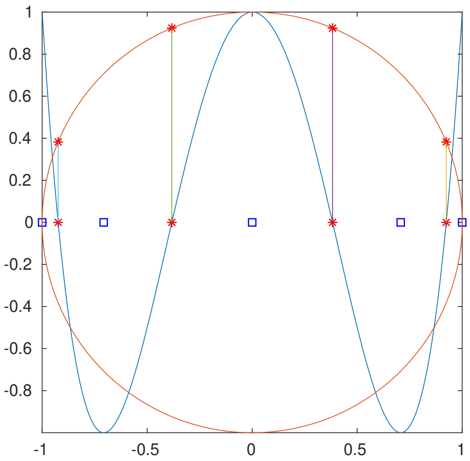
$$T''_n(x) = \frac{n^2\cos\left(n\arccos\left(x\right)\right)}{x^2-1} + \frac{n\,x\,\sin\left(n\arccos\left(x\right)\right)}{\left(1-x^2\right)^{3/2}};$$

$$T''_n(x'_k) \neq 0.$$

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 $k = 0, 1, \ldots, n$, $T_n(x'_k)$ attains its extreme values at x'_k since $T_n(x'_0) = 1$, $T_n(x'_1) = -1$, $\ldots$, and by (3.39) we have $|T_n(x)| \leq 1$. Clearly these are the only extrema of $T_n(x)$ on $[-1, 1]$. □



Exercise 3.8. Write a program to reproduce the above plot.

Theorem 3.32 (Chebyshev)**.** Denote by $\tilde{\mathbb{P}}_n$ the class of all polynomials of degree $n \in \mathbb{N}^+$ with leading coefficient 1. Then

$$\forall p \in \tilde{\mathbb{P}}_n, \qquad \max_{x\in[-1,1]}\left|\frac{T_n(x)}{2^{n-1}}\right| \leq \max_{x\in[-1,1]}|p(x)|. \tag{3.43}$$

Proof. By Theorem 3.31, $T_n(x)$ assumes its extrema $n+1$ times at the points x'_k defined in (3.42). Suppose (3.43) does not hold. Then Theorem 3.31 implies that

$$\exists p \in \tilde{\mathbb{P}}_n \;\text{ s.t. }\; \max_{x\in[-1,1]}|p(x)| < \frac{1}{2^{n-1}}. \tag{3.44}$$

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 (3.44), $Q(x)$ 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 (3.44). □

Corollary 3.33. For $n \in \mathbb{N}^+$, we have

$$\max_{x\in[-1,1]}|x^n + a_1x^{n-1} + \cdots + a_n| \geq \frac{1}{2^{n-1}}. \tag{3.45}$$

Corollary 3.34. Suppose polynomial interpolation is performed for f on the $n+1$ zeros of $T_{n+1}(x)$ as in Theorem 3.31. The Cauchy remainder in Theorem 3.6 satisfies

$$|R_n(f;x)| \leq \frac{1}{2^n(n+1)!}\max_{x\in[-1,1]}\left|f^{(n+1)}(x)\right|. \tag{3.46}$$

Proof. Theorem 3.6, Corollary 3.30, and Theorem 3.31 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 3.28 completes the proof as $|T_{n+1}| \leq 1$. □

Theorem 3.35 (Weierstrass approximation)**.** Every continuous function $f : [a, b] \to \mathbb{R}$ can be uniformly approximated as closely as desired by a polynomial function. More precisely, let $\mathbb{P}_n$ denote the polynomials of degree no more than n. Then we have

$$\begin{aligned}&\forall f \in \mathcal{C}[a,b],\ \forall\epsilon > 0,\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > N,\\ &\qquad\qquad \exists p_n \in \mathbb{P}_n \text{ s.t. } \forall x \in [a,b],\ \|p_n - f\| < \epsilon.\end{aligned} \tag{3.47}$$

Proof. Not required. □

3.8 Problems

3.8.1 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.
- For $x \in [x_0, x_1]$, 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 $\leq m$ that are non-negative on the real line,

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

Find $p \in \mathbb{P}_{2n}^+$ such that $p(x_i) = f_i$ for $i = 0, 1, \ldots, n$ where $f_i \geq 0$ 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 3.17 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 $|f^{(5)}(x)| \leq M$ 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

$$\min \max_{x \in [a,b]} |a_0 x^n + a_1 x^{n-1} + \cdots + a_n|,$$

where the minimum is taken over all $a_i \in \mathbb{R}$ and $a_0 \neq 0$.

X. Imitate the proof of Chebyshev Theorem.
Let $a > 1$ and denote $\mathbb{P}_n^a = \{p \in \mathbb{P}_n : p(a) = 1\}$. Define

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

where T_n is the Chebyshev polynomial of degree n. Clearly $\hat{p}_n(x) \in \mathbb{P}_n^a$. Define the *max-norm* of a function $f : \mathbb{R} \to \mathbb{R}$ as

$$\|f\|_\infty = \max_{x \in [-1,1]} |f(x)|.$$

Prove

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

3.8.2 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.

Chapter 4

Splines

4.1 Piecewise-polynomial splines

Definition 4.1. Given nonnegative integers n, k, and a strictly increasing sequence $\{x_i\}$ that partitions $[a,b]$,

$$a = x_1 < x_2 < \cdots < x_N = b, \tag{4.1}$$

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

$$\mathbb{S}_n^k = \left\{s : \; s \in \mathcal{C}^k[a,b]; \; \forall i \in [1, N-1], \; s\big|_{[x_i,x_{i+1}]} \in \mathbb{P}_n\right\}. \tag{4.2}$$

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

Notation 5. In Section 3, the polynomial degree is denoted by n for all methods. Here we use N to denote the number of knots for a spline.

Example 4.1. 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 4.2. 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

$$\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = 3\mu_i f[x_i, x_{i+1}] + 3\lambda_i f[x_{i-1}, x_i], \tag{4.3}$$

where

$$\mu_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}, \qquad \lambda_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}. \tag{4.4}$$

Proof. Denote $p_i(x) = s\big|_{[x_i,x_{i+1}]}$ 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

$$\begin{aligned} p_i(x) =& f_i + (x - x_i)m_i + (x-x_i)^2 \frac{K_i - m_i}{x_{i+1} - x_i} \\ &+ (x - x_i)^2 (x - x_{i+1}) \frac{m_i + m_{i+1} - 2K_i}{(x_{i+1} - x_i)^2}, \end{aligned} \tag{4.5}$$

or equivalently

$$\begin{cases} p_i(x) &= c_{i,0} + c_{i,1}(x-x_i) + c_{i,2}(x-x_i)^2 + c_{i,3}(x-x_i)^3, \\ c_{i,0} &= f_i, \\ c_{i,1} &= m_i, \\ c_{i,2} &= \frac{3K_i - 2m_i - m_{i+1}}{x_{i+1}-x_i}, \\ c_{i,3} &= \frac{m_i + m_{i+1} - 2K_i}{(x_{i+1}-x_i)^2}. \end{cases} \tag{4.6}$$

$s \in \mathcal{C}^2$ 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 (4.3). □

Definition 4.3. The method of *dynamic programming*, or *dynamic optimization*, solves a complex problem by breaking it down into a collection of simpler sub-problems, solving each of those sub-problems just once, and storing their solutions. When the same sub-problem occurs, instead of recomputing its solution, one simply looks up the previously computed solution, thereby saving computation time at the expense of an increase in storage space.

Lemma 4.4. 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

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = 6f[x_{i-1}, x_i, x_{i+1}] \tag{4.7}$$

where μ_i and λ_i are the same as those in (4.4).

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

$$s(x) = f_i + s'(x_i)(x - x_i) + \frac{M_i}{2}(x-x_i)^2 + \frac{s'''(x_i)}{6}(x-x_i)^3, \tag{4.8}$$

where $x \geq x_i$ and the derivatives should be interpreted as the right-hand derivatives. Differentiate (4.8) twice, set $x = x_{i+1}$, and we have

$$s'''(x_i) = \frac{M_{i+1} - M_i}{x_{i+1} - x_i}. \tag{4.9}$$

Substitute (4.9) into (4.8), set $x = x_{i+1}$, and we have

$$s'(x_i) = f[x_i, x_{i+1}] - \frac{1}{6}(M_{i+1} + 2M_i)(x_{i+1} - x_i). \tag{4.10}$$

Differentiate (4.8) 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 (4.8) yields

$$s'(x_i) = f[x_{i-1}, x_i] - \frac{1}{6}(M_{i-1} + 2M_i)(x_{i-1} - x_i). \quad (4.11)$$

The subtraction of (4.10) and (4.11) yields (4.7). □

Definition 4.5 (Types of splines)**.**

- 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 4.6. Denote $M_i = s''(f; x_i)$ for $s \in \mathbb{S}_3^2$ and we have

$$2M_1 + M_2 = 6f[x_1, x_1, x_2], \quad (4.12)$$

$$M_{N-1} + 2M_N = 6f[x_{N-1}, x_N, x_N]. \quad (4.13)$$

Proof. As for (4.12), 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

$$\begin{aligned} s_1(x) =& f[x_1] + f[x_1, x_1](x - x_1) \\ &+ \frac{M_1}{2}(x - x_1)^2 + \frac{M_2 - M_1}{6(x_2 - x_1)}(x - x_1)^3. \end{aligned} \quad (4.14)$$

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 (4.12). (4.13) can be proven similarly. □

Theorem 4.7. 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 4.2, 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 (4.3) into a linear system

$$\begin{bmatrix} 2 & \mu_2 & & & & & \\ \lambda_3 & 2 & \mu_3 & & & & \\ & & \ddots & & & & \\ & & \lambda_i & 2 & \mu_i & & \\ & & & & \ddots & & \\ & & & & \lambda_{N-2} & 2 & \mu_{N-2} \\ & & & & & \lambda_{N-1} & 2 \end{bmatrix} \begin{bmatrix} m_2 \\ m_3 \\ \vdots \\ m_i \\ \vdots \\ m_{N-2} \\ m_{N-1} \end{bmatrix} = \mathbf{b}, \quad (4.15)$$

where the vector $\mathbf{b}$ is determined from the known information. (4.4) 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 4.4 and 4.6, following arguments similar to the above. □

Example 4.2. 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 4.4 yields 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 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 (4.14). After this expression is obtained, we then evaluate it and obtain $s(5) \approx 1.60977$. In comparison, $\ln(5) \approx 1.60944$.

4.2 The minimum properties

Theorem 4.8 (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

$$\int_a^b [s''(x)]^2 \mathrm{d}x \le \int_a^b [g''(x)]^2 \mathrm{d}x, \tag{4.16}$$

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 4.9 (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

$$\int_a^b [s''(x)]^2 \mathrm{d}x \le \int_a^b [g''(x)]^2 \mathrm{d}x, \tag{4.17}$$

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

Proof. The proof is similar to that of Theorem 4.8. Although $\eta'(a) = \eta'(b) = 0$ does not hold, we do have $s''(a) = s''(b) = 0$. □

Lemma 4.10. 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

$$\forall x \in [a, b], \qquad |s''(x)| \le 3 \max_{x\in[a,b]} |f''(x)|. \tag{4.18}$$

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 4.4 and Corollary 3.17 that

$$\begin{aligned} &2M_j = 6f[x_{j-1}, x_j, x_{j+1}] - \mu_j M_{j-1} - \lambda_j M_{j+1} \\ \Rightarrow\ &2|M_j| \le 6\,|f[x_{j-1}, x_j, x_{j+1}]| + (\mu_j + \lambda_j)|M_j| \\ \Rightarrow\ &\exists \xi \in [x_{j-1}, x_{j+1}] \text{ s.t. } |M_j| \le 3\,|f''(\xi)| \\ \Rightarrow\ &|s''(x)| \le 3 \max_{x\in[a,b]} |f''(x)|. \end{aligned} \tag{4.19}$$

If $|s''(x)|$ attains its maximum at x_1 or x_N, (4.19) clearly holds for a cubic spline with specified second derivatives at these end points. Due to symmetry, it suffices to prove (4.19) for the complete spline 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 (4.12), 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 3.17, implies

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

This completes the proof. □

4.3 Error analysis

Theorem 4.11. 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

$$\forall j = 0, 1, 2, \quad \left|f^{(j)}(x) - s^{(j)}(x)\right| \le c_j h^{4-j} \max_{x\in[a,b]} \left|f^{(4)}(x)\right|, \tag{4.20}$$

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 prove the conclusion for other cases.

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

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

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 3.6) implies

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

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

$$|f''(x) - \hat{s}''(x)| \le \frac{h^2}{8} \max_{x\in[a,b]} |f^{(4)}(x)|. \tag{4.21}$$

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 4.10 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 (4.21), leads to (4.20) for $j = 2$:

$$\begin{aligned}|f''(x) - s''(x)| &\le |f''(x) - \hat{s}''(x)| + |\hat{s}''(x) - s''(x)| \\ &\le 4 \max_{x\in[a,b]} |f''(x) - \hat{s}''(x)| \\ &\le \frac{1}{2}h^2 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|. \end{aligned} \tag{4.22}$$

For $j = 0$, we have $f(x) - s(x) = 0$ for $x = x_i, x_{i+1}$. Then Rolle's theorem T0.34 implies $f'(\xi_i) - s'(\xi_i) = 0$ for some $\xi_i \in [x_i, x_{i+1}]$. It follows from the fundamental theorem of calculus 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 T0.56 and (4.22), 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 (4.20) 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 3.6 and the third step from (4.22). □

Exercise 4.3. Verify Theorem 4.11 using the results in Example 4.2.

4.4 B-Splines

Notation 6. 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 4.12. 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 (4.2) and Definition 0.69 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 4.4. The cubic splines in Definition 4.5, 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 4.5.

4.4.1 Truncated power functions

Definition 4.13. The *truncated power function* with exponent n is defined as

$$x_+^n = \begin{cases} x^n & \text{if } x \ge 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{4.23}$$

Example 4.5. According to Definition 4.13, we have

$$\forall t \in [a, b], \quad \int_a^b (t - x)_+^n \mathrm{d}x = \int_a^t (t - x)^n \mathrm{d}x = \frac{(t - a)^{n+1}}{n + 1}. \tag{4.24}$$

Lemma 4.14. The following is a basis of $\mathbb{S}_n^{n-1}(t_1, \ldots, t_N)$,

$$1, x, x^2, \ldots, x^n, (x-t_2)_+^n, (x-t_3)_+^n, \ldots, (x-t_{N-1})_+^n. \tag{4.25}$$

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

$$\sum_{i=0}^{n} a_i x^i + \sum_{j=2}^{N-1} a_{n+j}(x - t_j)_+^n = \mathbf{0}(x). \tag{4.26}$$

To satisfy (4.26) for all $x < t_2$, a_i must be 0 for each $i = 0, 1, \cdots, n$. To satisfy (4.26) 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 (4.25) are linearly independent by Definition 0.76. The proof is completed by Theorem 4.12, Lemma 0.86, and the fact that there are $n+N-1$ functions in (4.25). □

Corollary 4.15. Any $s \in \mathbb{S}_{n,N}^{n-1}$ can be expressed as

$$s(x) = \sum_{i=0}^{n} a_i(x - t_1)^i + \sum_{j=2}^{N-1} a_{n+j}(x - t_j)_+^n, \qquad x \in [t_1, t_N]. \tag{4.27}$$

Proof. By Lemma 4.14, it suffices to point out that $\mathrm{span}\{1, x, \ldots, x^n\} = \mathrm{span}\{1, (x - t_1), \ldots, (x - t_1)^n\}$. □

Example 4.6. (4.27) 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 (4.27).

4.4.2 The local support of B-splines

Definition 4.16. The *hat function* at t_i is

$$\hat{B}_i(x) = \begin{cases} \frac{x-t_{i-1}}{t_i-t_{i-1}} & x \in (t_{i-1}, t_i], \\ \frac{t_{i+1}-x}{t_{i+1}-t_i} & x \in (t_i, t_{i+1}], \\ 0 & \text{otherwise.} \end{cases} \tag{4.28}$$

Theorem 4.17. The hat functions form a basis of $\mathbb{S}_1^0$.

Proof. By Definition 4.16, we have

$$\hat{B}_i(t_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{4.29}$$

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 (4.29). Hence by Definition 0.76 the hat functions are linearly independent. It suffices to show that $\text{span}\{\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_N\} = \mathbb{S}_1^0$, which is true because

$$\forall s(x) \in \mathbb{S}_1^0, \ \ \exists s_B(x) = \sum_{i=1}^N s(t_i)\hat{B}_i(x) \text{ s.t. } s(x) = s_B(x).$$

On each interval $[t_i, t_{i+1}]$, (4.29) implies $s_B(t_i) = s(t_i)$ and $s_B(t_{i+1}) = s(t_{i+1})$. Hence $s_B(x) \equiv s(x)$ because they are both linear. Then Definition 0.79 completes the proof. □

Definition 4.18. *B-splines* are defined recursively by

$$B_i^{n+1}(x) = \frac{x - t_{i-1}}{t_{i+n} - t_{i-1}} B_i^n(x) + \frac{t_{i+n+1} - x}{t_{i+n+1} - t_i} B_{i+1}^n(x). \tag{4.30}$$

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

$$B_i^0(x) = \begin{cases} 1 & \text{if } x \in (t_{i-1}, t_i], \\ 0 & \text{otherwise.} \end{cases} \tag{4.31}$$

Example 4.7. The hat functions in Definition 4.16 are clearly the B-splines of degree one:

$$B_i^1 = \hat{B}_i. \tag{4.32}$$

In (4.30), B-splines of higher degrees are defined by generalizing the idea of hat functions.

Example 4.8. The quadratic B-splines $B_i^2(x) =$

$$\begin{cases} \frac{(x-t_{i-1})^2}{(t_{i+1}-t_{i-1})(t_i-t_{i-1})}, & x \in (t_{i-1}, t_i]; \\ \frac{(x-t_{i-1})(t_{i+1}-x)}{(t_{i+1}-t_{i-1})(t_{i+1}-t_i)} + \frac{(t_{i+2}-x)(x-t_i)}{(t_{i+2}-t_i)(t_{i+1}-t_i)}, & x \in (t_i, t_{i+1}]; \\ \frac{(t_{i+2}-x)^2}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & x \in (t_{i+1}, t_{i+2}]; \\ 0, & \text{otherwise.} \end{cases} \tag{4.33}$$

Definition 4.19. The *support* of a function $f : X \to \mathbb{R}$ is

$$\text{supp}(f) = \text{closure}\{x \in X \mid f(x) \neq 0\}. \tag{4.34}$$

Lemma 4.20. For $n \in \mathbb{N}^+$, the interval of support of B_i^n is $[t_{i-1}, t_{i+n}]$. Also, $\forall x \in (t_{i-1}, t_{i+n})$, $B_i^n(x) > 0$.

