Notes on Numerical Analysis
数值分析+微分方程数值解

Qinghai Zhang
张庆海

Spring 2021
2021年春季学期

Contents

1 Solving Nonlinear Equations **1**
- 1.1 The bisection method . . . 1
- 1.2 The signature of an algorithm . . . 1
- 1.3 Proof of correctness and simplification of algorithms . . . 1
- 1.4 Q-order convergence . . . 1
- 1.5 Newton's method . . . 2
- 1.6 The secant method . . . 3
- 1.7 Fixed-point iterations . . . 5
- 1.8 Problems . . . 6
  - 1.8.1 Theoretical questions . . . 6
  - 1.8.2 Programming assignments . . . 7

2 Computer Arithmetic **8**
- 2.1 Floating-point number systems . . . 8
- 2.2 Rounding error analysis . . . 9
  - 2.2.1 Rounding a single number . . . 9
  - 2.2.2 Binary floating-point operations . . . 10
  - 2.2.3 The propagation of rounding errors . . . 11
- 2.3 Accuracy and stability . . . 12
  - 2.3.1 Avoiding catastrophic cancellation . . . 12
  - 2.3.2 Backward stability and numerical stability . . . 13
  - 2.3.3 Condition numbers: scalar functions . . . 14
  - 2.3.4 Condition numbers: vector functions . . . 15
  - 2.3.5 Condition numbers: algorithms . . . 15
  - 2.3.6 Overall error of a computer solution . . . 16
- 2.4 Problems . . . 16
  - 2.4.1 Theoretical questions . . . 16
  - 2.4.2 Programming assignments . . . 17

3 Polynomial Interpolation **18**
- 3.1 The Vandermonde determinant . . . 18
- 3.2 The Cauchy remainder . . . 18
- 3.3 The Lagrange formula . . . 19
- 3.4 The Newton formula . . . 19
- 3.5 The Neville-Aitken algorithm . . . 21
- 3.6 Hermite interpolation . . . 22
- 3.7 The Chebyshev polynomials . . . 22
- 3.8 Problems . . . 24
  - 3.8.1 Theoretical questions . . . 24
  - 3.8.2 Programming assignments . . . 24

4 Splines **26**
- 4.1 Piecewise-polynomial splines . . . 26
- 4.2 The minimum properties . . . 27
- 4.3 Error analysis . . . 28
- 4.4 B-Splines . . . 29
  - 4.4.1 Truncated power functions . . . 29
  - 4.4.2 The local support of B-splines . . . 30
  - 4.4.3 Integrals and derivatives . . . 31
  - 4.4.4 Marsden's identity . . . 32
  - 4.4.5 Symmetric polynomials . . . 33
  - 4.4.6 B-splines indeed form a basis . . . 34

4.4.7 Cardinal B-splines . . . 34
4.5 Curve fitting via splines . . . 36
4.6 Problems . . . 36
4.6.1 Theoretical questions . . . 36
4.6.2 Programming assignments . . . 37

5 Approximation **38**
5.1 Orthonormal systems . . . 39
5.2 Fourier expansions . . . 40
5.3 The normal equations . . . 41
5.4 Discrete least squares (DLS) . . . 43
5.4.1 Reusing the formalism . . . 43
5.4.2 DLS via normal equations . . . 44
5.4.3 DLS via QR decomposition . . . 44
5.5 Problems . . . 45
5.5.1 Theoretical questions . . . 45
5.5.2 Programming assignments . . . 46

6 Numerical Integration **47**
6.1 Accuracy and convergence . . . 47
6.2 Newton-Cotes formulas . . . 48
6.3 Composite formulas . . . 48
6.4 Gauss formulas . . . 49
6.5 Problems . . . 50
6.5.1 Theoretical questions . . . 50

7 Initial Value Problems (IVPs) **52**
7.1 Mathematical foundation . . . 52
7.1.1 Operator norm . . . 52
7.1.2 Matrix exponential . . . 54
7.1.3 Lipschitz continuity . . . 55
7.1.4 Existence and uniqueness of solution . . . 56
7.1.5 Linear IVPs with constant matrices . . . 57
7.1.6 Jordan canonical form . . . 58
7.2 Basic numerical methods . . . 59
7.2.1 Truncation errors and solution errors . . . 59
7.2.2 Convergence of Euler's method . . . 59
7.2.3 Zero stability and absolute stability . . . 60
7.3 Linear multistep methods . . . 62
7.3.1 Classical formulas . . . 62
7.3.2 Consistency and accuracy . . . 62
7.3.3 Zero stability . . . 64
7.3.4 Linear difference equations . . . 65
7.3.5 Convergence . . . 65
7.3.6 Absolute stability . . . 67
7.3.7 The first Dahlquist barrier . . . 69
7.4 Runge-Kutta methods . . . 69
7.4.1 Classical formulas . . . 69
7.4.2 Consistency and convergence . . . 70
7.4.3 Absolute stability . . . 72
7.5 Stiff IVPs . . . 73
7.5.1 The notion of stiffness . . . 73
7.5.2 A-stability and L-stability . . . 73

8 Boundary Value Problems (BVPs) **75**
8.1 Finite difference (FD) methods . . . 75
8.2 Errors and consistency . . . 76
8.3 Convergence and stability . . . 77
8.3.1 Stability in the 2-norm . . . 77
8.3.2 Gaussian and Dirac delta functions . . . 78
8.3.3 Green's function . . . 78
8.3.4 Stability in the max-norm . . . 79
8.4 A solution via Green's function . . . 79
8.5 Other boundary conditions . . . 80
8.6 BVPs in two dimensions . . . 81

8.6.1 Kronecker product . . . 81
8.6.2 Convergence in the 2-norm . . . 82
8.6.3 Convergence in the max-norm via a discrete maximum principle . . . 83
8.6.4 Convergence on irregular domains . . . 84

9 Multigrid Methods **85**
9.1 The model problem . . . 85
9.1.1 The residual equation . . . 85
9.1.2 Fourier modes on Ω^h . . . 85
9.2 Classical iterative methods . . . 86
9.3 Key elements of multigrid . . . 86
9.3.1 Restriction and prolongation . . . 86
9.3.2 Two-grid correction . . . 87
9.3.3 Multigrid cycles . . . 87
9.4 Why multigrid methods work? . . . 88
9.4.1 The spectral picture . . . 88
9.4.2 The algebraic picture . . . 90
9.4.3 The optimal complexity of FMG . . . 91

10 Parabolic Problems **92**
10.1 Parabolic equations . . . 92
10.2 The method of lines (MOL) . . . 92
10.3 Accuracy and consistency . . . 93
10.4 Stability . . . 93
10.5 Convergence . . . 94
10.6 Von Neumann analysis . . . 94
10.7 Green's function of the heat equation in $(-\infty, +\infty)$ . . . 95

11 Hyperbolic Problems **97**
11.1 Classical MOLs . . . 97
11.1.1 The FTCS method . . . 98
11.1.2 The leapfrog method . . . 98
11.1.3 Lax-Friedrichs . . . 98
11.1.4 Lax-Wendroff . . . 99
11.1.5 The Upwind method . . . 99
11.1.6 The Beam-Warming method . . . 100
11.2 The CFL condition . . . 100
11.3 Modified equations . . . 101
11.4 Von Neumann analysis . . . 103

12 Fourth-order Finite Volume (FV) Methods **104**
12.1 The FV formulation . . . 104
12.2 Discrete operators . . . 105
12.3 Ghost cells . . . 107
12.4 FV methods for BVPs . . . 108
12.5 FV-MOL algorithms for the advection-diffusion equation . . . 108

13 FV Methods for the Incompressible Navier-Stokes Equations (INSE) **111**
13.1 Leray-Helmholtz Projection . . . 111
13.2 The approximate projection . . . 112
13.3 The INSE on periodic domains . . . 113

A Sets, Logic, and Functions **114**
A.1 First-order logic . . . 114
A.2 Ordered sets . . . 115
A.3 Functions . . . 116

B Linear Algebra **117**
B.1 Vector spaces . . . 117
B.1.1 Subspaces . . . 117
B.1.2 Span and linear independence . . . 118
B.1.3 Bases . . . 118
B.1.4 Dimension . . . 119
B.2 Linear maps . . . 119
B.2.1 Null spaces and ranges . . . 119
B.2.2 The matrix of a linear map . . . 119

B.2.3 Duality . . . 119
B.3 Eigenvalues, eigenvectors, and invariant subspaces . . . 122
B.3.1 Invariant subspaces . . . 122
B.3.2 Upper-triangular matrices . . . 122
B.3.3 Eigenspaces and diagonal matrices . . . 122
B.4 Inner product spaces . . . 123
B.4.1 Inner products . . . 123
B.4.2 Norms induced from inner products . . . 123
B.4.3 Norms and induced inner-products . . . 123
B.4.4 Orthonormal bases . . . 124
B.5 Operators on inner-product spaces . . . 125
B.5.1 Adjoint and self-adjoint operators . . . 125
B.5.2 Normal operators . . . 126
B.5.3 The spectral theorem . . . 127
B.5.4 Isometries . . . 127
B.5.5 The singular value decomposition . . . 127
B.6 Trace and determinant . . . 127

C Basic Analysis **129**
C.1 Sequences . . . 129
C.1.1 Convergence . . . 129
C.1.2 Limit points . . . 129
C.2 Series . . . 130
C.3 Continuous functions on $\mathbb{R}$ . . . 131
C.4 Differentiation of functions . . . 131
C.5 Taylor series . . . 132
C.6 Riemann integral . . . 133
C.7 Metric spaces . . . 133
C.8 Uniform convergence . . . 135
C.9 Vector calculus . . . 136

D Fourier Analysis of Linear PDEs **138**
D.1 Fourier transform . . . 138
D.2 Fourier analysis . . . 140

Preface

This book comes out from my teaching two courses at the school of mathematical sciences in Zhejiang University. The first six chapters come from the course "Numerical Analysis" (formerly "Numerical Approximation") in the fall semester of 2016 and in the spring semesters of 2018, 2019, and 2020. The other chapters are from the course "Numerical Solutions of Differential Equations" in the fall semester of 2020 and the spring semester of 2021.

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 1

Solving Nonlinear Equations

1.1 The bisection method

Algorithm 1.1. 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 : $f \in \mathcal{C}[a, b],$
$\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$
Output: c, h, k
Postconditions: $|f(c)| < \epsilon$ or $|h| < \delta$ or $k = M$

```
1  u ← f(a)
2  v ← f(b)
3  for k = 1 : M do
4  |   h ← b − a
5  |   c ← a + h/2
6  |   w ← f(c)
7  |   if |h| < δ or |w| < ϵ then
8  |   |   break
9  |   else if sgn(w) ≠ sgn(u) then
10 |   |   b ← c
11 |   |   v ← w
12 |   else
13 |   |   a ← c
14 |   |   u ← w
15 |   end
16 end
```

1.2 The signature of an algorithm

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

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

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

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

1.3 Proof of correctness and simplification of algorithms

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

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

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

Algorithm 1.9. A simplified bisection algorithm.

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

```
1  h ← b − a
2  u ← f(a)
3  for k = 1 : M do
4  |   h ← h/2
5  |   c ← a + h
6  |   w ← f(c)
7  |   if |h| < δ or |w| < ϵ then
8  |   |   break
9  |   else if sgn(w) = sgn(u) then
10 |   |   a ← c
11 |   end
12 end
```

1.4 Q-order convergence

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

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

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

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

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

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

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

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

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

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

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

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

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

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 1.12, both $\{a_n\}$ and $\{b_n\}$ converge. Also, $\lim(b_n - a_n) = \lim \frac{1}{2^n}(b_0 - a_0) = 0$, hence $\lim b_n = \lim a_n = \alpha$. By the given condition and the algorithm, the invariant $f(a_n)f(b_n) \le 0$ 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 $f(\alpha) = 0$. (1.6) is another important invariant that can be proven by induction. Comparing (1.6) to (1.2) yields convergence of the bisection method. Also, the convergence is linear with asymptotic factor as $c = \frac{1}{2}$. □

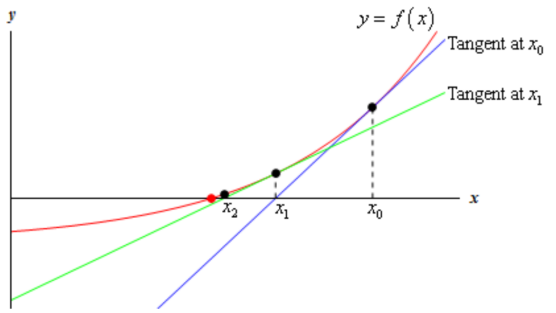
1.5 Newton's method

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

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

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

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



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

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

Proof. By Taylor's theorem (Theorem C.53) 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 (1.7), we have

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

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

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

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

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

and pick x_0 sufficiently close to α such that

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

(ii) $M\delta_0 < 1$.

The definition of M and (*) imply

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

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

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

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

which shows both (a) and (b) and completes the proof. □

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

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

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

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

Then $f' > 0$ and $f'' > 0$ further imply 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 1.12 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 (1.9), as in Theorem 1.15. □

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

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

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

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

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

1.6 The secant method

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

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

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

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

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

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

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

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

Proof. By Definition 1.20, we have

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

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

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

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

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

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

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

the second equation of which yields (1.14). □

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

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

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

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

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

Proof. Define a divided difference as

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

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

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

By (1.17) and the mean value theorem (Theorem C.44), there exists ξ_n between x_{n-1} and x_n such that

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

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

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

for some β between x_{n-1} and α. Compute the derivative of $g'(\beta)$ from (1.17), use the Lagrangian remainder Theorem C.53, and we have

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

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

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

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

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

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

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

and we have from Lemma 1.23

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

Pick x_0, x_1 such that

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

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

then an induction by the above equation shows that $E_n < \delta$, $ME_n < \eta$. To prove convergence, we write $ME_0 < \eta$, $ME_1 < \eta$, $ME_2 < ME_1 ME_0 < \eta^2$, $ME_3 < ME_2 ME_1 < \eta^3$, $\cdots$, $ME_{n+1} < ME_n ME_{n-1} < \eta^{q_n + q_{n-1}} = \eta^{q_{n+1}}$, 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 1.21, as $n \to \infty$ we have $q_n \to \frac{1.618^{n+1}}{\sqrt{5}}$ since $|r_1| \approx 0.618 < 1$. Hence $\lim_{n \to \infty} E_n = 0$.

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

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

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

To prove convergence rates, we define

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

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

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

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

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

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

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

which means

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, m_n \in (m_\alpha - \epsilon, m_\alpha + \epsilon)\,. \tag{1.25}$$

We define

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

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 (1.25)

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

and then

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

The proof is then completed by Definition 1.10. □

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

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

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

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

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

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

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

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

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

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

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

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

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. □

1.7 Fixed-point iterations

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

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

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

Proof. The function $f(x) = g(x) - x$ satisfies $f(a) \ge 0$ and $f(b) \le 0$. The proof is then completed by the intermediate value theorem (Theorem C.32). □

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

Theorem 1.30 (Brouwer's fixed point)**.** Any 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 1.31. Take two pieces of the same-sized paper and lay one on top of the other. Every point on the top sheet of paper is associated with some point right below it on the bottom sheet. Crumple the top sheet into a ball without ripping it. Place the crumpled ball on top of (and simultaneously within the realm of) the bottom sheet of paper. Use Theorem 1.30 to prove that there always exists some point in the crumpled ball that sits above the same point it sat above prior to crumpling.

Example 1.32. 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 1.33. A *fixed-point iteration* is a method for finding a fixed point of g with a formula of the form

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

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

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

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

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

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

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

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

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

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

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

By induction and the triangle inequality,

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

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 (1.31). □

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

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

Proof. The mean value theorem (Theorem C.44) implies that, for all $x, y \in [a, b]$, $|g(x) - g(y)| \le \lambda|x - y|$. Theorem 1.38 yields all the results except (1.32), which follows from

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

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

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

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

Corollary 1.41. Consider $g : [a, b] \to [a, b]$ with a fixed point $g(\alpha) = \alpha \in [a, b]$. The fixed-point iteration (1.29) converges to α with pth-order accuracy ($p > 1$, $p \in \mathbb{N}$) for any choice $x_0 \in [a, b]$ if

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

Proof. By Corollary 1.40, 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 C.41 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 1.42. The following method has third-order convergence for computing $\sqrt{R}$:

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

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

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

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

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

The rest follows from Corollary 1.41.

1.8 Problems

1.8.1 Theoretical questions

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

- What is the width of the interval at the nth step?
- What is the 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?

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

Prove that the sequence of values converges. (Hint: this can be interpreted as $x = \lim_{n\to\infty} x_n$, where $x_1 = \frac{1}{p}$, $x_2 = \cfrac{1}{p + \cfrac{1}{p}}$, $x_3 = \cfrac{1}{p + \cfrac{1}{p + \cfrac{1}{p}}}$, ..., 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:

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

1.8.2 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 2

Computer Arithmetic

2.1 Floating-point number systems

Definition 2.1. A *bit* is the basic unit of information in computing; it can have only one of two values 0 and 1.

Definition 2.2. 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 2.3. 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 2.4. 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 2.5 (Floating point numbers)**.** A *floating point number* (FPN) is a number of the form

$$x = \pm m \times \beta^e, \tag{2.1}$$

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

$$m = \left(d_0 + \frac{d_1}{\beta} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right), \tag{2.2}$$

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 2.6. 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 2.7. Convert 156 to binary number:

$$156 = (10011100)_2.$$

Example 2.8. 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 2.9 (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

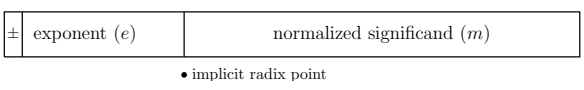
- the *base* (or radix) β;
- the *precision* (or significand digits) p;
- the *exponent range* $[L, U]$.

Definition 2.10. An FPN is *normalized* if its mantissa satisfies $1 \le m < \beta$.

Definition 2.11 (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,

$$\beta = 2,\ p = 23 + 1,\ e \in [-126, 127], \tag{2.3a}$$
$$\beta = 2,\ p = 52 + 1,\ e \in [-1022, 1023]. \tag{2.3b}$$

Example 2.12. 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 2.13. The *machine precision* of a normalized FPN system $\mathcal{F}$ is the distance between 1.0 and the next larger FPN in $\mathcal{F}$,

$$\epsilon_M := \beta^{1-p}. \tag{2.4}$$

Definition 2.14. The underflow limit (UFL) and the overflow limit (OFL) of a normalized FPN system $\mathcal{F}$ are respectively

$$\mathrm{UFL}(\mathcal{F}) := \min |\mathcal{F} \setminus \{0\}| = \beta^L, \tag{2.5}$$
$$\mathrm{OFL}(\mathcal{F}) := \max |\mathcal{F}| = \beta^U(\beta - \beta^{1-p}). \tag{2.6}$$

Example 2.15. 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 2.16 (Cardinality of $\mathcal{F}$)**.** For a normalized binary FPN system $\mathcal{F}$,

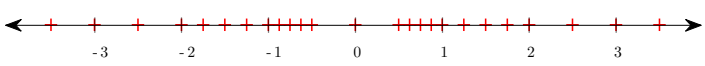
$$\#\mathcal{F} = 2^p(U - L + 1) + 1. \tag{2.7}$$

Proof. The cardinality can be proved by Axiom A.21. The factor 2^p comes from the sign bit and the mantissa. By Example 2.12, $U-L+1$ is the number of exponents represented in $\mathcal{F}$. The trailing "+1" in (2.7) accounts for the number 0. □

Definition 2.17. The *range* of a normalized FPN system is a subset of $\mathbb{R}$,

$$\mathcal{R}(\mathcal{F}) := \{x : x \in \mathbb{R}, \mathrm{UFL}(\mathcal{F}) \le |x| \le \mathrm{OFL}(\mathcal{F})\}. \tag{2.8}$$

Example 2.18. 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 2.19. Two normalized FPNs a, b are *adjacent* to each other in $\mathcal{F}$ iff

$$\forall c \in \mathcal{F} \setminus \{a, b\},\quad |a - b| < |a - c| + |c - b|. \tag{2.9}$$

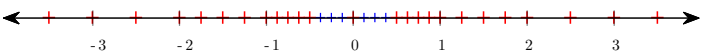
Lemma 2.20. Let a, b be two adjacent normalized FPNs satisfying $|a| < |b|$ and $ab > 0$. Then

$$\beta^{-1}\epsilon_M|a| < |a - b| \le \epsilon_M|a|. \tag{2.10}$$

Proof. Consider $a > 0$, then $\Delta a := b - a > 0$. By Definitions 2.5 and 2.10, $a = m \times \beta^e$ 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 2.21. The *subnormal* or *denormalized* numbers are FPNs of the form (2.1) with $e = L$ and $m \in (0, 1)$. A normalized FPN system can be *extended* by including the subnormal numbers.

Example 2.22. Add subnormal FPNs to the FPN system in Example 2.18 and we have the following plot.



2.2 Rounding error analysis

2.2.1 Rounding a single number

Definition 2.23 (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 2.24. 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 2.25. The *unit roundoff* of $\mathcal{F}$ is the number

$$\epsilon_u := \frac{1}{2}\epsilon_M = \frac{1}{2}\beta^{1-p}. \tag{2.11}$$

Theorem 2.26. For $x \in \mathcal{R}(\mathcal{F})$ as in (2.8), we have

$$\mathrm{fl}(x) = x(1 + \delta), \qquad |\delta| < \epsilon_u. \tag{2.12}$$

Proof. By Definition A.32, $\mathcal{R}(\mathcal{F})$ is a subset of $\mathbb{R}$ and is thus a chain. Therefore $\forall x \in \mathcal{R}(\mathcal{F})$, $\exists x_L, x_R \in \mathcal{F}$ s.t.

- x_L and x_R are adjacent,
- $x_L \le x \le x_R$.

If $x = x_L$ or x_R, then $\mathrm{fl}(x) - x = 0$ and (2.12) clearly holds. Otherwise $x_L < x < x_R$. Then Lemma 2.20 and Definitions 2.19 and 2.23 yield

$$|\mathrm{fl}(x)-x| \le \frac{1}{2}|x_R-x_L| \le \epsilon_u \min(|x_L|,|x_R|) < \epsilon_u|x|. \quad (2.13)$$

Hence $-\epsilon_u|x| < \mathrm{fl}(x) - x < \epsilon_u|x|$, which yields (2.12). □

Theorem 2.27. For $x \in \mathcal{R}(\mathcal{F})$, we have

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \qquad |\delta| \le \epsilon_u. \quad (2.14)$$

Proof. The proof is the same as that of Theorem 2.26, except that we replace the last inequality "$< \epsilon_u|x|$" in (2.13) by "$\le \epsilon_u|\mathrm{fl}(x)|$." Consequently, the equality in (2.14) holds when $x = \frac{1}{2}(x_L + x_R)$ and $\mathrm{fl}(x) = x_L$ has $m = 1.0$. □

Example 2.28. 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 2.8, we have

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

where the last bit of x_L must be 0 because the IEEE 754 standard states that 23 bits are reserved for the mantissa. It follows that

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

Thus Definition 2.23 implies $\mathrm{fl}(x) = x_R$.

2.2.2 Binary floating-point operations

Definition 2.29 (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 2.30. 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 2.31. 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 2.32. Repeat Example 2.30 with $b = 8.769 \times 10^4$, $b = -5.678 \times 10^0$, and $b = -5.678 \times 10^3$.

Lemma 2.33. For $a, b \in \mathcal{F}$, $a + b \in \mathcal{R}(\mathcal{F})$ implies

$$\mathrm{fl}(a + b) = (a + b)(1 + \delta), \qquad |\delta| < \epsilon_u. \quad (2.15)$$

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 (2.15) follows from Theorem 2.26. □

Definition 2.34 (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 2.35. Consider the calculation of $c := \mathrm{fl}(ab)$ with $a = 2.345 \times 10^4$ and $b = 6.789 \times 10^0$ in an FPN system $\mathcal{F} : (10, 4, -7, 8)$.

(i) $e_c \leftarrow 4$.

(ii) $M_c \leftarrow 15.9202$.

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

$$\mathrm{fl}(ab) = (ab)(1+\delta), \qquad |\delta| < \epsilon_u. \tag{2.16}$$

Proof. The error only comes from the round-off in steps (ii) and (v). Then (2.16) follows from Theorem 2.26. □

Definition 2.37 (Division of two FPNs)**.** Express $a, b \in \mathcal{F}$ as $a = M_a \times \beta^{e_a}$ and $b = M_b \times \beta^{e_b}$ where $M_a = \pm m_a$ and $M_b = \pm m_b$. The quotient $c = \mathrm{fl}\left(\frac{a}{b}\right) \in \mathcal{F}$ is calculated in a register of precision at least $2p+1$ as follows.

(i) If $m_b = 0$, return NaN; otherwise set $e_c \leftarrow e_a - e_b$.

(ii) Perform the division $M_c \leftarrow M_a/M_b$ in the register with rounding to nearest.

(iii) Normalization:

- If $|M_c| < 1$, set $M_c \leftarrow M_c\beta$, $e_c \leftarrow e_c - 1$.

(iv) Check range:

- return NaN if e_c overflows,
- return 0 if e_c underflows.

(v) Round M_c (to nearest) to precision p.