Proof. This is an easy induction by (4.31) and (4.30). □

Definition 4.21. 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

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y), \tag{4.35}$$

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

Example 4.9. $X = \mathcal{C}[a, b]$, 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 7. 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 4.22 (Leibniz formula)**.** For $k \in \mathbb{N}$, the kth divided difference of a product of two functions satisfies

$$[x_0, \ldots, x_k]fg = \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_i, \ldots, x_k]g. \tag{4.36}$$

Proof. The induction basis $k = 0$ holds because (4.36) reduces to $[x_0]fg = f(x_0)g(x_0)$. Now suppose (4.36) holds. For the induction step, we have from Theorem 3.14 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

$$[x_1, \ldots, x_{k+1}]fg = \sum_{i=0}^k [x_1, \ldots, x_{i+1}]f \cdot [x_{i+1}, \ldots, x_{k+1}]g$$

$$= S_1 + \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_{i+1}, \ldots, x_{k+1}]g, \text{ where}$$

$$S_1 = \sum_{i=0}^k (x_{i+1} - x_0) \cdot [x_0, \ldots, x_{i+1}]f \cdot [x_{i+1}, \ldots, x_{k+1}]g$$

$$= \sum_{i=1}^{k+1} (x_i - x_0) \cdot [x_0, \ldots, x_i]f \cdot [x_i, \ldots, x_{k+1}]g.$$

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

$$= -S_2 + \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot [x_{i+1}, \ldots, x_{k+1}]g, \text{ where}$$

$$S_2 = \sum_{i=0}^k [x_0, \ldots, x_i]f \cdot (x_{k+1} - x_i) \cdot [x_i, \ldots, x_{k+1}]g.$$

In the above derivation, we have applied Theorem 3.14 to go from the kth divided difference to the $(k+1)$th. Then

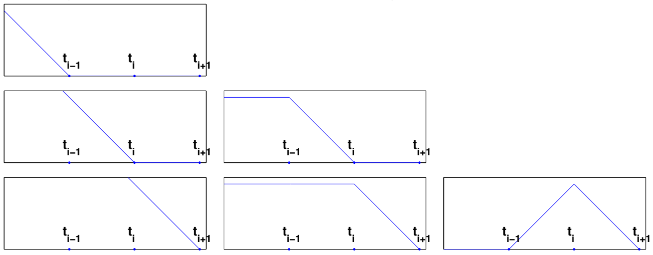
$$\begin{aligned}[x_0,\ldots,x_{k+1}]fg &= \frac{S_1+S_2}{x_{k+1}-x_0}\\ &= \sum_{i=0}^{k+1}[x_0,\ldots,x_i]f\cdot[x_i,\ldots,x_{k+1}]g,\end{aligned}$$

which completes the inductive proof. □

Example 4.10. 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 4.23 (B-splines as divided difference of truncated power functions)**.** For any $n\in\mathbb{N}$, we have

$$B_i^n(x) = (t_{i+n}-t_{i-1})\cdot[t_{i-1},\ldots,t_{i+n}](t-x)_+^n. \tag{4.37}$$

Proof. For $n=0$, (4.37) 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 (4.31). Hence the induction basis holds. Now assume the induction hypothesis (4.37) hold.

By Definition 4.13, $(t-x)_+^{n+1} = (t-x)(t-x)_+^n$. Then the application of Theorem 4.22 with $f=(t-x)$ and $g=(t-x)_+^n$ yields

$$\begin{aligned}&[t_{i-1},\ldots,t_{i+n}](t-x)_+^{n+1}\\ =&(t_{i-1}-x)\cdot[t_{i-1},\ldots,t_{i+n}](t-x)_+^n\\ &+[t_i,\ldots,t_{i+n}](t-x)_+^n.\end{aligned}\tag{4.38}$$

Definition 4.18 and the induction hypothesis yield

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

where the last step follows from (4.38). Similarly,

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

where the second last step follows from Theorem 3.14 and (4.38). The above arguments yield

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

which completes the inductive proof. □

4.4.3 Integrals and derivatives

Corollary 4.24 (Integrals of B-splines)**.** The average of a B-spline over its support only depends on its degree,

$$\frac{1}{t_{i+n}-t_{i-1}}\int_{t_{i-1}}^{t_{i+n}} B_i^n(x)\mathrm{d}x = \frac{1}{n+1}. \tag{4.39}$$

Proof. The left-hand side (LHS) of (4.39) is

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

where the first step follows from Theorem 4.23, the second step from the commutativity of integration and taking divided difference, the third step from (4.24), and the last step from Corollary 3.17. □

Theorem 4.25 (Derivatives of B-splines)**.** For $n \ge 2$, we have, $\forall x \in \mathbb{R}$,

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^n(x) = \frac{nB_i^{n-1}(x)}{t_{i+n-1} - t_{i-1}} - \frac{nB_{i+1}^{n-1}(x)}{t_{i+n} - t_i}. \tag{4.40}$$

For $n = 1$, (4.40) 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 (4.40) holds for all x except at the knots t_j. By (4.32), (4.28), and (4.31), we have

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

Hence the induction basis holds. Now suppose (4.40) holds $\forall x \in \mathbb{R} \setminus \{t_{i-1}, \ldots, t_{i+n}\}$. Differentiate (4.30), apply the induction hypothesis (4.40), and we have

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^{n+1}(x) = \frac{B_i^n(x)}{t_{i+n} - t_{i-1}} - \frac{B_{i+1}^n(x)}{t_{i+n+1} - t_i} + nC(x), \tag{4.41}$$

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 (4.30). Then (4.41) 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 (4.40) except at the knots. Since $B_i^1 = \hat{B}_i$ is continuous, an easy induction with (4.30) shows that B_i^n is continuous for all $n \ge 1$. Hence the right-hand side of (4.40) is continuous for all $n \ge 2$. Therefore, if $n \ge 2$, $\frac{\mathrm{d}}{\mathrm{d}x} B_i^n(x)$ exists for all $x \in \mathbb{R}$. This completes the proof. □

Corollary 4.26 (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 (4.32). The rest of the proof follows from (4.30) and Theorem 4.25 via an easy induction. □

4.4.4 Marsden's identity

Theorem 4.27 (Marsden's identity)**.** For any $n \in \mathbb{N}$,

$$(t - x)^n = \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n-1}) B_i^n(x), \tag{4.42}$$

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

Proof. For $n = 0$, (4.42) follows from Definition 4.18. Now suppose (4.42) holds. A linear interpolation of the linear function $f(t) = t - x$ is the function itself,

$$t - x = \frac{t - t_{i+n}}{t_{i-1} - t_{i+n}} (t_{i-1} - x) + \frac{t - t_{i-1}}{t_{i+n} - t_{i-1}} (t_{i+n} - x). \tag{4.43}$$

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 (4.43), the third step from replacing i with $i + 1$ in the second summation, and the last step from (4.30). □

Corollary 4.28 (Truncated power functions as summation of B-splines)**.** For any $j \in \mathbb{Z}$ and $n \in \mathbb{N}$,

$$(t_j - x)_+^n = \sum_{i=-\infty}^{j-n} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x). \tag{4.44}$$

Proof. We need to show that the RHS is $(t_j - x)^n$ if $x \le t_j$ and 0 otherwise. Set $t = t_j$ in (4.42) 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 $i = j - n + 1, \ldots, j$, the corresponding term in the summation is zero regardless of x; for each $i \ge j + 1$, Lemma 4.20 implies that $B_i^n(x) = 0$ for all $x \le t_j$. Hence

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

Otherwise $x > t_j$, then Lemma 4.20 implies $B_i^n(x) = 0$ for each $i \le j - n$. This completes the proof. □

4.4.5 Symmetric polynomials

Definition 4.29. 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,

$$\sigma_k(x_1,\ldots,x_n) = \sum_{1\le i_1<\cdots<i_k\le n} x_{i_1}x_{i_2}\cdots x_{i_k}. \qquad (4.45)$$

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

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

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

$$\tau_k(x_1,\ldots,x_n) = \sum_{1\le i_1\le\cdots\le i_k\le n} x_{i_1}x_{i_2}\cdots x_{i_k}. \qquad (4.46)$$

Example 4.11. $\sigma_2(x_1,x_2,x_3)=x_1x_2+x_1x_3+x_2x_3$. 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 4.30. For $k\le n$, the elementary symmetric polynomials satisfy a recursion,

$$\begin{aligned}&\sigma_{k+1}(x_1,\ldots,x_n,x_{n+1})\\ =&\sigma_{k+1}(x_1,\ldots,x_n)+x_{n+1}\sigma_k(x_1,\ldots,x_n). \end{aligned}\qquad (4.47)$$

Proof. The terms in $\sigma_{k+1}(x_1,\ldots,x_n,x_{n+1})$ 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 (4.45), group (a) must be $x_{n+1}\sigma_k(x_1,\ldots,x_n)$ and group (b) must be $\sigma_{k+1}(x_1,\ldots,x_n)$. □

Example 4.12. $\sigma_2(x_1,x_2,x_3)=x_1x_2+x_3(x_1+x_2)$.

Definition 4.31. The *generating function for the elementary symmetric polynomials* is

$$g_{\sigma,n}(z)=\prod_{i=1}^n(1+x_iz)=(1+x_1z)\cdots(1+x_nz) \qquad (4.48)$$

while that for the complete symmetric polynomials is

$$g_{\tau,n}(z)=\prod_{i=1}^n\frac{1}{1-x_iz}=\frac{1}{1-x_1z}\cdots\frac{1}{1-x_nz}. \qquad (4.49)$$

Lemma 4.32 (Generating elementary and complete symmetric polynomials)**.** The elementary and complete symmetric polynomials are related to their generating functions as

$$g_{\sigma,n}(z)=\sum_{k=0}^n\sigma_k(x_1,\ldots,x_n)z^k. \qquad (4.50)$$

$$g_{\tau,n}(z)=\sum_{k=0}^{+\infty}\tau_k(x_1,\ldots,x_n)z^k. \qquad (4.51)$$

Proof. With Lemma 4.30, we can prove (4.50) by an easy induction. For (4.51), (4.49) and the identity

$$\frac{1}{1-x}=\sum_{k=0}^{+\infty}x^k \qquad (4.52)$$

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 $x_1,x_2,\ldots,x_n$. Definition 4.29 then completes the proof. □

Example 4.13.

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

Lemma 4.33 (Recursive relations of complete symmetric polynomials)**.** The complete symmetric polynomials satisfy a recursion,

$$\begin{aligned}&\tau_{k+1}(x_1,\ldots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_1,\ldots,x_n)+x_{n+1}\tau_k(x_1,\ldots,x_n,x_{n+1}).\end{aligned}\qquad (4.53)$$

Proof. (4.49) implies

$$g_{\tau,n+1}=g_{\tau,n}+x_{n+1}zg_{\tau,n+1}. \qquad (4.54)$$

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 4.34 (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.

$$\begin{aligned}&\forall m\in\mathbb{N}^+,\ i\in\mathbb{N},\ \forall n=0,1,\ldots,m,\\ &\tau_{m-n}(x_i,\ldots,x_{i+n})=[x_i,\ldots,x_{i+n}]x^m.\end{aligned}\qquad (4.55)$$

Proof. By Lemma 4.33, we have

$$\begin{aligned}&(x_{n+1}-x_1)\tau_k(x_1,\ldots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_1,\ldots,x_n,x_{n+1})-\tau_{k+1}(x_1,\ldots,x_n)\\ &-x_1\tau_k(x_1,\ldots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_2,\ldots,x_n,x_{n+1})+x_1\tau_k(x_1,\ldots,x_n,x_{n+1})\\ &-\tau_{k+1}(x_1,\ldots,x_n)-x_1\tau_k(x_1,\ldots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_2,\ldots,x_n,x_{n+1})-\tau_{k+1}(x_1,\ldots,x_n).\end{aligned}\qquad (4.56)$$

The rest of the proof is an induction on n. For $n=0$, (4.55) reduces to

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

which is trivially true. Now suppose (4.55) holds for a nonnegative integer $n<m$. Then (4.56) and the induction hypothesis yield

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

which completes the proof. □

4.4.6 B-splines indeed form a basis

Theorem 4.35. Given any $k \in \mathbb{N}$, the monomial x^k can be expressed as a linear combination of B-splines for any fixed $n \geq k$, in the form

$$\binom{n}{k} x^k = \sum_{i=-\infty}^{+\infty} \sigma_k(t_i, \ldots, t_{i+n-1}) B_i^n(x), \qquad (4.57)$$

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

Proof. Lemma 4.32 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 (4.42) 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 4.36 (Partition of Unity)**.**

$$\forall n \in \mathbb{N}, \qquad \sum_{i=-\infty}^{+\infty} B_i^n = 1. \qquad (4.58)$$

Proof. Setting $k = 0$ in Theorem 4.35 yields (4.58). □

Theorem 4.37. The following list of B-splines is a basis of $\mathbb{S}_n^{n-1}(t_1, t_2, \ldots, t_N)$,

$$B_{2-n}^n(x), B_{3-n}^n(x), \ldots, B_N^n(x). \qquad (4.59)$$

Proof. It is easy to verify that

$$\forall t_i \in \mathbb{R}, \quad (x - t_i)_+^n = (x - t_i)^n - (-1)^n (t_i - x)_+^n. \quad (4.60)$$

Then it follows from Theorem 4.27 and Corollary 4.28 that each truncated power function $(x - t_i)_+^n$ can be expressed as a linear combination of B-splines. By Lemma 4.14, 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 4.35 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 4.20 that only those B-splines in the list of (4.59) 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 0.85, Theorem 4.12, and the fact that the length of the list (4.59) is also $n + N - 1$. □

4.4.7 Cardinal B-splines

Definition 4.38. The *cardinal B-spline* of degree n, denoted by $B_{i,\mathbb{Z}}^n$, is the B-spline in Definition 4.18 on the knot set $\mathbb{Z}$.

Corollary 4.39. Cardinal B-splines of the same degree are translates of one another, i.e.

$$\forall x \in \mathbb{R}, \qquad B_{i,\mathbb{Z}}^n(x) = B_{i+1,\mathbb{Z}}^n(x+1). \qquad (4.61)$$

Proof. The recurrence relation (4.30) reduces to

$$B_{i,\mathbb{Z}}^{n+1}(x) = \frac{x - i + 1}{n+1} B_{i,\mathbb{Z}}^n(x) + \frac{i + n + 1 - x}{n+1} B_{i+1,\mathbb{Z}}^n(x). \qquad (4.62)$$

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

Corollary 4.40. A cardinal B-spline is symmetric about the center of its interval of support, i.e.

$$\forall n > 0, \ \forall x \in \mathbb{R}, \qquad B_{i,\mathbb{Z}}^n(x) = B_{i,\mathbb{Z}}^n(2i + n - 1 - x). \ (4.63)$$

Proof. The proof is similar with that of Corollary 4.39. □

Example 4.14. For $t_i = i$, the quadratic B-spline in Example 4.8 simplifies to

$$B_{i,\mathbb{Z}}^2(x) = \begin{cases} \frac{(x-i+1)^2}{2}, & x \in (i-1, i]; \\ \frac{3}{4} - \left(x - (i + \frac{1}{2})\right)^2, & x \in (i, i+1]; \\ \frac{(i+2-x)^2}{2}, & x \in (i+1, i+2]; \\ 0, & \text{otherwise}. \end{cases} \qquad (4.64)$$

It is straightforward to verify Corollaries 4.39 and 4.40. It also follows from (4.64) that

$$B_{i,\mathbb{Z}}^2(j) = \begin{cases} \frac{1}{2} & j \in \{i, i+1\}; \\ 0, & j \in \mathbb{Z} \setminus \{i, i+1\}. \end{cases} \qquad (4.65)$$

Example 4.15. For $t_i = i$, the cubic cardinal B-spline is

$$B_{i,\mathbb{Z}}^3(x) = \begin{cases} \frac{(x-i+1)^3}{6}, & x \in (i-1, i]; \\ \frac{2}{3} - \frac{1}{2}(x - i + 1)(i + 1 - x)^2, & x \in (i, i+1]; \\ B_{i,\mathbb{Z}}^3(2i + 2 - x), & x \in (i+1, i+3); \\ 0, & \text{otherwise}. \end{cases} \qquad (4.66)$$

It follows that

$$B_{i,\mathbb{Z}}^3(j) = \begin{cases} \frac{1}{6}, & j \in \{i, i+2\}; \\ \frac{2}{3}, & j = i+1; \\ 0, & j \in \mathbb{Z} \setminus \{i, i+1, i+2\}. \end{cases} \qquad (4.67)$$

This illustrates Corollary 4.39 that cardinal B-splines have the same shape, i.e., they are invariant under integer translations.

Theorem 4.41. The cardinal B-spline of degree n can be explicitly expressed as

$$B^n_{i,\mathbb{Z}}(x) = \frac{1}{n!}\sum_{k=-1}^{n}(-1)^{n-k}\binom{n+1}{k+1}(k+i-x)^n_+. \quad (4.68)$$

Proof. Theorems 4.23, 3.22, and 3.21 yield

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

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

Corollary 4.42. The value of a cardinal B-spline at an integer j is

$$B^n_{i,\mathbb{Z}}(j) = \frac{1}{n!}\sum_{k=j-i+1}^{n}(-1)^{n-k}\binom{n+1}{k+1}(k+i-j)^n \quad (4.69)$$

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

Proof. This follows directly from Theorem 4.41 and Definition 4.13. □

Corollary 4.43 (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 $1, 2, \ldots, N$ with $S'(1) = f'(1)$ and $S'(N) = f'(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=-1}^{N} a_i B^3_{i,\mathbb{Z}}(x), \quad (4.70)$$

where

$$a_{-1} = a_1 - 2f'(1), \qquad a_N = a_{N-2} + 2f'(N), \quad (4.71)$$

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

$$\begin{aligned} \mathbf{b}^T =& [6f(1) + 2f'(1), 6f(2), \\ & \ldots, 6f(N-1), 6f(N) - 2f'(N)], \end{aligned}$$

$$M = \begin{bmatrix} 4 & 2 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 2 & 4 \end{bmatrix}.$$

Proof. By Theorem 4.37 and Lemma 4.20, we have

$$\begin{aligned} &\forall i = 1, 2, \ldots, N, \\ &\quad f(i) = a_{i-2}B^3_{i-2,\mathbb{Z}}(i) + a_{i-1}B^3_{i-1,\mathbb{Z}}(i) + a_i B^3_{i,\mathbb{Z}}(i). \end{aligned}$$

Then (4.67) yields

$$\forall i = 1, 2, \ldots, N, \qquad a_{i-2} + 4a_{i-1} + a_i = 6f(i), \quad (4.72)$$

which proves the middle $N-2$ equations of $M\mathbf{a} = \mathbf{b}$. By Theorem 4.25, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}B^n_{i,\mathbb{Z}}(x) = B^{n-1}_{i,\mathbb{Z}}(x) - B^{n-1}_{i+1,\mathbb{Z}}(x). \quad (4.73)$$

Differentiate (4.70), apply (4.73), set $x = 1$, apply (4.65) and we have the first identity in (4.71), which, together with (4.72), yields

$$4a_0 + 2a_1 = 2f'(1) + 6f(1);$$

this proves the first equation of $M\mathbf{a} = \mathbf{b}$. The last equation $M\mathbf{a} = \mathbf{b}$ and the second identity in (4.71) 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 (4.71). □

Corollary 4.44. 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 $i = 1, 2, \ldots, N-1$ with end conditions $S(1) = f(1)$ and $S(N) = f(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=0}^{N} a_i B^2_{i,\mathbb{Z}}(x), \quad (4.74)$$

where

$$a_0 = 2f(1) - a_1, \qquad a_N = 2f(N) - a_{N-1}, \quad (4.75)$$

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

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

$$M = \begin{bmatrix} 5 & 1 & & & \\ 1 & 6 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 6 & 1 \\ & & & 1 & 5 \end{bmatrix}.$$

Proof. It follows from Lemma 4.20 and Definition 4.38 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

$$f(t_i) = a_{i-1}B^2_{i-1,\mathbb{Z}}(t_i) + a_i B^2_{i,\mathbb{Z}}(t_i) + a_{i+1}B^2_{i+1,\mathbb{Z}}(t_i). \quad (4.76)$$

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 4.37! By Theorem 4.41, 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 4.40. Then Corollary 4.39 and (4.76) yield

$$a_{i-1}+6a_i+a_{i+1}=8f(t_i), \tag{4.77}$$

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 4.8 yields

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

and this proves the first identity in (4.75). Also, the above equation and (4.77) 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. □

4.5 Curve fitting via splines

Definition 4.45. 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 4.46. The *tangent vector of a curve* γ is its first derivative

$$\gamma':=\frac{\mathrm{d}\gamma}{\mathrm{d}s} \tag{4.78}$$

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

Definition 4.47. A *unit-speed curve* is a curve whose tangent vector has unit length at each of its points.