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

Lemma 2.38. For $a, b \in \mathcal{F}$, $\frac{a}{b} \in \mathcal{R}(\mathcal{F})$ implies

$$\mathrm{fl}\left(\frac{a}{b}\right) = \frac{a}{b}(1+\delta), \qquad |\delta| < \epsilon_u. \tag{2.17}$$

Proof. In the case of $|M_a| = |M_b|$, there is no rounding error in Definition 2.37 and (2.17) clearly holds. Hereafter we denote by M_{c1} and M_{c2} the results of steps (ii) and (v) in Definition 2.37, respectively.

In the case of $|M_a| > |M_b|$, the condition $a, b \in \mathcal{F}$, Definition 2.13, and $|M_a|, |M_b| \in [1, \beta)$ imply

$$\left|\frac{M_a}{M_b}\right| \geq \frac{\beta - \epsilon_M}{\beta - 2\epsilon_M} > 1 + \beta^{-1}\epsilon_M, \tag{2.18}$$

which further implies that the normalization step (iii) in Definition 2.37 is not invoked. By Definitions 2.23, 2.13, and 2.25, 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 2.37 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 (2.18) 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 2.37 step (iii) is invoked to yield

$$\begin{aligned} M_{c1} &= \frac{M_a}{M_b} + \delta_1, \qquad |\delta_1| < \beta^{-2}\epsilon_u\epsilon_M; \\ M_{c2} &= \beta M_{c1} + \delta_2, \qquad |\delta_2| < \epsilon_u \\ &= \beta\frac{M_a}{M_b}\left(1 + \frac{\beta\delta_1 + \delta_2}{\beta M_a/M_b}\right), \end{aligned}$$

where the denominator in the parentheses satisfies

$$\beta\left|\frac{M_a}{M_b}\right| \geq \frac{\beta}{\beta - \epsilon_M} = 1 + \frac{\epsilon_M}{\beta - \epsilon_M} > 1 + \beta^{-1}\epsilon_M.$$

Hence we have

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

Theorem 2.39 (Model of machine arithmetic)**.** Denote by $\mathcal{F}$ a normalized FPN system with precision p. For each arithmetic operation $\odot = +, -, \times, /$, we have

$$\forall a, b \in \mathcal{F},\ a \odot b \in \mathcal{R}(\mathcal{F}) \ \Rightarrow\ \mathrm{fl}(a \odot b) = (a \odot b)(1+\delta) \tag{2.19}$$

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 2.33, 2.36, and 2.38. □

2.2.3 The propagation of rounding errors

Theorem 2.40. If $\forall i = 0, 1, \cdots, n$, $a_i \in \mathcal{F}$, $a_i > 0$, then

$$\mathrm{fl}\left(\sum_{i=0}^{n} a_i\right) = (1+\delta_n)\sum_{i=0}^{n} a_i, \tag{2.20}$$

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 2.26 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

$$\begin{aligned} \forall k \in \mathbb{N},\ |\delta_{k+1}| &< \epsilon_u \sum_{i=0}^{k} (1+\epsilon_u)^i \qquad (2.21) \\ &= \epsilon_u \frac{(1+\epsilon_u)^{k+1} - 1}{1 + \epsilon_u - 1} = (1+\epsilon_u)^{k+1} - 1, \end{aligned}$$

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

Exercise 2.41. 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 2.40. Can you give some examples?

Exercise 2.42. 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 2.43. 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

$$1 - n\mu \le \prod_{i=1}^{n}(1+\delta_i) \le 1 + n\mu + (n\mu)^2, \qquad (2.22)$$

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

$$\exists \theta \in I_n \text{ s.t. } \prod_{i=1}^{n}(1+\delta_i) = 1 + \theta(n\mu + n^2\mu^2). \qquad (2.23)$$

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 (2.22). 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 (2.22).

Finally, (2.22) 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. □

2.3 Accuracy and stability

2.3.1 Avoiding catastrophic cancellation

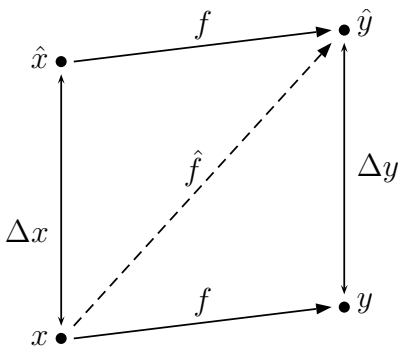
Definition 2.44. Let $\hat{x}$ be an approximation to $x \in \mathbb{R}$. The accuracy of $\hat{x}$ can be measured by its *absolute error*

$$E_{\mathrm{abs}}(\hat{x}) = |\hat{x} - x| \qquad (2.24)$$

and/or its *relative error*

$$E_{\mathrm{rel}}(\hat{x}) = \frac{|\hat{x} - x|}{|x|}. \qquad (2.25)$$

Definition 2.45. 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 2.46 (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 2.47 (Catastrophic cancellation)**.** For two real numbers $x, y \in \mathcal{R}(\mathcal{F})$, Theorems 2.26 and 2.39 imply

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

where $|\delta_i| \le \epsilon_u$. From Theorems 2.39 and 2.43, we know that *multiplication is accurate*:

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

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

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

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

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

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

Theorem 2.48 (Loss of most significant digits)**.** Suppose $x, y \in \mathcal{F}$, $x > y > 0$, and

$$\beta^{-t} \le 1 - \frac{y}{x} \le \beta^{-s}. \tag{2.26}$$

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 2.29 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 2.49. Catastrophic cancellation should be avoided whenever possible.

Example 2.50. 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

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

2.3.2 Backward stability and numerical stability

Definition 2.51 (Backward stability)**.** An algorithm $\hat{f}(x)$ for computing $y = f(x)$ is *backward stable* if its backward error is small for all x, i.e.

$$\begin{aligned}&\forall x \in \text{dom}(f),\ \exists \hat{x} \in \text{dom}(f),\ \text{ s.t.} \\ &\hat{f}(x) = f(\hat{x}) \ \Rightarrow\ E_{\text{rel}}(\hat{x}) \le c\epsilon_u,\end{aligned} \tag{2.27}$$

where c is a small constant.

Definition 2.52. An algorithm $\hat{f}(x_1, x_2)$ for computing $y = f(x_1, x_2)$ is *backward stable* if

$$\begin{aligned}&\forall (x_1, x_2) \in \text{dom}(f),\ \exists (\hat{x}_1, \hat{x}_2) \in \text{dom}(f) \text{ s.t.} \\ &\hat{f}(x_1, x_2) = f(\hat{x}_1, \hat{x}_2) \ \Rightarrow\ \begin{cases} E_{\text{rel}}(\hat{x}_1) \le c_1\epsilon_u, \\ E_{\text{rel}}(\hat{x}_2) \le c_2\epsilon_u, \end{cases}\end{aligned} \tag{2.28}$$

where c_1, c_2 are two small constants.

Corollary 2.53. For $f(x_1, x_2) = x_1 - x_2$, $x_1, x_2 \in \mathcal{R}(\mathcal{F})$, the algorithm $\hat{f}(x_1, x_2) = \text{fl}(\text{fl}(x_1) - \text{fl}(x_2))$ is backward stable.

Proof. We have $\hat{f}(x_1, x_2) = (\text{fl}(x_1) - \text{fl}(x_2))(1 + \delta_3)$ from Theorem 2.39. Then Theorem 2.26 implies

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

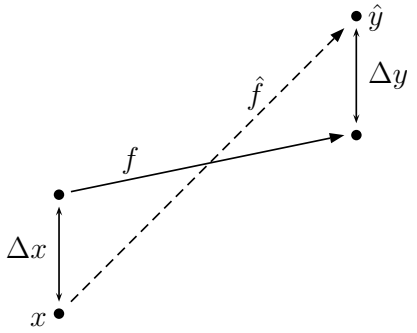
Take $\hat{x}_1$ and $\hat{x}_2$ to be the two terms in the above line and we have

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

Then Definition 2.52 completes the proof. □

Example 2.54. For $f(x) = 1 + x$, $x \in (0, \text{OFL})$, show that the algorithm $\hat{f}(x) = \text{fl}(1.0 + \text{fl}(x))$ is not backward stable.

We prove a stronger statement that implies the negation of (2.27). For each $x \in (0, \epsilon_u)$, Definition 2.23 yields $\hat{f}(x) = 1.0$. Then $\hat{f}(x) = f(\hat{x})$ implies $\hat{x} = 0$, which further implies $E_{\text{rel}}(\hat{x}) = 1$.



Definition 2.55. An algorithm $\hat{f}(x)$ for computing $y = f(x)$ is *stable* or *numerically stable* iff

$$\forall x \in \mathrm{dom}(f),\ \exists \hat{x} \in \mathrm{dom}(f) \text{ s.t. } \begin{cases} \left|\frac{\hat{f}(x)-f(\hat{x})}{f(\hat{x})}\right| \le c_f \epsilon_u, \\ E_{\mathrm{rel}}(\hat{x}) \le c\epsilon_u, \end{cases} \tag{2.29}$$

where c_f, c are two small constants.

Corollary 2.56. If an algorithm is backward stable, then it is numerically stable.

Proof. By Definition 2.51, $f(\hat{x}) = \hat{f}(x)$, hence $c_f = 0$. The other condition also follows trivially. □

Example 2.57. 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 2.25 and 2.17) yield $x \in \mathcal{R}(\mathcal{F})$. By Theorem 2.26, $\hat{f}(x) = (1 + x(1 + \delta_1))(1 + \delta_2)$, 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 $x > 0$.

2.3.3 Condition numbers: scalar functions

Definition 2.58. 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,

$$C_f(x) = \left|\frac{xf'(x)}{f(x)}\right|. \tag{2.30}$$

Definition 2.59. 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 2.60. Definition 2.58 yields

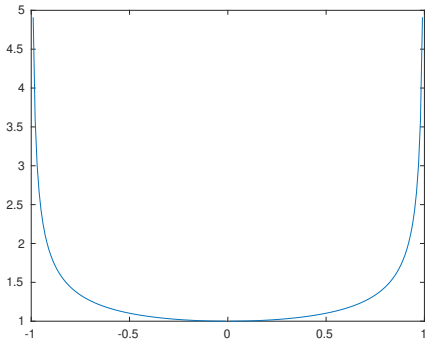
$$E_{\mathrm{rel}}(\hat{y}) \lessapprox C_f E_{\mathrm{rel}}(\hat{x}). \tag{2.31}$$

The approximation mark "≈" refers to the fact that the quadratic term $(\Delta x)^2$ has been ignored. As one way to interpret (2.31) and to understand Definition 2.58, *the computed solution to an ill-conditioned problem may have a large forward error.*

Example 2.61. For the function $f(x) = \arcsin(x)$, its condition number, according to Definition 2.58, 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 2.62. 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

$$h \approx -\epsilon \frac{g(r)}{f'(r)}. \tag{2.32}$$

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 2.63 (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 2.62, 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.

2.3.4 Condition numbers: vector functions

Definition 2.64. The *condition number of a vector function* $\mathbf{f}:\mathbb{R}^m\to\mathbb{R}^n$ is

$$\operatorname{cond}_{\mathbf{f}}(\mathbf{x})=\frac{\|\mathbf{x}\|\,\|\nabla\mathbf{f}\|}{\|\mathbf{f}(\mathbf{x})\|},\tag{2.33}$$

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

Example 2.65. 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 2.64 yields

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

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

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

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

$$\|A\|:=\max_{\|\mathbf{u}\|\neq0}\frac{\|A\mathbf{u}\|}{\|\mathbf{u}\|}.\tag{2.34}$$

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

$$\operatorname{cond}A=\left\|A\right\|\left\|A^{-1}\right\|.\tag{2.35}$$

Definition 2.66. The *componentwise condition number* of a vector function $\mathbf{f}:\mathbb{R}^m\to\mathbb{R}^n$ is

$$\operatorname{cond}_{\mathbf{f}}(\mathbf{x})=\|A(\mathbf{x})\|,\tag{2.36}$$

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

$$a_{ij}(\mathbf{x})=\left|\frac{x_j\frac{\partial f_i}{\partial x_j}}{f_i(\mathbf{x})}\right|.\tag{2.37}$$

Example 2.67. 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 2.66 clearly captures the fact that $x_1\pm x_2\approx0$ 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 2.64 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 $x_1\pm x_2\approx0$ yields $C_1\approx2$. 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 2.68. The *Hilbert matrix* $H_n\in\mathbb{R}^{n\times n}$ is

$$h_{i,j}=\frac{1}{i+j-1}.\tag{2.38}$$

Definition 2.69. The *Vandermonde matrix* $V_n\in\mathbb{R}^{n\times n}$ is

$$v_{i,j}=t_j^{i-1},\tag{2.39}$$

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

2.3.5 Condition numbers: algorithms

Definition 2.70. 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

$$\forall\mathbf{x}\in\mathcal{F}^m,\ \exists\mathbf{x}_A\in\mathbb{R}^m\text{ s.t. }\mathbf{f}_A(\mathbf{x})=\mathbf{f}(\mathbf{x}_A),\tag{2.40}$$

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

$$\operatorname{cond}_A(\mathbf{x})=\frac{1}{\epsilon_u}\inf_{\{\mathbf{x}_A\}}\frac{\|\mathbf{x}_A-\mathbf{x}\|}{\|\mathbf{x}\|}.\tag{2.41}$$

Example 2.71. 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 $|\delta|\le5\epsilon_u$. 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

$$\begin{aligned}E_{\text{rel}}(x_A)&=\left|\frac{x^{1+\delta}-x}{x}\right|=|x^\delta-1|=|e^{\delta\ln x}-1|\\&\approx|\delta\ln x|\le5|\ln x|\epsilon_u.\end{aligned}$$

Hence A is well conditioned except when $x\to0^+$.

Theorem 2.72. Suppose a smooth function $f:\mathbb{R}\to\mathbb{R}$ is approximated by an algorithm $A:\mathcal{F}\to\mathcal{F}$, producing $f_A(x)=f(x)(1+\delta(x))$ where $|\delta(x)|\le\varphi(x)\epsilon_u$. If $\operatorname{cond}_f(x)$ is bounded, then $\forall x\in\mathcal{F}$,

$$\operatorname{cond}_A(x)\le\frac{\varphi(x)}{\operatorname{cond}_f(x)}.\tag{2.42}$$

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

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

Take inf with respect to all x_A's, take sup with respect to x, and we have (2.42). □

Example 2.73. 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 2.72 to analyze the condition of the algorithm

$$f_A = \text{fl}\left[\frac{\text{fl}\big(1-\text{fl}(\cos x)\big)}{\text{fl}(\sin x)}\right] \tag{2.43}$$

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

By Definition 2.58, it is easy to compute that

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

Furthermore, by Theorem 2.39 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

$$\text{cond}_A(x) \le \frac{\sin x}{x}\left(3+\frac{\cos x}{1-\cos x}\right).$$

Hence, $\text{cond}_A(x) \to +\infty$ as $x \to 0$. On the other hand, $\text{cond}_A(x) \to \frac{6}{\pi}$ as $x \to \frac{\pi}{2}$.

Exercise 2.74. Repeat Example 2.73 for $f(x) = \frac{\sin x}{1+\cos x}$ on the same interval.

2.3.6 Overall error of a computer solution

Theorem 2.75. Consider using normalized FPN arithmetics to solve a math problem

$$\mathbf{f}: \mathbb{R}^m \to \mathbb{R}^n, \qquad \mathbf{y} = \mathbf{f}(\mathbf{x}). \tag{2.44}$$

Denote the computer input and output as

$$\mathbf{x}^* \approx \mathbf{x}, \qquad \mathbf{y}_A^* = \mathbf{f}_A(\mathbf{x}^*), \tag{2.45}$$

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

$$E_{\text{rel}}(\mathbf{y}_A^*) \lessapprox E_{\text{rel}}(\mathbf{x}^*)\text{cond}_{\mathbf{f}}(\mathbf{x}) + \epsilon_u \text{cond}_{\mathbf{f}}(\mathbf{x}^*)\,\text{cond}_A(\mathbf{x}^*), \tag{2.46}$$

where the relative error is defined in (2.25).

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 (2.31), the first term is

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

By (2.31) and Definition 2.70, the second term is

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

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

2.4 Problems

2.4.1 Theoretical questions

I. Convert the decimal integer 477 to a normalized FPN with $\beta = 2$.

II. Convert the decimal fraction 3/5 to a normalized FPN with $\beta = 2$.

III. Let $x = \beta^e$, $e \in \mathbb{Z}$, $L < e < U$ be a normalized FPN in $\mathbb{F}$ and $x_L, x_R \in \mathbb{F}$ the two normalized FPNs adjacent to x such that $x_L < x < x_R$. Prove $x_R - x = \beta(x - x_L)$.

IV. By reusing your result of II, find out the two normalized FPNs adjacent to $x = 3/5$ under the IEEE 754 single-precision protocol. What is $\text{fl}(x)$ and the relative roundoff error?

V. If the IEEE 754 single-precision protocol did not round off numbers to the nearest, but simply dropped excess bits, what would the unit roundoff be?

VI. How many bits of precision are lost in the subtraction $1 - \cos x$ when $x = \frac{1}{4}$?

VII. Suggest at least two ways to compute $1 - \cos x$ to avoid catastrophic cancellation caused by subtraction.

VIII. What are the condition numbers of the following functions? Where are they large?

- $(x-1)^\alpha$,
- $\ln x$,
- e^x,
- $\arccos x$.

IX. Consider the function $f(x) = 1 - e^{-x}$ for $x \in [0, 1]$.

- Show that $\text{cond}_f(x) \le 1$ for $x \in [0, 1]$.
- Let A be the algorithm that evaluates $f(x)$ for the machine number $x \in \mathbb{F}$. Assume that the exponential function is computed with relative error within machine roundoff. Estimate $\text{cond}_A(x)$ for $x \in [0, 1]$.
- Plot $\text{cond}_f(x)$ and $\text{cond}_A(x)$ as a function of x on $[0, 1]$. Discuss your results.

X. The math problem of root finding for a polynomial

$$q(x) = \sum_{i=0}^{n} a_i x^i, \qquad a_n = 1, a_0 \neq 0, a_i \in \mathbb{R} \tag{2.47}$$

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?

2.4.2 Programming assignments

A. Print values of the functions in (2.48) 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.

$$f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1 \tag{2.48a}$$

$$g(x) = (((((((x - 8)x + 28)x - 56)x + 70)x - 56)x + 28)x - 8)x + 1 \tag{2.48b}$$

$$h(x) = (x - 1)^8 \tag{2.48c}$$

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

3.1 The Vandermonde determinant

Definition 3.3. 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 3.4. 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 3.5 (Uniqueness of polynomial interpolation)**.** Given distinct points $x_0, x_1, \ldots, x_n \in \mathbb{C}$ and corresponding values $f_0, f_1, \ldots, f_n \in \mathbb{C}$. Denote by $\mathbb{P}_n$ the class of polynomials of degree at most n. There exists a unique polynomial $p_n(x) \in \mathbb{P}_n$ such that

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

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 3.4, the determinant of the system is $\prod_{i>j}(x_i - x_j)$. The proof is completed by the distinctness of the points and Cramer's rule. □

3.2 The Cauchy remainder

Theorem 3.6 (Generalized Rolle)**.** Let $n \geq 2$. Suppose that $f \in \mathcal{C}^{n-1}[a, b]$ and $f^{(n)}(x)$ exists at each point of (a, b). Suppose that $f(x_0) = f(x_1) = \cdots = f(x_n) = 0$ for $a \leq x_0 < x_1 < \cdots < x_n \leq b$. Then there is a point $\xi \in (x_0, x_n)$ such that $f^{(n)}(\xi) = 0$.

Proof. Applying Rolle's theorem (Theorem C.43) 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 3.7 (Cauchy remainder of polynomial interpolation)**.** Let $f \in \mathcal{C}^n[a,b]$ and suppose that $f^{(n+1)}(x)$ exists at each point of (a,b). Let $p_n(f;x)$ denote the unique polynomial in $\mathbb{P}_n$ that coincides with f at $x_0, x_1, \ldots, x_n$. Define

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

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 $x \neq x_0, x_1, \ldots, x_n$ and define

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

and a function of t

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

The function $W(t)$ vanishes at $t = x_0, x_1, \ldots, x_n$. In addition $W(x) = 0$. By Theorem 3.6, $W^{(n+1)}(\xi) = 0$ for some $\xi \in (a,b)$, i.e.

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

Hence $K(x) = f^{(n+1)}(\xi)/(n+1)!$ and (3.6) holds. □

Corollary 3.8. Suppose $f(x) \in \mathcal{C}^{n+1}[a,b]$. Then

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

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

Example 3.9. A value for $\arcsin(0.5335)$ is obtained by interpolating linearly between the values for $x = 0.5330$ and $x = 0.5340$. Estimate the error committed.

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

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

Since the third derivative is positive over $[0.5330, 0.5340]$. The maximum value of f'' occurs at 0.5340. By Corollary 3.8 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 3.10. To interpolate given values $f_0, f_1, \ldots, f_n$ at distinct points $x_0, x_1, \ldots, x_n$, the *Lagrange formula* is

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

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

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

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

Example 3.11. For $i = 0, 1, 2$, we are given $x_i = 1, 2, 4$ and $f(x_i) = 8, 1, 5$, respectively. The Lagrangian formula generates $p_2(x) = 3x^2 - 16x + 21$.

Lemma 3.12. Define a symmetric polynomial

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

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

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

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

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

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

Proof. By Theorems 3.5 and 3.7, for each $q(x) \in \mathbb{P}_n$ we have $p_n(q;x) \equiv q(x)$. Interpolating the constant function $f(x) \equiv 1$ with the Lagrange formula yields (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 3.14 (Divided difference and the Newton formula)**.** The *Newton formula* for interpolating the values $f_0, f_1, \ldots, f_n$ at distinct points $x_0, x_1, \ldots, x_n$ is

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

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 3.15. 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 3.5. □

Corollary 3.16. 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 3.5 implies that the two polynomials in (3.8) and (3.14) are the same. Then the first equality follows from (3.9) and Definition 3.14, while the second equality follows from Lemma 3.12. □

Theorem 3.17. 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 3.14, $f[x_1, x_2, \ldots, x_k]$ 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 3.18. The kth divided difference ($k \in \mathbb{N}^+$) on the *table of divided differences*

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	$f[x_3]$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

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

x	0	1	2	3
$f(x)$	6	−3	−6	9

By Definition 3.18, we can construct the following table of divided difference,

0	6			
1	−3	−9		
2	−6	−3	3	
3	9	15	9	2

(3.18)

By Definition 3.14, the interpolating polynomial is generated from the main diagonal and the first column of the above table as follows.

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

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

Exercise 3.20. Redo Example 3.11 with the Newton formula.

Theorem 3.21. 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 3.22. Suppose $f \in \mathcal{C}^n[a, b]$ and $f^{(n+1)}(x)$ exists at each point of (a, b). If $a = x_0 < x_1 < \cdots < x_n = b$ and $x \in [a, b]$, then

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

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

Proof. This follows from Theorems 3.21 and 3.7. □

Corollary 3.23. If $x_0 < x_1 < \cdots < x_n$ and $f \in \mathcal{C}^n[x_0, x_n]$, we have

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

Proof. Set $x = x_{n+1}$ in Corollary 3.22, 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 3.24. 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 3.25. 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 3.26. 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 3.27. 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 3.16, the second from (3.29), and the last from Theorem 3.26. □

Theorem 3.28 (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 3.21, apply Theorem 3.27, 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 3.29. 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

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

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

Example 3.30. To estimate $f(x)$ for $x = \frac{3}{2}$ directly from the table in Example 3.19, 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 3.19. In contrast, the calculation and layout of the two tables are distinct.

3.6 Hermite interpolation

Definition 3.31. 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

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

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

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

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

Theorem 3.33. For the Hermite interpolation problem in Definition 3.31, 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

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

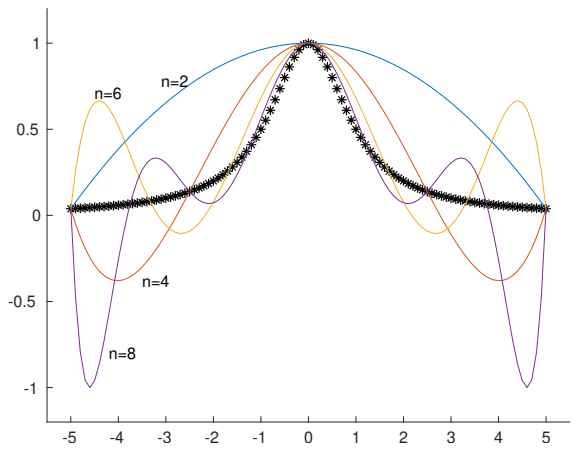
Proof. The proof is similar to that of Theorem 3.7. Pay attention to the difference caused by the multiple roots of the polynomial $\prod_{i=0}^{k} (x - x_i)^{m_i+1}$. □

3.7 The Chebyshev polynomials

Example 3.34 (Runge phenomenon)**.** The points $x_0, x_1, \ldots, x_n$ in Theorem 3.5 are usually given *a priori*, e.g., as uniformly distributed over the interval $[x_0, x_n]$. As n increases, the degree of the interpolating polynomial also increases. Ideally we would like to have

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

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

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

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

Definition 3.35. The *Chebyshev polynomial* of degree n of the first kind is a polynomial $T_n : [-1, 1] \to [-1, 1]$,

$$T_n(x) = \cos(n \arccos x). \qquad (3.39)$$

Theorem 3.36.