Definition 4.48. 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 4.49. The *arc-length* of a curve starting at the point $\gamma(t_0)$ is defined as

$$s_\gamma(t)=\int_{t_0}^{t}\|\gamma'(u)\|_2\mathrm{d}u. \tag{4.79}$$

Definition 4.50. A map $X\mapsto Y$ is a *homeomorphism* if it is continuous and bijective and its inverse is continuous; then the two sets X and Y are said to be *homeomorphic.*

Definition 4.51. 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 4.52. A reparametrization of a regular curve is unit-speed if and only if it is based on the arc-length.

Definition 4.53. 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 4.54. 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 4.55. For a unit-speed curve γ, its *signed curvature* is defined as

$$\kappa_s:=\gamma''\cdot\mathbf{n}_s. \tag{4.80}$$

Definition 4.56. The *cumulative chordal length*s associated with a sequence of n points

$$\{\mathbf{x}_i\in\mathbb{R}^{\mathrm{D}}:i=1,2,\ldots,n\} \tag{4.81}$$

are the n real numbers,

$$t_i=\begin{cases}0, & i=1;\\ t_{i-1}+\|\mathbf{x}_i-\mathbf{x}_{i-1}\|_2, & i>1,\end{cases} \tag{4.82}$$

where $\|\cdot\|_2$ denotes the Euclidean 2-norm.

4.6 Problems

4.6.1 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 $s\in\mathbb{S}^1_2$.

(a) Why an additional condition is needed to determine s uniquely?

(b) Define $m_i=s'(x_i)$ and $p_i=s\big|_{[x_i,x_{i+1}]}$. 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 $B^2_i(x)$.

(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 $B_1^2(x)$ for $t_i = i$.

VI. Verify Theorem 4.23 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.

4.6.2 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 Corollaries 4.43 and 4.44, 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 Corollary 4.43 and $t_i = i - \frac{11}{2}$, $i = 1, \ldots, 10$ for Corollary 4.44, 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:

$$x^2 + \left(\frac{3}{2}y - \sqrt{|x|}\right)^2 = 3. \tag{4.83}$$

Write a program to plot the heart. The parameter of the curve should be the *cumulative chordal length* defined in (4.82). 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 (4.37) 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 4.10.

Chapter 5

Multivariate Interpolation

Definition 5.1. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ denote N distinct points in $\mathbb{R}^{\mathrm{D}}$, and $\phi_1, \phi_2, \ldots, \phi_N$ denote N linearly independent continuous functions $\mathbb{R}^{\mathrm{D}} \mapsto \mathbb{R}$. The *multivariate interpolation problem of a given function* $f : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ seeks $a_1, a_2, \ldots, a_N \in \mathbb{R}$ such that

$$\forall j = 1, 2, \ldots, N, \qquad \sum_{i=1}^{N} a_i \phi_i(\mathbf{x}_j) = f(\mathbf{x}_j). \tag{5.1}$$

Definition 5.2. The *sites* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ of the multivariate interpolation problem are said to be *poised* with respect to the basis functions $\phi_1, \phi_2, \ldots, \phi_N$ iff the *sample matrix*

$$M = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_N(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_N(\mathbf{x}_N) \end{bmatrix} \tag{5.2}$$

is non-singular.

Theorem 5.3. The multivariate interpolation problem has a unique solution if and only if its sites are poised.

Example 5.1. Suppose that the values of a function $f : \mathbb{R}^2 \to \mathbb{R}$ are known at the sites $(1,0), (-1,0), (0,1)$, and $(0,-1)$. For the basis functions $1, x, y, xy$, the sample matrix

$$M = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix}$$

is clearly singular, and hence this multivariate interpolation problem does not admit a unique solution.

5.1 Rectangular grids

Theorem 5.4 (Lagrange formula for rectangular grids)**.** Given two subsets of $\mathbb{R}$ as $X = \{x_0, x_1, \ldots, x_m\}$ and $Y = \{y_0, y_1, \ldots, y_n\}$, the multivariate interpolation problem on the rectangular grid $X \times Y$ with the set of basis functions

$$\Phi = \{x^i y^j : i = 0, 1, \ldots, m;\ j = 0, 1, \ldots, n\} \tag{5.3}$$

is solved by the unique solution

$$p(x,y) = \sum_{i=0}^{m} \sum_{j=0}^{n} f(x_i, y_j) L_i(x) M_j(y), \tag{5.4}$$

where $L_i(x)$ and $M_j(y)$ are the elementary Lagrange interpolation polynomials defined in (3.9).

Proof. Define a *blending function*

$$\xi(x,y) = \sum_{i=0}^{m} f(x_i, y) L_i(x) \tag{5.5}$$

and apply the uniqueness of interpolating polynomials (Theorem 3.4) and the Lagrange formula (Definition 3.8) dimension-by-dimension in a recursive manner. □

Corollary 5.5. The unique solution in Theorem 5.4 can also be expressed via divided differences as

$$p(x,y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \pi_i(x) \pi_j(y) [x_0, \ldots, x_i][y_0, \ldots, y_j] f(x,y), \tag{5.6}$$

where π is defined in (3.10), and each divided difference acts on the function f with the other coordinate fixed.

Proof. Theorem 5.4 and the definition of divided differences (Definition 3.11) yield (5.6). □

Example 5.2. Find the unique polynomial that interpolates the following data on a square grid.

(x,y)	$(-1,-1)$	$(-1,1)$	$(1,-1)$	$(1,1)$
$f(x,y)$	1	5	-5	3

To apply (5.6), we calculate

$$\begin{aligned} [-1,1]_x f(x,-1) &= \frac{-5-1}{1+1} = -3, \\ [-1,1]_y f(-1,y) &= \frac{5-1}{1+1} = 2, \\ [-1,1]_x f(x,1) &= \frac{3-5}{1+1} = -1 \\ [-1,1]_x [-1,1]_y f(x,y) &= [-1,1]_x \frac{f(x,1) - f(x,-1)}{1+1} \\ &= \frac{-1+3}{1+1} = 1. \end{aligned}$$

It follows that the unique interpolating polynomial is

$$\begin{aligned}p(x,y) &= 1-3(x+1)+2(y+1)+(x+1)(y+1)\\ &= 1-2x+3y+xy.\end{aligned}$$

Lemma 5.6. For $k,\ell\in\mathbb{N}^+$, define

$$\left(\Delta_x^k\Delta_y^\ell\right)f(x,y)=\Delta_x^k(\Delta_y^\ell f(x,y)). \tag{5.7}$$

Then the two difference operators Δ_x^k and Δ_y^ℓ commute,

$$\Delta_x^k\Delta_y^\ell f(x,y)=\Delta_y^\ell\Delta_x^k f(x,y). \tag{5.8}$$

Proof. (5.8) follows from (5.7) and the definition of forward differences (Definition 3.19). □

Corollary 5.7. Consider a rectangular grid with uniform spacing along each dimension,

$$\begin{aligned}\forall i&=0,1,\ldots,m, & x_i&=x_0+ih_x;\\ \forall j&=0,1,\ldots,n, & y_j&=y_0+jh_y.\end{aligned}$$

The unique solution in Theorem 5.4 can also be expressed via forward differences as

$$p(x_0+sh_x,y_0+th_y)=\sum_{i=0}^{m}\sum_{j=0}^{n}\binom{s}{i}\binom{t}{j}\Delta_x^i\Delta_y^j f(x_0,y_0), \tag{5.10}$$

where $\binom{s}{i}$ and $\binom{t}{j}$ are defined in (3.31).

Proof. (5.10) follows from Newton's forward difference formula (Theorem 3.23), Lemma 5.6, and the structure of rectangular grids. □

5.2 Triangular grids

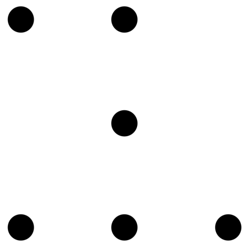
5.2.1 Triangular lattices in the plane

Definition 5.8. A *triangular lattice of degree n in two dimensions* is a set of isolated points in $\mathbb{R}^2$,

$$\mathcal{T}_2^n=\{(x_i,y_j): i,j\ge 0, i+j\le n\}, \tag{5.11}$$

where x_i's are $n+1$ distinct x-coordinates and y_j's are $n+1$ distinct y-coordinates.

Example 5.3. The constraints in (5.11) are not on the coordinates, but on their *indices*. Hence a triangular lattice might have a shape that does not look like a triangle.



For example, the above triangular lattice

$$\mathcal{T}_2^2=\{(0,0),(0,2),(1,0),(1,1),(1,2),(2,0)\}$$

has distinct coordinates $x_0=1$, $x_1=0$, $x_2=2$ and $y_0=0$, $y_1=2$, $y_2=1$.

Theorem 5.9. A triangular lattice $\mathcal{T}_2^n$ is poised with respect to bivariate polynomials of degree no more than n

$$\Phi_2^n=\{1,x,y,x^2,xy,y^2,\ldots,x^n,x^{n-1}y,\ldots,xy^{n-1},y^n\}; \tag{5.12}$$

the corresponding sample matrix M_2 satisfies

$$\det M_2=C\psi_n(x)\psi_n(y), \tag{5.13}$$

where C is a nonzero constant and $\psi_n(x)$ is a polynomial in terms of the $n+1$ distinct coordinates x_i's,

$$\psi_n(x):=\prod_{i=1}^{n}\prod_{\ell=0}^{i-1}(x_i-x_\ell)^{n+1-i}. \tag{5.14}$$

Proof. For any fixed i,ℓ with $i>\ell$, replacing (x_i,y_j) with (x_ℓ,y_j) in $\mathcal{T}_2^n$ makes the corresponding sample matrix singular for each $j=0,1,\ldots,n-i$; furthermore, j cannot exceed $n-i$ because (5.11) dictates $i+j\le n$. Therefore the number of this type of replacements is $n-i+1$, and hence the exponent of (x_i-x_ℓ) in (5.14) is $n-i+1$.

Now we vary ℓ while keeping i fixed. Since there are i indices less than i, the term $\prod_{\ell=0}^{i-1}(x_i-x_\ell)^{n+1-i}$ contributes to a total degree of $i(n+1-i)$ in terms of the $n+1$ coordinates $x_0,\ldots,x_n$. It follows that the total degree of $\psi_n(x)$ is

$$\sum_{i=1}^{n}i(n+1-i)=(n+1)\sum_{i=1}^{n}i-\sum_{i=1}^{n}i^2=\frac{n(n+1)(n+2)}{6}, \tag{5.15}$$

where the second equality is proven by an easy induction.

Similarly, $\det M_2$ must contain a factor of $\psi_n(y)$, of which the total degree is also (5.15).

From the other viewpoint of Definition 0.146, the determinant of the sample matrix M_2 in Definition 5.2 is also a polynomial in terms of the variables $x_0,x_1,\ldots,x_n$ and $y_0,y_1,\ldots,y_n$, with each monomial being a product of all basis functions in (5.12) evaluated at some point (x_i,y_j). Hence the total degree of $\det M_2$ in the variables $x_0,x_1,\ldots,x_n$ and $y_0,y_1,\ldots,y_n$ is

$$\sum_{i=1}^{n}i(i+1)=\frac{n(n+1)(n+2)}{3}, \tag{5.16}$$

where i refers to the degree of a monomial and $i+1$ the number of monomials of degree i, c.f. (5.12).

The proof is completed by the fact that if two complete polynomials have the same variables, the same total degree, and the same factors in terms of the same variables, then one is a constant multiple of the other. □

Corollary 5.10. A polynomial of the form

$$p_n(x,y)=\sum_{k=0}^{n}\sum_{r=0}^{k}c_{r,k-r}x^ry^{k-r} \tag{5.17}$$

uniquely interpolates any continuous function f on $\mathcal{T}_2^n$.

Proof. This follows from Theorems 5.9 and 5.4. □

Theorem 5.11. For any scalar function f whose domain includes $\mathcal{T}_2^n$, we have,

$$\forall m = 0, 1, \ldots, n, \quad f(x, y) = p_m(x, y) + r_m(x, y), \tag{5.18}$$

where the polynomial $p_m(x, y)$ interpolates $f(x, y)$ on $\mathcal{T}_2^m$ with $r_m(x, y)$ being the remainder,

$$p_m(x, y) = \begin{cases} [x_0][y_0]f = f(x_0, y_0), & m = 0; \\ p_{m-1}(x, y) + q_m(x, y), & m > 0, \end{cases} \tag{5.19a}$$

$$q_m(x, y) = \sum_{k=0}^{m} \pi_k(x)\pi_{m-k}(y)[x_0, \ldots, x_k][y_0, \ldots, y_{m-k}]f, \tag{5.19b}$$

$$\begin{aligned} r_m(x, y) = &\sum_{k=0}^{m} \pi_{k+1}(x)\pi_{m-k}(y)[x, x_0, \ldots, x_k][y_0, \ldots, y_{m-k}]f \\ &+ \pi_{m+1}(y)[x][y, y_0, \ldots, y_m]f. \end{aligned} \tag{5.19c}$$

Proof. The polynomial $p_m(x, y)$ clearly interpolates $f(x, y)$ on $\mathcal{T}_2^m$ because, for each $(x_i, y_j) \in \mathcal{T}_2^m$, all the $m + 2$ terms of $r_m(x, y)$ in (5.19c) are identically zero. The total degree of $p_m(x, y)$ is m while that of $r_m(x, y)$ is $m + 1$. It is easily verified that

$$f(x, y) = f(x_0, y_0) + (x - x_0)[x, x_0][y_0]f + (y - y_0)[x][y, y_0]f.$$

Hence (5.18) and (5.19) hold for the induction basis $m = 0$. Assume that (5.18) holds for $m \geq 0$. For the inductive step, we define

$$\begin{aligned} S_1 &= \sum_{k=0}^{m} \pi_{k+1}(x)\pi_{m-k}(y)[x_0, \ldots, x_{k+1}][y_0, \ldots, y_{m-k}]f, \\ S_2 &= \sum_{k=0}^{m} \pi_{k+2}(x)\pi_{m-k}(y)[x, x_0, \ldots, x_{k+1}][y_0, \ldots, y_{m-k}]f, \\ T_1 &= \pi_{m+1}(y)[x_0][y_0, \ldots, y_{m+1}]f, \\ T_2 &= \pi_1(x)\pi_{m+1}(y)[x, x_0][y_0, \ldots, y_{m+1}]f, \\ T_3 &= \pi_{m+2}(y)[x][y, y_0, \ldots, y_{m+1}]f. \end{aligned}$$

Utilizing (3.17), it is not difficult to prove

$$\begin{aligned} S_1 + T_1 &= q_{m+1}(x, y) = p_{m+1}(x, y) - p_m(x, y), \\ S_2 + T_2 + T_3 &= r_{m+1}(x, y), \\ r_m(x, y) &= r_{m+1}(x, y) + q_{m+1}(x, y). \end{aligned}$$

Hence we have

$$r_{m+1}(x, y) + p_{m+1}(x, y) = r_m(x, y) + p_m(x, y) = f(x, y),$$

which completes the inductive proof. □

Corollary 5.12. The interpolating polynomial on $\mathcal{T}_2^n$ in Theorem 5.11 can also be expressed as

$$p(x, y) = \sum_{m=0}^{n} \sum_{k=0}^{m} \pi_k(x)\pi_{m-k}(y)[x_0, \ldots, x_k][y_0, \ldots, y_{m-k}]f. \tag{5.20}$$

Proof. This follows directly from (5.19a) and (5.19b). □

The rest of this section concerns a special type of triangular grids.

Corollary 5.13. For the *principal lattice*

$$\mathcal{P}_2^n = \{(i, j) \in \mathbb{N}^2 : i + j \leq n\}, \tag{5.21}$$

the unique interpolating polynomial in Theorem 5.11 can be expressed as

$$p(x, y) = \sum_{m=0}^{n} \sum_{k=0}^{m} \binom{x}{k}\binom{y}{m-k} \Delta_x^k \Delta_y^{m-k} f(0, 0). \tag{5.22}$$

Proof. This follows from Corollary 5.12 and Theorem 3.22. □

Theorem 5.14 (Lagrange formula for principal lattices)**.** The unique interpolation polynomial on the principal lattice (5.21) can be expressed as

$$p_n(x, y) = \sum_{i,j} f(i, j) L_{i,j}(x, y), \tag{5.23}$$

where $(i, j) \in \mathcal{P}_2^n$ and the fundamental polynomial is

$$\begin{aligned} L_{i,j}(x, y) &= \prod_{s=0}^{i-1} \frac{x - s}{i - s} \prod_{s=0}^{j-1} \frac{y - s}{j - s} \prod_{s=i+j+1}^{n} \frac{x + y - s}{i + j - s} \\ &= \binom{x}{i}\binom{y}{j}\binom{n - x - y}{n - i - j}. \end{aligned} \tag{5.24}$$

Proof. Clearly we only need to show

$$L_{i,j}(x, y) = \begin{cases} 1 & \text{if } x = i, y = j; \\ 0 & \text{otherwise.} \end{cases}$$

It is trivial to verify $L_{i,j}(i, j) = 1$ in (5.24). As for the second clause, the following families of straight lines

- $x = 0, 1, \ldots, i - 1$,
- $y = 0, 1, \ldots, j - 1$,
- $x + y = i + j + 1, i + j + 2, \ldots, n$,

contains all sites of $\mathcal{P}_2^n \setminus \{(i, j)\}$, hence at least one factor in (5.24) is zero at any site. □

Example 5.4. The principal lattice $\mathcal{P}_2^2$ contains six sites and the corresponding interpolating polynomial is

$$\begin{aligned} p_2(x, y) = &\frac{1}{2}(x + y - 1)(x + y - 2)f(0, 0) + xyf(1, 1) \\ &- x(x + y - 2)f(1, 0) + \frac{1}{2}x(x - 1)f(2, 0) \\ &- y(x + y - 2)f(0, 1) + \frac{1}{2}y(y - 1)f(0, 2) \end{aligned}$$

The evaluation of $p_2(x, y)$ at the centroid of the triangle with vertices $(0, 0)$, $(2, 0)$, and $(0, 2)$ yields

$$p_2\left(\frac{2}{3}, \frac{2}{3}\right) = \frac{1}{3}(4\alpha - \beta),$$

where β is the mean of the values of f on the triangle vertices and α that on the other sites.