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

Proof. By trigonometric identities, we have

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

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

Corollary 3.37. 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 3.38. $T_n(x)$ 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, $\forall k = 1, 2, \ldots, n-1$,

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

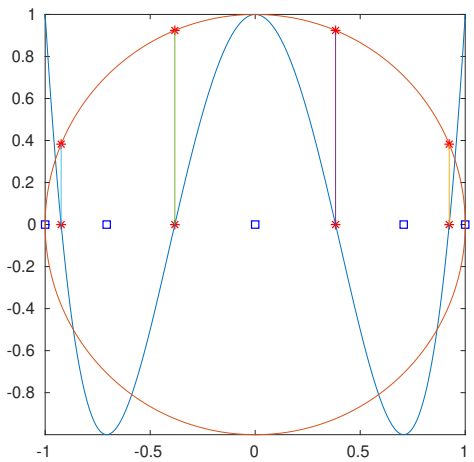
$$T''_n(x) = \frac{n^2\cos(n\arccos(x))}{x^2-1} + \frac{n\,x\,\sin(n\arccos(x))}{(1-x^2)^{3/2}};$$

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

Hence a Taylor expansion of T_n yields

$$T_n(x'_k+\delta) = T_n(x'_k) + \frac{1}{2}T''_n(x'_k)\delta^2 + O(\delta^3),$$

and T_n must attain local extremes at each x'_k. For $k = 0, 1, \ldots, n$, $T_n(x'_k)$ attains its extreme values at x'_k since $T_n(x'_0) = 1$, $T_n(x'_1) = -1$, ..., 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 3.39. Write a program to reproduce the above plot.

Theorem 3.40 (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 3.38, $T_n(x)$ assumes its extrema $n+1$ times at the points x'_k defined in (3.42). Suppose (3.43) does not hold. Then Theorem 3.38 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 3.41. 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 3.42. Suppose polynomial interpolation is performed for f on the $n+1$ zeros of $T_{n+1}(x)$ as in Theorem 3.38. The Cauchy remainder in Theorem 3.7 satisfies

$$|R_n(f;x)| \leq \frac{1}{2^n(n+1)!}\max_{x\in[-1,1]}\left|f^{(n+1)}(x)\right|. \tag{3.46}$$

Proof. Theorem 3.7, Corollary 3.37, and Theorem 3.38 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 3.35 completes the proof as $|T_{n+1}| \leq 1$. □

Theorem 3.43 (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

$$\begin{aligned}&\forall f \in \mathcal{C}[a,b],\ \forall\epsilon > 0,\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > N,\\ &\qquad \exists p_n \in \mathbb{P}_n \text{ s.t. } \forall x \in [a,b],\ \|p_n - f\| < \epsilon.\end{aligned} \tag{3.47}$$

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,

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

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 3.22 we know

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

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

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

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

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

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

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

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

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

VII. Define forward difference by

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

and backward difference by

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

Prove

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

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

$$\mathbb{S}_n^k = \big\{ s : \ s \in \mathcal{C}^k[a,b];\ \forall i \in [1, N-1],\ s|_{[x_i, x_{i+1}]} \in \mathbb{P}_n \big\}. \tag{4.2}$$

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

Notation 1. 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 4.2. As an extreme, $\mathbb{S}_n^n = \mathbb{P}_n$, i.e. all the pieces of $s \in \mathbb{S}_n^n$ belong to a single polynomial. On the other end, $\mathbb{S}_1^0$ is the class of piecewise linear interpolating functions. The most popular splines are the cubic splines in $\mathbb{S}_3^2$.

Lemma 4.3. Denote $m_i = s'(f; x_i)$ for $s \in \mathbb{S}_3^2$. Then, for each $i = 2, 3, \ldots, N-1$, we have

$$\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = 3\mu_i f[x_i, x_{i+1}] + 3\lambda_i f[x_{i-1}, x_i], \tag{4.3}$$

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 $p_i(x) = s|_{[x_i, x_{i+1}]}$ 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

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}	$\frac{m_{i+1} - K_i}{x_{i+1} - x_i}$	$\frac{m_i + m_{i+1} - 2K_i}{(x_{i+1} - x_i)^2}$

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 4.4. 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 4.5. 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 4.6 (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 4.7. 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

$$\begin{aligned} s_1(x) =& f[x_1] + f[x_1, x_1](x - x_1) \\ &+ \frac{M_1}{2}(x - x_1)^2 + \frac{M_2 - M_1}{6(x_2 - x_1)}(x - x_1)^3. \quad (4.14) \end{aligned}$$

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 4.8. 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 4.3, s is uniquely determined if all the m_i's are uniquely determined on all intervals. For a complete cubic spline we already have $m_1 = f'(a)$ and $m_N = f'(b)$. Assemble (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 4.5 and 4.7, following arguments similar to the above. □

Example 4.9. 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 4.5 and Lemma 4.7 yield a linear system

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

where elements in the 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 4.10 (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 4.11 (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 4.10. Although $\eta'(a) = \eta'(b) = 0$ does not hold, we do have $s''(a) = s''(b) = 0$. □

Lemma 4.12. 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 4.5 and Corollary 3.22 that

$$\begin{aligned}
&2M_j = 6f[x_{j-1}, x_j, x_{j+1}] - \mu_j M_{j-1} - \lambda_j M_{j+1} \\
\Rightarrow\ &2|M_j| \le 6\,|f[x_{j-1}, x_j, x_{j+1}]| + (\mu_j + \lambda_j)|M_j| \\
\Rightarrow\ &\exists \xi \in [x_{j-1}, x_{j+1}] \ \text{ s.t. } |M_j| \le 3\,|f''(\xi)| \\
\Rightarrow\ &|s''(x)| \le 3 \max_{x\in[a,b]} |f''(x)|\,.
\end{aligned} \tag{4.19}$$

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 3.22, 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 4.13. 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 3.7) implies

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

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 4.12 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 C.43 implies $f'(\xi_i) - s'(\xi_i) = 0$ for some $\xi_i \in [x_i, x_{i+1}]$. It follows from the second fundamental theorem of calculus (Theorem C.66) that

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

which, together with the integral mean value theorem C.64 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 3.7 and the third step from (4.22). □

Exercise 4.14. Verify Theorem 4.13 using the results in Example 4.9.

4.4 B-Splines

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

Theorem 4.15. 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 B.2 that $\mathbb{S}_n^{n-1}(t_1, t_2, \cdots, t_N)$ is indeed a linear space. Note that the additive identity is the zero function not the number zero. One polynomial of degree n is determined by $n + 1$ coefficients. The $N - 1$ intervals lead to $(N - 1)(n + 1)$ coefficients. At each of the $N - 2$ interval knots, the smoothness condition requires that the 0th, 1st, ..., $(n - 1)$th derivatives of adjacent polynomials match. Hence the dimension is $(N - 1)(n + 1) - n(N - 2) = n + N - 1$. □

Example 4.16. The cubic splines in Definition 4.6, 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 4.6.

4.4.1 Truncated power functions

Definition 4.17. 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 4.18. According to Definition 4.17, 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 4.19. 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 B.24. The proof is completed by Theorem 4.15, Lemma B.39, and the fact that there are $n + N - 1$ functions in (4.25). □

Corollary 4.20. 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 4.19, 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 4.21. (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 4.22. 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 4.23. The hat functions form a basis of $\mathbb{S}_1^0$.

Proof. By Definition 4.22, 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 B.24 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

$$\forall s(x) \in \mathbb{S}_1^0, \quad \exists s_B(x) = \sum_{i=1}^N s(t_i)\hat{B}_i(x) \text{ s.t. } s(x) = s_B(x).$$

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 B.31 completes the proof. □

Definition 4.24. *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 4.25. The hat functions in Definition 4.22 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 4.26. 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 4.27. 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 4.28. 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 4.29. 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 4.30. $X = \mathcal{C}[a, b]$, then the elements of X are functions continuous over $[a, b]$.

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

are both linear functionals over X.

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

Theorem 4.31 (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 3.17 that

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

By the induction hypothesis, we have

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

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

In the above derivation, we have applied Theorem 3.17 to go from the kth divided difference to the $(k+1)$th. Then

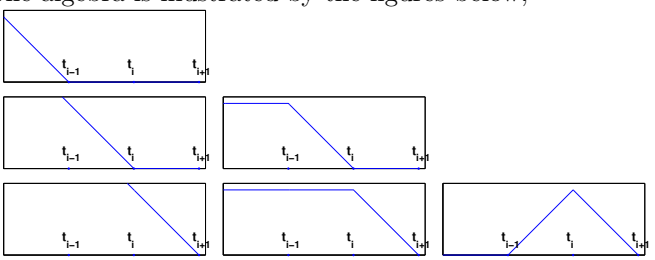
$$\begin{aligned}[x_0,\dots,x_{k+1}]fg &= \frac{S_1+S_2}{x_{k+1}-x_0}\\ &= \sum_{i=0}^{k+1}[x_0,\dots,x_i]f\cdot[x_i,\dots,x_{k+1}]g,\end{aligned}$$

which completes the inductive proof. □

Example 4.32. 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 4.33 (B-splines as divided difference of truncated power functions)**.** For any $n\in\mathbb{N}$, we have

$$B_i^n(x) = (t_{i+n}-t_{i-1})\cdot[t_{i-1},\dots,t_{i+n}](t-x)_+^n. \tag{4.37}$$

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 4.17, $(t-x)_+^{n+1}=(t-x)(t-x)_+^n$. Then the application of Theorem 4.31 with $f=(t-x)$ and $g=(t-x)_+^n$ yields

$$\begin{aligned}&[t_{i-1},\dots,t_{i+n}](t-x)_+^{n+1}\\ =&(t_{i-1}-x)\cdot[t_{i-1},\dots,t_{i+n}](t-x)_+^n\\ &+[t_i,\dots,t_{i+n}](t-x)_+^n.\end{aligned} \tag{4.38}$$

Definition 4.24 and the induction hypothesis yield

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

where the last step follows from (4.38). Similarly,

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

where the second last step follows from Theorem 3.17 and (4.38). The above arguments yield

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

which completes the inductive proof. □

4.4.3 Integrals and derivatives

Corollary 4.34 (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

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

where the first step follows from Theorem 4.33, the second step from the commutativity of integration and taking divided difference, the third step from (4.24), and the last step from Corollary 3.22. □

Theorem 4.35 (Derivatives of B-splines)**.** For $n \geq 2$, we have, $\forall x \in \mathbb{R}$,

$$\frac{\mathrm{d}}{\mathrm{d}x} B_i^n(x) = \frac{nB_i^{n-1}(x)}{t_{i+n-1} - t_{i-1}} - \frac{nB_{i+1}^{n-1}(x)}{t_{i+n} - t_i}. \tag{4.40}$$

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

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

Hence the induction basis holds. Now suppose (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 $n \geq 1$. Hence the right-hand side of (4.40) is continuous for all $n \geq 2$. Therefore, if $n \geq 2$, $\frac{\mathrm{d}}{\mathrm{d}x} B_i^n(x)$ exists for all $x \in \mathbb{R}$. This completes the proof. □

Corollary 4.36 (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 4.35 via an easy induction. □

4.4.4 Marsden's identity

Theorem 4.37 (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 4.24. 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 4.38 (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 $x \leq t_j$ 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 $i \geq j + 1$, Lemma 4.28 implies that $B_i^n(x) = 0$ for all $x \leq t_j$. Hence

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

Otherwise $x > t_j$, then Lemma 4.28 implies $B_i^n(x) = 0$ for each $i \leq j - n$. This completes the proof. □

4.4.5 Symmetric polynomials

Definition 4.39. The *elementary symmetric polynomial* of degree k in n variables is the sum of all products of k distinct variables chosen from the n variables,

$$\sigma_k(x_1,\dots,x_n)=\sum_{1\le i_1<\cdots<i_k\le n} x_{i_1}x_{i_2}\cdots x_{i_k}. \tag{4.45}$$

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

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

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

$$\tau_k(x_1,\dots,x_n)=\sum_{1\le i_1\le\cdots\le i_k\le n} x_{i_1}x_{i_2}\cdots x_{i_k}. \tag{4.46}$$

Example 4.40. $\sigma_2(x_1,x_2,x_3)=x_1x_2+x_1x_3+x_2x_3$. 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 4.41. For $k\le n$, the elementary symmetric polynomials satisfy a recursion,

$$\begin{aligned}&\sigma_{k+1}(x_1,\dots,x_n,x_{n+1})\\ =&\sigma_{k+1}(x_1,\dots,x_n)+x_{n+1}\sigma_k(x_1,\dots,x_n).\end{aligned} \tag{4.47}$$

Proof. The terms in $\sigma_{k+1}(x_1,\dots,x_n,x_{n+1})$ can be assorted into two groups: (a) those that contain the factor x_{n+1} and (b) those that do not. By the symmetry in (4.45), group (a) must be $x_{n+1}\sigma_k(x_1,\dots,x_n)$ and group (b) must be $\sigma_{k+1}(x_1,\dots,x_n)$. □

Example 4.42. $\sigma_2(x_1,x_2,x_3)=x_1x_2+x_3(x_1+x_2)$.

Definition 4.43. The *generating function for the elementary symmetric polynomials* is

$$g_{\sigma,n}(z)=\prod_{i=1}^{n}(1+x_iz)=(1+x_1z)\cdots(1+x_nz) \tag{4.48}$$

while that for the complete symmetric polynomials is

$$g_{\tau,n}(z)=\prod_{i=1}^{n}\frac{1}{1-x_iz}=\frac{1}{1-x_1z}\cdots\frac{1}{1-x_nz}. \tag{4.49}$$

Lemma 4.44 (Generating elementary and complete symmetric polynomials)**.** The elementary and complete symmetric polynomials are related to their generating functions as

$$g_{\sigma,n}(z)=\sum_{k=0}^{n}\sigma_k(x_1,\dots,x_n)z^k. \tag{4.50}$$

$$g_{\tau,n}(z)=\sum_{k=0}^{+\infty}\tau_k(x_1,\dots,x_n)z^k. \tag{4.51}$$

Proof. With Lemma 4.41, we can prove (4.50) by an easy induction. For (4.51), (4.49) and the identity

$$\frac{1}{1-x}=\sum_{k=0}^{+\infty}x^k \tag{4.52}$$

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 $x_1,x_2,\dots,x_n$. Definition 4.39 then completes the proof. □

Example 4.45.

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

Lemma 4.46 (Recursive relations of complete symmetric polynomials)**.** The complete symmetric polynomials satisfy a recursion,

$$\begin{aligned}&\tau_{k+1}(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_1,\dots,x_n)+x_{n+1}\tau_k(x_1,\dots,x_n,x_{n+1}).\end{aligned} \tag{4.53}$$

Proof. (4.49) implies

$$g_{\tau,n+1}=g_{\tau,n}+x_{n+1}zg_{\tau,n+1}. \tag{4.54}$$

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 4.47 (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.

$$\begin{aligned}&\forall m\in\mathbb{N}^+,\ i\in\mathbb{N},\ \forall n=0,1,\dots,m,\\ &\tau_{m-n}(x_i,\dots,x_{i+n})=[x_i,\dots,x_{i+n}]x^m.\end{aligned} \tag{4.55}$$

Proof. By Lemma 4.46, we have

$$\begin{aligned}&(x_{n+1}-x_1)\tau_k(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_1,\dots,x_n,x_{n+1})-\tau_{k+1}(x_1,\dots,x_n)\\ &-x_1\tau_k(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_2,\dots,x_n,x_{n+1})+x_1\tau_k(x_1,\dots,x_n,x_{n+1})\\ &-\tau_{k+1}(x_1,\dots,x_n)-x_1\tau_k(x_1,\dots,x_n,x_{n+1})\\ =&\tau_{k+1}(x_2,\dots,x_n,x_{n+1})-\tau_{k+1}(x_1,\dots,x_n).\end{aligned} \tag{4.56}$$

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 non-negative integer $n<m$. Then (4.56) and the induction hypothesis yield

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

which completes the proof. □

4.4.6 B-splines indeed form a basis

Theorem 4.48. 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

$$\binom{n}{k} x^k = \sum_{i=-\infty}^{+\infty} \sigma_k(t_i, \ldots, t_{i+n-1}) B_i^n(x), \tag{4.57}$$

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 4.44 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 4.49 (Partition of Unity)**.**

$$\forall n \in \mathbb{N}, \qquad \sum_{i=-\infty}^{+\infty} B_i^n = 1. \tag{4.58}$$

Proof. Setting $k = 0$ in Theorem 4.48 yields (4.58). □

Theorem 4.50. The following list of B-splines is a basis of $\mathbb{S}_n^{n-1}(t_1, t_2, \ldots, t_N)$,

$$B_{2-n}^n(x), B_{3-n}^n(x), \ldots, B_N^n(x). \tag{4.59}$$

Proof. It is easy to verify that

$$\forall t_i \in \mathbb{R}, \quad (x - t_i)_+^n = (x - t_i)^n - (-1)^n (t_i - x)_+^n. \tag{4.60}$$

Then it follows from Theorem 4.37 and Corollary 4.38 that each truncated power function $(x - t_i)_+^n$ can be expressed as a linear combination of B-splines. By Lemma 4.19, 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 4.48 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 4.28 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 B.38, Theorem 4.15, and the fact that the length of the list (4.59) is also $n + N - 1$. □

4.4.7 Cardinal B-splines

Definition 4.51. The *cardinal B-spline* of degree n, denoted by $B_{i,\mathbb{Z}}^n$, is the B-spline in Definition 4.24 on the knot set $\mathbb{Z}$.

Corollary 4.52. Cardinal B-splines of the same degree are translates of one another, i.e.

$$\forall x \in \mathbb{R}, \qquad B_{i,\mathbb{Z}}^n(x) = B_{i+1,\mathbb{Z}}^n(x + 1). \tag{4.61}$$

Proof. The recurrence relation (4.30) reduces to

$$B_{i,\mathbb{Z}}^{n+1}(x) = \frac{x - i + 1}{n + 1} B_{i,\mathbb{Z}}^n(x) + \frac{i + n + 1 - x}{n + 1} B_{i+1,\mathbb{Z}}^n(x). \tag{4.62}$$

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

Corollary 4.53. A cardinal B-spline is symmetric about the center of its interval of support, i.e.

$$\forall n > 0, \ \forall x \in \mathbb{R}, \qquad B_{i,\mathbb{Z}}^n(x) = B_{i,\mathbb{Z}}^n(2i + n - 1 - x). \tag{4.63}$$

Proof. The proof is similar with that of Corollary 4.52. □

Example 4.54. For $t_i = i$, the quadratic B-spline in Example 4.26 simplifies to

$$B_{i,\mathbb{Z}}^2(x) = \begin{cases} \frac{(x-i+1)^2}{2}, & x \in (i-1, i]; \\ \frac{3}{4} - \left(x - (i + \frac{1}{2})\right)^2, & x \in (i, i+1]; \\ \frac{(i+2-x)^2}{2}, & x \in (i+1, i+2]; \\ 0, & \text{otherwise.} \end{cases} \tag{4.64}$$

It is straightforward to verify Corollaries 4.52 and 4.53. It also follows from (4.64) that

$$B_{i,\mathbb{Z}}^2(j) = \begin{cases} \frac{1}{2} & j \in \{i, i+1\}; \\ 0, & j \in \mathbb{Z} \setminus \{i, i+1\}. \end{cases} \tag{4.65}$$

Example 4.55. For $t_i = i$, the cubic cardinal B-spline is

$$B_{i,\mathbb{Z}}^3(x) = \begin{cases} \frac{(x-i+1)^3}{6}, & x \in (i-1, i]; \\ \frac{2}{3} - \frac{1}{2}(x - i + 1)(i + 1 - x)^2, & x \in (i, i+1]; \\ B_{i,\mathbb{Z}}^3(2i + 2 - x), & x \in (i+1, i+3); \\ 0, & \text{otherwise.} \end{cases} \tag{4.66}$$

It follows that

$$B_{i,\mathbb{Z}}^3(j) = \begin{cases} \frac{1}{6}, & j \in \{i, i+2\}; \\ \frac{2}{3}, & j = i + 1; \\ 0, & j \in \mathbb{Z} \setminus \{i, i+1, i+2\}. \end{cases} \tag{4.67}$$

This illustrates Corollary 4.52 that cardinal B-splines have the same shape, i.e., they are invariant under integer translations.

Theorem 4.56. 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 4.33, 3.27, and 3.26 yield

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

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

Corollary 4.57. 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 4.56 and Definition 4.17. □

Corollary 4.58 (Unique interpolation by complete cubic cardinal B-splines)**.** There is a unique B-spline $S(x) \in \mathbb{S}^2_3$ that interpolates $f(x)$ at $1, 2, \dots, N$ with $S'(1) = f'(1)$ and $S'(N) = f'(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=-1}^{N} a_i B^3_{i,\mathbb{Z}}(x), \quad (4.70)$$

where

$$a_{-1} = a_1 - 2f'(1), \qquad a_N = a_{N-2} + 2f'(N), \quad (4.71)$$

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

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

Proof. By Theorem 4.50 and Lemma 4.28, we have

$$\begin{aligned} &\forall i = 1, 2, \dots, N, \\ &\quad f(i) = a_{i-2}B^3_{i-2,\mathbb{Z}}(i) + a_{i-1}B^3_{i-1,\mathbb{Z}}(i) + a_i B^3_{i,\mathbb{Z}}(i). \end{aligned}$$

Then (4.67) yields

$$\forall i = 1, 2, \dots, N, \qquad a_{i-2} + 4a_{i-1} + a_i = 6f(i), \quad (4.72)$$

which proves the middle $N-2$ equations of $M\mathbf{a} = \mathbf{b}$. By Theorem 4.35, 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 4.59. There is a unique B-spline $S(x) \in \mathbb{S}^1_2$ that interpolates $f(x)$ at $t_i = i + \frac{1}{2}$ for each $i = 1, 2, \dots, N-1$ with end conditions $S(1) = f(1)$ and $S(N) = f(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=0}^{N} a_i B^2_{i,\mathbb{Z}}(x), \quad (4.74)$$

where

$$a_0 = 2f(1) - a_1, \qquad a_N = 2f(N) - a_{N-1}, \quad (4.75)$$

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

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

Proof. It follows from Lemma 4.28 and Definition 4.51 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 4.50! By Theorem 4.56, 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 4.53. Then Corollary 4.52 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 4.26 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 4.60. 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 4.61. 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 4.62. A *unit-speed curve* is a curve whose tangent vector has unit length at each of its points.

Definition 4.63. 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 4.64. 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 4.65. 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 4.66. 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 4.67. A reparametrization of a regular curve is unit-speed if and only if it is based on the arc-length.

Definition 4.68. A *closed curve* is (the image of) a continuous map $\mathring{\gamma} : [0, 1] \to \mathbb{R}^2$ that satisfies $\mathring{\gamma}(0) = \mathring{\gamma}(1)$. If the restriction of $\mathring{\gamma}$ to $[0, 1)$ is further injective, then the closed curve is a *simple closed curve* or *Jordan curve*.

Definition 4.69. The *signed unit normal* of a curve, denoted by $\mathbf{n}_s$, is the unit vector obtained by rotating its unit tangent vector counterclockwise by $\frac{\pi}{2}$.

Definition 4.70. For a unit-speed curve γ, its *signed curvature* is defined as

$$\kappa_s := \gamma'' \cdot \mathbf{n}_s. \tag{4.80}$$

Definition 4.71. 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 $p_i = s|_{[x_i, x_{i+1}]}$. 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 4.33 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 4.58 and 4.59, 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 4.58 and $t_i = i - \frac{11}{2}$, $i = 1, \ldots, 10$ for Corollary 4.59, 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 4.32.

Chapter 5

Approximation

Definition 5.1. 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

$$\forall \varphi \in X, \qquad \|f - \hat{\varphi}\| \leq \|f - \varphi\|. \tag{5.1}$$

Example 5.2. The Chebyshev Theorem 3.40 can be restated in the format of Definition 5.1 as follows. As in Example B.23, denote by $\mathbb{P}_n(\mathbb{R})$ the set of all polynomials with coefficients in $\mathbb{R}$ and degree at most n. For $Y = \mathbb{P}_n(\mathbb{R})$, and $X = \mathbb{P}_{n-1}(\mathbb{R})$, the best approximation to $f(x) = -x^n$ in Y from X with respect to the max-norm $\|\cdot\|_\infty$

$$\|g\|_\infty = \max_{x \in [-1,1]} |g(x)| \tag{5.2}$$

is $\hat{\varphi} = \frac{T_n}{2^{n-1}} - x^n$, where T_n is Chebyshev polynomial of degree n. Clearly $\hat{\varphi}$ satisfies (5.1).

Example 5.3. 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 $X = \operatorname{span}\{1, x, x^2, \ldots\}$ is a problem of linear approximation, where n can be any positive integer and the norm can be the max-norm (5.2), the 1-norm

$$\|g\|_1 := \int_{-1}^{+1} |g(x)| \mathrm{d}x, \tag{5.3}$$

or the 2-norm

$$\|g\|_2 := \left(\int_{-1}^{+1} |g(x)|^2 \mathrm{d}x \right)^{\frac{1}{2}}. \tag{5.4}$$

The three different norms are motivated differently: the max-norm corresponds to the min-max error, the 1-norm is related to the area bounded between $g(x)$ and the x-axis, and the 2-norm is related to the Euclidean distance, c.f. Section 5.4.

Example 5.4. 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 4.71. Denote by $\operatorname{Int}(\gamma)$ 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 $\mathcal{S}_1 := \operatorname{Int}(\gamma)$ and $\mathcal{S}_2 := \operatorname{Int}(p)$ can be defined as

$$\|\mathcal{S}_1 \oplus \mathcal{S}_2\|_1 := \int_{\mathcal{S}_1 \oplus \mathcal{S}_2} \mathrm{d}\mathbf{x},$$

where

$$\mathcal{S}_1 \oplus \mathcal{S}_2 := \mathcal{S}_1 \cup \mathcal{S}_2 \setminus (\mathcal{S}_1 \cap \mathcal{S}_2)$$

is the exclusive disjunction of S_1 and S_2.

The minimization of this area difference can be formulated by a best approximation problem based on the 1-norm.

Theorem 5.5. Suppose X is a finite-dimensional subspace of a normed space $(Y, \|\cdot\|)$. Then we have

$$\forall y \in Y,\ \exists \hat{\varphi} \in X \text{ s.t. } \forall \varphi \in X,\ \|\hat{\varphi} - y\| \leq \|\varphi - y\|. \tag{5.5}$$

Proof. For a given $y \in Y$, define a closed ball

$$B_y := \{x \in X : \|x\| \leq 2\|y\|\}.$$

Clearly $0 \in B_y$, and the distance from y to B_y is

$$\operatorname{dist}(y, B_y) := \inf_{x \in B_y} \|y - x\| \leq \|y - 0\| = \|y\|.$$

By definition, any $z \in X$, $z \notin B_y$ must satisfy $\|z\| > 2\|y\|$, and thus

$$\|z - y\| \geq \|z\| - \|y\| > \|y\|.$$

Therefore, if a best approximation to y exists, it must be in B_y. As a subspace of X, B_y is finite dimensional, closed, and bounded, hence B_y is compact. The extreme value theorem states that a continuous scalar function attains its minimum and maximum on a compact set. A norm is a continuous function, hence the function $d : B_y \to \mathbb{R}^+ \cup \{0\}$ given by $d(x) = \|x - y\|$ must attain its minimum on B_y. □

Definition 5.6 (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 5.7. For a *weight function* $\rho(x) \in L[a,b]$, define

$$L^2_\rho[a,b] := \left\{ f(x) \in L[a,b] : \rho(x)|f(x)|^2 \in L[a,b] \right\}. \tag{5.6}$$

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

$$\langle u, v \rangle = \int_a^b \rho(t) u(t) \overline{v(t)} \mathrm{d}t \tag{5.7}$$

is an inner product space over $\mathbb{R}$; the set $L^2_\rho[a,b]$ with

$$\|u\|_2 = \left(\int_a^b \rho(t)|u(t)|^2 \mathrm{d}t\right)^{\frac{1}{2}} \tag{5.8}$$

is a normed vector space over $\mathbb{R}$.

Proof. This follows from Definitions B.2, B.103, and B.108. □

Definition 5.8. The *least-square approximation* on $L^2_\rho[a,b]$ is a best approximation problem with the norm in (5.1) set to that in (5.8).

5.1 Orthonormal systems

Definition 5.9. A subset S of an inner product space X is called *orthonormal* if

$$\forall u, v \in S, \qquad \langle u, v\rangle = \begin{cases} 0 & \text{if } u \neq v, \\ 1 & \text{if } u = v. \end{cases} \tag{5.9}$$

Example 5.10. The standard basis vectors in $\mathbb{R}^n$ are orthonormal.

Example 5.11. The Chebyshev polynomials of the first kind as in Definition 3.35 are orthogonal with respect to (5.7) where $a = -1, b = 1, \rho = \frac{1}{\sqrt{1-x^2}}$. However, they do not satisfy the second case in (5.9).

Theorem 5.12. Any finite set of nonzero orthogonal elements $u_1, u_2, \ldots, u_n$ is linearly independent.

Proof. This is easily proven by contradiction using Definitions B.24 and 5.9. □

Definition 5.13. The *Gram-Schmidt process* takes in a finite or infinite independent list $(u_1, u_2, \ldots)$ and output two other lists $(v_1, v_2, \ldots)$ and $(u_1^*, u_2^*, \ldots)$ by

$$v_{n+1} = u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u_k^*\rangle\, u_k^*, \tag{5.10a}$$

$$u_{n+1}^* = v_{n+1}/\|v_{n+1}\|, \tag{5.10b}$$

with the recursion basis as $v_1 = u_1,\ u_1^* = v_1/\|v_1\|$.

Theorem 5.14. For a finite or infinite independent list $(u_1, u_2, \ldots)$, the Gram-Schmidt process yields constants

$$\begin{matrix} a_{11} & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \\ \vdots & & \end{matrix}$$

such that $a_{kk} = \frac{1}{\|v_k\|} > 0$ and the elements $u_1^*, u_2^*, \ldots$

$$\begin{aligned} u_1^* &= a_{11}u_1 \\ u_2^* &= a_{21}u_1 + a_{22}u_2 \\ u_3^* &= a_{31}u_1 + a_{32}u_2 + a_{33}u_3 \\ &\vdots \end{aligned} \tag{5.11}$$

are orthonormal.

Proof. By Definition 5.13, the formulae (5.10) can be rewritten in the form of (5.11). It is clear from (5.10b) 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 B.103 and the third step from u_1^* being normal. The inductive step also holds because for any $j < n+1$ we have

$$\begin{aligned} \left\langle v_{n+1}, u_j^*\right\rangle &= \left\langle u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u_k^*\rangle\, u_k^*, u_j^*\right\rangle \\ &= \left\langle u_{n+1}, u_j^*\right\rangle - \sum_{k=1}^{n} \langle u_{n+1}, u_k^*\rangle \left\langle u_k^*, u_j^*\right\rangle \\ &= \left\langle u_{n+1}, u_j^*\right\rangle - \left\langle u_{n+1}, u_j^*\right\rangle = 0, \end{aligned}$$

where the third step follows from the induction hypothesis, i.e., $\left\langle u_k^*, u_j^*\right\rangle$ 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 (5.11), the third step from (5.10a) and the conclusion just proved, the fourth step from (5.10b), and the last step from Definitions B.103 and B.108. □

Corollary 5.15. For a finite or infinite independent list $(u_1, u_2, \ldots)$, we can find constants

$$\begin{matrix} b_{11} & & \\ b_{21} & b_{22} & \\ b_{31} & b_{32} & b_{33} \\ \vdots & & \end{matrix}$$

and an orthonormal list $(u_1^*, u_2^*, \ldots)$ such that $b_{ii} > 0$ and

$$\begin{aligned} u_1 &= b_{11}u_1^* \\ u_2 &= b_{21}u_1^* + b_{22}u_2^* \\ u_3 &= b_{31}u_1^* + b_{32}u_2^* + b_{33}u_3^* \\ &\vdots \end{aligned} \tag{5.12}$$

Proof. This follows from (5.11) and that a lower-triangular matrix with positive diagonal elements is invertible. □

Corollary 5.16. In Theorem 5.14, we have $\langle u_n^*, u_i\rangle = 0$ for each $i = 1, 2, \ldots, n-1$.

Proof. By Corollary 5.15, each u_i can be expressed as

$$u_i = \sum_{k=1}^{i} b_{ik}u_k^*.$$

Inner product the above equation with u_n^*, apply the orthogonal conditions, and we reach the conclusion. □

Definition 5.17. Using the Gram-Schmidt orthonormalizing process with the inner product (5.7), 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 5.18. We compute the first 3 Legendre polynomials using the Gram-Schmidt process.

$$u_1 = 1,\ v_1 = 1, \|v_1\|^2 = \int_{-1}^{+1} \mathrm{d}x = 2,\ \ u_1^* = \frac{1}{\sqrt{2}}.$$

$$u_2 = x,\ v_2 = x - \left\langle x, \frac{1}{\sqrt{2}} \right\rangle \frac{1}{\sqrt{2}} = x,\ \|v_2\|^2 = \frac{2}{3},$$

$$u_2^* = \sqrt{\frac{3}{2}} x.$$

$$v_3 = x^2 - \left\langle x^2, \sqrt{\frac{3}{2}} x \right\rangle \sqrt{\frac{3}{2}} x - \left\langle x^2, \frac{1}{\sqrt{2}} \right\rangle \frac{1}{\sqrt{2}} = x^2 - \frac{1}{3},$$

$$\|v_3\|^2 = \int_{-1}^{+1} \left(x^2 - \frac{1}{3}\right)^2 \mathrm{d}x = \frac{8}{45},$$

$$u_3^* = \frac{3}{4}\sqrt{10}\left(x^2 - \frac{1}{3}\right).$$

5.2 Fourier expansions

Definition 5.19. 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

$$w \sim \sum_n \langle w, u_n^* \rangle\, u_n^*, \tag{5.13}$$

where the constants $\langle w, u_n^* \rangle$ are known as the *Fourier coefficients* of w and the term $\langle w, u_n^* \rangle\, u_n^*$ the *projection* of w on u_n^*. The *error of the Fourier expansion* of w with respect to $(u_1^*, u_2^*, \ldots)$ is simply $\sum_n \langle w, u_n^* \rangle\, u_n^* - w$.

Example 5.20. With the Euclidean inner product in Definition B.107, 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 5.22.

Exercise 5.21. With the following orthonormal list in $L^2_{\rho=1}[-\pi, \pi]$,

$$\frac{1}{\sqrt{2\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos x}{\sqrt{\pi}}, \ldots, \frac{\sin(nx)}{\sqrt{\pi}}, \frac{\cos(nx)}{\sqrt{\pi}}, \ldots, \tag{5.14}$$

derive the *Fourier series* of a function $f(x)$ as

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{+\infty} (a_k \cos kx + b_k \sin kx), \tag{5.15}$$

where the coefficients are

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx \mathrm{d}x,\ b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx \mathrm{d}x.$$

Theorem 5.22. Let $u_1, u_2, \ldots, u_n$ be linearly independent and let u_i^* be the u_i's orthonormalized by the Gram-Schmidt process. If $w = \sum_{i=1}^n a_i u_i$, then

$$w = \sum_{i=1}^n \langle w, u_i^* \rangle\, u_i^*, \tag{5.16}$$

i.e. w is equal to its Fourier expansion.

Proof. By the condition $w = \sum_{i=1}^n a_i u_i$ and Corollary 5.15, 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 5.23 (Minimum properties of Fourier expansions)**.** Let $u_1^*, u_2^*, \ldots$ be an orthonormal system and let w be arbitrary. Then

$$\left\| w - \sum_{i=1}^N \langle w, u_i^* \rangle\, u_i^* \right\| \le \left\| w - \sum_{i=1}^N a_i u_i^* \right\|, \tag{5.17}$$

for any selection of constants $a_1, a_2, \cdots, a_N$.

Proof. With the shorthand notation $\sum_i = \sum_{i=1}^N$, we deduce

from Definition B.103 and properties of inner products

$$\begin{aligned}
&\left\| w - \sum_i a_i u_i^* \right\|^2 = \left\langle w - \sum_i a_i u_i^*, w - \sum_i a_i u_i^* \right\rangle \\
&= \langle w, w \rangle - \left\langle w, \sum_i a_i u_i^* \right\rangle - \left\langle \sum_i a_i u_i^*, w \right\rangle \\
&\quad + \left\langle \sum_i a_i u_i^*, \sum_i a_i u_i^* \right\rangle \\
&= \langle w, w \rangle - \sum_i \overline{a_i} \langle w, u_i^* \rangle - \sum_i a_i \langle u_i^*, w \rangle \\
&\quad + \sum_i \sum_j a_i \overline{a_j} \langle u_i^*, u_j^* \rangle \\
&= \langle w, w \rangle - \sum_i \overline{a_i} \langle w, u_i^* \rangle - \sum_i a_i \langle u_i^*, w \rangle + \sum_i |a_i|^2 \\
&\quad - \sum_i \langle u_i^*, w \rangle \langle w, u_i^* \rangle + \sum_i \langle u_i^*, w \rangle \langle w, u_i^* \rangle \\
&\quad = \|w\|^2 - \sum_i |\langle w, u_i^* \rangle|^2 + \sum_i |a_i - \langle w, u_i^* \rangle|^2 , \qquad (5.18)
\end{aligned}$$

where "$|\cdot|$" denotes the modulus of a complex number. The first two terms are independent of a_i. Therefore $\|w - \sum_i a_i u_i^*\|^2$ is minimized only when $a_i = \langle w, u_i^* \rangle$. □

Corollary 5.24. Let $(u_1, u_2, \ldots, u_n)$ be an independent list. The fundamental problem of linearly approximating an arbitrary vector w is solved by the best approximation $\hat{\varphi} = \sum_k \langle w, u_k^* \rangle u_k^*$ where u_k^*'s are the u_k's orthonormalized by the Gram-Schmidt process. The error norm is

$$\|w - \hat{\varphi}\|^2 := \min_{a_k} \left\| w - \sum_{k=1}^n a_k u_k \right\|^2 = \|w\|^2 - \sum_{k=1}^n |\langle w, u_k^* \rangle|^2 . \qquad (5.19)$$

Proof. This follows directly from (5.18). □

Corollary 5.25 (Bessel inequality)**.** If $u_1^*, u_2^*, \ldots, u_N^*$ are orthonormal, then, for an arbitrary w,

$$\sum_{i=1}^N |\langle w, u_i^* \rangle|^2 \le \|w\|^2. \qquad (5.20)$$

Proof. This follows directly from Corollary 5.24 and the real positivity of a norm. □

Corollary 5.26. The Gram-Schmidt process in Definition 5.13 satisfies

$$\forall n \in \mathbb{N}^+, \quad \|v_{n+1}\|^2 = \|u_{n+1}\|^2 - \sum_{k=1}^n |\langle u_{n+1}, u_k^* \rangle|^2 . \qquad (5.21)$$

Proof. By (5.10a), 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 5.24, identifying w with u_{n+1} completes the proof. □

Example 5.27. Consider the problem in Example 5.3 in the sense of least square approximation with the weight function $\rho = 1$. It is equivalent to

$$\min_{a_i} \int_{-1}^{+1} \left(e^x - \sum_{i=0}^n a_i x^i \right)^2 \mathrm{d}x. \qquad (5.22)$$

For $n = 1, 2$, use the Legendre polynomials derived in Example 5.18:

$$u_1^* = \frac{1}{\sqrt{2}}, \ u_2^* = \sqrt{\frac{3}{2}} x, \ u_3^* = \frac{1}{4}\sqrt{10}(3x^2 - 1),$$

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

$$\hat{\varphi}_n = \begin{cases} \frac{1}{2e}(e^2 - 1) + \frac{3}{e} x & n = 1; \\ \hat{\varphi}_1 + \frac{5}{4e}(e^2 - 7)(3x^2 - 1) & n = 2. \end{cases} \qquad (5.23)$$

5.3 The normal equations

Theorem 5.28. Let $u_1, u_2, \ldots, u_n \in X$ be linearly independent and let u_i^* be the u_i's orthonormalized by the Gram-Schmidt process. Then, for any element w,

$$\forall j = 1, 2, \ldots, n, \qquad \left(w - \sum_{k=1}^n \langle w, u_k^* \rangle u_k^* \right) \perp u_j^*, \qquad (5.24)$$

where "$\perp$" denotes orthogonality.

Proof. Take the inner product of the two vectors and apply the conditions on orthonormal systems. □

Corollary 5.29. Let $u_1, u_2, \ldots, u_n \in X$ be linearly independent. If $\hat{\varphi} = \sum_{k=1}^n a_k u_k$ is the best linear approximant to w, then

$$\forall j = 1, 2, \ldots, n, \qquad (w - \hat{\varphi}) \perp u_j. \qquad (5.25)$$

Proof. Since $\hat{\varphi} = \sum_{k=1}^n a_k u_k$ is the best linear approximant to w, Theorem 5.23 implies that

$$\sum_{k=1}^n a_k u_k = \sum_{k=1}^n \langle w, u_k^* \rangle u_k^*.$$

Corollary 5.15 and Theorem 5.28 complete the proof. □

Definition 5.30. Let $u_1, u_2, \ldots, u_n$ be a sequence of elements in an inner product space. The $n \times n$ matrix

$$\begin{aligned}
G &= G(u_1, u_2, \cdots, u_n) = (\langle u_i, u_j \rangle) \\
&= \begin{bmatrix} \langle u_1, u_1 \rangle & \langle u_1, u_2 \rangle & \ldots & \langle u_1, u_n \rangle \\ \langle u_2, u_1 \rangle & \langle u_2, u_2 \rangle & \ldots & \langle u_2, u_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, u_1 \rangle & \langle u_n, u_2 \rangle & \ldots & \langle u_n, u_n \rangle \end{bmatrix}
\end{aligned} \qquad (5.26)$$

is the *Gram matrix* of $u_1, u_2, \dots, u_n$. Its determinant

$$g = g(u_1, u_2, \dots, u_n) = \det(\langle u_i, u_j \rangle) \tag{5.27}$$

is the *Gram determinant*.

Lemma 5.31. Let $w_i = \sum_{j=1}^n a_{ij} u_j$ for $i = 1, 2, \dots, n$. Let $A = (a_{ij})$ and its conjugate transpose $A^H = (\overline{a_{ji}})$. Then we have

$$G(w_1, w_2, \dots, w_n) = AG(u_1, u_2, \dots, u_n)A^H \tag{5.28}$$

and

$$g(w_1, w_2, \dots, w_n) = |\det A|^2 g(u_1, u_2, \dots, u_n). \tag{5.29}$$

Proof. The inner product of u_i and w_j yields

$$\begin{aligned}
&\begin{bmatrix} \langle u_1, w_1\rangle & \langle u_1, w_2\rangle & \dots & \langle u_1, w_n\rangle \\ \langle u_2, w_1\rangle & \langle u_2, w_2\rangle & \dots & \langle u_2, w_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, w_1\rangle & \langle u_n, w_2\rangle & \dots & \langle u_n, w_n\rangle \end{bmatrix} \\
=&\begin{bmatrix} \langle u_1, u_1\rangle & \langle u_1, u_2\rangle & \dots & \langle u_1, u_n\rangle \\ \langle u_2, u_1\rangle & \langle u_2, u_2\rangle & \dots & \langle u_2, u_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, u_1\rangle & \langle u_n, u_2\rangle & \dots & \langle u_n, u_n\rangle \end{bmatrix} \begin{bmatrix} \overline{a_{11}} & \dots & \overline{a_{n1}} \\ \overline{a_{12}} & \dots & \overline{a_{n2}} \\ \vdots & \ddots & \vdots \\ \overline{a_{1n}} & \dots & \overline{a_{nn}} \end{bmatrix} \\
=&G(u_1, u_2, \dots, u_n)A^H.
\end{aligned}$$

Therefore (5.28) holds since

$$\begin{aligned}
G(w_1, w_2, \dots, w_n) &= \begin{bmatrix} \langle w_1, w_1\rangle & \langle w_1, w_2\rangle & \dots & \langle w_1, w_n\rangle \\ \langle w_2, w_1\rangle & \langle w_2, w_2\rangle & \dots & \langle w_2, w_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle w_n, w_1\rangle & \langle w_n, w_2\rangle & \dots & \langle w_n, w_n\rangle \end{bmatrix} \\
&= \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} \langle u_1, w_1\rangle & \langle u_1, w_2\rangle & \dots & \langle u_1, w_n\rangle \\ \langle u_2, w_1\rangle & \langle u_2, w_2\rangle & \dots & \langle u_2, w_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n, w_1\rangle & \langle u_n, w_2\rangle & \dots & \langle u_n, w_n\rangle \end{bmatrix} \\
&= AG(u_1, u_2, \dots, u_n)A^H.
\end{aligned}$$

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 yields

$$\overline{\det A} = \overline{\det A^T} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^n \overline{a_{\sigma_i, i}} = \det A^H.$$

Take the determinant of (5.28), apply the identity $\det(AB) = \det(A)\det(B)$, and we have (5.29). □

Theorem 5.32. For nonzero elements $u_1, u_2, \dots, u_n \in X$, we have

$$0 \le g(u_1, u_2, \dots, u_n) \le \prod_{k=1}^n \|u_k\|^2, \tag{5.30}$$

where the lower equality holds if and only if $u_1, u_2, \dots, u_n$ are linearly dependent and the upper equality holds if and only if they are orthogonal.

Proof. Suppose $u_1, u_2, \dots, u_n$ are linearly dependent. Then we can find constants $c_1, c_2, \dots, c_n$ such that $\sum_{i=1}^n c_i u_i = \mathbf{0}$ with at least one constant c_j being nonzero. Construct vectors

$$w_k = \begin{cases} \sum_{i=1}^n c_i u_i = \mathbf{0}, & k = j; \\ u_k, & k \ne j. \end{cases}$$

We have $g(w_1, w_2, \dots, w_n) = 0$ 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 5.31 yields $g(u_1, u_2, \dots, u_n) = 0$.

Now suppose $u_1, u_2, \dots, u_n$ are linearly independent. Theorem 5.14 yields constants a_{ij} such that $a_{kk} > 0$ and the following vectors are orthonormal:

$$u_k^* = \sum_{i=1}^k a_{ki} u_i.$$

Then Definition 5.30 implies $g(u_1^*, u_2^*, \dots, u_n^*) = 1$. Also, we have $\det(a_{ij}) = \prod_{k=1}^n a_{kk}$ because the matrix (a_{ij}) is triangular. It then follows from Lemma 5.31 that

$$g(u_1, u_2, \dots, u_n) = \prod_{k=1}^n \frac{1}{a_{kk}^2} > 0. \tag{5.31}$$

Since the list of vectors $(u_1, u_2, \dots, u_n)$ is either dependent or independent, the arguments so far show that $g(u_1, u_2, \dots, u_n) = 0$ if and only if $u_1, u_2, \dots, u_n$ are linearly dependent.

Suppose $u_1, u_2, \dots, u_n$ are orthogonal. By Definition 5.30, $G(u_1, u_2, \dots, u_n)$ is a diagonal matrix with $\|u_k\|^2$ on the diagonals. Hence the orthogonality of u_k's implies

$$g(u_1, u_2, \dots, u_n) = \prod_{k=1}^n \|u_k\|^2. \tag{5.32}$$

For the converse statement, suppose (5.32) holds. Then $u_1, u_2, \dots, u_n$ must be independent because otherwise it would contradict the lower equality proved as above. Apply the Gram-Schmidt process to $(u_1, u_2, \dots, u_n)$ and we know from Theorem 5.14 that $\frac{1}{a_{kk}} = \|v_k\|$. Set the length of the list in Theorem 5.14 to $1, 2, \dots, n$ and we know from (5.31) and (5.32) that

$$\forall k = 1, 2, \dots, n, \qquad \|u_k\|^2 = \|v_k\|^2. \tag{5.33}$$

Then Corollary 5.26 and (5.33) imply

$$\forall k = 1, 2, \ldots, n, \qquad \sum_{j=1}^{k-1} \left|\langle u_k, u_j^* \rangle\right|^2 = 0,$$

which further implies

$$\forall k = 1, 2, \ldots, n,\ \forall j = 1, 2, \ldots, k-1, \qquad \langle u_k, u_j^* \rangle = 0,$$

which, together with Corollary 5.15, implies the orthogonality of u_k's. Finally, we remark that the maximum of $g(u_1, u_2, \ldots, u_n)$ is indeed $\prod_{k=1}^n \|u_k\|^2$ because of (5.31), $\frac{1}{a_{kk}} = \|v_k\|$, and Corollary 5.26. □

Theorem 5.33. Let $\hat{\varphi} = \sum_{i=1}^n a_i u_i$ be the best approximation to w constructed from the list of independent vectors $(u_1, u_2, \ldots, u_n)$. Then the coefficients

$$\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$$

are uniquely determined from the linear system of *normal equations*,

$$G(u_1, u_2, \ldots, u_n)^T \mathbf{a} = \mathbf{c}, \tag{5.34}$$

where $\mathbf{c} = [\langle w, u_1 \rangle, \langle w, u_2 \rangle, \ldots, \langle w, u_n \rangle]^T$.

Proof. Corollary 5.29 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 (5.34). The uniqueness of the coefficients follows from Theorem 5.32 and Cramer's rule. □

Example 5.34. Solve Example 5.27 by normal equations.