Theorem 5.15 (Neville-Aitken formula for principal lattices)**.** Define $p_0^{[i,j]} = f(i,j)$ and denote by $p_k^{[i,j]}(x,y)$ the unique interpolating polynomial of total degree k for the function $f(x,y)$ on the principal lattice

$$\mathcal{P}_k^{[i,j]} = \{(i+r, j+s) : r,s \ge 0,\ r+s \le k\}. \tag{5.25}$$

Then, for $k \ge 0$ and $i,j \ge 0$, we have

$$\begin{aligned} p_{k+1}^{[i,j]}(x,y) = &\frac{i+j+k+1-x-y}{k+1} p_k^{[i,j]}(x,y) \\ &+ \frac{x-i}{k+1} p_k^{[i+1,j]}(x,y) + \frac{y-j}{k+1} p_k^{[i,j+1]}(x,y). \end{aligned} \tag{5.26}$$

Proof. The induction basis $k=0$ clearly holds because $\mathcal{P}_0^{[i,j]} = \{(i,j)\}$. Suppose $p_k^{[i,j]}(x,y)$ interpolates $f(x,y)$ on $\mathcal{P}_k^{[i,j]}$ for some $k \ge 0$ and all $i,j \in \mathbb{N}$. By (5.25), we have

$$\begin{aligned} \mathcal{I}_p :=& \mathcal{P}_k^{[i,j]} \cap \mathcal{P}_k^{[i+1,j]} \cap \mathcal{P}_k^{[i,j+1]} \\ =& \{(i+r, j+s) : r,s \ge 1, r+s \le k\}. \end{aligned}$$

The induction hypothesis implies that, $\forall (\ell, m) \in \mathcal{I}_p$,

$$p_k^{[i,j]}(\ell,m) = p_k^{[i+1,j]}(\ell,m) = p_k^{[i,j+1]}(\ell,m) = f(\ell,m),$$

which, together with (5.26), yields

$$\forall (\ell,m) \in \mathcal{I}_p, \qquad p_{k+1}^{[i,j]}(\ell,m) = f(\ell,m).$$

Similarly, we have

$$\begin{aligned} \forall m-j = 1,2,\ldots,k,\ \ & p_k^{[i,j]}(i,m) = p_k^{[i,j+1]}(i,m) = f(i,m), \\ \forall \ell-i = 1,2,\ldots,k,\ \ & p_k^{[i,j]}(\ell,j) = p_k^{[i+1,j]}(\ell,j) = f(\ell,j), \\ \forall (\ell,m) \in \mathcal{L},\ \ & p_k^{[i+1,j]}(\ell,m) = p_k^{[i,j+1]}(\ell,m) = f(\ell,m), \end{aligned}$$

where $\mathcal{L} = \{(i+r, j+s) : r,s \ge 0, r+s = k+1\}$.

It follows from the above three equations and (5.26) that $\forall y-j = 1,2,\ldots,k$, $\forall x-i = 1,2,\ldots,k$,

$$\begin{aligned} \ell = i \Rightarrow\ \ & p_{k+1}^{[i,j]}(\ell,y) = p_k^{[i,j]}(\ell,y) = f(\ell,y); \\ m = j \Rightarrow\ \ & p_{k+1}^{[i,j]}(x,m) = p_k^{[i,j]}(x,m) = f(x,m), \\ & p_{k+1}^{[i,j]}(i,j) = p_k^{[i,j]}(i,j) = f(i,j), \end{aligned}$$

and $\forall (\ell,m) \in \mathcal{L}$,

$$\begin{aligned} p_{k+1}^{[i,j]}(\ell,m) &= \frac{\ell-i}{k+1} p_k^{[i+1,j]}(\ell,m) + \frac{m-j}{k+1} p_k^{[i,j+1]}(\ell,m) \\ &= \frac{\ell-i}{k+1} f(\ell,m) + \frac{m-j}{k+1} f(\ell,m) \\ &= \frac{\ell+m-i-j}{k+1} f(\ell,m) = f(\ell,m). \end{aligned}$$

Hence $p_{k+1}^{[i,j]}(x,y)$ also interpolates $f(x,y)$ on $\mathcal{P}_{k+1}^{[i,j]} \setminus \mathcal{I}_p$ as any site in it satisfies $\ell = i$ or $m = j$ or $(\ell,m) \in \mathcal{L}$. In summary, $p_{k+1}^{[i,j]}(x,y)$ interpolates $f(x,y)$ on $\mathcal{P}_{k+1}^{[i,j]}$.

The total degree of $p_k^{[i,j]}(x,y)$ being k can also be proved by an easy induction. □

Example 5.5. Use formula (5.26) to obtain the last equation in Example 5.4.

5.2.2 Triangular lattices in D dimensions

Notation 8. The first $n+1$ natural numbers is denoted by

$$\mathbb{Z}_n := \{0, 1, \ldots, n\} \tag{5.27}$$

and the first n positive integers by

$$\mathbb{Z}_n^+ := \{1, \ldots, n\}. \tag{5.28}$$

Definition 5.16. A subset $\mathcal{T}_{\mathrm{D}}^n$ of $\mathbb{R}^{\mathrm{D}}$ is called a *triangular lattice of degree n in* D *dimensions* if there exists $n+1$ coordinates for each of the D dimensions,

$$\begin{bmatrix} p_{1,0} & p_{1,1} & \cdots & p_{1,n} \\ p_{2,0} & p_{2,1} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{\mathrm{D},0} & p_{\mathrm{D},1} & \cdots & p_{\mathrm{D},n} \end{bmatrix} \in \mathbb{R}^{\mathrm{D}\times(n+1)}, \tag{5.29}$$

such that

$$\mathcal{T}_{\mathrm{D}}^n = \left\{ (p_{1,k_1}, p_{2,k_2}, \ldots, p_{D,k_D}) \in \mathbb{R}^{\mathrm{D}} :\ k_i \in \mathbb{Z}_n;\ \sum_{i=1}^{\mathrm{D}} k_i \le n \right\}, \tag{5.30}$$

where $p_{i,j}$ denotes the jth coordinate of the ith variable p_i.

Example 5.6. For D = 2, Definition 5.16 reduces to Definition 5.8 since (5.30) simplifies to

$$\mathcal{T}_2^n = \{(p_{1,k_1}, p_{2,k_2}) :\ k_1, k_2 \ge 0;\ k_1 + k_2 \le n\},$$

which is the same as (5.11).

Lemma 5.17. The cardinality of a triangular lattice is

$$\#\mathcal{T}_{\mathrm{D}}^n = \binom{n+\mathrm{D}}{\mathrm{D}} = \sum_{i=0}^{n} \#\mathcal{T}_{\mathrm{D}}^{=i} = \sum_{i=0}^{n} \binom{i+\mathrm{D}-1}{\mathrm{D}-1}, \tag{5.31}$$

where

$$\mathcal{T}_{\mathrm{D}}^{=m} := \left\{ (p_{1,k_1}, p_{2,k_2}, \ldots, p_{\mathrm{D},k_{\mathrm{D}}}) : k_i \ge 0; \sum_{i=1}^{\mathrm{D}} k_i = m \right\} \tag{5.32}$$

and its cardinality is

$$\#\mathcal{T}_{\mathrm{D}}^{=m} = \#\mathcal{T}_{\mathrm{D}}^{m} - \#\mathcal{T}_{\mathrm{D}}^{m-1} = \binom{m+\mathrm{D}-1}{\mathrm{D}-1}. \tag{5.33}$$

Proof. As the only nontrivial constraint on $\mathcal{T}_{\mathrm{D}}^n$ in (5.30), the sum of the D non-negative integers $i_1, \ldots, i_{\mathrm{D}}$ cannot exceed n. Hence $\#\mathcal{T}_{\mathrm{D}}^n$ equals the number of possibilities of placing n indistinguishable balls into D + 1 urns, with the first D urns corresponding to the D dimensions, respectively, and the last urn accounting for the deficit of $\sum_k i_k$ from n. This ball-urn problem is further equivalent to choosing D balls from $n+\mathrm{D}$ balls in a row because the chosen D balls divide the rest n balls into D+1 consecutive arrays of balls, each of which corresponds to an urn. This proves the first equality in (5.31).

As for the second equality in (5.31), definitions (5.30) and (5.32) implies $\mathcal{T}_{\mathrm{D}}^n = \cup_{m=0}^n \mathcal{T}_{\mathrm{D}}^{=m}$ and that two subsets $\mathcal{T}_{\mathrm{D}}^{=m}$ and $\mathcal{T}_{\mathrm{D}}^{=n}$ are disjoint if and only if $m \ne n$.

An easy induction shows (5.33), which further implies the third equality in (5.31). □

Definition 5.18. The D-*variate polynomials of degree no more than* n form the set

$$\Phi_{\mathrm{D}}^{n}=\left\{p_1^{e_1}p_2^{e_2}\dots p_{\mathrm{D}}^{e_{\mathrm{D}}}: e_i\ge 0; \sum_{i=1}^{\mathrm{D}} e_i\le n\right\}, \tag{5.34}$$

and the D-*variate polynomials of degree* m form the set

$$\Phi_{\mathrm{D}}^{=m}=\left\{p_1^{e_1}p_2^{e_2}\dots p_{\mathrm{D}}^{e_{\mathrm{D}}}: e_i\ge 0; \sum_{i=1}^{\mathrm{D}} e_i= m\right\}. \tag{5.35}$$

Lemma 5.19. The results on triangular lattices in Lemma 5.17 also hold for sets of multivariate polynomials. More precisely, (5.31) and (5.33) still hold when the symbol $\mathcal{T}$ is replaced with the symbol Φ.

Proof. This follows from a natural bijection between Φ_{D}^n and $\mathcal{T}_{\mathrm{D}}^n$, the restriction of which is also a bijection between $\Phi_{\mathrm{D}}^{=n}$ and $\mathcal{T}_{\mathrm{D}}^{=n}$. □

Lemma 5.20. For any positive integers D and n, we have

$$(\mathrm{D}+1)\sum_{j=1}^{n} j\binom{n-j+\mathrm{D}}{\mathrm{D}}=\sum_{j=1}^{n} j\binom{j+\mathrm{D}}{\mathrm{D}}. \tag{5.36}$$

Proof. For $n=1$, both sides reduce to D+1. Suppose (5.36) holds. Then the inductive step also holds because

$$\begin{aligned}&(\mathrm{D}+1)\sum_{j=1}^{n+1} j\binom{n+1-j+\mathrm{D}}{\mathrm{D}}\\ =&(\mathrm{D}+1)\sum_{i=0}^{n}(i+1)\binom{n-i+\mathrm{D}}{\mathrm{D}}\\ =&(\mathrm{D}+1)\sum_{i=0}^{n} i\binom{n-i+\mathrm{D}}{\mathrm{D}}+(\mathrm{D}+1)\sum_{j=0}^{n}\binom{j+\mathrm{D}}{\mathrm{D}}\\ =&\sum_{j=1}^{n} j\binom{j+\mathrm{D}}{\mathrm{D}}+(\mathrm{D}+1)\binom{n+\mathrm{D}+1}{\mathrm{D}+1}\\ =&\sum_{j=1}^{n+1} j\binom{j+\mathrm{D}}{\mathrm{D}},\end{aligned}$$

where the first two steps follow from variable substitutions, the third step from the induction hypothesis and the well-known identity $\sum_{j=0}^{n}\binom{\mathrm{D}+j}{\mathrm{D}}=\binom{\mathrm{D}+n+1}{\mathrm{D}+1}$, and the last step from

$$\begin{aligned}&(\mathrm{D}+1)\binom{n+\mathrm{D}+1}{\mathrm{D}+1}=(\mathrm{D}+1)\frac{(n+\mathrm{D}+1)!}{(\mathrm{D}+1)!n!}\\ =&(n+1)\frac{(n+\mathrm{D}+1)!}{\mathrm{D}!(n+1)!}=(n+1)\binom{n+\mathrm{D}+1}{\mathrm{D}}.\end{aligned}$$ □

Theorem 5.21. A triangular lattice $\mathcal{T}_{\mathrm{D}}^n$ is poised with respect to the D-variate polynomials of degree no more than n; the corresponding sample matrix M_{D} satisfies

$$\det M_{\mathrm{D}}=C\prod_{k=1}^{\mathrm{D}}\psi_n(p_k) \tag{5.37}$$

where C is a nonzero constant and $\psi_n(p_k)$ is a polynomial in terms of the $n+1$ distinct coordinates of the variable p_k,

$$\psi_n(p_k):=\prod_{i_k=1}^{n}\prod_{\ell=0}^{i_k-1}(p_{k,i_k}-p_{k,\ell})^{\alpha(i_k)}; \tag{5.38}$$

$$\alpha(i_k):=\binom{n-i_k+\mathrm{D}-1}{\mathrm{D}-1}. \tag{5.39}$$

Proof. We follow the steps in the proof of Theorem 5.9, with more complicated book-keeping on the combinatorics.

Consider the kth variable p_k. For any fixed i_k and ℓ with $i_k>\ell$, replacing the coordinate p_{k,i_k} with $p_{k,\ell}$ in a point

$$\mathbf{p}_k:=(p_{1,i_1},\dots,p_{k,i_k},\dots,p_{\mathrm{D},i_{\mathrm{D}}})\in\mathcal{T}_{\mathrm{D}}^n \tag{5.40}$$

makes the corresponding sample matrix M_{D} singular. Furthermore, when the coordinate index of the kth variable p_k is fixed at i_k in $\mathcal{T}_{\mathrm{D}}^n$, the cardinality of $\mathcal{T}_{\mathrm{D}}^n$ reduces to $\#\mathcal{T}_{\mathrm{D}-1}^{n-i_k}$, which, by Lemma 5.17, must equal $\alpha(i_k)$. In other words,

$$\#\left\{(p_{1,i_1},\dots,p_{k,i_k},\dots,p_{\mathrm{D},i_{\mathrm{D}}})\in\mathcal{T}_{\mathrm{D}}^n:\ i_k \text{ is fixed}\right\}$$

equals the cardinality of a triangular lattice of degree $n-i_k$ in $\mathrm{D}-1$ dimensions because an index of i_k has been consumed from the total index n and one of the D dimensions has already been fixed; see Lemma 5.25. Therefore, the number of possible $\mathbf{p}_k$'s that admit the replacement of p_{k,i_k} with $p_{k,\ell}$ is $\alpha(i_k)$, and this justifies the exponent of $(p_{k,i_k}-p_{k,\ell})$ in (5.38).

Now we vary ℓ while keeping i_k fixed. Since there are i_k indices less than i_k, the term $\prod_{\ell=0}^{i_k-1}(p_{k,i_k}-p_{k,\ell})^{\alpha(i_k)}$ contributes to a total degree $i_k\alpha(i_k)$ in terms of the $n+1$ coordinates $p_{k,0},\dots,p_{k,n}$. It follows that $\psi_n(p_k)$ must be a factor of $\det M_{\mathrm{D}}$ and the total degree of $\psi_n(p_k)$ is

$$\sum_{i_k=1}^{n} i_k\alpha(i_k)=\sum_{j=1}^{n} j\binom{n-j+\mathrm{D}-1}{\mathrm{D}-1}.$$

Similarly, $\det M_{\mathrm{D}}$ must contain a factor of $\psi_n(p_j)$ for each variable p_j, $j=1,\dots,\mathrm{D}$. Hence the total degree of $\det M_{\mathrm{D}}$ is at least

$$\xi:=\mathrm{D}\sum_{j=1}^{n} j\binom{n-j+\mathrm{D}-1}{\mathrm{D}-1}.$$

From the other viewpoint of Definition 0.146, the determinant of the sample matrix M_{D} is also a polynomial in terms of the coordinates $p_{1,0},p_{1,1},\dots,p_{1,n}$, ..., $p_{\mathrm{D},0},p_{\mathrm{D},1},\dots,p_{\mathrm{D},n}$, with each monomial being a product of all basis functions in (5.34) evaluated at some point $(p_{1,i_1},p_{2,i_2},\dots,p_{n,i_n})$. By Lemma 5.19 and (5.33), the total degree of $\det M_{\mathrm{D}}$ equals

$$\eta:=\sum_{i=1}^{n} i\binom{i+\mathrm{D}-1}{\mathrm{D}-1},$$

where i refers to the degree of monomials in the subset $\Phi_{\mathrm{D}}^{=i}$ and $\binom{i+\mathrm{D}-1}{\mathrm{D}-1}$ the cardinality of $\Phi_{\mathrm{D}}^{=i}$.

Lemma 5.20 implies $\xi=\eta$. Hence the terms in $\prod_{k=1}^{\mathrm{D}}\psi_n(p_k)$ constitute all the non-constant factors of $\det M_{\mathrm{D}}$, which yields (5.37). □

5.3 Poised-lattice generation (PLG)

5.3.1 Formulating the PLG problem

Notation 9. For a *fixed* coordinate system of $\mathbb{Z}_n^{\mathrm{D}}$, the *set of all triangular lattices of degree* n in $\mathbb{Z}_n^{\mathrm{D}}$ is denoted by

$$\mathcal{X} := \left\{\mathcal{T}_{\mathcal{D}}^n : \mathcal{T}_{\mathcal{D}}^n \subset \mathbb{Z}_n^{\mathrm{D}}\right\}, \tag{5.41}$$

where the D-dimensional cube of size $n+1$ is denoted by

$$\mathbb{Z}_n^{\mathrm{D}} := (\mathbb{Z}_n)^{\mathrm{D}} = \{0, 1, \ldots, n\}^{\mathrm{D}}. \tag{5.42}$$

Definition 5.22. Given $K \subseteq \mathbb{Z}_n^{\mathrm{D}}$ and $\mathbf{q} \in K$, the *poised-lattice generation problem* seeks $\mathcal{T} \in \mathcal{X}$ such that $\mathbf{q} \in \mathcal{T}$ and $\mathcal{T} \subseteq K$.