To find the best approximation $\hat{\varphi} = a_0 + a_1 x + a_2 x^2$ to e^x from the linearly independent list $(1, x, x^2)$, we first construct the Gram matrix from (5.26), (5.7), and $\rho = 1$:

$$G(1, x, x^2) = \begin{bmatrix} \langle 1, 1 \rangle & \langle 1, x \rangle & \langle 1, x^2 \rangle \\ \langle x, 1 \rangle & \langle x, x \rangle & \langle x, x^2 \rangle \\ \langle x^2, 1 \rangle & \langle x^2, x \rangle & \langle x^2, x^2 \rangle \end{bmatrix} = \begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix}.$$

We then calculate the vector

$$\mathbf{c} = \begin{bmatrix} \langle e^x, 1 \rangle \\ \langle e^x, x \rangle \\ \langle e^x, x^2 \rangle \end{bmatrix} = \begin{bmatrix} e - 1/e \\ 2/e \\ e - 5/e \end{bmatrix}.$$

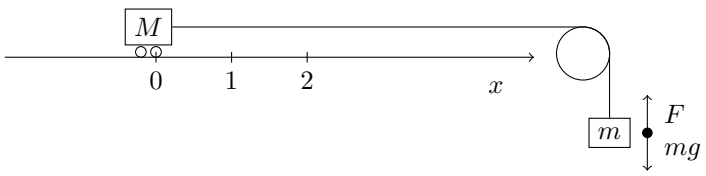
The normal equations then yields

$$a_0 = \frac{3(11 - e^2)}{4e},\ a_1 = \frac{3}{e},\ a_2 = \frac{15(e^2 - 7)}{4e}.$$

With these values, it is easily verified that the best approximation $\hat{\varphi} = a_0 + a_1 x + a_2 x^2$ equals that in (5.23).

5.4 Discrete least squares (DLS)

Example 5.35 (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."

5.4.1 Reusing the formalism

Definition 5.36. Define a function $\lambda : \mathbb{R} \to \mathbb{R}$

$$\lambda(t) = \begin{cases} 0 & \text{if } t \in (-\infty, a), \\ \int_a^t \rho(\tau)\mathrm{d}\tau & \text{if } t \in [a, b], \\ \int_a^b \rho(\tau)\mathrm{d}\tau & \text{if } t \in (b, +\infty). \end{cases} \tag{5.35}$$

Then a corresponding *continuous measure* $\mathrm{d}\lambda$ can be defined as

$$\mathrm{d}\lambda = \begin{cases} \rho(t)\mathrm{d}t & \text{if } t \in [a, b], \\ 0 & \text{otherwise}, \end{cases} \tag{5.36}$$

where the *support of the continuous measure* $\mathrm{d}\lambda$ is the interval $[a, b]$.

Definition 5.37. 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 5.38. The *Heaviside function* is the truncated power function with exponent 0,

$$H(x) = x_+^0 = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{5.37}$$

Lemma 5.39. For a function $u : \mathbb{R} \to \mathbb{R}$, define

$$\lambda(t) = \sum_{i=1}^{N} \rho_i H(t - t_i), \tag{5.38}$$

and we have

$$\int_{\mathbb{R}} u(t)\mathrm{d}\lambda = \sum_{i=1}^{N} \rho_i u(t_i). \tag{5.39}$$

Proof. The *Dirac Delta function*, $\delta(x)$, is roughly a generalized function that satisfies

$$\delta(x) = \begin{cases} +\infty & x = 0, \\ 0 & x \neq 0. \end{cases} \tag{5.40}$$

Note: the above definition of $\delta(x)$ is heuristic. A rigorous one should employ the concept of measures.

Useful properties of $\delta(x)$ include

$$\int_{-\infty}^{+\infty} \delta(x)\mathrm{d}x = 1, \tag{5.41}$$

$$\int_0^x \delta(t)\mathrm{d}t = H(x), \tag{5.42}$$

$$\int_{-\infty}^{+\infty} f(t)\delta(t - t_0)\mathrm{d}t = f(t_0). \tag{5.43}$$

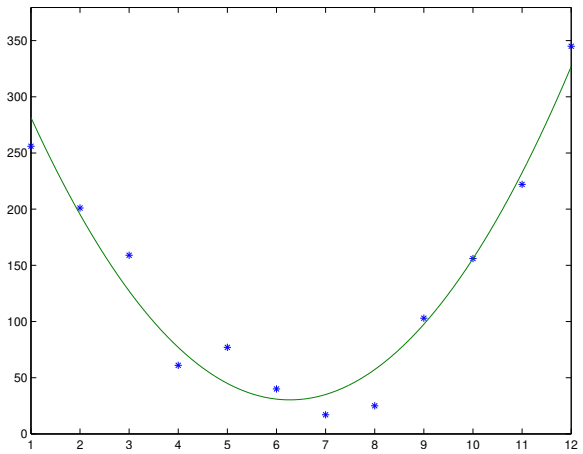
Then (5.38), (5.42), and (5.43) yield

$$\int_{\mathbb{R}} u(t)\mathrm{d}\lambda = \int_{\mathbb{R}} \sum_{i=1}^{N} \rho_i \delta(t - t_i) u(t)\mathrm{d}t = \sum_{i=1}^{N} \rho_i u(t_i). \qquad \square$$

5.4.2 DLS via normal equations

Example 5.40. 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 5.34, we have

$$G(1, x, x^2) = \begin{bmatrix} \langle 1,1\rangle & \langle 1,x\rangle & \langle 1,x^2\rangle \\ \langle x,1\rangle & \langle x,x\rangle & \langle x,x^2\rangle \\ \langle x^2,1\rangle & \langle x^2,x\rangle & \langle x^2,x^2\rangle \end{bmatrix} = \begin{bmatrix} 12 & 78 & 650 \\ 78 & 650 & 6084 \\ 650 & 6084 & 60710 \end{bmatrix},$$

$$\mathbf{c} = \begin{bmatrix} \langle y,1\rangle \\ \langle y,x\rangle \\ \langle y,x^2\rangle \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{12} y_i \\ \sum_{i=1}^{12} y_i x_i \\ \sum_{i=1}^{12} y_i x_i^2 \end{bmatrix} = \begin{bmatrix} 1662 \\ 11392 \\ 109750 \end{bmatrix}.$$

Then the normal equations yield

$$\mathbf{a} = G^{-1}\mathbf{c} = [386.00,\ -113.43,\ 9.04]^T.$$

The corresponding polynomial is plotted in the figure.

5.4.3 DLS via QR decomposition

Definition 5.41. A matrix $A \in \mathbb{R}^{n\times n}$ is *orthogonal* iff $A^T A = I$.

Definition 5.42. A matrix A is *upper triangular* iff

$$\forall i, j, \qquad i > j \ \Rightarrow\ a_{i,j} = 0.$$

Similarly, a matrix A is *lower triangular* iff

$$\forall i, j, \qquad i < j \ \Rightarrow\ a_{i,j} = 0.$$

Theorem 5.43 (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 5.15, the Gram-Schmidt process determines a unique orthogonal matrix $A_r^* = [u_1^*,\ u_2^*,\ \ldots,\ u_r^*] \in \mathbb{R}^{m\times r}$ and a unique upper triangular matrix such that

$$A_r = A_r^* \begin{bmatrix} b_{11} & b_{21} & \ldots & b_{r1} \\ & b_{22} & \ldots & b_{r2} \\ & & \ddots & \vdots \\ & & & b_{rr} \end{bmatrix}. \tag{5.44}$$

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 (5.44) and show that the QR form of (5.44) is maintained. For those zero column vectors in (S-1), we have

$$\begin{aligned} A_\xi =& [\xi_1\ \ldots\ \xi_{k_1-1}\ u_1\ u_2\ \ldots\ u_r] \\ =& A_r^* \begin{bmatrix} 0 & \ldots & 0 & b_{11} & b_{21} & \ldots & b_{r1} \\ 0 & \ldots & 0 & & b_{22} & \ldots & b_{r2} \\ \vdots & \ddots & \vdots & & & \ddots & \vdots \\ 0 & \ldots & 0 & & & & b_{rr} \end{bmatrix}. \end{aligned} \tag{5.45}$$

For each ξ_ℓ with $\ell \in R_j$ in (S-2), we have

$$\begin{aligned} & [u_1,\ u_2,\ \ldots,\ u_{j-1},\ \xi_\ell] \\ =& \begin{bmatrix} u_1^*,\ u_2^*,\ \ldots,\ u_{j-1}^* \end{bmatrix} \begin{bmatrix} b_{11} & \ldots & b_{j-1,1} & c_{\ell,1} \\ & \ddots & \vdots & \vdots \\ & & b_{j-1,j-1} & c_{\ell,j-1} \end{bmatrix}, \end{aligned} \tag{5.46}$$

where $\xi_\ell = c_{\ell,1}u_1^* + \ldots + c_{\ell,j-1}u_{j-1}^*$. With (5.45) as the induction basis and (5.46) 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 5.41 and 5.9 complete the proof. Otherwise $r < m$ and the proof is completed by the well-known fact in linear algebra that a list of orthonormal vectors can be extended to an orthonormal basis. □

Lemma 5.44. An orthogonal matrix preserves the 2-norm of the vectors it acts on.

Proof. Definition 5.41 yields

$$\forall \mathbf{x} \in \mathrm{dom}(Q), \qquad \|Q\mathbf{x}\|_2^2 = \mathbf{x}^T Q^T Q \mathbf{x} = \mathbf{x}^T \mathbf{x} = \|x\|_2^2. \quad \square$$

Theorem 5.45. Consider an over-determined linear system $A\mathbf{x} = \mathbf{b}$ where $A \in \mathbb{R}^{m\times n}$ and $m \ge n$. The discrete linear least square problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|Ax - b\|_2^2$$

is solved by $\mathbf{x}^*$ satisfying

$$R_1 \mathbf{x}^* = \mathbf{c}, \tag{5.47}$$

where $R_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{c} \in \mathbb{R}^n$ result from the QR factorization of A:

$$Q^T A = R = \begin{bmatrix} R_1 \\ \mathbf{0} \end{bmatrix}, \qquad Q^T \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{r} \end{bmatrix}. \tag{5.48}$$

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 5.44. □

5.5 Problems

5.5.1 Theoretical questions

I. Fill in the details for the proof of Theorem 5.7.

II. Consider the Chebyshev polynomials of the first kind.

(a) Show that they are orthogonal on $[-1, 1]$ with respect to the inner product in Theorem 5.7 with the weight function $\rho(x) = \frac{1}{\sqrt{1-x^2}}$.

(b) Normalize the first three Chebyshev polynomials to arrive at an orthonormal system.

III. Least-square approximation of a continuous function. Approximate the circular arc given by the equation $y(x) = \sqrt{1-x^2}$ for $x \in [-1, 1]$ by a quadratic polynomial with respect to the inner product in Theorem 5.7.

(a) $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ with Fourier expansion,

(b) $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ with normal equations.

IV. Discrete least square via orthonormal polynomials. Consider the example on the table of sales record in Example 5.40.

(a) Starting from the independent list $(1, x, x^2)$, construct orthonormal polynomials by the Gram-Schmidt process using

$$\langle u(t), v(t)\rangle = \sum_{i=1}^{N} \rho(t_i) u(t_i) v(t_i) \tag{5.49}$$

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?

5.5.2 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 6

Numerical Integration

Definition 6.1. A *weighted quadrature formula* $I_n(f)$ for a function $f \in L[a,b]$ is a formula

$$I_n(f) := \sum_{k=1}^{n} w_k f(x_k) \tag{6.1}$$

that approximates the definite integral of f on $[a,b]$

$$I(f) := \int_a^b f(x)\rho(x)\mathrm{d}x, \tag{6.2}$$

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 6.2. If a and/or b are infinite, $I(f)$ and $I_n(f)$ in (6.1) may still be well defined if the *moment of weight function*

$$\mu_j := \int_a^b x^j \rho(x)\mathrm{d}x \tag{6.3}$$

exists and is finite for all $j \in \mathbb{N}$.

6.1 Accuracy and convergence

Definition 6.3. The *remainder*, or *error*, of $I_n(f)$ is

$$E_n(f) := I(f) - I_n(f). \tag{6.4}$$

$I_n(f)$ is said to be *convergent* for $\mathcal{C}[a,b]$ iff

$$\forall f \in \mathcal{C}[a,b], \qquad \lim_{n\to+\infty} I_n(f) = I(f). \tag{6.5}$$

Definition 6.4. A subset $\mathbb{V} \subset \mathcal{C}[a,b]$ is *dense* in $\mathcal{C}[a,b]$ iff

$$\forall f \in \mathcal{C}[a,b], \forall \epsilon > 0, \exists f_\epsilon \in \mathbb{V}, \text{ s.t. } \max_{x\in[a,b]} |f(x) - f_\epsilon(x)| \le \epsilon. \tag{6.6}$$

Theorem 6.5. Let $\{I_n(f) : n \in \mathbb{N}^+\}$ be a sequence of quadrature formulas that approximate $I(f)$, where I_n and $I(f)$ are defined in (6.1) and (6.2). Let $\mathbb{V}$ be a dense subset of $\mathcal{C}[a,b]$. $I_n(f)$ is convergent for $\mathcal{C}[a,b]$ if and only if

(a) $\forall f \in \mathbb{V}$, $\lim_{n\to+\infty} I_n(f) = I(f)$,

(b) $B := \sup_{n\in\mathbb{N}^+} \sum_{k=0}^{n} |w_k| < +\infty$.

Proof. For necessity, it is trivial to deduce (a) from (6.5). In contrast, it is highly nontrivial to deduce (b) from (6.5). 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 (6.6) holds, define $K := \max_{x\in[a,b]} |f(x) - f_\epsilon(x)|$. Then we have

$$\begin{aligned}
|E_n(f)| \le& |I(f) - I(f_\epsilon)| + |I(f_\epsilon) - I_n(f_\epsilon)| + |I_n(f_\epsilon) - I_n(f)| \\
=& \left| \int_a^b [f(x) - f_\epsilon(x)]\, \rho(x)\mathrm{d}x \right| \\
&+ |I(f_\epsilon) - I_n(f_\epsilon)| + \left| \sum_{k=1}^{n} w_k \left[f(x_k) - f_\epsilon(x_k) \right] \right| \\
\le& K \left[\int_a^b \rho(x)\mathrm{d}x + \sum_{k=1}^{n} |w_k| \right] + |I(f_\epsilon) - I_n(f_\epsilon)|,
\end{aligned}$$

where the first step follows from the triangular inequality, the second from Definition 6.1, and the third from the integral mean value theorem C.64. 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 6.6 (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

$$\forall x \in [a,b], \qquad |f(x) - p_n(x)| \le \epsilon. \tag{6.7}$$

Proof. Not required. □

Definition 6.7. A weighted quadrature formula (6.1) has (polynomial) *degree of exactness* d_E iff

$$\begin{cases} \forall f \in \mathbb{P}_{d_E}, & E_n(f) = 0, \\ \exists g \in \mathbb{P}_{d_E+1}, \text{ s.t.} & E_n(g) \neq 0, \end{cases} \tag{6.8}$$

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

Lemma 6.8. Let $x_1, \ldots, x_n$ be given as distinct nodes of $I_n(f)$. If $d_E \geq n-1$, then its weights can be deduced as

$$\forall k = 1, \ldots, n, \qquad w_k = \int_a^b \rho(x)\ell_k(x)\mathrm{d}x, \tag{6.9}$$

where $\ell_k(x)$ is the fundamental polynomial for pointwise interpolation in (3.9) applied to the given nodes,

$$\ell_k(x) := \prod_{i\neq k; i=1}^{n} \frac{x - x_i}{x_k - x_i}. \tag{6.10}$$

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 3.5). Then we have

$$\begin{aligned}\sum_{k=1}^{n} w_k p_{n-1}(x_k) &= \int_a^b p_{n-1}(f;x)\rho(x)\mathrm{d}x \\ = \int_a^b \sum_{k=1}^{n} \{\ell_k(x) f(x_k)\}\, \rho(x)\mathrm{d}x &= \sum_{k=1}^{n} w_k f(x_k),\end{aligned}$$

where the first step follows from $d_E \geq n-1$ and the second step from the interpolation conditions (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 4.22) for each x_k. □

6.2 Newton-Cotes formulas

Definition 6.9. A *Newton-Cotes formula* is a formula (6.1) based on approximating $f(x)$ by interpolating it on uniformly spaced nodes $x_1, \ldots, x_n \in [a,b]$.

Definition 6.10. *The trapezoidal rule* is a formula (6.1) based on approximating $f(x)$ by the straight line that connects $(a, f(a))^T$ and $(b, f(b))^T$. In particular, for $\rho(x) \equiv 1$, it is simply

$$I^T(f) = \frac{b-a}{2}[f(a) + f(b)]. \tag{6.11}$$

Example 6.11. Derive the trapezoidal rule for the weight function $\rho(x) = x^{-1/2}$ on the interval $[0,1]$. Note that one cannot apply (6.11) to $\rho(x)f(x)$ because $\rho(0) = \infty$. (6.9) 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

$$I^T(f) = \frac{2}{3}[2f(0) + f(1)]. \tag{6.12}$$

Theorem 6.12. For $f \in \mathcal{C}^2[a,b]$ with weight function $\rho(x) \equiv 1$, the remainder of the trapezoidal rule satisfies

$$\exists \zeta \in [a,b] \text{ s.t. } E^T(f) = -\frac{(b-a)^3}{12} f''(\zeta). \tag{6.13}$$

Proof. By Theorem 3.5, the interpolating polynomial $p_1(f;x)$ is unique. Then we have

$$\begin{aligned} E^T(f) &= -\int_a^b \frac{f''(\xi(x))}{2}(x-a)(b-x)\mathrm{d}x \\ &= -\frac{f''(\zeta)}{2}\int_a^b (x-a)(b-x)\mathrm{d}x = -\frac{(b-a)^3}{12} f''(\zeta), \end{aligned}$$

where the first step follows from Theorem 3.7 and the second step from the integral mean value theorem (Theorem C.64). Here we can apply Theorem C.64 because

$$w(x) = (x-a)(b-x)$$

is always positive on (a,b). Also note that ξ is a function of x while ζ is a constant depending only on f, a, and b. □

Definition 6.13. *Simpson's rule* is a formula (6.1) 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

$$I^S(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]. \tag{6.14}$$

Theorem 6.14. For $f \in \mathcal{C}^4[a,b]$ with weight function $\rho(x) \equiv 1$, the remainder of Simpson's rule satisfies

$$\exists \zeta \in [a,b] \text{ s.t. } E^S(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta). \tag{6.15}$$

Proof. It is difficult to imitate the proof of Theorem 6.12, since $(x-a)(x-b)(x-\frac{a+b}{2})$ changes sign over $[a,b]$ and the integral mean value theorem is not applicable. To overcome this difficulty, we can formulate the interpolation via a Hermite problem so that Theorem C.64 can be applied. See problem I in Section 6.5 for the main steps. □

Example 6.15. Consider the integral

$$I = \int_{-4}^{4} \frac{\mathrm{d}x}{1+x^2} = 2\tan^{-1}(4) = 2.6516\cdots \tag{6.16}$$

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 6.9.

6.3 Composite formulas

Definition 6.16. The *composite trapezoidal rule* for approximating $I(f)$ in (6.2) with $\rho(x) \equiv 1$ is

$$I_n^T(f) = \frac{h}{2}f(x_0) + h\sum_{k=1}^{n-1} f(x_k) + \frac{h}{2}f(x_n), \tag{6.17}$$

where $h = \frac{b-a}{n}$ and $x_k = a + kh$.

Theorem 6.17. For $f \in \mathcal{C}^2[a,b]$, the remainder of the composite trapezoidal rule satisfies

$$\exists \xi \in (a,b) \text{ s.t. } E_n^T(f) = -\frac{b-a}{12}h^2 f''(\xi). \tag{6.18}$$

Proof. Apply Theorem 6.12 to the subintervals, sum up the errors, and we have

$$E_n^T(f) = -\frac{b-a}{12}h^2\left[\frac{1}{n}\sum_{k=0}^{n-1} f''(\xi_k)\right]. \tag{6.19}$$

$f \in \mathcal{C}^2[a,b]$ implies $f'' \in \mathcal{C}[a,b]$. The proof is completed by (6.19) and the intermediate value Theorem C.32. □

Definition 6.18. The *composite Simpson's rule* for approximating $I(f)$ in (6.2) with $\rho(x) \equiv 1$ is

$$I_n^S(f) = \frac{h}{3}\big[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots + 4f(x_{n-1}) + f(x_n)\big], \tag{6.20}$$

where $h = \frac{b-a}{n}$, $x_k = a + kh$, and n is even.

Theorem 6.19. For $f \in \mathcal{C}^4[a,b]$ and $n \in 2\mathbb{N}^+$, the remainder of the composite Simpson's rule satisfies

$$\exists \xi \in (a,b) \text{ s.t. } E_n^S(f) = -\frac{b-a}{180}h^4 f^{(4)}(\xi). \tag{6.21}$$

Proof. Exercise. □

6.4 Gauss formulas

Lemma 6.20. 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

$$p = qs + r. \tag{6.22}$$

Proof. Rewrite (6.22) as

$$\sum_{i=0}^{n+m} p_i x^i = \left(\sum_{i=0}^{m} q_i x^i\right)\left(\sum_{i=0}^{n} s_i x^i\right) + \sum_{i=0}^{n-1} r_i x^i. \tag{6.23}$$

Since monomials are linearly independent, (6.23) 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 (6.23). □

Definition 6.21. The *node polynomial* associated with the nodes x_k's of a weighted quadrature formula is

$$v_n(x) = \prod_{k=1}^{n}(x - x_k). \tag{6.24}$$

Theorem 6.22. Suppose a quadrature formula (6.1) 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,

$$\forall p \in \mathbb{P}_{j-1}, \qquad \int_a^b v_n(x)p(x)\rho(x)\mathrm{d}x = 0. \tag{6.25}$$

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 (6.24).

To prove the sufficiency, we must show that $E_n(p) = 0$ for any $p \in \mathbb{P}_{n+j-1}$. Lemma 6.20 yields

$$\forall p \in \mathbb{P}_{n+j-1},\ \exists q \in \mathbb{P}_{j-1},\ r \in \mathbb{P}_{n-1},\ \text{s.t. } p = qv_n + r. \tag{6.26}$$

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 (6.26), the second from (6.25), the third from the condition of $d_E \ge n-1$, the fourth from (6.26), and the last from (6.24). □

Definition 6.23. A *Gaussian quadrature formula* is a formula (6.1) whose nodes are the zeros of the polynomial $v_n(x)$ in (6.24) that satisfies (6.25) for $j = n$.

Corollary 6.24. A Gauss formula has $d_E = 2n - 1$.

Proof. The index j in (6.25) 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 6.22 is optimal: the formula (6.1) achieves the highest degree of exactness $2n-1$. From an algebraic viewpoint, the $2n$ degrees of freedom of nodes and weights in (6.1) determine a polynomial of degree at most $2n-1$. The rest follows from Theorem 6.22. □

Corollary 6.25. Weights of a Gauss formula $I_n(f)$ are

$$\forall k = 1, \cdots, n, \quad w_k = \int_a^b \frac{v_n(x)}{(x - x_k)v_n'(x_k)}\rho(x)\mathrm{d}x, \tag{6.27}$$

where $v_n(x)$ is the node polynomial that defines $I_n(f)$.

Proof. This follows from Lemma 6.8; also see (3.11). □

Example 6.26. 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 (6.9), but it is simpler to set up a linear system of equations by exploiting Corollary 6.24, 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

$$I_2^G(f) = \left(1 + \frac{1}{3}\sqrt{\frac{5}{6}}\right) f\left(\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}\right) + \left(1 - \frac{1}{3}\sqrt{\frac{5}{6}}\right) f\left(\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}\right). \tag{6.28}$$

Theorem 6.27. 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 $\left|\left\langle P_{n-j}, v_j^2 \right\rangle\right| > 0$, which contradicts the orthogonality of $p_n(x)$ and $v_j(x)$. □

Corollary 6.28. All nodes of a Gauss formula are real, distinct, and contained in (a, b).

Proof. This follows from Definition 6.23 and Theorem 6.27. □

Lemma 6.29. Gauss formulas have positive weights.

Proof. For each $j = 1, 2, \ldots, n$, the definition of $\ell_j(x)$ in (6.10) 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 (6.10), second step from $d_E = 2n - 1$ and the last step from the conditions on ρ. □

Lemma 6.30. A Gauss formula satisfies

$$\sum_{k=1}^n w_k = \mu_0 \in (0, +\infty).$$

Proof. This follows from setting $j = 0$ in (6.3) and applying the condition on ρ in Definition 6.1. □

Theorem 6.31. Gauss formulas are convergent for $\mathcal{C}[a, b]$.

Proof. Denote by $\mathbb{P}$ the set of real polynomials. Theorem 6.6 states that $\mathbb{P}$ is dense in $\mathcal{C}[a, b]$, i.e. condition (a) in Theorem 6.5 holds. Condition (b) also holds because of Lemma 6.30, (6.3), and $\rho \in L[a, b]$. The rest of the proof follows from Theorem 6.5. □

Theorem 6.32. For $f \in \mathcal{C}^{2n}[a, b]$, the remainder of a Gauss formula $I_n(f)$ satisfies

$$\exists \xi \in [a, b] \text{ s.t. } E_n^G(f) = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b \rho(x)v_n^2(x)\mathrm{d}x, \tag{6.29}$$

where v_n is the node polynomial that defines I_n.

Proof. Not required. □

6.5 Problems

6.5.1 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 $I = 0.596347361\cdots$)

Chapter 7

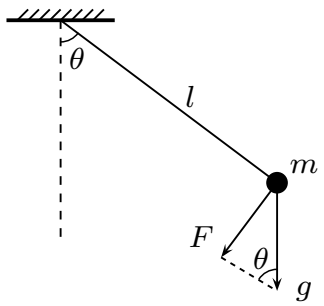
Initial Value Problems (IVPs)

Definition 7.1. A *system of ordinary differential equations* (ODEs) of dimension N is a set of differential equations of the form

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t), t), \tag{7.1}$$

where t is time, $\mathbf{u} \in \mathbb{R}^N$ is the evolutionary variable, and the RHS function has the signature $\mathbf{f} : \mathbb{R}^N \times (0, +\infty) \to \mathbb{R}^N$. In particular, (7.1) is an ODE for $N = 1$.

Definition 7.2. A system of ODEs is *linear* if its RHS function can be expressed as $\mathbf{f}(\mathbf{u}, t) = \alpha(t)\mathbf{u} + \boldsymbol{\beta}(t)$, and *nonlinear* otherwise; it is *homogeneous* if it is linear and $\boldsymbol{\beta}(t) = \mathbf{0}$; it is *autonomous* if $\mathbf{f}$ does not depend on t explicitly; and *nonautonomous* otherwise.



Example 7.3. For the simple pendulum shown above, the moment of inertial and the torque are

$$I = m\ell^2, \ \tau = -mg\ell \sin\theta,$$

and the equation of motion can be derived from Newton's second law $\tau = I\theta''(t)$ as

$$\theta''(t) = -\frac{g}{\ell} \sin\theta, \tag{7.2}$$

which admits a unique solution if we impose two initial conditions

$$\theta(0) = \theta_0, \ \theta'(0) = \omega_0.$$

Alternatively, (7.2) can be derived by the consideration that the total energy remains a constant with respect to time.

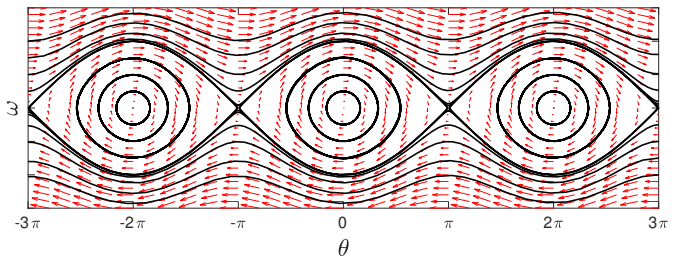
$$L = \frac{1}{2} m(\ell\theta')^2 + mg\ell(1 - \cos\theta);$$

$$\frac{\mathrm{d}L}{\mathrm{d}t} = 0 \Rightarrow \ m\ell^2\theta'\theta'' + mg\ell\theta' \sin\theta = 0.$$

The ODE (7.2) is second-order, nonlinear, and autonomous; it can be reduced to a first-order system as follows,

$$\omega = \theta', \ \mathbf{u} = \begin{pmatrix} \theta \\ \omega \end{pmatrix} \ \Rightarrow \ \mathbf{u}'(t) = \mathbf{f}(\mathbf{u}) := \begin{pmatrix} \omega \\ -\frac{g}{\ell} \sin\theta \end{pmatrix}.$$

See the following plot for some solutions.



Definition 7.4. Given $T > 0$, $\mathbf{f} : \mathbb{R}^N \times [0, T] \to \mathbb{R}^N$, and $\mathbf{u}_0 \in \mathbb{R}^N$, the *initial value problem* (IVP) is to find $\mathbf{u}(t) \in \mathcal{C}^1$ such that

$$\begin{cases} \mathbf{u}(0) & = \mathbf{u}_0, \\ \mathbf{u}'(t) & = \mathbf{f}(\mathbf{u}(t), t), \ \forall t \in [0, T]. \end{cases} \tag{7.3}$$

Definition 7.5. The IVP in Definition 7.4 is *well-posed* if

(i) it admits a unique solution for any fixed $t > 0$,

(ii) $\exists c > 0$, $\hat{\epsilon} > 0$ s.t. $\forall \epsilon < \hat{\epsilon}$, the perturbed IVP

$$\mathbf{v}' = \mathbf{f}(\mathbf{v}, t) + \boldsymbol{\delta}(t), \qquad \mathbf{v}(0) = \mathbf{u}_0 + \boldsymbol{\epsilon}_0 \tag{7.4}$$

satisfies

$$\forall t \in [0, T], \begin{cases} \|\boldsymbol{\epsilon}_0\| < \epsilon \\ \|\boldsymbol{\delta}(t)\| < \epsilon \end{cases} \ \Rightarrow \ \|\mathbf{u}(t) - \mathbf{v}(t)\| \le c\epsilon. \tag{7.5}$$

7.1 Mathematical foundation

7.1.1 Operator norm

Lemma 7.6. The length-scaling factor of a linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ upon any vector is bounded, i.e.,

$$\forall T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m), \ \exists M \in \mathbb{F} \text{ s.t. } \forall \mathbf{x} \in \mathbb{F}^n, \ |T\mathbf{x}| \le M|\mathbf{x}|, \tag{7.6}$$

where $|\cdot|$ denotes the Euclidean 2-norm in (B.35).

Proof. Since $\mathbf{x} = \sum_j x_j \mathbf{e}_j$ and T is a linear map, we have

$$\begin{aligned} |T\mathbf{x}| &= \left| \sum_j x_j T\mathbf{e}_j \right| \le \sum_j |x_j T\mathbf{e}_j| = \sum_j |x_j| |T\mathbf{e}_j| \\ &\le |\mathbf{x}| \sum_j |T\mathbf{e}_j| , \end{aligned}$$

where the second step follows from (NRM-4) in Definition B.114, the third step from (NRM-3) in Definition B.114, and the last step from $|x_j| \le |\mathbf{x}|$. The proof is completed by setting $M := \sum_j |T\mathbf{e}_j|$. □

Corollary 7.7. Any linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is uniformly continuous on $\mathbb{F}^n$.

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, we have

$$|T\mathbf{x} - T\mathbf{y}| = |T(\mathbf{x} - \mathbf{y})| \le M|\mathbf{x} - \mathbf{y}|.$$

Setting $\delta = \epsilon / M$ in Definition C.78. □

Definition 7.8. The *operator norm* of a linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is the non-negative number

$$\|T\| := \inf \{ M : \forall \mathbf{x} \in \mathbb{F}^n, |T\mathbf{x}| \le M|\mathbf{x}| \} . \tag{7.7}$$

Exercise 7.9. Verify that (7.7) is indeed a norm in the sense of Definition B.114.

Corollary 7.10. $\forall \mathbf{x} \in \mathbb{F}^n,\ |T\mathbf{x}| \le \|T\||\mathbf{x}|$.

Proof. This follows directly from (7.7). □

Corollary 7.11. $\|T\| = \sup_{|\mathbf{x}| \le 1} |T\mathbf{x}| = \sup_{|\mathbf{x}| = 1} |T\mathbf{x}|$.

Proof. Since T is a linear map and $|\cdot|$ is a norm, we have

$$|T(c\mathbf{x})| = |cT\mathbf{x}| = |c||T\mathbf{x}|.$$

Hence the inequality $|T\mathbf{x}| \le M|\mathbf{x}|$ in (7.7) holds for all $\mathbf{x} \ne 0$ if and only if it holds for all $\mathbf{x}$ with $|\mathbf{x}| \in (0, 1]$, if and only if it holds for all $\mathbf{x}$ with $|\mathbf{x}| = 1$. □

Corollary 7.12. The composition of two linear maps $S \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ and $T \in \mathcal{L}(\mathbb{F}^m, \mathbb{F}^k)$ satisfies

$$\|TS\| \le \|T\|\|S\|. \tag{7.8}$$

Proof. By Corollary 7.10, we have

$$|(TS)(\mathbf{x})| = |T(S\mathbf{x})| \le \|T\||S\mathbf{x}| \le \|T\|\|S\||\mathbf{x}|.$$

Taking supremum of the above for $|\mathbf{x}| \le 1$ and applying Corollary 7.11 yield (7.8). □

Corollary 7.13. The identity function $I \in \mathcal{L}(\mathbb{F}^n)$ satisfies $\|I\| = 1$.

Proof. This follows directly from (7.7). □

Exercise 7.14. Verify that the space $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ becomes a metric space if we define the metric as $d(T, S) = \|T - S\|$.

Exercise 7.15. For $T \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^m)$, suppose $T\mathbf{e}_j \in \mathbb{R}^m$ for each standard basis vectors, i.e., $j = 1, \dots, n$. Prove that T carries $\mathbb{R}^n$ into $\mathbb{R}^m$ and $\|T\|$ is consistently defined in the sense that

$$\|T\| = \sup_{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x}| \le 1} |T\mathbf{x}| = \sup_{\mathbf{z} \in \mathbb{C}^n; |\mathbf{z}| \le 1} |T\mathbf{z}|. \tag{7.9}$$

Definition 7.16. The *Hilbert-Schmidt norm* of a linear map $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is the non-negative number

$$|T| := \left(\sum_{j=1}^{n} |T\mathbf{e}_j|^2 \right)^{\frac{1}{2}} \tag{7.10}$$

where $\mathbf{e}_1, \dots, \mathbf{e}_n$ is the standard basis in Definition B.32.

Exercise 7.17. Verify that (7.10) is indeed a norm in the sense of Definition B.114.

Corollary 7.18. The matrix A of $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ satisfies

$$|A| := \left(\sum_{i,j} |a_{ij}|^2 \right)^{\frac{1}{2}} \tag{7.11}$$

Proof. Since A is a linear map, we rewrite (7.10) as

$$|A|^2 = |A\mathbf{e}_1|^2 + |A\mathbf{e}_2|^2 + \cdots + |A\mathbf{e}_n|^2.$$

The proof is completed by the fact that $|\cdot|$ denotes the Euclidean 2-norm. □

Corollary 7.19. $\forall \mathbf{x} \in \mathbb{F}^n,\ |T\mathbf{x}| \le |T||\mathbf{x}|$.

Proof. We have

$$\begin{aligned} |T\mathbf{x}| &= \left| \sum_j x_j T\mathbf{e}_j \right| \le \sum_j |x_j| \, |T\mathbf{e}_j| \\ &\le \left(\sum_j |x_j|^2 \right)^{\frac{1}{2}} \left(\sum_j |T\mathbf{e}_j|^2 \right)^{\frac{1}{2}} = |T||\mathbf{x}|, \end{aligned}$$

where the first inequality follows from (NRM-3,4) in Definition B.114, the second inequality from the Cauchy-Schwarz inequality (B.43), and the last step from Definition 7.16. □

Corollary 7.20. The composition of two linear maps $S \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ and $T \in \mathcal{L}(\mathbb{F}^m, \mathbb{F}^k)$ satisfies

$$|TS| \le |T||S|. \tag{7.12}$$

Exercise 7.21. Prove Corollary 7.20.

Corollary 7.22. The identity function $I \in \mathcal{L}(\mathbb{F}^n)$ satisfies $|I| = \sqrt{n}$.

Proof. This follows directly from (7.10). □

Theorem 7.23. The operator norm and the Hilbert-Schmidt norm on $\mathcal{L}(\mathbb{F}^n)$ are related by

$$\|T\| \le |T| \le \sqrt{n} \|T\|. \tag{7.13}$$

Proof. Take supremum of Corollary 7.19, apply Corollary 7.11, and we have $\|T\| \le |T|$. $|T| \le \sqrt{n}\|T\|$ is given by

$$|T|^2 = \sum_j |T\mathbf{e}_j|^2 \le \sum_j \|T\|^2 |\mathbf{e}_j|^2 = n\|T\|^2,$$

where the inequality follows from Corollary 7.10. □

Corollary 7.24. Let d_1 and d_2 denote the metrics induced from the operator norm and the Hilbert-Schmidt norm, respectively. The identity map

$$I : (\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m), d_1) \to (\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m), d_2)$$

is uniformly continuous and has a uniformly continuous inverse.

Proof. This follows directly from Theorem 7.23 and Definition C.78. □

7.1.2 Matrix exponential

Definition 7.25. The *matrix exponential* e^A of a complex matrix $A \in \mathbb{C}^{n\times n}$ is

$$e^A := \sum_{N=0}^{\infty} \frac{1}{N!} A^N = I + A + \frac{1}{2!}A^2 + \cdots \tag{7.14}$$

Lemma 7.26. The series in (7.14) is *entry-by-entry* convergent.

Proof. Apply the operator norm in Definition 7.8 to (7.14) and we have

$$\left\|e^A\right\| = \left\|\sum_{N=0}^{\infty} \frac{1}{N!} A^N\right\| \le \sum_{N=0}^{\infty} \frac{1}{N!} \left\|A^N\right\| \le \sum_{N=0}^{\infty} \frac{1}{N!} \|A\|^N,$$

where the second inequality follows from Corollary 7.12. The ratio test in Theorem C.26 implies that the convergence of the last series in the operator norm, which, by Theorem 7.23, is equivalent to the convergence in the Hilbert-Schmidt norm, which, by (7.11), is equivalent to the entry-by-entry convergence. □

Theorem 7.27. Interpreted as a function $f : \mathbb{R}^{2n^2} \to \mathbb{R}^{2n^2}$, the matrix exponential $A \mapsto e^A$ is $\mathcal{C}^\infty$. In other words, every partial derivative of f to any order is entry-by-entry uniformly convergent to some continuous function.

Proof. Denote by E_j, $j = 1, \ldots, 2n^2$, an n-by-n complex matrix that has 1 or $\mathbf{i}$ in one entry and 0 in all other entries. By Definition C.112, the partial derivative of f in the direction of E_j is

$$\frac{\partial f}{\partial E_j}(A) = \frac{\mathrm{d}}{\mathrm{d}t} f(A + tE_j)\bigg|_{t=0}.$$

By (7.14) and Definition C.18, the sequence associated with the series $f(A)$ is $\left\{\frac{1}{N!}A^N\right\}$. Hence $\frac{\partial f}{\partial E_j}(A)$ is the sum of derivatives of all terms in the sequence; by the chain rule, each derivative is of the form

$$\frac{1}{N!}\sum_{i=1}^{N} g_1(A)\cdots g_{i-1}(A)\ \frac{\mathrm{d}}{\mathrm{d}t} g_i(A + tE_j)\bigg|_{t=0}\ g_{i+1}(A)\cdots g_N(A),$$

where each g_i is A. Taking further partial derivatives preserves this general form, except that g_i is either A or E_j and that the number of products to be summed up is increased.

The kth-order partial derivative of the Nth term $\frac{1}{N!}A^N$ in the sequence is a sum of N^k products, each product consisting of N terms and each term is either A or E_j satisfying

$$\max(\|A\|, \|E_j\|) \le M := \max(\|A\|, 1).$$

Hence we have, for any fixed $k \in \mathbb{N}$,

$$\left\|\frac{\partial^k}{\partial E_{j_1}\cdots\partial E_{j_k}}\left(\frac{1}{N!}A^N\right)\right\| \le \frac{N^k M^N}{N!}.$$

By the ratio test, the series $\frac{\partial^k f}{\partial E_{j_1}\cdots\partial E_{j_k}}(A)$ uniformly converges entry-by-entry to some function. The rest of the proof follows from Theorem C.101 and Lemma 7.26. □

Example 7.28. For the real skew-symmetric matrix

$$A = \begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix},$$

we have

$$e^A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Indeed, define

$$I_2 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad J_2 := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

and we have

$$A^{4n} = \theta^{4n} I_2,\ A^{4n+2} = -\theta^{4n+2} I_2,$$
$$A^{4n+1} = \theta^{4n+1} J_2,\ A^{4n+3} = -\theta^{4n+3} J_2.$$

It follows that

$$e^A = \cos\theta I_2 + \sin\theta J_2.$$

Lemma 7.29. If two matrices X and Y commute, then

$$e^X e^Y = e^{X+Y}. \tag{7.15}$$

Proof. By rearranging double summations, we have

$$\begin{aligned} e^X e^Y &= \left(\sum_{r=0}^{\infty}\frac{1}{r!}X^r\right)\left(\sum_{s=0}^{\infty}\frac{1}{s!}Y^s\right) = \sum_{r,s\in\mathbb{N}}\frac{1}{r!s!}X^r Y^s \\ &= \sum_{N=0}^{\infty}\sum_{k=0}^{N}\frac{X^k Y^{N-k}}{k!(N-k)!} = \sum_{N=0}^{\infty}\frac{1}{N!}\sum_{k=0}^{N}\binom{N}{k}X^k Y^{N-k} \\ &= \sum_{N=0}^{\infty}\frac{1}{N!}(X+Y)^N = e^{X+Y}, \end{aligned}$$

where the commutativity of X and Y ensures the validity of the last two steps. □

Example 7.30. If two matrices X and Y do not commute, then Lemma 7.29 does not hold, e.g.,

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},\ Y = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Corollary 7.31. The matrix e^X is nonsingular for any $X \in \mathbb{C}^{n\times n}$.

Proof. This follows from setting $Y = -X$ in (7.15) and taking determinant of both sides. □

Lemma 7.32. $\frac{\mathrm{d}}{\mathrm{d}t}(e^{tX}) = Xe^{tX}$.

Proof. By (7.14) and Theorem 7.27, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{tX}\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{N=0}^{\infty}\frac{1}{N!}(tX)^N = X\sum_{N=1}^{\infty}\frac{1}{(N-1)!}(tX)^{N-1}\\ &= Xe^{tX}.\end{aligned}$$

□

Lemma 7.33. For any nonsingular matrix W, we have

$$e^{W^{-1}XW} = W^{-1}e^XW. \tag{7.16}$$

Proof. By (7.14), we have

$$\begin{aligned}e^{W^{-1}XW} &= \sum_{N=0}^{\infty}\frac{1}{N!}(W^{-1}XW)^N = \sum_{N=0}^{\infty}\frac{1}{N!}W^{-1}X^NW\\ &= W^{-1}e^XW.\end{aligned}$$

□

Definition 7.34. Two matrices A and B are *similar* if there exists a nonsingular matrix S such that

$$B = S^{-1}AS, \tag{7.17}$$

and $S^{-1}AS$ is called a *similarity transformation* of A.

Lemma 7.35. Two similar matrices A and B have the same set of eigenvalues.

Proof. Let $(\lambda, \mathbf{u})$ be an eigen-pair of B, i.e.,

$$B\mathbf{u} = \lambda\mathbf{u}.$$

Combine it with (7.17), and we have

$$AS\mathbf{u} = SB\mathbf{u} = \lambda S\mathbf{u},$$

and thus λ is an eigenvalue of A with corresponding eigenvector $S\mathbf{u}$. □

Definition 7.36. $A \in \mathbb{C}^{m\times m}$ is *diagonalizable* if there exists a similarity transformation that maps A to a diagonal matrix Λ, i.e.,

$$\exists \text{ invertible } R \text{ s.t. } R^{-1}AR = \Lambda. \tag{7.18}$$

Corollary 7.37. For a diagonalizable matrix $A = R\Lambda R^{-1}$, we have

$$e^A = Re^\Lambda R^{-1}. \tag{7.19}$$

Proof. This follows directly from Lemma 7.33. □

Lemma 7.38. If $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $A \in \mathbb{C}^{n\times n}$, then $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are eigenvalues of e^A. Furthermore, if $\mathbf{u}$ is an eigenvector of A for λ_i, then $\mathbf{u}$ is an eigenvector of e^A for e^{λ_i}.

Proof. By the Schur Theorem B.134, there exist an invertible matrix P and an upper triangular matrix T such that

$$A = P^{-1}TP.$$

Then Lemma 7.33 yields

$$e^A = e^{P^{-1}TP} = P^{-1}e^TP,$$

where e^T, by Definition 7.25, is an upper triangular matrix with its diagonal entries as $e^{\lambda_1}, \ldots, e^{\lambda_n}$. If $\mathbf{u}$ is an eigenvector of A for the eigenvalue λ, then $\mathbf{u}$ is an eigenvector of A^n for the eigenvalue λ^n, the rest follows from Definition 7.25. □

Theorem 7.39. $\det e^X = e^{\text{Trace } X}$.

Exercise 7.40. Prove Theorem 7.39.

7.1.3 Lipschitz continuity

Definition 7.41. A function $\mathbf{f} : \mathbb{R}^N \times [0, +\infty) \to \mathbb{R}^N$ is *Lipschitz continuous* in its first variable over some domain

$$\mathcal{D} = \{(\mathbf{u}, t) : \|\mathbf{u} - \mathbf{u}_0\| \le a, t \in [0, T]\} \tag{7.20}$$

if

$$\exists L \ge 0 \text{ s.t. } \forall(\mathbf{u}, t), (\mathbf{v}, t) \in \mathcal{D},\ \ \|\mathbf{f}(\mathbf{u}, t) - \mathbf{f}(\mathbf{v}, t)\| \le L\|\mathbf{u} - \mathbf{v}\|. \tag{7.21}$$

Example 7.42. If $\mathbf{f}(\mathbf{u}, t) = \mathbf{f}(t)$, then $L = 0$.

Example 7.43. If $\mathbf{f} \notin \mathcal{C}^0$, then $\mathbf{f}$ is not Lipschitz continuous.

Definition 7.44. A subset $S \subset \mathbb{R}^n$ is *star-shaped* with respect to a point $p \in S$ if for each $x \in S$ the line segment from p to x lies in S.

Theorem 7.45. Let $S \subset \mathbb{R}^n$ be star-shaped with respect to $p = (p_1, p_2, \ldots, p_n) \in S$. For a continuously differentiable function $f : S \to \mathbb{R}^m$, there exist continuously differentiable functions $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_n(\mathbf{x})$ such that

$$f(\mathbf{x}) = f(p) + \sum_{j=1}^{n}(x_j - p_j)g_j(\mathbf{x}),\quad g_j(p) = \frac{\partial f}{\partial x_j}(p). \tag{7.22}$$

Proof. Since S is star-shaped, for any given $\mathbf{y} \in S$ and $t \in [0, 1]$, $f(\mathbf{x})$ is defined for $\mathbf{x} = p + t(\mathbf{y} - p)$. Then the chain rule yields

$$\frac{\mathrm{d}}{\mathrm{d}t}f(\mathbf{x}) = \sum_i \frac{\partial f}{\partial x_i}\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_i (y_i - p_i)\frac{\partial f}{\partial x_i}(\mathbf{x}).$$

An integration with respect to t from 0 to 1 leads to

$$\begin{aligned}f(\mathbf{y}) - f(p) &= \sum_i (y_i - p_i)g_i(\mathbf{y}),\\ g_i(\mathbf{y}) &= \int_0^1 \frac{\partial f}{\partial x_i}(p + t(\mathbf{y} - p))\mathrm{d}t,\end{aligned}$$

where the function $g_i(p) = \frac{\partial f}{\partial x_i}(p)$. □

Lemma 7.46. If $\mathbf{f}(\mathbf{u},t):\mathcal{D}\to\mathbb{R}^m$ is continuously differentiable on some compact convex set $\mathcal{D}\subseteq\mathbb{R}^{n+1}$, then $\mathbf{f}$ is Lipschitz continuous in $\mathbf{u}$ on $\mathcal{D}$.

Proof. For a fixed t, the matrix form of Theorem 7.45 yields

$$f(\mathbf{u},t)-f(\mathbf{v},t)=G(\mathbf{u},\mathbf{v})(\mathbf{u}-\mathbf{v}).$$

Take the Euclidean 2-norm and we have

$$|f(\mathbf{u},t)-f(\mathbf{v},t)|=|G(\mathbf{u},\mathbf{v})(\mathbf{u}-\mathbf{v})|\le|G(\mathbf{u},\mathbf{v})||\mathbf{u}-\mathbf{v}|,$$

where the last step follows from Corollary 7.19. Each entry in G is a continuous function $\mathcal{D}\times\mathcal{D}\to\mathbb{R}$ and is bounded since the compactness of $\mathcal{D}$ implies that $\mathcal{D}\times\mathcal{D}$ is compact. Hence $|G(\mathbf{u},\mathbf{v})|$ is bounded and this completes the proof. □

7.1.4 Existence and uniqueness of solution

Lemma 7.47. Let (M,ρ) denote a complete metric space and $\phi:M\to M$ a contractive mapping in the sense that

$$\exists c\in[0,1)\text{ s.t. }\forall\eta,\zeta\in M,\ \rho(\phi(\eta),\phi(\zeta))\le c\rho(\eta,\zeta).\quad(7.23)$$

Then there exists a unique $\xi\in M$ such that $\phi(\xi)=\xi$.

Theorem 7.48 (Fundamental theorem of ODEs)**.** If $\mathbf{f}(\mathbf{u}(t),t)$ is Lipschitz continuous in $\mathbf{u}$ and continuous in t over some region $\mathcal{D}=\{(\mathbf{u},t):\|\mathbf{u}-\mathbf{u}_0\|\le a,t\in[0,T]\}$, then there is a unique solution to the IVP as in Definition 7.4 at least up to time $T^*=\min(T,\frac{a}{S})$ where $S=\max_{(\mathbf{u},t)\in\mathcal{D}}\|\mathbf{f}(\mathbf{u},t)\|$.

Proof. It suffices to prove the case of $a=+\infty$ since the minimum ensures that the solution $\mathbf{u}(t)$ remains in the domain $\mathcal{D}$ where the Lipschitz continuity holds.

Let (M,ρ) denote the complete metric space of continuous functions $\mathbf{u}:[0,T]\to\mathbb{R}^N$ such that $\mathbf{u}(0)=\mathbf{u}_0$. The metric is defined by

$$\rho(\mathbf{u},\mathbf{v})=\sup_{t\in[0,T]}\exp(-Kt)\|\mathbf{u}(t)-\mathbf{v}(t)\|,$$

where $K>L$.