5.3.2 A group action on triangular lattices

Definition 5.23. The *principal lattice* of degree n in $\mathbb{N}^{\mathrm{D}}$ is

$$\mathcal{P}_{\mathrm{D}}^n = \left\{(j_1, \ldots, j_{\mathrm{D}}) \in \mathbb{N}^{\mathrm{D}} : \sum_{k=1}^{\mathrm{D}} j_k \le n\right\}. \tag{5.43}$$

Example 5.7. The principal lattice of degree 2 in two dimensions is

$$\mathcal{P}_2^2 = \{(0,0), (0,1), (0,2), (1,0), (1,1), (2,0)\}. \tag{5.44}$$

Definition 5.24. The *m-slice of a subset* $T \subseteq \mathbb{Z}_n^{\mathrm{D}}$ across the ith dimension is a subset of T defined as

$$L_{i,m}(T) = \{\mathbf{p} \in T : p_i = m\}. \tag{5.45}$$

Lemma 5.25. Any m-slice of a triangular lattice of degree n in D dimensions is a triangular lattice of degree $n-m$ in $\mathrm{D}-1$ dimensions.

Proof. By Definition 5.16, the triangular lattice is

$$\mathcal{T}_{\mathrm{D}}^n = \left\{(p_{1,j_1}, p_{2,j_2}, \ldots, p_{D,j_D}) : \ j_k \ge 0; \ \sum_{k=1}^{\mathrm{D}} j_k \le n\right\},$$

where each variable p_i has exactly $n+1$ coordinates. By Definition 5.24, we have

$$L_{i,m}(\mathcal{T}) = \left\{ \left(p_{1,j_1}, \ldots, p_{i-1,j_{i-1}}, m, p_{i+1,j_{i+1}}, \ldots, p_{D,j_D}\right) : \right. \\ \left. j_k \ge 0; \ \sum_{k\ne i, k=1}^{\mathrm{D}} j_k \le n-m\right\},$$

which, by Definition 5.16, is a triangular lattice of degree $n-m$ in $\mathrm{D}-1$ dimensions after renumbering the $n-m+1$ coordinates for each dimension $k \ne i$. □

Definition 5.26. A D-*permutation of degree* n, written

$$A = (a_1, a_2, \ldots, a_{\mathrm{D}})^T,$$

is a map $A : \mathbb{Z}_n^{\mathrm{D}} \to \mathbb{Z}_n^{\mathrm{D}}$ defined as

$$A\mathbf{p} = \left(a_1(p_1), a_2(p_2), \cdots, a_{\mathrm{D}}(p_{\mathrm{D}})\right)^T, \tag{5.46}$$

where each $a_i : \mathbb{Z}_n \to \mathbb{Z}_n$ is a permutation.

Notation 10. Denote by G the *set of all* D-*permutations*.

Definition 5.27. The *multiplication of two* D-*permutations* is a binary operation $\cdot : G \times G \to G$ given by

$$A \cdot B = (a_1 \circ b_1, a_2 \circ b_2, \ldots, a_{\mathrm{D}} \circ b_{\mathrm{D}})^T \tag{5.47}$$

where "$\circ$" denotes function composition.

Definition 5.28. The *inverse of a* D-*permutation* A is a unitary operation $^{-1} : G \to G$ such that A^{-1} satisfies

$$A^{-1} \cdot A = A \cdot A^{-1} = E = (e_1, e_2, \ldots, e_{\mathrm{D}})^T, \tag{5.48}$$

where E denotes the distinguished D-permutation with each constituting permutation e_i as the identity map on $\mathbb{Z}_n$.

Lemma 5.29. The following algebra is a group,

$$(G, \cdot, ^{-1}, E). \tag{5.49}$$

Proof. It is straightforward to verify the conditions of a group from Definitions 5.26, 5.27, and 5.28. □

Lemma 5.30. For any $A \in G$, the map σ_A given by

$$\forall \mathcal{T} \in \mathcal{X}, \ \sigma_A(\mathcal{T}) = A\mathcal{T} := \left\{A\mathbf{p} : \mathbf{p} \in T\right\} \tag{5.50}$$

is a permutation of $\mathcal{X}$. In other words, D-permutations map triangular lattices to triangular lattices.

Proof. Since $\mathcal{T}$ is a triangular lattice, we know from Definition 5.16 that there exist $n+1$ coordinates for each of the D dimensions such that *indices* of the D constituting coordinates of each point $\mathbf{p} \in \mathcal{T}$ add up to no more than n. The action of A upon $\mathcal{T}$ in (5.46) can be undone by applying A^{-1}; this means that for $A\mathcal{T}$ there exists a renumbering (specified by A^{-1}) of the coordinates along each axis such that $A\mathcal{T}$ is a triangular lattice. □

Lemma 5.31. The set of triangular lattices $\mathcal{X}$ is a G-set with its group action $G \times \mathcal{X} \to \mathcal{X}$ given by $\sigma_A(\mathcal{T})$ in (5.50).

Proof. By Lemma 5.30, $\bullet(A, \mathcal{T}) = \sigma_A(\mathcal{T})$ indeed has the signature $G \times \mathcal{X} \to \mathcal{X}$. By Definition (5.48), we have

$$\forall \mathcal{T} \in \mathcal{X}, \quad E\mathcal{T} = \mathcal{T}.$$

In addition, for any $A, B \in G$ and any $\mathcal{T} \in \mathcal{X}$, we have

$$\begin{aligned}(A \cdot B)\mathcal{T} &= \left\{(A \cdot B)\mathbf{p} : \mathbf{p} \in \mathcal{T}\right\} \\ &= \left\{\left((a_1 \circ b_1)(p_1), \ldots, (a_{\mathrm{D}} \circ b_{\mathrm{D}})(p_{\mathrm{D}})\right)^T : \mathbf{p} \in \mathcal{T}\right\} \\ &= \left\{\left(a_1(b_1(p_1)), \ldots, a_{\mathrm{D}}(b_{\mathrm{D}}(p_{\mathrm{D}}))\right)^T : B\mathbf{p} \in \mathcal{T}\right\} \\ &= A(B\mathcal{T}),\end{aligned}$$

where the first step follows from (5.50), the second from (5.46) and (5.47), the third from Lemma 5.30, and the last from (5.50). The proof is completed by Definition 0.129. □

Definition 5.32. The *restoration of a triangular lattice* $\mathcal{T}_{\mathrm{D}}^n$ is a D-permutation $R_{\mathcal{T}} = (r_1, r_2, \ldots, r_{\mathrm{D}})^T$ such that

$$\begin{aligned}&\forall i = 1, 2, \ldots, \mathrm{D}, \ \forall m \in \mathbb{Z}_n, \\ &r_i(m) = \#\left\{j \in \mathbb{Z}_n : \#L_{i,j}(\mathcal{T}) > \#L_{i,m}(\mathcal{T})\right\}.\end{aligned} \tag{5.51}$$

Lemma 5.33. The restoration of a triangular lattice $\mathcal{T}_{\mathrm{D}}^n$ is a bijection that maps $\mathcal{T}_{\mathrm{D}}^n$ to the principal lattice $\mathcal{P}_{\mathrm{D}}^n$, i.e.

$$R_{\mathcal{T}_{\mathrm{D}}^n}\mathcal{T}_{\mathrm{D}}^n = \mathcal{P}_{\mathrm{D}}^n. \tag{5.52}$$

Proof. Lemma 5.25 and Lemma 5.17 imply that the slices of $\mathcal{T}$ along each axis have pairwise distinct cardinalities. By Definition 5.32, the cardinalities of rearranged slices along the ith dimension are not changed by any constituting permutations except r_i. By Lemma 5.30, $R_{\mathcal{T}}\mathcal{T}$ is also a triangular lattice. Furthermore, cardinalities of the m-slices of $R_{\mathcal{T}}\mathcal{T}$ decrease strictly monotonically as m increases. There is only one such triangular lattice, namely $\mathcal{P}_{\mathrm{D}}^n$ in (5.43). Finally, $R_{\mathcal{T}}$ is a bijection because each constituting permutation is a bijection. □

Definition 5.34. The *formation of a triangular lattice* $\mathcal{T}$ is the inverse of its restoration, i.e.,

$$A_{\mathcal{T}} = R_{\mathcal{T}}^{-1}. \tag{5.53}$$

Lemma 5.35. The formation of a triangular lattice $\mathcal{T}_{\mathrm{D}}^n$ is a bijection that maps the principal lattice $\mathcal{P}_{\mathrm{D}}^n$ to $\mathcal{T}_{\mathrm{D}}^n$,

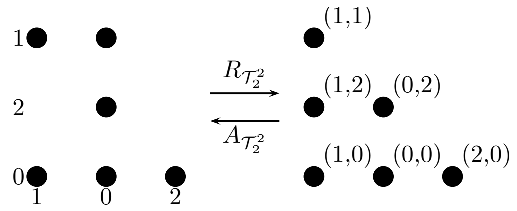
$$\mathcal{T}_{\mathrm{D}}^n = A_{\mathcal{T}_{\mathrm{D}}^n}\mathcal{P}_{\mathrm{D}}^n. \tag{5.54}$$

Proof. This follows directly from Definitions 5.32 and 5.34 and Lemma 5.33. □

Example 5.8. For the triangular lattice $\mathcal{T}_2^2$ in Example 5.3, its formation $A_{\mathcal{T}_2^2}$ equals its restoration $R_{\mathcal{T}_2^2}$,

$$\begin{cases} r_1 : 0 \mapsto 1; 1 \mapsto 0; 2 \mapsto 2, \\ r_2 : 0 \mapsto 0; 1 \mapsto 2; 2 \mapsto 1. \end{cases} \tag{5.55}$$

The processes of restoration and formation are shown below.



On the left, the integers below the lattice are $r_1([0,1,2])$, i.e., the numbers of vertical slices with cardinalities larger than the current slice, and the integers to the left of the lattice are $r_2([0,1,2])$, i.e., the numbers of horizontal slices with cardinalities larger than the current slice. On the right, the multiindex close to a dot is the preimage of the dot under $R_{\mathcal{T}_2^2}$. The following table illustrates that the restoration is indeed a bijection.

$\mathbf{p} \in \mathcal{P}_2^2$	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(2,0)
$\mathbf{p} \in \mathcal{T}_2^2$	(1,0)	(1,2)	(1,1)	(0,0)	(0,2)	(2,0)

Corollary 5.36. Denote by $S_{\mathcal{X}}$ the symmetric group on $\mathcal{X}$. The map $\phi : G \to S_{\mathcal{X}}$ defined by

$$\phi(A) = \sigma_A \tag{5.56}$$

is a monomorphism.

Proof. By Theorem 0.130 and Lemma 5.31, ϕ is a homomorphism. We still need to show that ϕ is injective. Suppose there exists $A \neq B$ in G such that $\phi(A) = \phi(B)$. Then

$$\forall \mathcal{T} \in \mathcal{X},\ A\mathcal{T} = \sigma_A(\mathcal{T}) = \phi(A) = \phi(B) = \sigma_B(\mathcal{T}) = B\mathcal{T},$$

which contradicts $A \neq B$. □

Corollary 5.37. The PLG problem $(K, \mathbf{q})$ in Definition 5.22 is solved by a triangular lattice $\mathcal{T}$ if and only if its formation $A_{\mathcal{T}}$ satisfies

$$\exists \mathbf{p} \in \mathcal{P}_{\mathrm{D}}^n,\ \text{s.t. } A_{\mathcal{T}}\mathbf{p} = \mathbf{q}; \tag{5.57a}$$

$$\forall \mathbf{p} \in \mathcal{P}_{\mathrm{D}}^n,\ A_{\mathcal{T}}\mathbf{p} \in K. \tag{5.57b}$$

Furthermore, all solutions to the PLG problem can be obtained by enumerating the formations.

Proof. The first sentence follows from Definition 5.22, Lemma 5.33, and Definition 5.34. By Lemma 5.35, any triangular lattice $\mathcal{T}$ determines a D-permutation, namely its formation, that generates $\mathcal{T}$ from the principal lattice. Hence all triangular lattices are generated by enumerating formations. □

5.3.3 Partitioning the principal lattice

Definition 5.38. A function $c : \mathbb{Z}_{\mathrm{D}}^+ \to \mathbb{Z}_n$ is called a *column-pick map* for a multiindex $(\ell, m) \in \mathbb{Z}_{\mathrm{D}}^+ \times \mathbb{Z}_n$ if

$$c(i) = \begin{cases} \leq m & \text{if } i < \ell; \\ m & \text{if } i = \ell; \\ < m & \text{if } i > \ell. \end{cases} \tag{5.58}$$

The set of all column-pick maps for a given multiindex (ℓ, m) is denoted by $C_{(\ell,m)}$.

Definition 5.39. The *test set* at $(\ell, m) \in \mathbb{Z}_{\mathrm{D}}^+ \times \mathbb{Z}_n$ is a set of D-dimensional multiindices,

$$W_{(\ell,m)} = \left\{ (c(1), \ldots, c(\mathrm{D})) \in \mathbb{Z}_n^{\mathrm{D}} : c \in C_{(\ell,m)}; \sum_{i=1}^{\mathrm{D}} c(i) \leq n \right\}. \tag{5.59}$$

In particular, $W_{(\ell,m)} = \emptyset$ if $C_{(\ell,m)} = \emptyset$.

Example 5.9. The test-set partition of $\mathcal{P}_2^2$ in (5.44) is

$$\begin{aligned} &W_{(1,1)} = \{(1,0)\},\ W_{(1,2)} = \{(2,0)\}, \\ &W_{(2,0)} = \{(0,0)\},\ W_{(2,1)} = \{(0,1), (1,1)\}, W_{(2,2)} = \{(0,2)\}. \end{aligned}$$

Lemma 5.40. Test sets form a partition of the principal lattice $\mathcal{P}_{\mathrm{D}}^n$. More precisely, we have

$$\bigcup_{\ell \in \mathbb{Z}_{\mathrm{D}}^+, m \in \mathbb{Z}_n} W_{(\ell,m)} = \mathcal{P}_{\mathrm{D}}^n; \tag{5.60a}$$

$$(\ell, m) \neq (i, j) \ \Leftrightarrow\ W_{(\ell,m)} \cap W_{(i,j)} = \emptyset. \tag{5.60b}$$

Proof. Suppose $\mathbf{p} \in W_{(\ell,m)}$ for some (ℓ, m). Then $\mathbf{p} \in \mathcal{P}_{\mathrm{D}}^n$ must hold because of the condition $\sum_{i=1}^{\mathrm{D}} c(i) \leq n$ in (5.59).

Conversely, let m be the largest coordinate of $\mathbf{p} \in \mathcal{P}_{\mathrm{D}}^n$ and ℓ be the largest dimension index of $\mathbf{p}$ such that $p_\ell = m$. Then (5.58) and (5.59) imply $\mathbf{p} \in W_{(\ell,m)}$.

Suppose for some $(i,j) \neq (\ell,m)$ we also have $\mathbf{p} \in W_{(i,j)}$. Since $\mathbf{p}$ has only one largest coordinate (that is assumed to be m), we must have $j = m$, which implies $i \neq \ell$. Because ℓ is the largest dimension index satisfying $c(\ell) = m$, we have $i < \ell$, which contradicts the third branch of (5.58). □

5.3.4 Partitioning actions of D-permutations

Definition 5.41. A *partial function* from Y to Z is a function $Y' \to Z$ on some $Y' \subseteq Y$.

Definition 5.42. The (ℓ,m)*-partial* D*-permutation* of a D-permutation A, denoted by $A^{(\ell,m)}$, is a partial function on the test set $W_{(\ell,m)}$ that satisfies

$$\forall \mathbf{p} \in W_{(\ell,m)},\ A^{(\ell,m)}\mathbf{p} = A\mathbf{p}. \tag{5.61}$$

Definition 5.43. The *linear ordering of integer pairs* on a grid $\mathbb{Z}^+_{\mathrm{D}} \times \mathbb{Z}_n$ is the column-wise ordering of the grid, i.e., the total ordering obtained by stacking all the columns of the grid into one column. More precisely, this ordering is a bijection s that maps a pair $(i,j) \in \mathbb{Z}^+_{\mathrm{D}} \times \mathbb{Z}_n$ to a scalar index $k \in \mathbb{Z}^+_{\mathrm{D}(n+1)}$,

$$s(i,j) = i + j\mathrm{D}; \tag{5.62}$$

$$s^{-1}(k) = \left(1 + (k-1) \bmod \mathrm{D},\ \left\lfloor \frac{k-1}{\mathrm{D}} \right\rfloor\right), \tag{5.63}$$

where $\lfloor \cdot \rfloor : \mathbb{Q} \to \mathbb{N}$ is the floor operator.

Corollary 5.44. If $C_{(\ell,m)}$ is nonempty, any column-pick map $c \in C_{(\ell,m)}$ satisfies

$$\forall i \in \mathbb{Z}^+_{\mathrm{D}},\ s(i,c(i)) \leq s(\ell,m). \tag{5.64}$$

Proof. This follows from (5.58) and Definition 5.43. □

Notation 11. In matrix notation, a D-permutation is

$$A = \begin{bmatrix} a_1(0) & a_1(1) & \ldots & a_1(n) \\ a_2(0) & a_2(1) & \ldots & a_2(n) \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mathrm{D}}(0) & a_{\mathrm{D}}(1) & \ldots & a_{\mathrm{D}}(n) \end{bmatrix}, \tag{5.65}$$

which means that the (i,j)th-element of A is

$$A(i,j) = a_i(j). \tag{5.66}$$

Note that the row index i of the matrix starts from 1 while the column index j starts from 0. By Corollary 5.44, the matrix of a partial D-permutation is simply

$$A^{(\ell,m)}(i,j) = \begin{cases} A(i,j) & \text{if } s(i,j) \leq s(\ell,m); \\ -1 & \text{otherwise,} \end{cases} \tag{5.67}$$

where s is defined in (5.62), and "-1" indicates undefined behavior. Since s is a bijection, it also makes sense to write

$$\forall t = s(\ell,m),\ A^{(t)} := A^{(\ell,m)}. \tag{5.68}$$

Lemma 5.45. A triangular lattice $\mathcal{T}^n_{\mathrm{D}}$ has the partition

$$\mathcal{T}^n_{\mathrm{D}} = \bigcup_{\ell \in \mathbb{Z}^+_{\mathrm{D}}, m \in \mathbb{Z}_n} A^{(\ell,m)}_{\mathcal{T}} W_{(\ell,m)}, \tag{5.69}$$

where the terms $A^{(\ell,m)}_{\mathcal{T}} W_{(\ell,m)}$ are pairwise disjoint.