For a given $\mathbf{u}\in M$, define $\boldsymbol{\phi}(\mathbf{u})$ as the solution $\mathbf{U}$ on $[0,T]$ to the IVP in Definition 7.4, which is solvable by integration as

$$\boldsymbol{\phi}(\mathbf{u})(t)=\mathbf{u}_0+\int_0^t\mathbf{f}(\mathbf{u}(s),s)\mathrm{d}s.$$

$\boldsymbol{\phi}$ is a contractive mapping because $\forall\mathbf{u},\mathbf{v}\in M$,

$$\begin{aligned}
&\rho(\boldsymbol{\phi}(\mathbf{u}),\boldsymbol{\phi}(\mathbf{v}))\\
=&\sup_{t\in[0,T]}\exp(-Kt)\left\|\int_0^t\big(\mathbf{f}(\mathbf{u}(s),s)-\mathbf{f}(\mathbf{v}(s),s)\big)\mathrm{d}s\right\|\\
\le&\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\|\mathbf{f}(\mathbf{u}(s),s)-\mathbf{f}(\mathbf{v}(s),s)\|\,\mathrm{d}s\\
\le&L\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\|\mathbf{u}(s)-\mathbf{v}(s)\|\,\mathrm{d}s\\
\le&L\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\exp(-Ks)\,\|\mathbf{u}(s)-\mathbf{v}(s)\|\exp(Ks)\mathrm{d}s\\
\le&L\rho(\mathbf{u},\mathbf{v})\sup_{t\in[0,T]}\exp(-Kt)\int_0^t\exp(Ks)\mathrm{d}s\\
\le&\frac{L}{K}\rho(\mathbf{u},\mathbf{v}).
\end{aligned}$$

The rest follows from Lemma 7.47. □

Theorem 7.49. If $\mathbf{f}(\mathbf{u},t)$ is Lipschitz continuous in $\mathbf{u}$ and continuous in t on $\mathcal{D}=\{(\mathbf{u},t):\mathbf{u}\in\mathbb{R}^N,t\in[0,T]\}$, then the IVP in Definition 7.4 is well-posed for all initial data.

Proof. Theorem 7.48 has already established the existence and uniqueness of the solution of the IVP. It remains to prove that the solution is overly sensitive neither to the initial condition nor to the RHS $f(u,t)$. To this end, we consider two IVPs

$$\begin{cases}v'=f(v,t)\\v(0)=v_0\end{cases}\quad\text{and}\quad\begin{cases}w'=f(w,t)+\delta(t)\\w(0)=w_0\end{cases}$$

where v_0, w_0 and $\delta(t)$ satisfy the conditions

$$(*):\ \begin{cases}\forall t\in[0,T]\ \ \|\delta(t)\|<\epsilon\\\|v_0-w_0\|<\epsilon\end{cases},$$

with ϵ being a small positive constant. Obviously, the two IVPs are respectively solved by

$$\begin{aligned}v(t)&=v_0+\int_0^t f(v(s),s)\mathrm{d}s,\\w(t)&=w_0+\int_0^t[f(w(s),s)+\delta(s)]\mathrm{d}s.\end{aligned}$$

Thus we have

$$v(t)-w(t)=v_0-w_0+\int_0^t[f(v(s),s)-f(w(s),s)-\delta(s)]\mathrm{d}s.$$

Take the 2-norm on both sides and we have

$$\begin{aligned}
&\|v(t)-w(t)\| \\
=&\|v_0-w_0+\int_0^t [f(v(s),s)-f(w(s),s)-\delta(s)]\mathrm{d}s\| \\
\le&\|v_0-w_0\|+\left\|\int_0^t [f(v(s),s)-f(w(s),s)-\delta(s)]\mathrm{d}s\right\| \\
\le&\|v_0-w_0\|+\int_0^t \|f(v(s),s)-f(w(s),s)-\delta(s)\|\mathrm{d}s \\
\le&\|v_0-w_0\|+\int_0^t \|f(v(s),s)-f(w(s),s)\|+\|\delta(s)\|\mathrm{d}s \\
\le&\epsilon+\int_0^t L\|v(s)-w(s)\|\mathrm{d}s+\int_0^t \epsilon\mathrm{d}s \\
=&(1+t)\epsilon+\int_0^t L\|v(s)-w(s)\|\mathrm{d}s,
\end{aligned}$$

where the fifth step follows from condition $(*)$. To proceed with

$$(**):\ \|v(t)-w(t)\|\le(1+t)\epsilon+\int_0^t L\|v(s)-w(s)\|\mathrm{d}s,$$

we define a function $h:[0,T]\to\mathbb{R}$ as

$$h(s)=e^{-sL}\int_0^s L\|v(r)-w(r)\|\mathrm{d}r$$

and the derivative $h'(s)$ is

$$\begin{aligned}
&-Le^{-sL}\int_0^s L\|v(r)-w(r)\|\mathrm{d}r+e^{-sL}L\|v(s)-w(s)\| \\
=&Le^{-sL}\left(\|v(s)-w(s)\|-\int_0^s L\|v(r)-w(r)\|\mathrm{d}r\right),
\end{aligned}$$

which implies

$$\begin{aligned}
&e^{-tL}\int_0^t L\|v(r)-w(r)\|\mathrm{d}r \\
=&h(t)=h(t)-h(0)=\int_0^t h'(s)\mathrm{d}s \\
=&\int_0^t Le^{-sL}\left(\|v(s)-w(s)\|-\int_0^s L\|v(r)-w(r)\|\mathrm{d}r\right)\mathrm{d}s \\
\le&\int_0^t Le^{-sL}\left((1+s)\epsilon+\int_0^s L\|v(r)-w(r)\|\mathrm{d}r\right)\mathrm{d}s \\
&-\int_0^t Le^{-sL}\int_0^s L\|v(r)-w(r)\|\mathrm{d}r\mathrm{d}s \\
=&\int_0^t Le^{-sL}(1+s)\epsilon\mathrm{d}s \\
\Longrightarrow&\int_0^t L\|v(r)-w(r)\|\mathrm{d}r\le\int_0^t Le^{L(t-s)}(1+s)\epsilon\mathrm{d}s.
\end{aligned}$$

where the first step follows from the definition of $h(t)$, the second from $h(0)=0$ and the fifth from $(*)$. Substitute the above inequality to $(**)$ and we have

$$\begin{aligned}
\|v(t)-w(t)\|&\le(1+t)\epsilon+\int_0^t Le^{L(t-s)}(1+s)\epsilon\mathrm{d}s \\
&\le\left(1+T+TLe^{LT}(1+T)\right)\epsilon=:C\epsilon.
\end{aligned}$$

□

Example 7.50. Consider $N=1$, $u'(t)=\sqrt{u(t)}$, $u(0)=0$.

$$\lim_{u\to0}f'(u)=\lim_{u\to0}\frac{1}{2\sqrt{u}}=+\infty.$$

Hence $f(u)$ is not Lipschitz continuous near $u=0$. However, $u(t)\equiv0$ and $u(t)=\frac{1}{4}t^2$ are both solutions. Hence the Lipschitz condition is not necessary for existence.

Example 7.51. Consider the IVP $u'(t)=u^2$, $u_0=\eta>0$. The slope $f'(u)=2u$ goes to $+\infty$ as $u\to+\infty$. So there is no unique solution on $[0,+\infty)$, but there might exist T^* such that unique solutions are guaranteed on $[0,T^*]$.

In fact, $u(t)=\frac{1}{\eta^{-1}-t}$ is a solution, but blows up at $t=1/\eta$. According to Theorem 7.48, $f(u)=u^2$ and we would like to maximize a/S. Since $S=\max_{\mathcal{D}}|f(u)|=(\eta+a)^2$, it is equivalent to find $\min_{a>0}(\eta+a)^2/a$:

$$(\eta+a)^2/a=2\eta+a+\eta^2\frac{1}{a}\ge2\eta+2\sqrt{\eta^2}=4\eta.$$

Hence $T^*=\frac{1}{4\eta}$. The estimation of T^* in Theorem 7.48 is thus quite conservative for this case.

Example 7.52. For the simple pendulum in Example 7.3, we have

$$|\sin\theta-\sin\theta^*|\le|\theta-\theta^*|\le\|\mathbf{u}-\mathbf{u}^*\|_\infty$$

since $\cos\theta^*\le1$. In addition, we have $|\omega-\omega^*|\le\|\mathbf{u}-\mathbf{u}^*\|_\infty$.

$$\begin{aligned}
&\|\mathbf{f}(\mathbf{u})-\mathbf{f}(\mathbf{u}^*)\|_\infty=\max\left(|\omega-\omega^*|,\frac{g}{\ell}|\sin\theta-\sin\theta^*|\right) \\
\le&\max(\frac{g}{\ell},1)\|\mathbf{u}-\mathbf{u}^*\|_\infty.
\end{aligned}$$

Therefore, $\mathbf{f}$ is Lipschitz continuous with $L=\max(g/\ell,1)$.

Exercise 7.53. Show that the solution of IVP is not overly sensitive to the initial condition of problems that satisfy a Lipschitz condition. In other words, if both $\mathbf{v}$ and $\mathbf{w}$ satisfy the *same* IVP with initial condition $\mathbf{v}(a)=\mathbf{v}_0$ and $\mathbf{w}(a)=\mathbf{w}_0$, then we have

$$\|\mathbf{v}(t)-\mathbf{w}(t)\|\le\|\mathbf{v}_0-\mathbf{w}_0\|\exp(L(t-a)).$$

7.1.5 Linear IVPs with constant matrices

Theorem 7.54 (Duhamel's principle)**.** For a linear IVP

$$\mathbf{u}'(t)=A\mathbf{u}+\mathbf{g}(t) \tag{7.24}$$

with a time-independent matrix A, the solution is

$$\mathbf{u}(t)=e^{tA}\mathbf{u}_0+\int_0^t e^{(t-\tau)A}\mathbf{g}(\tau)\mathrm{d}\tau. \tag{7.25}$$

Proof. This solution follows from Lemma 7.32 and Leibniz's formula

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\int_{a(x)}^{b(x)}f(x,y)\mathrm{d}y=&\int_a^b\frac{\partial}{\partial x}f(x,y)\mathrm{d}y-f(x,a)\frac{\mathrm{d}a}{\mathrm{d}x} \\
&+f(x,b)\frac{\mathrm{d}b}{\mathrm{d}x}.
\end{aligned}$$

□

Example 7.55. Many linear problems are naturally formulated in the form of a single high-order ODE

$$v^{(m)}(t) - \sum_{j=1}^{m} c_j v^{(m-j)}(t) = \phi(t). \tag{7.26}$$

By setting $u_j(t) = v^{(j-1)}(t)$ and $\mathbf{u} = [u_1, u_2, \ldots, u_m]^T$, we can rewrite (7.26) as

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{g}(t), \tag{7.27}$$

where $\mathbf{g}(t) = [0, \ldots, 0, \phi(t)]^T$ and

$$a_{ij} = \begin{cases} 1 & \text{if } i = j-1, \\ c_{m+1-j} & \text{if } i = m, \\ 0 & \text{otherwise.} \end{cases}$$

If A is diagonalizable, say $A = X^{-1}\Lambda X$, then we can define $\mathbf{v} = X\mathbf{u}$ and rewrite (7.24) as

$$\mathbf{v}'(t) = \Lambda\mathbf{v} + X\mathbf{g}(t).$$

Thus the linear system can be decoupled into a number of scalar IVPs, each of which has its solution from Theorem 7.54.

Example 7.56. The matrix of the linear IVP system

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \tag{7.28}$$

is not diagonalizable. For v we have $v(t) = v(0)e^{\lambda t}$. For u we consider the form of the solution

$$u = (E + Ft)e^{\lambda t};$$

then (7.28) yields $E = u(0)$ and $F = v(0)$, i.e.,

$$u(t) = u(0)e^{\lambda t} + tv(t).$$

7.1.6 Jordan canonical form

Theorem 7.57 (Factorization of a polynomial over $\mathbb{C}$)**.** If $p \in \mathcal{P}(\mathbb{C})$ is a nonconstant polynomial, then p has a unique factorization (except for the order of the factors) of the form

$$p(z) = c(z - \lambda_1) \cdots (z - \lambda_m), \tag{7.29}$$

where $c, \lambda_1, \ldots, \lambda_m \in \mathbb{C}$.

Definition 7.58. Let $A \in \mathbb{C}^{m\times m}$, then the *characteristic polynomial* of A is

$$p_A(z) = \det(zI - A). \tag{7.30}$$

Lemma 7.59. Let $A \in \mathbb{C}^{m\times m}$, then λ is an eigenvalue of A iff λ is a root of the characteristic polynomial of A.

Exercise 7.60. Show that

$$p_M(z) = z^s + \sum_{j=0}^{s-1} \alpha_j z^j$$

is the characteristic polynomial of

$$M = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{s-2} & -\alpha_{s-1} \end{bmatrix} \in \mathbb{C}^{s\times s}. \tag{7.31}$$

Definition 7.61. If the characteristic polynomial $p_A(z)$ has a factor $(z-\lambda)^n$, then λ is said to have *algebraic multiplicity* $m_a(\lambda) = n$.

Definition 7.62. Let λ be an eigenvalue of $A \in \mathbb{C}^{m\times m}$, the *eigenspace* of A corresponding to λ is

$$\begin{aligned} \mathcal{N}(A - \lambda I) &= \{\mathbf{u} \in \mathbb{C}^m : (A - \lambda I)\mathbf{u} = \mathbf{0}\} \\ &= \{\mathbf{u} \in \mathbb{C}^m : A\mathbf{u} = \lambda\mathbf{u}\}. \end{aligned} \tag{7.32}$$

The dimension of $\mathcal{N}(A - \lambda I)$ is the *geometric multiplicity* $m_g(\lambda)$ of λ.

Lemma 7.63. Geometric multiplicity and algebraic multiplicity satisfy

$$1 \le m_g(\lambda) \le m_a(\lambda). \tag{7.33}$$

Definition 7.64. An eigenvalue λ of A is *defective* if

$$m_g(\lambda) < m_a(\lambda). \tag{7.34}$$

A is *defective* if A has one or more defective eigenvalues.

Lemma 7.65. A nondefective matrix A is diagonalizable, i.e.,

$$\exists \text{ nonsingular } R \text{ s.t. } R^{-1}AR = \Lambda \text{ is diagonal.} \tag{7.35}$$

Definition 7.66. A *Jordan block* of order k has the form

$$J(\lambda, k) = \lambda I_k + S_k, \tag{7.36}$$

where

$$(S_k)_{i,j} = \begin{cases} 1, & i = j-1, \\ 0, & \text{otherwise.} \end{cases}$$

Example 7.67. The Jordan blocks of orders 1, 2, and 3 are

$$J(\lambda, 1) = \lambda, \quad J(\lambda, 2) = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \quad J(\lambda, 3) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}.$$

Theorem 7.68 (Jordan canonical form)**.** Every square matrix A can be expressed as

$$A = RJR^{-1}, \tag{7.37}$$

where R is invertible and J is a block diagonal matrix of the form

$$J = \begin{bmatrix} J(\lambda_1, k_1) & & & \\ & J(\lambda_2, k_2) & & \\ & & \ddots & \\ & & & J(\lambda_s, k_s) \end{bmatrix}. \tag{7.38}$$

Each $J(\lambda_i, k_i)$ is a Jordan block of some order k_i and $\sum_{i=1}^{s} k_i = m$. If λ is an eigenvalue of A with algebraic multiplicity m_a and geometric multiplicity m_g, then λ appears in m_g blocks and the sum of the orders of these blocks is m_a.

7.2 Basic numerical methods

Notation 4. To numerically solve the IVP (7.3), we are given initial data $\mathbf{U}^0 = \mathbf{u}_0$, and want to compute approximations $\mathbf{U}^1, \mathbf{U}^2, \ldots$ such that

$$\mathbf{U}^n \approx \mathbf{u}(t_n),$$

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

Definition 7.69. The *(forward) Euler's method* solves the IVP (7.3) by

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^n, t_n), \tag{7.39}$$

which is based on replacing $\mathbf{u}'(t_n)$ with the forward difference $(\mathbf{U}^{n+1} - \mathbf{U}^n)/k$ and $\mathbf{u}(t_n)$ with $\mathbf{U}^n$ in $\mathbf{f}(\mathbf{u}, t)$.

Definition 7.70. The *backward Euler's method* solves the IVP (7.3) by

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^{n+1}, t_{n+1}), \tag{7.40}$$

which is based on replacing $\mathbf{u}'(t_{n+1})$ with the backward difference $(\mathbf{U}^{n+1} - \mathbf{U}^n)/k$ and $\mathbf{u}(t_{n+1})$ with $\mathbf{U}^{n+1}$ in $\mathbf{f}(\mathbf{u}, t)$.

Definition 7.71. The *trapezoidal method* is

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \frac{k}{2}\left(\mathbf{f}(\mathbf{U}^n, t_n) + \mathbf{f}(\mathbf{U}^{n+1}, t_{n+1})\right). \tag{7.41}$$

Definition 7.72. The *midpoint (or leapfrog) method* is

$$\mathbf{U}^{n+1} = \mathbf{U}^{n-1} + 2k\mathbf{f}(\mathbf{U}^n, t_n). \tag{7.42}$$

Example 7.73. Applying Euler's method (7.39) with step size $k = 0.2$ to solve the IVP

$$u'(t) = u, \quad u(0) = 1, \quad t \in [0, 1],$$

yields the following table:

n	U^n	$kf(U^n, t_n)$
0	1	0.2
1	1.2	$0.2 \times 1.2 = 0.24$
2	1.44	$0.2 \times 1.44 = 0.288$
3	1.728	$0.2 \times 1.728 = 0.3456$
4	2.0736	$0.2 \times 2.0736 = 0.41472$
5	2.48832	

7.2.1 Truncation errors and solution errors

Definition 7.74. The *local truncation error* (LTE) is the error caused by replacing continuous derivatives with finite difference formulas.

Example 7.75. The LTE of the leapfrog method is

$$\begin{aligned}\boldsymbol{\tau}^n &= \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1})}{2k} - \mathbf{f}(\mathbf{u}(t_n), t_n) \\ &= \left[\mathbf{u}'(t_n) + \frac{1}{6}k^2\mathbf{u}'''(t_n) + O(k^4)\right] - \mathbf{u}'(t_n) \\ &= \frac{1}{6}k^2\mathbf{u}'''(t_n) + O(k^4).\end{aligned}$$

Definition 7.76. For a numerical method of the form

$$\mathbf{U}^{n+1} = \boldsymbol{\phi}(\mathbf{U}^{n+1}, \mathbf{U}^n, \ldots, \mathbf{U}^{n-m}),$$

the *one-step error* is defined by

$$\mathcal{L}^n := \mathbf{u}(t_{n+1}) - \boldsymbol{\phi}(\mathbf{u}(t_{n+1}), \mathbf{u}(t_n), \ldots, \mathbf{u}(t_{n-m})). \tag{7.43}$$

In other words, $\mathcal{L}^n$ is the error that would be introduced in one time step if the past values $\mathbf{U}^n, \mathbf{U}^{n-1}, \ldots$ were all taken to be the exact values from $\mathbf{u}(t)$.

Example 7.77. The one-step error of the leapfrog method is

$$\begin{aligned}\mathcal{L}^n &= \mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1}) - 2k\mathbf{f}(\mathbf{u}(t_n), t_n) \\ &= \frac{1}{3}k^3\mathbf{u}'''(t_n) + O(k^5) \\ &= 2k\boldsymbol{\tau}^n.\end{aligned}$$

Definition 7.78. The *solution error* of a numerical method for solving the IVP in Definition 7.4 is

$$\mathbf{E}^N := \mathbf{U}^{T/k} - \mathbf{u}(T); \qquad \mathbf{E}^n = \mathbf{U}^n - \mathbf{u}(t_n). \tag{7.44}$$

7.2.2 Convergence of Euler's method

Definition 7.79. A numerical method is *convergent* iff its application to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\forall T > 0, \qquad \lim_{\substack{k \to 0 \\ Nk = T}} \mathbf{U}^N = \mathbf{u}(T). \tag{7.45}$$

Lemma 7.80. Consider a linear IVP (7.46) of the form

$$\begin{cases} \mathbf{u}'(t) = \lambda\mathbf{u}(t) + \mathbf{g}(t), \\ \mathbf{u}(0) = \mathbf{u}_0, \end{cases} \tag{7.46}$$

where λ is either a scalar or a diagonal matrix. The solution error and the LTE of Euler's method satisfy

$$\mathbf{E}^{n+1} = (1 + k\lambda)\mathbf{E}^n - k\boldsymbol{\tau}^n. \tag{7.47}$$

Proof. By Definition 7.74, we have

$$\begin{aligned}\boldsymbol{\tau}^n &= \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)}{k} - \mathbf{u}'(t_n) \\ &= \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)}{k} - (\lambda\mathbf{u}(t_n) + \mathbf{g}(t_n)),\end{aligned}$$

and therefore

$$\mathbf{u}(t_{n+1}) = (1 + k\lambda)\mathbf{u}(t_n) + k\mathbf{g}(t_n) + k\boldsymbol{\tau}^n.$$

Euler's method applied to the linear IVP (7.46) reads

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k(\lambda\mathbf{U}^n + \mathbf{g}(t_n)) = (1 + k\lambda)\mathbf{U}^n + k\mathbf{g}(t_n).$$

Subtracting the above two equations yields (7.47). □

Lemma 7.81. For the linear IVP (7.46), the solution errors and the LTEs of Euler's method satisfy

$$\mathbf{E}^n = (1 + k\lambda)^n\mathbf{E}^0 - k\sum_{m=1}^{n}(1 + k\lambda)^{n-m}\boldsymbol{\tau}^{m-1}. \tag{7.48}$$

Proof. We proceed by induction on n.

The induction basis holds because of (7.47). Suppose (7.48) holds for all integers no greater than n. Then for $n+1$, we have

$$\begin{aligned}\mathbf{E}^{n+1} &= (1+k\lambda)\mathbf{E}^n - k\boldsymbol{\tau}^n \\ &= (1+k\lambda)^{n+1}\mathbf{E}^0 - k\sum_{m=1}^{n+1}(1+k\lambda)^{n+1-m}\boldsymbol{\tau}^{m-1},\end{aligned}$$

where the first equality follows from (7.47) and the second from the induction hypothesis. □

Theorem 7.82. Euler's method is convergent for solving the linear IVP (7.46).

Proof. We have

$$|1+k\lambda| \le 1+|k\lambda| \le e^{k|\lambda|},$$

and hence for $m < n \le T/k$

$$(1+k\lambda)^{n-m} \le e^{(n-m)k|\lambda|} \le e^{nk|\lambda|} \le e^{|\lambda|T},$$

then Lemma 7.81 yields

$$\begin{aligned}\|\mathbf{E}^n\| &\le e^{|\lambda|T}\left(\|\mathbf{E}^0\| + k\sum_{m=1}^{n}\|\boldsymbol{\tau}^{m-1}\|\right) \\ &\le e^{|\lambda|T}\left(\|\mathbf{E}^0\| + nk\max_{1\le m\le n}\|\boldsymbol{\tau}^{m-1}\|\right).\end{aligned}$$

The LTE of Euler's method is

$$\boldsymbol{\tau}^n = \frac{\mathbf{u}(t_{n+1})-\mathbf{u}(t_n)}{k} - \mathbf{u}'(t_n) = \frac{1}{2}k\mathbf{u}''(t_n) + O(k^2),$$

hence

$$\|\mathbf{E}^N\| \le e^{|\lambda|T}(\|\mathbf{E}^0\| + TO(k)) = O(k),$$

where we have assumed that $\|\mathbf{E}^0\| = O(k)$. □

Lemma 7.83. Consider a nonlinear IVP of the form

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t), t),$$

where $\mathbf{f}(\mathbf{u}, t)$ is continuous in t and is Lipschitz continuous in $\mathbf{u}$ with L as the Lipschitz constant. Euler's method satisfies

$$\|\mathbf{E}^{n+1}\| \le (1+kL)\|\mathbf{E}^n\| + k\|\boldsymbol{\tau}^n\|. \tag{7.49}$$

Proof. The definition of the LTE yields

$$\boldsymbol{\tau}^n = \frac{\mathbf{u}(t_{n+1})-\mathbf{u}(t_n)}{k} - \mathbf{f}(\mathbf{u}(t_n), t_n),$$

and hence

$$\mathbf{u}(t_{n+1}) = \mathbf{u}(t_n) + k\mathbf{f}(\mathbf{u}(t_n), t_n) + k\boldsymbol{\tau}^n,$$

the Euler's method is

$$\mathbf{U}^{n+1} = \mathbf{U}^n + k\mathbf{f}(\mathbf{U}^n, t_n),$$

subtracting the above two equations gives

$$\mathbf{E}^{n+1} = \mathbf{E}^n + k\left(\mathbf{f}(\mathbf{U}^n, t_n) - \mathbf{f}(\mathbf{u}(t_n), t_n)\right) - k\boldsymbol{\tau}^n,$$

the triangle inequality and Lipschitz continuity of $\mathbf{f}$ yield

$$\begin{aligned}\|\mathbf{E}^{n+1}\| &\le \|\mathbf{E}^n\| + k\|\mathbf{f}(\mathbf{U}^n, t_n) - \mathbf{f}(\mathbf{u}(t_n), t_n)\| + k\|\boldsymbol{\tau}^n\| \\ &\le (1+kL)\|\mathbf{E}^n\| + k\|\boldsymbol{\tau}^n\|.\end{aligned}$$

□

Theorem 7.84. For the nonlinear IVP in Lemma 7.83, Euler's method is convergent.