Proof. By Lemma 5.40, Lemma 5.33, and (5.61), we have

$$\begin{aligned} \mathcal{T}^n_{\mathrm{D}} &= A_{\mathcal{T}}\mathcal{P}^n_{\mathrm{D}} = A_{\mathcal{T}} \bigcup_{(\ell,m)} W_{(\ell,m)} = \bigcup_{(\ell,m)} A_{\mathcal{T}} W_{(\ell,m)} \\ &= \bigcup_{(\ell,m)} A^{(\ell,m)}_{\mathcal{T}} W_{(\ell,m)}. \end{aligned}$$

(5.60b) yields the pairwise disjointness of the terms. □

Example 5.10. The triangular lattice in Example 5.3 is

$$\mathcal{T} = \{(0,0), (0,2), (1,0), (1,1), (1,2), (2,0)\} \tag{5.70}$$

and its formation is

$$A_{\mathcal{T}} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix}. \tag{5.71}$$

Following Examples 5.8 and 5.9, $\mathcal{T}$ is partitioned into

$$\begin{aligned} &A^{(1,1)}_{\mathcal{T}} W_{(1,1)} = \{(0,0)\},\ A^{(1,2)}_{\mathcal{T}} W_{(1,2)} = \{(2,0)\}, \\ &A^{(2,0)}_{\mathcal{T}} W_{(2,0)} = \{(1,0)\},\ A^{(2,1)}_{\mathcal{T}} W_{(2,1)} = \{(1,2),(0,2)\}, \\ &A^{(2,2)}_{\mathcal{T}} W_{(2,2)} = \{(1,1)\}. \end{aligned}$$

Theorem 5.46. The PLG problem $(K, \mathbf{q})$ in Definition 5.22 is solved by a triangular lattice $\mathcal{T} = A_{\mathcal{T}}\mathcal{P}^n_{\mathrm{D}}$ if and only if its formation $A_{\mathcal{T}}$ satisfies

$$\forall (\ell,m) \in \mathbb{Z}^+_{\mathrm{D}} \times \mathbb{Z}_n,\ A^{(\ell,m)}_{\mathcal{T}} W_{(\ell,m)} \subset K; \tag{5.72a}$$

$$\exists \mathbf{p} \in \mathcal{P}^n_{\mathrm{D}},\ \text{s.t. } A_{\mathcal{T}}\mathbf{p} = \mathbf{q}. \tag{5.72b}$$

Furthermore, an enumeration based on the partial D-permutations finds all solutions to the PLG problem.

Proof. This follows directly from Lemma 5.45 and Corollary 5.37. □

5.3.5 Depth-first search and backtracking

Definition 5.47. *Depth-first search (DFS)* is an algorithm for traversing a graph $G = (V, E)$; it starts at a given node $v \in V$ and explores as far as possible along each edge before backtracking. A recursive version is as follows.

```
Procedure DFS($G$, $v$)
    Input: A finite graph $G = (V, E)$ and a node
           $v \in V$
    Preconditions: $G$ is connected and initially all
                   nodes in $V$ are "undiscovered."
    Side effects   : all nodes in $V$ are "discovered."
label $v$ as "discovered"
for each edge $(v, w) \in E$ do
    if $w$ is not discovered then
        DFS($G$, $w$)
    end
end
```

Definition 5.48 (Backtracking)**.** Denote by P a problem that admits an incremental assemblage of the solutions and hence a spanning-tree organization of the solution space. *Backtracking* is a generic algorithm that solves P by calling `BackTrack`$(P, \texttt{root}(P), \{\})$,

```
Procedure BackTrack(P, r, T)
    Input: r is a starting node in the solution space
           and T is a set of the solutions
    Preconditions: Initially T is an empty set
    Side effects  : T contains all valid solutions

if accept(P, r) then
    T = T ∪ {r}
    if stopAfterAccept(P, r) then return
else if reject(P, r) then
    return
end
s ← first(P, r)
while s is not null do
    BackTrack(P, s, T)
    s ← next(P, r, s)
end
```

where the details of P are given in the following user-defined subroutines,

- `root`(P): return the root node of the spanning tree of the solution space,
- `accept`$(P, \mathbf{r})$: return true if and only if $\mathbf{r}$ is already a solution; return false otherwise,
- `stopAfterAccept`$(P, \mathbf{r})$: return true if and only if the valid solution $\mathbf{r}$ can never be extended to another valid solution; return false otherwise,
- `reject`$(P, \mathbf{r})$: return true if and only if the partial candidate (or node) $\mathbf{r}$ can never be completed to a valid solution; return false otherwise,
- `first`$(P, \mathbf{r})$: return the first extension of $\mathbf{r}$ in the spanning tree if $\mathbf{r}$ is extendable; return null otherwise,
- `next`$(P, \mathbf{r}, \mathbf{s})$: return the next extension of $\mathbf{r}$ after $\mathbf{s}$ if $\mathbf{r}$ is still extendable; return null otherwise.

5.3.6 Backtracking for PLG

Definition 5.49. For a PLG problem $(K, \mathbf{q})$, the *test ordering* "<" of a subset $J_i \subseteq \mathbb{Z}_n$ along the ith dimension is a total order on J_i determined first by the cardinality of the slices, and then by breaking any tie with the distance to $\mathbf{q}$ along the ith dimension. More precisely, for any distinct $j, k \in J_i$, we say that k is *greater than* j, written $k > j$, or that j is *less than* k, written $j < k$, if and only if

$$\begin{aligned}&\big(\#L_{i,j}(K) < \#L_{i,k}(K)\big)\\ &\vee\big(\#L_{i,j}(K) = \#L_{i,k}(K) \;\;\wedge\;\; |j - q_i| > |k - q_i|\big), \end{aligned}\qquad (5.73)$$

where q_i is the ith coordinate of $\mathbf{q}$. In particular, the element in J_i that is greater and less than all other elements in J_i is denoted by $\max J_i$ and $\min J_i$, respectively.

Definition 5.50 (Backtracking for PLG)**.** The following recursive algorithm finds all solutions to a PLG problem.

```
Input: the degree $n \in \mathbb{N}^+$, the dimensionality
       $\mathrm{D} \in \mathbb{N}^+$, the search domain $K \subseteq \mathbb{Z}_n^{\mathrm{D}}$, and
       the starting point $\mathbf{q} \in K$
Output: a set $\mathcal{U}$ of D-permutations
Postconditions: $\{A\mathcal{P}_{\mathrm{D}}^n : A \in \mathcal{U}\}$ is the set of all
                solutions of the PLG problem

$\mathcal{U}$ ← an empty set of D-permutations
$A^{(r)}$ ← root$(n, \mathrm{D})$
BackTrack $\big((K, \mathbf{q}), A^{(r)}, \mathcal{U}\big)$
```

where the procedures in Definition 5.48 are as follows.

- `root`(n, D): set $A^{(0)}$ to a D-by-$(n+1)$ matrix of constant -1 following Notation 11; return $A^{(0)}$.
- `accept`$((K, \mathbf{q}), A^{(t)})$: return false if $t < \mathrm{D}(n+1)$; otherwise return true if and only if
$$A^{(t)} W_{(\mathrm{D},n)} \subset K \text{ and } \mathbf{q} \in A^{(t)} \mathcal{P}_{\mathrm{D}}^n.$$
- `stopAfterAccept`$((K, \mathbf{q}), A^{(t)})$: return true.
- `reject`$((K, \mathbf{q}), A^{(t)})$: if $t = \mathrm{D}(n+1)$, return the negation of `accept`$((K, \mathbf{q}), A^{(t)})$; otherwise return the result of the logical statement
$$\exists \mathbf{p} \in W_{(\ell,m)} \text{ s.t. } A^{(t)}\mathbf{p} \notin K,$$
where $(\ell, m) = s^{-1}(t)$.
- `first`$((K, \mathbf{q}), A^{(t)})$: let $(\ell, m) = s^{-1}(t+1)$ and initialize
$$B^{(t+1)} \leftarrow A^{(t)},$$
$$J_\ell := \mathbb{Z}_n \backslash \Big\{ B^{(t+1)}(\ell, j) : j = 0, 1, \ldots, m-1 \Big\}. \quad (5.74)$$
Set $B^{(t+1)}(\ell, m) \leftarrow \max J_\ell$ as in Definition 5.49 and return $B^{(t+1)}$.
- `next`$((K, \mathbf{q}), A^{(t)}, B^{(t+1)})$: let $(\ell, m) = s^{-1}(t+1)$ and compute J_ℓ by (5.74). Return null if $B^{(t+1)}(\ell, m)$ equals $\min J_\ell$ as in Definition 5.49. Otherwise initialize $C^{(t+1)} \leftarrow B^{(t+1)}$, set $C^{(t+1)}(\ell, m)$ to be the element in J_ℓ that is *immediately* smaller than $B^{(t+1)}(\ell, m)$, and return $C^{(t+1)}$.

5.4 Programming assignments

A. Write a program to implement the recursive algorithm in Definition 5.50 and test your implement for $\mathrm{D} = 2, 3$ and $n = 2, 3, 4, 5$. In designing the tests $(K, \mathbf{q})$, use a simply connected polygon/polyhedron to derive K by declaring that any point covered by the polygon/polyhedron is not available. Display the generated poised lattices for both two and three dimensions.

Chapter 6

Approximation

Definition 6.1. Given a normed vector space Y of functions and its subspace $X \subseteq Y$. A function $\hat{\varphi} \in X$ is called the *best approximation* to $f \in Y$ from X with respect to the norm $\|\cdot\|$ iff

$$\forall \varphi \in X, \qquad \|f - \hat{\varphi}\| \le \|f - \varphi\|. \tag{6.1}$$

Example 6.1. The Chebyshev Theorem 3.32 can be restated in the format of Definition 6.1 as follows. As in Example 0.50, 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 $\|\cdot\|_\infty$

$$\|g\|_\infty = \max_{x\in[-1,1]} |g(x)| \tag{6.2}$$

is $\hat{\varphi} = \frac{T_n}{2^{n-1}} - x^n$, where T_n is Chebyshev polynomial of degree n. Clearly $\hat{\varphi}$ satisfies (6.1).

Example 6.2. 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 (6.2), the 1-norm

$$\|g\|_1 := \int_{-1}^{+1} |g(x)| \mathrm{d}x, \tag{6.3}$$

or the 2-norm

$$\|g\|_2 := \left(\int_{-1}^{+1} |g(x)|^2 \mathrm{d}x\right)^{\frac{1}{2}}. \tag{6.4}$$

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 6.4.

Example 6.3. 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 4.56. 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 6.2. Suppose X is a finite-dimensional subspace of a normed space $(Y, \|\cdot\|)$. Then we have

$$\forall y \in Y,\ \exists \hat{\varphi} \in X \text{ s.t. } \forall \varphi \in X,\ \|\hat{\varphi} - y\| \le \|\varphi - y\|. \tag{6.5}$$

Proof. For a given $y \in Y$, define a closed ball

$$B_y := \{x \in X : \|x\| \le 2\|y\|\}.$$

Clearly $0 \in B_y$, and the distance from y to B_y is

$$\text{dist}(y, B_y) := \inf_{x\in B_y} \|y - x\| \le \|y - 0\| = \|y\|.$$

By definition, any $z \in X$, $z \notin B_y$ must satisfy $\|z\| > 2\|y\|$, and thus

$$\|z - y\| \ge \|z\| - \|y\| > \|y\|.$$

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. □

Definition 6.3 (L^p functions)**.** Let $p > 0$. The class of functions $f(x)$ which are measurable and for which $|f(x)|^p$ is Lebesgue integrable over $[a,b]$ is known as $L^p[a,b]$. If $p = 1$, the class is denoted by $L[a,b]$.

Theorem 6.4. For a *weight function* $\rho(x) \in L[a,b]$, define

$$L^2_\rho[a,b] := \left\{f(x) \in L[a,b] : \rho(x)|f(x)|^2 \in L[a,b]\right\}. \tag{6.6}$$

Then $L^2_\rho[a,b]$ is a vector space. If we further require that $\forall x \in (a,b), \rho(x) > 0$, then the vector space $L^2_\rho[a,b]$ with

$$\langle u, v\rangle = \int_a^b \rho(t) u(t) \overline{v(t)} \mathrm{d}t \tag{6.7}$$

is an inner product space over $\mathbb{R}$; the set $L^2_\rho[a,b]$ with

$$\|u\|_2 = \left(\int_a^b \rho(t)|u(t)|^2 \mathrm{d}t\right)^{\frac{1}{2}} \tag{6.8}$$

is a normed vector space over $\mathbb{R}$.

Proof. This follows from Definitions 0.69, 0.87, and 0.89. □

Definition 6.5. The *least-square approximation* on $L^2_\rho[a,b]$ is a best approximation problem with the norm in (6.1) set to that in (6.8).

6.1 Orthonormal systems

Definition 6.6. A subset S of an inner product space X is called *orthonormal* if

$$\forall u, v \in S, \qquad \langle u, v\rangle = \begin{cases} 0 & \text{if } u \neq v, \\ 1 & \text{if } u = v. \end{cases} \tag{6.9}$$

Example 6.4. The standard basis vectors in $\mathbb{R}^n$ are orthonormal.

Example 6.5. The Chebyshev polynomials of the first kind as in Definition 3.28 are orthogonal with respect to (6.7) where $a = -1, b = 1, \rho = \frac{1}{\sqrt{1-x^2}}$. However, they do not satisfy the second case in (6.9).

Theorem 6.7. 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 0.76 and 6.6. □

Definition 6.8. 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

$$v_{n+1} = u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u_k^*\rangle u_k^*, \tag{6.10a}$$

$$u_{n+1}^* = v_{n+1}/\|v_{n+1}\|, \tag{6.10b}$$

with the recursion basis as $v_1 = u_1,\ u_1^* = v_1/\|v_1\|$.

Theorem 6.9. For a finite or infinite independent list $(u_1, u_2, \ldots)$, the Gram-Schmidt process yields constants

$$\begin{array}{lll} a_{11} & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \\ \vdots & & \end{array}$$

such that $a_{kk} = \frac{1}{\|v_k\|} > 0$ and the elements $u_1^*, u_2^*, \ldots$

$$\begin{aligned} u_1^* &= a_{11}u_1 \\ u_2^* &= a_{21}u_1 + a_{22}u_2 \\ u_3^* &= a_{31}u_1 + a_{32}u_2 + a_{33}u_3 \\ &\ \vdots \end{aligned} \tag{6.11}$$

are orthonormal.

Proof. By Definition 6.8, the formulae (6.10) can be rewritten in the form of (6.11). It is clear from (6.10b) that u_{n+1}^* is normal. We show u_{n+1}^* is orthogonal to u_n^*, u_{n-1}^*, …, u_1^* by induction. 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 0.87 and the third step from u_1^* being normal. The inductive step also holds because for any $j < n+1$ we have

$$\begin{aligned} \langle v_{n+1}, u_j^*\rangle &= \left\langle u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u_k^*\rangle u_k^*, u_j^* \right\rangle \\ &= \langle u_{n+1}, u_j^*\rangle - \sum_{k=1}^{n} \langle u_{n+1}, u_k^*\rangle \langle u_k^*, u_j^*\rangle \\ &= \langle u_{n+1}, u_j^*\rangle - \langle u_{n+1}, u_j^*\rangle = 0, \end{aligned}$$

where the third step follows from the induction hypothesis, i.e., $\langle u_k^*, u_j^*\rangle$ is 1 if $k = j$ and 0 otherwise. It remains to show $a_{kk} = \frac{1}{\|v_k\|}$, which holds because

$$\begin{aligned} 1 = \langle u_n^*, u_n^*\rangle &= \langle a_{nn}u_n, u_n^*\rangle + \left\langle \sum_{i=1}^{n-1} a_{ni}u_i, u_n^* \right\rangle \\ &= a_{nn}\langle v_n, u_n^*\rangle = a_{nn}\left\langle v_n, \frac{v_n}{\|v_n\|}\right\rangle = a_{nn}\|v_n\|, \end{aligned}$$

where the second step follows from the nth equation of (6.11), the third step from (6.10a) and the conclusion just proved, the fourth step from (6.10b), and the last step from Definitions 0.87 and 0.89. □

Corollary 6.10. For a finite or infinite independent list $(u_1, u_2, \ldots)$, we can find constants

$$\begin{array}{lll} b_{11} & & \\ b_{21} & b_{22} & \\ b_{31} & b_{32} & b_{33} \\ \vdots & & \end{array}$$

and an orthonormal list $(u_1^*, u_2^*, \ldots)$ such that $b_{ii} > 0$ and

$$\begin{aligned} u_1 &= b_{11}u_1^* \\ u_2 &= b_{21}u_1^* + b_{22}u_2^* \\ u_3 &= b_{31}u_1^* + b_{32}u_2^* + b_{33}u_3^* \\ &\ \vdots \end{aligned} \tag{6.12}$$

Proof. This follows from (6.11) and that a lower-triangular matrix with positive diagonal elements is invertible. □

Corollary 6.11. In Theorem 6.9, we have $\langle u_n^*, u_i\rangle = 0$ for each $i = 1, 2, \ldots, n-1$.

Proof. By Corollary 6.10, 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 6.12. Using the Gram-Schmidt orthonormalizing process with the inner product (6.7), 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 6.6. We compute the first 3 Legendre polynomials using the Gram-Schmidt process.

$$
\begin{aligned}
& u_1 = 1,\ v_1 = 1, \|v_1\|^2 = \int_{-1}^{+1} \mathrm{d}x = 2,\ \ u_1^* = \frac{1}{\sqrt{2}}.\\
& u_2 = x,\ v_2 = x - \left\langle x, \frac{1}{\sqrt{2}} \right\rangle \frac{1}{\sqrt{2}} = x,\ \|v_2\|^2 = \frac{2}{3},\\
& u_2^* = \sqrt{\frac{3}{2}}x.\\
& v_3 = x^2 - \left\langle x^2, \sqrt{\frac{3}{2}}x \right\rangle \sqrt{\frac{3}{2}}x - \left\langle x^2, \frac{1}{\sqrt{2}} \right\rangle \frac{1}{\sqrt{2}} = x^2 - \frac{1}{3},\\
& \|v_3\|^2 = \int_{-1}^{+1} \left(x^2 - \frac{1}{3}\right)^2 \mathrm{d}x = \frac{8}{45},\\
& u_3^* = \frac{3}{4}\sqrt{10}\left(x^2 - \frac{1}{3}\right).
\end{aligned}
$$

6.2 Fourier expansions

Definition 6.13. 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

$$w \sim \sum_n \langle w, u_n^* \rangle u_n^*, \tag{6.13}$$

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 6.7. With the Euclidean inner product in Definition 0.88, 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

$$\langle w, u_1^* \rangle u_1^* = (a,0,0)^T,\ \ \langle w, u_2^* \rangle u_2^* = (0,b,0)^T.$$

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 6.14.

Exercise 6.8. With the following orthonormal list in $L^2_{\rho=1}[-\pi, \pi]$,

$$\frac{1}{\sqrt{2\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos x}{\sqrt{\pi}}, \ldots, \frac{\sin(nx)}{\sqrt{\pi}}, \frac{\cos(nx)}{\sqrt{\pi}}, \ldots, \tag{6.14}$$

derive the *Fourier series* of a function $f(x)$ as

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{+\infty}(a_k \cos kx + b_k \sin kx), \tag{6.15}$$

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 6.14. 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

$$w = \sum_{i=1}^n \langle w, u_i^* \rangle u_i^*, \tag{6.16}$$

i.e. w is equal to its Fourier expansion.

Proof. By the condition $w = \sum_{i=1}^n a_i u_i$ and Corollary 6.10, we can express w as a linear combination of u_i^*'s,

$$w = \sum_{i=1}^n c_i u_i^*.$$

Then the orthogonality of u_i^*'s implies

$$\forall k = 1, 2, \cdots, n, \qquad \langle u_k^*, w \rangle = c_k,$$

which completes the proof. □

Theorem 6.15 (Minimum properties of Fourier expansions)**.** Let $u_1^*, u_2^*, \ldots$ be an orthonormal system and let w be arbitrary. Then

$$\left\| w - \sum_{i=1}^N \langle w, u_i^* \rangle u_i^* \right\| \le \left\| w - \sum_{i=1}^N a_i u_i^* \right\|, \tag{6.17}$$

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 Definition 0.87 and properties of inner products

$$\begin{aligned}
&\left\| w-\sum_i a_i u_i^* \right\|^2 = \left\langle w-\sum_i a_i u_i^*, w-\sum_i a_i u_i^* \right\rangle \\
&= \langle w,w\rangle - \left\langle w, \sum_i a_i u_i^* \right\rangle - \left\langle \sum_i a_i u_i^*, w \right\rangle \\
&\quad + \left\langle \sum_i a_i u_i^*, \sum_i a_i u_i^* \right\rangle \\
&= \langle w,w\rangle - \sum_i \overline{a_i}\left\langle w,u_i^*\right\rangle - \sum_i a_i \left\langle u_i^*, w\right\rangle \\
&\quad + \sum_i \sum_j a_i \overline{a_j} \left\langle u_i^*, u_j^*\right\rangle \\
&= \langle w,w\rangle - \sum_i \overline{a_i}\left\langle w,u_i^*\right\rangle - \sum_i a_i \left\langle u_i^*, w\right\rangle + \sum_i |a_i|^2 \\
&\quad - \sum_i \left\langle u_i^*, w\right\rangle \left\langle w, u_i^*\right\rangle + \sum_i \left\langle u_i^*, w\right\rangle \left\langle w, u_i^*\right\rangle \\
&= \|w\|^2 - \sum_i \left|\left\langle w,u_i^*\right\rangle\right|^2 + \sum_i \left|a_i - \left\langle w,u_i^*\right\rangle\right|^2,
\end{aligned} \tag{6.18}$$

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 6.16. 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

$$\|w-\hat{\varphi}\|^2 := \min_{a_k} \left\| w - \sum_{k=1}^{n} a_k u_k \right\|^2 = \|w\|^2 - \sum_{k=1}^{n} \left|\langle w, u_k^* \rangle\right|^2 . \tag{6.19}$$

Proof. This follows directly from (6.18). □

Corollary 6.17 (Bessel inequality)**.** If $u_1^*, u_2^*, \ldots, u_N^*$ are orthonormal, then, for an arbitrary w,

$$\sum_{i=1}^{N} \left|\langle w, u_i^* \rangle\right|^2 \le \|w\|^2. \tag{6.20}$$

Proof. This follows directly from Corollary 6.16 and the real positivity of a norm. □

Corollary 6.18. The Gram-Schmidt process in Definition 6.8 satisfies

$$\forall n \in \mathbb{N}^+, \quad \|v_{n+1}\|^2 = \|u_{n+1}\|^2 - \sum_{k=1}^{n} \left|\langle u_{n+1}, u_k^* \rangle\right|^2 . \tag{6.21}$$

Proof. By (6.10a), 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 6.16, identifying w with u_{n+1} completes the proof. □

Example 6.9. Consider the problem in Example 6.2 in the sense of least square approximation with the weight function $\rho = 1$. It is equivalent to

$$\min_{a_i} \int_{-1}^{+1} \left(e^x - \sum_{i=0}^{n} a_i x^i \right)^2 \mathrm{d}x. \tag{6.22}$$

For $n = 1, 2$, use the Legendre polynomials derived in Example 6.6:

$$u_1^* = \frac{1}{\sqrt{2}}, \; u_2^* = \sqrt{\frac{3}{2}}x, \; u_3^* = \frac{1}{4}\sqrt{10}(3x^2-1),$$

and we have the Fourier coefficients of e^x as

$$\begin{aligned}
b_0 &= \int_{-1}^{+1} \frac{1}{\sqrt{2}} e^x \mathrm{d}x = \frac{1}{\sqrt{2}}\left(e - \frac{1}{e}\right), \\
b_1 &= \int_{-1}^{+1} \sqrt{\frac{3}{2}} x e^x \mathrm{d}x = \sqrt{6} e^{-1}, \\
b_2 &= \int_{-1}^{+1} \frac{1}{4}\sqrt{10}(3x^2-1) e^x \mathrm{d}x = \frac{\sqrt{10}}{2}\left(e - \frac{7}{e}\right).
\end{aligned}$$

The minimizing polynomials are thus

$$\hat{\varphi}_n = \begin{cases} \frac{1}{2e}(e^2-1) + \frac{3}{e}x & n = 1; \\ \hat{\varphi}_1 + \frac{5}{4e}(e^2-7)(3x^2-1) & n = 2. \end{cases} \tag{6.23}$$

6.3 The normal equations

Theorem 6.19. 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,

$$\forall j = 1, 2, \ldots, n, \qquad \left(w - \sum_{k=1}^{n} \langle w, u_k^* \rangle u_k^* \right) \perp u_j^*, \tag{6.24}$$

where "$\perp$" denotes orthogonality.

Proof. Take the inner product of the two vectors and apply the conditions on orthonormal systems. □

Corollary 6.20. 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

$$\forall j = 1, 2, \ldots, n, \qquad (w - \hat{\varphi}) \perp u_j. \tag{6.25}$$

Proof. Since $\hat{\varphi} = \sum_{k=1}^{n} a_k u_k$ is the best linear approximant to w, Theorem 6.15 implies that

$$\sum_{k=1}^{n} a_k u_k = \sum_{k=1}^{n} \langle w, u_k^* \rangle u_k^*.$$

Corollary 6.10 and Theorem 6.19 complete the proof. □

Definition 6.21. Let $u_1, u_2, \ldots, u_n$ be a sequence of elements in an inner product space. The $n \times n$ matrix

$$\begin{aligned}
G &= G(u_1, u_2, \cdots, u_n) = (\langle u_i, u_j \rangle) \\
&= \begin{bmatrix} \langle u_1, u_1 \rangle & \langle u_1, u_2 \rangle & \ldots & \langle u_1, u_n \rangle \\ \langle u_2, u_1 \rangle & \langle u_2, u_2 \rangle & \ldots & \langle u_2, u_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, u_1 \rangle & \langle u_n, u_2 \rangle & \ldots & \langle u_n, u_n \rangle \end{bmatrix}
\end{aligned} \tag{6.26}$$

is the *Gram matrix* of $u_1, u_2, \ldots, u_n$. Its determinant

$$g = g(u_1, u_2, \ldots, u_n) = \det(\langle u_i, u_j\rangle) \tag{6.27}$$

is the *Gram determinant.*

Lemma 6.22. 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

$$G(w_1, w_2, \ldots, w_n) = AG(u_1, u_2, \ldots, u_n)A^H \tag{6.28}$$

and

$$g(w_1, w_2, \ldots, w_n) = |\det A|^2 g(u_1, u_2, \ldots, u_n). \tag{6.29}$$

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 (6.28) 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 (0.86) yields

$$\overline{\det A} = \overline{\det A^T} = \sum_{\sigma\in S_n} \mathrm{sgn}(\sigma)\prod_{i=1}^n \overline{a_{\sigma_i, i}} = \det A^H.$$

Take the determinant of (6.28), apply the identity $\det(AB) = \det(A)\det(B)$, and we have (6.29). □

Theorem 6.23. For nonzero elements $u_1, u_2, \ldots, u_n \in X$, we have

$$0 \le g(u_1, u_2, \ldots, u_n) \le \prod_{k=1}^n \|u_k\|^2, \tag{6.30}$$

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 $c_1, c_2, \ldots, c_n$ such that $\sum_{i=1}^n c_iu_i = \mathbf{0}$ with at least one constant c_j being nonzero. Construct vectors

$$w_k = \begin{cases}\sum_{i=1}^n c_iu_i = \mathbf{0}, & k = j;\\ u_k, & k \ne j.\end{cases}$$

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, we 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 each M_i has a row of all zeros. Then Lemma 6.22 yields $g(u_1, u_2, \ldots, u_n) = 0$.

Now suppose $u_1, u_2, \ldots, u_n$ are linearly independent. Theorem 6.9 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 6.21 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 6.22 that

$$g(u_1, u_2, \ldots, u_n) = \prod_{k=1}^n \frac{1}{a_{kk}^2} > 0. \tag{6.31}$$

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 6.21, $G(u_1, u_2, \ldots, u_n)$ is a diagonal matrix with $\|u_k\|^2$ on the diagonals. Hence the orthogonality of u_k's implies

$$g(u_1, u_2, \ldots, u_n) = \prod_{k=1}^n \|u_k\|^2. \tag{6.32}$$

For the converse statement, suppose (6.32) holds. Then $u_1, u_2, \ldots, u_n$ must be independent because otherwise it would contradict the lower equality proved as above. Apply the Gram-Schmidt process to $(u_1, u_2, \ldots, u_n)$ and we know from Theorem 6.9 that $\frac{1}{a_{kk}} = \|v_k\|$. Set the length of the list in Theorem 6.9 to $1, 2, \ldots, n$ and we know from (6.31) and (6.32) that

$$\forall k = 1, 2, \ldots, n, \qquad \|u_k\|^2 = \|v_k\|^2. \tag{6.33}$$

Then Corollary 6.18 and (6.33) imply

$$\forall k = 1, 2, \ldots, n, \qquad \sum_{j=1}^{k-1} \left|\left\langle u_k, u_j^* \right\rangle\right|^2 = 0,$$

which further implies

$$\forall k = 1, 2, \ldots, n,\ \forall j = 1, 2, \ldots, k-1, \qquad \left\langle u_k, u_j^* \right\rangle = 0,$$

which, together with Corollary 6.10, 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 (6.31), $\frac{1}{a_{kk}} = \|v_k\|$, and Corollary 6.18. □

Theorem 6.24. 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*,

$$G(u_1, u_2, \ldots, u_n)^T\, \mathbf{a} = \mathbf{c}, \tag{6.34}$$

where $\mathbf{c} = [\langle w, u_1\rangle, \langle w, u_2\rangle, \ldots, \langle w, u_n\rangle]^T$.

Proof. Corollary 6.20 yields

$$\langle w, u_j\rangle = \sum_{k=1}^n a_k \langle u_k, u_j\rangle,$$

which is simply the jth equation of (6.34). The uniqueness of the coefficients follows from Theorem 6.23 and Cramer's rule. □

Example 6.10. Solve Example 6.9 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 (6.26), (6.7), and $\rho = 1$:

$$G(1, x, x^2) = \begin{bmatrix} \langle 1, 1\rangle & \langle 1, x\rangle & \left\langle 1, x^2\right\rangle \\ \langle x, 1\rangle & \langle x, x\rangle & \left\langle x, x^2\right\rangle \\ \left\langle x^2, 1\right\rangle & \left\langle x^2, x\right\rangle & \left\langle x^2, x^2\right\rangle \end{bmatrix} = \begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix}.$$

We then calculate the vector

$$\mathbf{c} = \begin{bmatrix} \langle e^x, 1\rangle \\ \langle e^x, x\rangle \\ \left\langle e^x, x^2\right\rangle \end{bmatrix} = \begin{bmatrix} e - 1/e \\ 2/e \\ e - 5/e \end{bmatrix}.$$

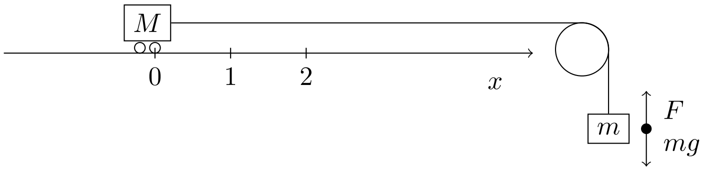
The normal equations then yields

$$a_0 = \frac{3(11 - e^2)}{4e},\ a_1 = \frac{3}{e},\ a_2 = \frac{15(e^2 - 7)}{4e}.$$

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 (6.23).

6.4 Discrete least squares (DLS)

Example 6.11 (An experiment on Newton's second law by 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.



Neglecting the friction of the track and the pulley system, we have from Newton's second law

$$m_j g = (m_j + M)a = (m_j + M)\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}.$$

A series of experiments can be designed to test the hypothesis of 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,

$$\min \sum_j (F_j - f(a_j))^2.$$

One verifies Newton's second law by showing that the data fitting result c_1 is very close to M. Note that the expressions in steps (ii) and (v) justify the name "least squares."

6.4.1 Reusing the formalism

Definition 6.25. Define a function $\lambda : \mathbb{R} \to \mathbb{R}$

$$\lambda(t) = \begin{cases} 0 & \text{if } t \in (-\infty, a), \\ \int_a^t \rho(\tau)\mathrm{d}\tau & \text{if } t \in [a, b], \\ \int_a^b \rho(\tau)\mathrm{d}\tau & \text{if } t \in (b, +\infty). \end{cases} \tag{6.35}$$

Then a corresponding *continuous measure* $\mathrm{d}\lambda$ can be defined as

$$\mathrm{d}\lambda = \begin{cases} \rho(t)\mathrm{d}t & \text{if } t \in [a, b], \\ 0 & \text{otherwise}, \end{cases} \tag{6.36}$$

where the *support of the continuous measure* $\mathrm{d}\lambda$ is the interval $[a, b]$.

Definition 6.26. 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 $\{t_1, t_2, \ldots, t_N\}$.

Definition 6.27. The *Heaviside function* is the truncated power function with exponent 0,

$$H(x) = x_+^0 = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{6.37}$$

Lemma 6.28. For a function $u : \mathbb{R} \to \mathbb{R}$, define

$$\lambda(t) = \sum_{i=1}^{N} \rho_i H(t - t_i), \tag{6.38}$$

and we have

$$\int_{\mathbb{R}} u(t)\mathrm{d}\lambda = \sum_{i=1}^{N} \rho_i u(t_i). \tag{6.39}$$

Proof. The *Dirac Delta function*, $\delta(x)$, is roughly a generalized function that satisfies

$$\delta(x) = \begin{cases} +\infty & x = 0, \\ 0 & x \neq 0. \end{cases} \tag{6.40}$$

Note: the above definition of $\delta(x)$ is heuristic. A rigorous one should employ the concept of measures.

Useful properties of $\delta(x)$ include

$$\int_{-\infty}^{+\infty} \delta(x)\mathrm{d}x = 1, \tag{6.41}$$

$$\int_{0}^{x} \delta(t)\mathrm{d}t = H(x), \tag{6.42}$$

$$\int_{-\infty}^{+\infty} f(t)\delta(t - t_0)\mathrm{d}t = f(t_0). \tag{6.43}$$

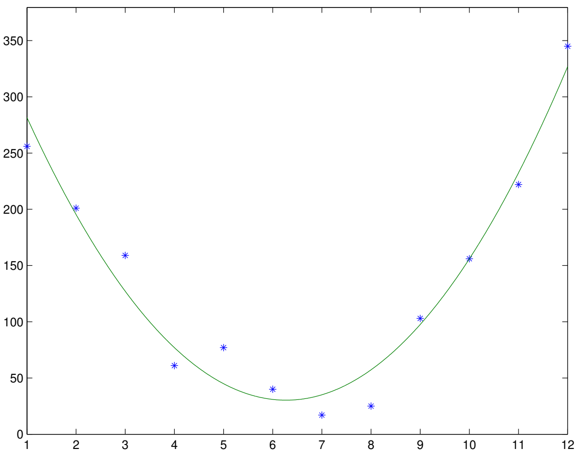
Then (6.38), (6.42), and (6.43) yield

$$\int_{\mathbb{R}} u(t)\mathrm{d}\lambda = \int_{\mathbb{R}} \sum_{i=1}^{N} \rho_i \delta(t - t_i) u(t)\mathrm{d}t = \sum_{i=1}^{N} \rho_i u(t_i). \qquad \square$$

6.4.2 DLS via normal equations

Example 6.12. 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 the following error,

$$\sum_{i=1}^{12} \left(y_i - \sum_{j=0}^{2} a_j x_i^j \right)^2 .$$

Reusing the procedures in Example 6.10, we have

$$\begin{aligned} G(1, x, x^2) &= \begin{bmatrix} \langle 1, 1 \rangle & \langle 1, x \rangle & \langle 1, x^2 \rangle \\ \langle x, 1 \rangle & \langle x, x \rangle & \langle x, x^2 \rangle \\ \langle x^2, 1 \rangle & \langle x^2, x \rangle & \langle x^2, x^2 \rangle \end{bmatrix} \\ &= \begin{bmatrix} 12 & 78 & 650 \\ 78 & 650 & 6084 \\ 650 & 6084 & 60710 \end{bmatrix}, \\ \mathbf{c} = \begin{bmatrix} \langle y, 1 \rangle \\ \langle y, x \rangle \\ \langle y, x^2 \rangle \end{bmatrix} &= \begin{bmatrix} \sum_{i=1}^{12} y_i \\ \sum_{i=1}^{12} y_i x_i \\ \sum_{i=1}^{12} y_i x_i^2 \end{bmatrix} = \begin{bmatrix} 1662 \\ 11392 \\ 109750 \end{bmatrix}. \end{aligned}$$

Then the normal equations yield

$$\mathbf{a} = G^{-1}\mathbf{c} = [386.00, \; -113.43, \; 9.04]^T.$$

The corresponding polynomial is plotted in the figure.

6.4.3 DLS via QR decomposition

Definition 6.29. A matrix $A \in \mathbb{R}^{n \times n}$ is *orthogonal* iff $A^T A = I$.

Definition 6.30. 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 6.31 (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 $\forall \ell < k_1,\ \xi_\ell = \mathbf{0}$.

(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}.

By Corollary 6.10, 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

$$A_r = A_r^* \begin{bmatrix} b_{11} & b_{21} & \ldots & b_{r1} \\ & b_{22} & \ldots & b_{r2} \\ & & \ddots & \vdots \\ & & & b_{rr} \end{bmatrix}. \tag{6.44}$$

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 (6.44) and show that the QR form of (6.44) is maintained. For those zero column vectors in (S-1), we have

$$\begin{aligned} A_\xi =& [\xi_1\ \ldots\ \xi_{k_1-1}\ u_1\ u_2\ \ldots\ u_r] \qquad (6.45) \\ =& A_r^* \begin{bmatrix} 0 & \ldots & 0 & b_{11} & b_{21} & \ldots & b_{r1} \\ 0 & \ldots & 0 & & b_{22} & \ldots & b_{r2} \\ \vdots & \ddots & \vdots & & & \ddots & \vdots \\ 0 & \ldots & 0 & & & & b_{rr} \end{bmatrix}. \end{aligned}$$

For each ξ_ℓ with $\ell \in R_j$ in (S-2), we have

$$\begin{aligned} & [u_1,\ u_2,\ \ldots,\ u_{j-1},\ \xi_\ell] \qquad (6.46) \\ =& \left[u_1^*,\ u_2^*,\ \ldots,\ u_{j-1}^*\right] \begin{bmatrix} b_{11} & \ldots & b_{j-1,1} & c_{\ell,1} \\ & \ddots & \vdots & \vdots \\ & & b_{j-1,j-1} & c_{\ell,j-1} \end{bmatrix}, \end{aligned}$$

where $\xi_\ell = c_{\ell,1}u_1^* + \ldots + c_{\ell,j-1}u_{j-1}^*$. With (6.45) as the induction basis and (6.46) 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 6.29 and 6.6 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 6.32. An orthogonal matrix preserves the 2-norm of the vectors it acts on.

Proof. Definition 6.29 yields

$$\forall \mathbf{x} \in \operatorname{dom}(Q), \qquad \|Q\mathbf{x}\|_2^2 = \mathbf{x}^T Q^T Q \mathbf{x} = \mathbf{x}^T\mathbf{x} = \|x\|_2^2. \quad \square$$

Theorem 6.33. 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

$$R_1 \mathbf{x}^* = \mathbf{c}, \tag{6.47}$$

where $R_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{c} \in \mathbb{R}^n$ result from the QR factorization of A:

$$Q^T A = R = \begin{bmatrix} R_1 \\ \mathbf{0} \end{bmatrix}, \qquad Q^T\mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{r} \end{bmatrix}. \tag{6.48}$$

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 6.32. □

6.5 Problems

6.5.1 Theoretical questions

I. Fill in the details for the proof of Theorem 6.4.

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 6.4 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 6.4.

(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 6.12.

(a) Starting from the independent list $(1, x, x^2)$, construct orthonormal polynomials by the Gram-Schmidt process using

$$\langle u(t), v(t)\rangle = \sum_{i=1}^N \rho(t_i)u(t_i)v(t_i) \tag{6.49}$$

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?

6.5.2 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.

Chapter 7

Numerical Integration

Definition 7.1. A *weighted quadrature formula* $I_n(f)$ for a function $f \in L[a,b]$ is a formula

$$I_n(f) := \sum_{k=1}^{n} w_k f(x_k) \tag{7.1}$$

that approximates the definite integral of f on $[a,b]$

$$I(f) := \int_a^b f(x)\rho(x)\mathrm{d}x, \tag{7.2}$$

where the weight function $\rho \in L[a,b]$ 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 *abscissa*, and the multiplier w_k's are called *weights* or *coefficients*.

Example 7.1. If a and/or b are infinite, $I(f)$ and $I_n(f)$ in (7.1) may still be well defined if the *moment of weight function*

$$\mu_j := \int_a^b x^j \rho(x)\mathrm{d}x \tag{7.3}$$

exists and is finite for all $j \in \mathbb{N}$.

7.1 Accuracy and convergence

Definition 7.2. The *remainder*, or *error*, of $I_n(f)$ is

$$E_n(f) := I(f) - I_n(f). \tag{7.4}$$

$I_n(f)$ is said to be *convergent* for $\mathcal{C}[a,b]$ iff

$$\forall f \in \mathcal{C}[a,b], \qquad \lim_{n\to+\infty} I_n(f) = I(f). \tag{7.5}$$

Definition 7.3. A subset $\mathbb{V} \subset \mathcal{C}[a,b]$ is *dense* in $\mathcal{C}[a,b]$ iff

$$\forall f \in \mathcal{C}[a,b], \forall \epsilon > 0, \exists f_\epsilon \in \mathbb{V}, \text{ s.t. } \max_{x\in[a,b]} |f(x) - f_\epsilon(x)| \le \epsilon. \tag{7.6}$$

Theorem 7.4. 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 (7.1) and (7.2). 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) $B := \sup_{n\in\mathbb{N}^+} \sum_{k=0}^{n} |w_k| < +\infty$.

Proof. For necessity, it is trivial to deduce (a) from (7.5). In contrast, it is highly nontrivial to deduce (b) from (7.5). This is an example of the principle of uniform boundedness, the proof of which is out of scope of this course. See a standard text on functional analysis, e.g. [Cryer, 1982, p. 121].

For the 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 (7.6) holds, define $K := \max_{x\in[a,b]} |f(x) - f_\epsilon(x)|$. Then we have

$$\begin{aligned}
|E_n(f)| \le& |I(f) - I(f_\epsilon)| + |I(f_\epsilon) - I_n(f_\epsilon)| + |I_n(f_\epsilon) - I_n(f)| \\
=& \left| \int_a^b [f(x) - f_\epsilon(x)]\, \rho(x)\mathrm{d}x \right| \\
&+ |I(f_\epsilon) - I_n(f_\epsilon)| + \left| \sum_{k=1}^{n} w_k\, [f(x_k) - f_\epsilon(x_k)] \right| \\
\le& K \left[\int_a^b \rho(x)\mathrm{d}x + \sum_{k=1}^{n} |w_k| \right] + |I(f_\epsilon) - I_n(f_\epsilon)|,
\end{aligned}$$

where the first step follows from the triangular inequality, the second from Definition 7.1, and the third from the integral mean value theorem 0.56. The terms inside the brackets is bounded because of $\rho \in L[a,b]$ and condition (b). By condition (a), $|I(f_\epsilon) - I_n(f_\epsilon)|$ can be made arbitrarily small. The proof is completed by the fact that K can also be arbitrarily small. □

Theorem 7.5 (Weierstrass)**.** The set of polynomials is dense in $\mathcal{C}[a,b]$. In other words, for any given $f(x) \in \mathcal{C}[a,b]$ and given $\epsilon > 0$, one can find a polynomial $p_n(x)$ (of sufficiently high degree) such that

$$\forall x \in [a,b], \qquad |f(x) - p_n(x)| \le \epsilon. \tag{7.7}$$

Proof. Not required. □

Definition 7.6. A weighted quadrature formula (7.1) has (polynomial) *degree of exactness* d_E iff

$$\left\{ \begin{array}{rl} \forall f \in \mathbb{P}_{d_E}, & E_n(f) = 0, \\ \exists g \in \mathbb{P}_{d_E+1}, \text{ s.t.} & E_n(g) \ne 0, \end{array} \right. \tag{7.8}$$

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

Lemma 7.7. Let $x_1, \ldots, x_n$ be given as distinct nodes of $I_n(f)$. If $d_E \ge n-1$, then its weights can be deduced as

$$\forall k = 1, \ldots, n, \qquad w_k = \int_a^b \rho(x)\ell_k(x)\mathrm{d}x, \tag{7.9}$$

where $\ell_k(x)$ is the fundamental polynomial for pointwise interpolation in (3.9) applied to the given nodes,

$$\ell_k(x) := \prod_{i\neq k; i=1}^{n} \frac{x-x_i}{x_k-x_i}. \tag{7.10}$$

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 3.4). Then we have

$$\begin{aligned}&\sum_{k=1}^{n} w_k p_{n-1}(x_k) = \int_a^b p_{n-1}(f;x)\rho(x)\mathrm{d}x\\ =&\int_a^b \sum_{k=1}^{n}\left\{\ell_k(x)f(x_k)\right\}\rho(x)\mathrm{d}x = \sum_{k=1}^{n} w_k f(x_k),\end{aligned}$$

where the first step follows from $d_E \ge n-1$ and the second step from the interpolation conditions (3.4), 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 4.16) for each x_k. □

7.2 Newton-Cotes formulas

Definition 7.8. A *Newton-Cotes formula* is a formula (7.1) based on approximating $f(x)$ by interpolating it on uniformly spaced nodes $x_1, \ldots, x_n \in [a,b]$.

Definition 7.9. *The trapezoidal rule* is a formula (7.1) based on approximating $f(x)$ by the straight line that connects $(a, f(a))^T$ and $(b, f(b))^T$. In particular, for $\rho(x) \equiv 1$, it is simply

$$I^T(f) = \frac{b-a}{2}[f(a)+f(b)]. \tag{7.11}$$

Example 7.2. Derive the trapezoidal rule for the weight function $\rho(x) = x^{-1/2}$ on the interval $[0,1]$. Note that one cannot apply (7.11) to $\rho(x)f(x)$ because $\rho(0) = \infty$. (7.9) 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

$$I^T(f) = \frac{2}{3}[2f(0)+f(1)]. \tag{7.12}$$

Theorem 7.10. For $f \in \mathcal{C}^2[a,b]$ with weight function $\rho(x) \equiv 1$, the remainder of the trapezoidal rule satisfies

$$\exists \zeta \in [a,b] \text{ s.t. } E^T(f) = -\frac{(b-a)^3}{12}f''(\zeta). \tag{7.13}$$

Proof. By Theorem 3.4, 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 3.6 and the second step from the integral mean value theorem (Theorem 0.56). Here we can apply Theorem 0.56 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 7.11. *Simpson's rule* is a formula (7.1) based on approximating $f(x)$ by a quadratic polynomial that goes through $(a, f(a))^T$, $(b, f(b))^T$, and $(\frac{a+b}{2}, f(\frac{a+b}{2}))^T$. For $\rho(x) \equiv 1$, it is simply

$$I^S(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]. \tag{7.14}$$

Theorem 7.12. For $f \in \mathcal{C}^4[a,b]$ with weight function $\rho(x) \equiv 1$, the remainder of Simpson's rule satisfies

$$\exists \zeta \in [a,b] \text{ s.t. } E^S(f) = -\frac{(b-a)^5}{2880}f^{(4)}(\zeta). \tag{7.15}$$

Proof. It is difficult to imitate the proof of Theorem 7.10, 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 0.56 can be applied. See problem I in Section 7.5 for the main steps. □

Example 7.3. Consider the integral

$$I = \int_{-4}^{4} \frac{\mathrm{d}x}{1+x^2} = 2\tan^{-1}(4) = 2.6516\cdots \tag{7.16}$$

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

Note $n-1$ is the number of sub-intervals that partition $[a,b]$ in Definition 7.8.

7.3 Composite formulas

Definition 7.13. The *composite trapezoidal rule* for approximating $I(f)$ in (7.2) with $\rho(x) \equiv 1$ is

$$I_n^T(f) = \frac{h}{2}f(x_0) + h\sum_{k=1}^{n-1} f(x_k) + \frac{h}{2}f(x_n), \tag{7.17}$$

where $h = \frac{b-a}{n}$ and $x_k = a + kh$.

Theorem 7.14. For $f \in \mathcal{C}^2[a,b]$, the remainder of the composite trapezoidal rule satisfies

$$\exists \xi \in (a,b) \text{ s.t. } E_n^T(f) = -\frac{b-a}{12}h^2 f''(\xi). \tag{7.18}$$

Proof. Apply Theorem 7.10 to the subintervals, sum up the errors, and we have

$$E_n^T(f) = -\frac{b-a}{12}h^2 \left[\frac{1}{n}\sum_{k=0}^{n-1} f''(\xi_k)\right]. \tag{7.19}$$

$f \in \mathcal{C}^2[a,b]$ implies $f'' \in \mathcal{C}[a,b]$. The proof is completed by (7.19) and the intermediate value Theorem 0.32. □

Definition 7.15. The *composite Simpson's rule* for approximating $I(f)$ in (7.2) with $\rho(x) \equiv 1$ is

$$\begin{aligned} I_n^S(f) = \frac{h}{3}\big[&f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) \\ &+ \cdots + 4f(x_{n-1}) + f(x_n)\big], \end{aligned} \tag{7.20}$$

where $h = \frac{b-a}{n}$, $x_k = a + kh$, and n is even.

Theorem 7.16. For $f \in \mathcal{C}^4[a,b]$ and $n \in 2\mathbb{N}^+$, the remainder of the composite Simpson's rule satisfies

$$\exists \xi \in (a,b) \text{ s.t. } E_n^S(f) = -\frac{b-a}{180}h^4 f^{(4)}(\xi). \tag{7.21}$$

Proof. Exercise. □

7.4 Gauss formulas

Lemma 7.17. Let $n, m \in \mathbb{N}^+$ and $m \le n$. 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

$$p = qs + r. \tag{7.22}$$

Proof. Rewrite (7.22) as

$$\sum_{i=0}^{n+m} p_i x^i = \left(\sum_{i=0}^{m} q_i x^i\right)\left(\sum_{i=0}^{n} s_i x^i\right) + \sum_{i=0}^{n-1} r_i x^i. \tag{7.23}$$

Since monomials are linearly independent, (7.23) 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 (7.23). □

Definition 7.18. The *node polynomial* associated with the nodes x_k's of a weighted quadrature formula is

$$v_n(x) = \prod_{k=1}^{n}(x - x_k). \tag{7.24}$$

Theorem 7.19. Suppose a quadrature formula (7.1) has $d_E \ge n-1$. Then it can be improved to have $d_E \ge n+j-1$ where $j \in (0,n]$ by and only by imposing the additional conditions on its node polynomial and weight function,

$$\forall p \in \mathbb{P}_{j-1}, \qquad \int_a^b v_n(x)p(x)\rho(x)\mathrm{d}x = 0. \tag{7.25}$$

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 $d_E \ge n+j-1$ and $v_n(x)p(x) \in \mathbb{P}_{n+j-1}$, and the second step from (7.24).

To prove the sufficiency, we must show that $E_n(p) = 0$ for any $p \in \mathbb{P}_{n+j-1}$. Lemma 7.17 yields

$$\forall p \in \mathbb{P}_{n+j-1},\ \exists q \in \mathbb{P}_{j-1},\ r \in \mathbb{P}_{n-1}, \text{ s.t. } p = qv_n + r. \tag{7.26}$$

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 (7.26), the second from (7.25), the third from the condition of $d_E \ge n-1$, the fourth from (7.26), and the last from (7.24). □

Definition 7.20. A *Gaussian quadrature formula* is a formula (7.1) whose nodes are the zeros of the polynomial $v_n(x)$ in (7.24) that satisfies (7.25) for $j = n$.

Corollary 7.21. A Gauss formula has $d_E = 2n-1$.

Proof. The index j in (7.25) 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 7.19 is optimal: the formula (7.1) achieves the highest degree of exactness $2n-1$. From an algebraic viewpoint, the $2n$ degrees of freedom of nodes and weights in (7.1) determine a polynomial of degree at most $2n-1$. The rest follows from Theorem 7.19. □

Corollary 7.22. Weights of a Gauss formula $I_n(f)$ are

$$\forall k = 1, \cdots, n, \quad w_k = \int_a^b \frac{v_n(x)}{(x-x_k)v_n'(x_k)}\rho(x)\mathrm{d}x, \tag{7.27}$$

where $v_n(x)$ is the node polynomial that defines $I_n(f)$.

Proof. This follows from Lemma 7.7; also see (3.11). □

Example 7.4. 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

$$\int_0^1 (c_0 - c_1 x + x^2)x^{-1/2}\mathrm{d}x = \frac{2}{5} + 2c_0 - \frac{2}{3}c_1 = 0,$$
$$\int_0^1 x(c_0 - c_1 x + x^2)x^{-1/2}\mathrm{d}x = \frac{2}{7} + \frac{2}{3}c_0 - \frac{2}{5}c_1 = 0.$$

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 (7.9), but it is simpler to set up a linear system of equations by exploiting Corollary 7.21, i.e. Gauss quadrature is exactly for all constants and linear polynomials,

$$w_1 + w_2 = \int_0^1 x^{-1/2}\mathrm{d}x = 2,$$
$$x_1 w_1 + x_2 w_2 = \int_0^1 x x^{-1/2}\mathrm{d}x = \frac{2}{3},$$

which yields

$$w_1 = \frac{-2x_2 + \frac{2}{3}}{x_1 - x_2}, \quad w_2 = \frac{2x_1 - \frac{2}{3}}{x_1 - x_2}.$$

The desired two-point Gauss formula is thus

$$\begin{aligned} I_2^G(f) = &\left(1 + \frac{1}{3}\sqrt{\frac{5}{6}}\right) f\left(\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}\right) \\ &+ \left(1 - \frac{1}{3}\sqrt{\frac{5}{6}}\right) f\left(\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}\right). \end{aligned} \tag{7.28}$$

Theorem 7.23. 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 $\int_a^b \rho(x)p_n(x)\mathrm{d}x = \langle p_n, p_0\rangle \neq 0$. But this contradicts orthogonality. Hence there exists $x_1 \in [a, b]$ such that $p_n(x_1) = 0$.

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 $\left|\langle P_{n-j}, v_j^2\rangle\right| > 0$, which contradicts the orthogonality of $p_n(x)$ and $v_j(x)$. □

Corollary 7.24. All nodes of a Gauss formula are real, distinct, and contained in (a, b).

Proof. This follows from Definition 7.20 and Theorem 7.23. □

Lemma 7.25. Gauss formulas have positive weights.

Proof. For each $j = 1, 2, \ldots, n$, the definition of $\ell_j(x)$ in (7.10) implies $\ell_j^2 \in \mathbb{P}_{2n-2}$, then we have

$$w_j = \sum_{k=1}^n w_k \ell_j^2(x_k) = \int_a^b \rho(x)\ell_j^2(x)\mathrm{d}x > 0,$$

where the first step follows from (7.10), second step from $d_E = 2n - 1$ and the last step from the conditions on ρ. □

Lemma 7.26. A Gauss formula satisfies

$$\sum_{k=1}^n w_k = \mu_0 \in (0, +\infty).$$

Proof. This follows from setting $j = 0$ in (7.3) and applying the condition on ρ in Definition 7.1. □

Theorem 7.27. Gauss formulas are convergent for $\mathcal{C}[a, b]$.

Proof. Denote by $\mathbb{P}$ the set of real polynomials. Theorem 7.5 states that $\mathbb{P}$ is dense in $\mathcal{C}[a, b]$, i.e. condition (a) in Theorem 7.4 holds. Condition (b) also holds because of Lemma 7.26, (7.3), and $\rho \in L[a, b]$. The rest of the proof follows from Theorem 7.4. □

Theorem 7.28. For $f \in \mathcal{C}^{2n}[a, b]$, the remainder of a Gauss formula $I_n(f)$ satisfies

$$\exists \xi \in [a, b] \text{ s.t. } E_n^G(f) = \frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b \rho(x)v_n^2(x)\mathrm{d}x, \tag{7.29}$$

where v_n is the node polynomial that defines I_n.

Proof. Not required. □

7.5 Problems

7.5.1 Theoretical questions

I. Simpson's rule.

(a) Show that on $[-1, 1]$ Simpson's rule can be obtained as follows

$$\int_{-1}^1 y(t)\mathrm{d}t = \int_{-1}^1 p_3(y; -1, 0, 0, 1; t)\mathrm{d}t + E^S(y),$$

where $y \in \mathcal{C}^4[-1, 1]$ and $p_3(y; -1, 0, 0, 1; t)$ is the interpolation polynomial of y with interpolation conditions $p_3(-1) = y(-1)$, $p_3(0) = y(0)$, $p_3'(0) = y'(0)$, 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 6 correct decimal places, i.e. the absolute error is no greater 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 $I = 0.596347361\cdots$)

Bibliography

C. W. Cryer. *Numerical Functional Analysis*. Monographs on Numerical Analysis. Oxford University Press, 1982. ISBN:9780198534105.