Proof. Follow the procedures of linear IVPs to show that

$$\|\mathbf{E}^N\| \le e^{LT}T\|\boldsymbol{\tau}\| = O(k) \text{ as } k \to 0.$$

□

7.2.3 Zero stability and absolute stability

Definition 7.85. A numerical method is *stable* or *zero-stable* iff its application to any IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t yields

$$\forall T > 0, \qquad \lim_{\substack{k\to 0 \\ Nk=T}} \|\mathbf{U}^N\| < \infty. \tag{7.50}$$

Example 7.86. Consider the scalar IVP

$$u'(t) = \lambda(u - \cos t) - \sin t,$$

with $\lambda = -2100$ and $u(0) = 1$. The exact solution is clearly

$$u(t) = \cos t.$$

The following table shows the error at time $T = 2$ when Euler's method is used with various values of k.

k	$E(T)$
2.00e-4	1.98e-8
4.00e-4	3.96e-8
8.00e-4	7.92e-8
9.50e-4	3.21e-7
9.76e-4	5.88e+35
1.00e-3	1.45e+76

The first three lines confirm the first-order accuracy of Euler's method, but something dramatic happens between $k =$ 9.76e-4 and $k =$ 9.50e-4. What's going on?

Definition 7.87. The Euler's method

$$U^{n+1} = (1+k\lambda)U^n$$

for solving the scalar IVP

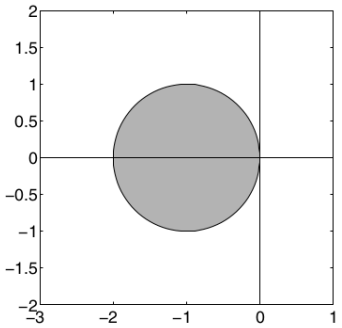
$$u'(t) = \lambda u(t) \tag{7.51}$$

is *absolutely stable* or has *eigenvalue stability* if

$$|1+k\lambda| \le 1. \tag{7.52}$$

Definition 7.88. The *region of absolute stability* for Euler's method applied to (7.51) is the set of all points

$$\{z \in \mathbb{C} : |1+z| \le 1\}. \tag{7.53}$$

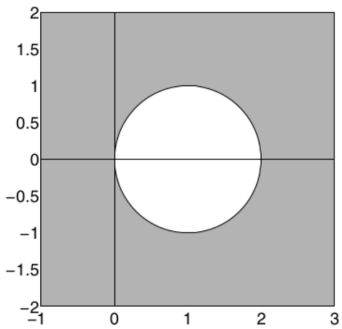


Example 7.89. The backward Euler's method applied to (7.51) reads

$$U^{n+1} = U^n + k\lambda U^{n+1} \Rightarrow U^{n+1} = \frac{1}{1-k\lambda}U^n.$$

Hence the region of absolute stability for backward Euler's method is

$$\{z \in \mathbb{C} : |1-z| \geq 1\}. \tag{7.54}$$



Lemma 7.90. Consider an autonomous, homogeneous, and linear system of IVPs

$$\mathbf{u}'(t) = A\mathbf{u} \tag{7.55}$$

where $\mathbf{u} \in \mathbb{R}^N$, $N > 1$, and A is diagonalizable with eigenvalues as λ_i's. Euler's method is absolutely stable if each $z_i := k\lambda_i$ is within the stability region (7.53).

Proof. Applying Euler's method to (7.55) gives

$$\mathbf{U}^{n+1} = \mathbf{U}^n + kA\mathbf{U}^n = (I + kA)\mathbf{U}^n.$$

Since A is diagonalizable, we have $AR = R\Lambda$ where R contains the eigenvectors of A that span $\mathbb{R}^N$. Then

$$R^{-1}\mathbf{U}^{n+1} = R^{-1}(I + kA)RR^{-1}\mathbf{U}^n.$$

Set $\mathbf{V} := R^{-1}\mathbf{U}$ and we have

$$\mathbf{V}^{n+1} = (I + k\Lambda)\mathbf{V}^n.$$

After advancing $\mathbf{V}^0$ to $\mathbf{V}^n$, we use $\mathbf{U}^n = R\mathbf{V}^n$ to recover the solution of (7.55). □

Definition 7.91. The *law of mass action* states that the rate of a chemical reaction is proportional to the product of the concentration of the reacting substances, with each concentration raised to a power equal to the coefficient that occurs in the reaction.

Example 7.92. For the reaction

$$\alpha\mathrm{A} + \beta\mathrm{B} \underset{\mathrm{k_-}}{\overset{\mathrm{k_+}}{\rightleftharpoons}} \sigma\mathrm{S} + \tau\mathrm{T},$$

the forward reaction rate is $k_+[A]^\alpha[B]^\beta$ and the backward reaction rate is $k_-[S]^\sigma[T]^\tau$.

Example 7.93. Consider

$$\mathrm{A} + \mathrm{B} \underset{\mathrm{c_2}}{\overset{\mathrm{c_1}}{\rightleftharpoons}} \mathrm{AB}$$

with $c_1, c_2 > 0$. Let

$$\mathbf{u} := \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} [A] \\ [B] \\ [AB] \end{bmatrix}.$$

Then we have

$$\begin{aligned} u_1' &= -c_1u_1u_2 + c_2u_3; \\ u_2' &= -c_1u_1u_2 + c_2u_3; \\ u_3' &= c_1u_1u_2 - c_2u_3, \end{aligned}$$

which can be written more compactly as

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}).$$

Let $\mathbf{v}(t) := \mathbf{u}(t) - \bar{\mathbf{u}}$ with $\bar{\mathbf{u}}$ independent of time. Then

$$\begin{aligned} \mathbf{v}'(t) = \mathbf{u}'(t) &= \mathbf{f}(\mathbf{u}(t)) = \mathbf{f}(\mathbf{v} + \bar{\mathbf{u}}) \\ &= \mathbf{f}(\bar{\mathbf{u}}) + \mathbf{f}'(\bar{\mathbf{u}})\mathbf{v}(t) + O(\|\mathbf{v}\|^2), \end{aligned}$$

and hence

$$\mathbf{v}'(t) = A\mathbf{v}(t) + \mathbf{b},$$

where $A = \mathbf{f}'(\bar{\mathbf{u}})$ is the Jacobian matrix, i.e.,

$$A = \begin{bmatrix} -c_1u_2 & -c_1u_1 & c_2 \\ -c_1u_2 & -c_1u_1 & c_2 \\ c_1u_2 & c_1u_1 & -c_2 \end{bmatrix},$$

with eigenvalues $\lambda_1 = -c_1(u_1 + u_2) - c_2$ and $\lambda_2 = \lambda_3 = 0$. As the total concentration of species A and B, the time-dependent quantity $u_1 + u_2 \in \mathbb{R}^+$ is bounded by the constant $M_0 := u_1(0) + u_2(0) + 2u_3(0)$. Therefore, the condition of absolute stability for Euler's method

$$|1 + \lambda_1 k| \leq 1$$

can always be satisfied by setting $k < \frac{2}{c_1M_0 + c_2}$, c.f. $\lambda_1 < 0$.

Example 7.94. For the reaction

$$\mathrm{A} \underset{\mathrm{c_2}}{\overset{\mathrm{c_1}}{\rightleftharpoons}} \mathrm{B},$$

we obtain the linear IVPs

$$\begin{cases} u_1' = -c_1u_1 + c_2u_2; \\ u_2' = c_1u_1 - c_2u_2. \end{cases}$$

Formula 7.95. A general way of reducing an IVP

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}, t)$$

to a collection of scalar, linear model problems of the form

$$w_i'(t) = \lambda_i w_i(t), \quad i = 1, 2, \cdots, n$$

consists of steps as follows.

(a) Linearization: at the neighborhood of a particular solution $\mathbf{u}^*(t)$, we write

$$\mathbf{u}(t) = \mathbf{u}^*(t) + (\delta\mathbf{u})(t)$$

and apply Taylor expansion

$$\mathbf{f}(\mathbf{u}, t) = \mathbf{f}(\mathbf{u}^*, t) + J(t)\delta\mathbf{u} + o(\|\delta\mathbf{u}\|)$$

to obtain

$$(\delta\mathbf{u})'(t) = J(t)(\delta\mathbf{u}).$$

(b) Freezing coefficients: set

$$A = J(t^*),$$

where t^* is the particular time of interest.

(c) Diagonalization: assume A is diagonalizable by V and we write

$$(\delta\mathbf{u})'(t) = V(V^{-1}AV)V^{-1}(\delta\mathbf{u}).$$

Define $\mathbf{w} := V^{-1}(\delta\mathbf{u})$ and we have a collection of decoupled scalar IVPs,

$$\mathbf{w}'(t) = \Lambda\mathbf{w}(t),$$

where $\Lambda = V^{-1}AV$ is a diagonal matrix.

7.3 Linear multistep methods

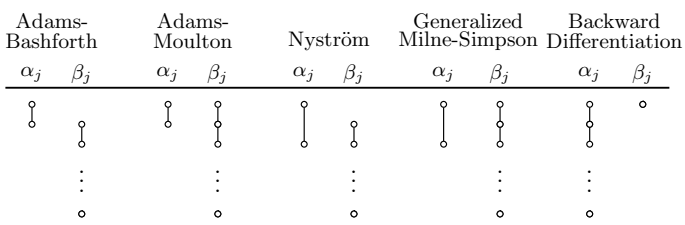
Definition 7.96. For solving the IVP (7.3), an *s-step linear multistep method* (LMM) has the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = k\sum_{j=0}^{s} \beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{7.56}$$

where $\alpha_s = 1$ is assumed WLOG.

Definition 7.97. An LMM (7.56) is *explicit* if $\beta_s = 0$; otherwise it is *implicit*.

7.3.1 Classical formulas



Definition 7.98. An *Adams formula* is an LMM of the form

$$\mathbf{U}^{n+s} = \mathbf{U}^{n+s-1} + k\sum_{j=0}^{s} \beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{7.57}$$

where β_j's are chosen to maximize the order of accuracy.

Definition 7.99. An *Adams-Bashforth formula* is an Adams formula with $\beta_s = 0$. An *Adams-Moulton formula* is an Adams formula with $\beta_s \neq 0$.

Example 7.100. Euler's method is the 1-step Adams-Bashforth formula with

$$s = 1,\ \alpha_1 = 1,\ \alpha_0 = -1,\ \beta_1 = 0,\ \beta_0 = 1.$$

Example 7.101. The trapezoidal method is a 1-step Adams-Moulton formula with

$$s = 1,\ \alpha_1 = 1,\ \alpha_0 = -1,\ \beta_1 = \beta_0 = \frac{1}{2}.$$

Another 1-step Adams-Moulton formula is the backward Euler's method.

Definition 7.102. A *Nyström formula* is an LMM of the form

$$\mathbf{U}^{n+s} = \mathbf{U}^{n+s-2} + k\sum_{j=0}^{s-1} \beta_j \mathbf{f}(\mathbf{U}^{n+j}, t_{n+j}), \tag{7.58}$$

where β_j's are chosen to give order s.

Example 7.103. The midpoint method is the 2-step Nyström formula with

$$s = 2,\ \alpha_2 = 1,\ \alpha_1 = 0,\ \alpha_0 = -1,\ \beta_1 = 2,\ \beta_0 = 0.$$

Definition 7.104. A *backward differentiation formula* (BDF) is an LMM of the form

$$\sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = k\beta_s \mathbf{f}(\mathbf{U}^{n+s}, t_{n+s}), \tag{7.59}$$

where α_j's are chosen to give order s.

Example 7.105. The backward Euler's method is the 1-step BDF with

$$s = 1,\ \alpha_1 = \beta_1 = 1,\ \alpha_0 = -1.$$

7.3.2 Consistency and accuracy

Definition 7.106. The *characteristic polynomials* or *generating polynomials* for the LMM (7.56) are

$$\rho(\zeta) = \sum_{j=0}^{s} \alpha_j \zeta^j; \qquad \sigma(\zeta) = \sum_{j=0}^{s} \beta_j \zeta^j. \tag{7.60}$$

Example 7.107. The forward Euler's method (7.39) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = 1, \tag{7.61}$$

while the backward Euler's method (7.40) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = \zeta. \tag{7.62}$$

Example 7.108. The trapezoidal method (7.41) has

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = \frac{1}{2}(\zeta + 1), \tag{7.63}$$

and the midpoint method (7.42) has

$$\rho(\zeta) = \zeta^2 - 1, \qquad \sigma(\zeta) = 2\zeta. \tag{7.64}$$

Notation 5. Denote by Z a *time shift operator* that acts on both discrete functions according to

$$Z\mathbf{U}^n = \mathbf{U}^{n+1} \tag{7.65}$$

and on continuous functions according to

$$Z\mathbf{u}(t) = \mathbf{u}(t+k). \tag{7.66}$$

Definition 7.109. The *difference operator of an LMM* is an operator $\mathcal{L}$ on the linear space of continuously differentiable functions given by

$$\mathcal{L} = \rho(Z) - k\mathcal{D}\sigma(Z), \tag{7.67}$$

where $\mathcal{D}\mathbf{u}(t_n) = \mathbf{u}_t(t_n) := \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}(t_n)$, Z is the time shift operator, and ρ, σ are characteristic polynomials for the LMM.

Lemma 7.110. The one-step error of the LMM (7.56) is

$$\mathcal{L}\mathbf{u}(t_n) = C_0\mathbf{u}(t_n) + C_1 k\mathbf{u}_t(t_n) + C_2 k^2\mathbf{u}_{tt}(t_n) + \cdots, \tag{7.68}$$

where

$$\begin{aligned} C_0 &= \textstyle\sum_{j=0}^s \alpha_j \\ C_1 &= \textstyle\sum_{j=0}^s (j\alpha_j - \beta_j) \\ C_2 &= \textstyle\sum_{j=0}^s \left(\frac{1}{2}j^2\alpha_j - j\beta_j\right) \\ &\vdots \\ C_q &= \textstyle\sum_{j=0}^s \left(\frac{1}{q!}j^q\alpha_j - \frac{1}{(q-1)!}j^{q-1}\beta_j\right). \end{aligned} \tag{7.69}$$

Proof. By definition of the one-step error (7.43), we have

$$\begin{aligned} \mathcal{L}\mathbf{u}(t_n) &= \sum_{j=0}^s \alpha_j \mathbf{u}(t_{n+j}) - k\sum_{j=0}^s \beta_j \mathbf{f}(\mathbf{u}(t_{n+j}), t_{n+j}) \\ &= \sum_{j=0}^s \alpha_j \mathbf{u}(t_{n+j}) - k\sum_{j=0}^s \beta_j \mathbf{u}'(t_{n+j}). \end{aligned}$$

Taylor's theorem yields

$$\begin{aligned} \mathbf{u}(t_{n+j}) &= \mathbf{u}(t_n) + jk\mathbf{u}'(t_n) + \frac{1}{2}(jk)^2\mathbf{u}''(t_n) + \cdots \\ \mathbf{u}'(t_{n+j}) &= \mathbf{u}'(t_n) + jk\mathbf{u}''(t_n) + \frac{1}{2}(jk)^2\mathbf{u}'''(t_n) + \cdots \end{aligned}$$

Substitution of the above into $\mathcal{L}\mathbf{u}(t_n)$ yields (7.68). □

Definition 7.111. An LMM has *order of accuracy* p if

$$\forall \mathbf{u} \in \mathcal{C}^{p+1}, \qquad \mathcal{L}\mathbf{u}(t_n) = \Theta(k^{p+1}) \text{ as } k \to 0, \tag{7.70}$$

i.e., if in (7.69) we have $C_0 = C_1 = \cdots = C_p = 0$ and $C_{p+1} \neq 0$. Then C_{p+1} is called the *error constant*.

Definition 7.112. An LMM is *preconsistent* if $\rho(1) = 0$, i.e. $\sum_{i=0}^s \alpha_i = 0$ or $\sum_{i=0}^{s-1} \alpha_i = -1$.

Definition 7.113. An LMM is *consistent* if it has order of accuracy $p \geq 1$.

Example 7.114. For Euler's method, the coefficients C_j's in (7.69) can be computed directly from Example 7.100 as $C_0 = C_1 = 0, C_2 = \frac{1}{2}, C_3 = \frac{1}{6}$.

Exercise 7.115. Compute the first five coefficients C_j's of the trapezoidal rule and the midpoint rule from Examples 7.101 and 7.103.

Example 7.116. A necessary condition for $\|\mathbf{E}^n\| = O(k)$ is $\|\mathcal{L}\mathbf{u}(t_n)\| = O(k^2)$ since there are $\frac{T}{k}$ time steps, and hence the first two terms in (7.68) should be zero, i.e.,

$$\sum_{j=0}^s \alpha_j = 0, \qquad \sum_{j=0}^s j\alpha_j = \sum_{j=0}^s \beta_j, \tag{7.71}$$

which is equivalent to

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1). \tag{7.72}$$

Second-order accuracy further requires

$$\frac{1}{2}\sum_{j=0}^s j^2\alpha_j = \sum_{j=0}^s j\beta_j.$$

In general, pth-order accuracy requires (7.71) and

$$\forall q = 2, \ldots, p, \quad \sum_{j=0}^s \frac{1}{q!}j^q\alpha_j = \sum_{j=0}^s \frac{1}{(q-1)!}j^{q-1}\beta_j. \tag{7.73}$$

Exercise 7.117. Express conditions of $\mathcal{L} = O(k^3)$ using characteristic polynomials.

Exercise 7.118. Derive coefficients of LMMs shown below by the method of undetermined coefficients and a programming language with symbolic computation such as `Matlab`.

Adams-Bashforth formulas in Definition 7.99

s	p	β_s	β_{s-1}	β_{s-2}	β_{s-3}	β_{s-4}
1	1	0	1			
2	2	0	$\frac{3}{2}$	$-\frac{1}{2}$		
3	3	0	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	
4	4	0	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$

Adams-Moulton formulas in Definition 7.99

s	p	β_s	β_{s-1}	β_{s-2}	β_{s-3}	β_{s-4}
1	1	1				
1	2	$\frac{1}{2}$	$\frac{1}{2}$			
2	3	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		
3	4	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	
4	5	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

Backward differentiation formulas in Definition 7.104

s	p	α_s	α_{s-1}	α_{s-2}	α_{s-3}	α_{s-4}	β_s
1	1	1	-1				1
2	2	1	$-\frac{4}{3}$	$\frac{1}{3}$			$\frac{2}{3}$
3	3	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$		$\frac{6}{11}$
4	4	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	$\frac{12}{25}$

Example 7.119. To derive coefficients of the 2nd-order Adams-Bashforth formula, we interpolate $\mathbf{f}(t)$ by a linear polynomial

$$q(t) = \mathbf{f}^{n+1} - \frac{t_{n+1} - t}{k}(\mathbf{f}^{n+1} - \mathbf{f}^n)$$

and then calculate

$$\mathbf{U}^{n+2} - \mathbf{U}^{n+1} = \int_{t_{n+1}}^{t_{n+2}} q(t)\mathrm{d}t = \frac{3}{2}k\mathbf{f}^{n+1} - \frac{1}{2}k\mathbf{f}^{n}.$$

Lemma 7.120. An LMM with $\sigma(1) \neq 0$ has order of accuracy p if and only if

$$\frac{\rho(e^{\kappa})}{\sigma(e^{\kappa})} = \kappa + \Theta(\kappa^{p+1}) \qquad \text{as } \kappa \to 0. \tag{7.74}$$

where $\kappa := k\mathcal{D}$.

Proof. By Taylor's theorem,

$$\mathbf{u}(t_{n+1}) = \mathbf{u}(t_n) + k\mathbf{u}_t(t_n) + \frac{1}{2}k^2\mathbf{u}_{tt}(t_n) + \cdots$$

By Notation 5, we also have $\mathbf{u}(t_{n+1}) = Z\mathbf{u}(t_n)$. A comparison of the two equalities yields

$$Z = 1 + (k\mathcal{D}) + \frac{1}{2!}(k\mathcal{D})^2 + \cdots + \frac{1}{n!}(k\mathcal{D})^n + \cdots = e^{k\mathcal{D}},$$

where the last step follows from Definition 7.25. Set $\kappa = k\mathcal{D}$ and we have from Definition 7.111,

$$\mathcal{L} = \rho(e^{\kappa}) - \kappa\sigma(e^{\kappa}) = \Theta(k^{p+1}),$$

Since $\sigma(e^{\kappa})$ is an analytic function of k and is nonzero at $k = 0$, we divide it on both sides to get (7.74). □

Theorem 7.121. An LMM with $\sigma(1) \neq 0$ has order of accuracy p if and only if

$$\begin{aligned} \tfrac{\rho(z)}{\sigma(z)} &= \log z + \Theta\left((z-1)^{p+1}\right) \\ &= \left[(z-1) - \tfrac{1}{2}(z-1)^2 + \tfrac{1}{3}(z-1)^3 - \cdots\right] \\ &\quad +\Theta((z-1)^{p+1}). \end{aligned} \tag{7.75}$$

as $z \to 1$.

Proof. We only prove the case of scalar IVPs. To get from (7.74) to the first equality, we make the change of variables $z = e^{\kappa}, \kappa = \log z$, noting that $\Theta(\kappa^{p+1})$ as $\kappa \to 0$ has the same meaning as $\Theta\left((z-1)^{p+1}\right)$ as $z \to 1$ since $e^{\kappa} = 1$ and $\mathrm{d}(e^{\kappa})/\mathrm{d}\kappa \neq 0$ at $\kappa = 0$. The second equality is just the usual Taylor series for $\log z$ at 1. □

Example 7.122. The trapezoidal rule has $\rho(z) = z - 1$ and $\sigma(z) = \frac{1}{2}(z+1)$. A comparison of (7.75) with the expansion

$$\frac{\rho(z)}{\sigma(z)} = \frac{z-1}{\frac{1}{2}(z+1)} = (z-1)\left[1 - \frac{z-1}{2} + \frac{(z-1)^2}{4} - \cdots\right]$$

confirms that the trapezoidal rule has order 2 with error constant $-\frac{1}{12}$.

Exercise 7.123. For the third-order BDF in Definition 7.104 and Exercise 7.118, derive its characteristic polynomials and apply Theorem 7.121 to verify that the order of accuracy is indeed 3.

Exercise 7.124. Prove that an s-step LMM has order of accuracy p if and only if, when applied to an ODE $u_t = q(t)$, it gives exact results whenever q is a polynomial of degree $< p$, but not whenever q is a polynomial of degree p. Assume arbitrary continuous initial data u_0 and exact numerical initial data $v^0, \cdots, v^{s-1}$.

7.3.3 Zero stability

Example 7.125 (A consistent but unstable LMM)**.** The LMM

$$\mathbf{U}^{n+2} - 3\mathbf{U}^{n+1} + 2\mathbf{U}^{n} = -k\mathbf{f}(\mathbf{U}^n, t_n) \tag{7.76}$$

has a one-step error given by

$$\begin{aligned} \mathcal{L}\mathbf{u}(t_n) &= \mathbf{u}(t_{n+2}) - 3\mathbf{u}(t_{n+1}) + 2\mathbf{u}(t_n) + k\mathbf{u}'(t_n) \\ &= \frac{1}{2}k^2\mathbf{u}''(t_n) + O(k^3), \end{aligned}$$

so the method is consistent with first-order accuracy. But the solution error may not exhibit first order accuracy, or even convergence. Consider the trivial IVP

$$u'(t) = 0, \qquad u(0) = 0,$$

with solution $u(t) \equiv 0$. The LMM (7.76) reads in this case

$$U^{n+2} = 3U^{n+1} - 2U^n \Rightarrow U^{n+2} - U^{n+1} = 2(U^{n+1} - U^n),$$

and therefore

$$U^n = 2U^0 - U^1 + 2^n(U^1 - U^0).$$

If we take $U^0 = 0$ and $U^1 = k$, then

$$U^n = k(2^n - 1) = k(2^{T/k} - 1) \to +\infty \text{ as } k \to 0.$$

Definition 7.126. An s-step LMM is *stable* or *zero-stable* if all solutions $\{\mathbf{U}^n\}$ of the recurrence

$$\rho(Z)\mathbf{U}^n = \sum_{j=0}^{s} \alpha_j \mathbf{U}^{n+j} = \mathbf{0} \tag{7.77}$$

are bounded as $n \to +\infty$.

Theorem 7.127. An LMM is zero-stable if and only if all the roots of $\rho(z)$ satisfy $|z| \leq 1$, and any root with $|z| = 1$ is simple.

Proof. WLOG, we only prove the case of scalar IVPs; see Hairer et al. [1993] for the vector case. For a scalar IVP, we write (7.77) as $U^{n+s} + \sum_{j=0}^{s-1} \alpha_j U^{n+j} = 0$, and this s-step recurrence formula can be expressed as a one-step matrix operation

$$\mathbf{V}^{n+1} = M\mathbf{V}^n,$$

where M is the *companion matrix* (7.31) and

$$\mathbf{V}^n = (U^n, U^{n+1}, \cdots, U^{n+s-1})^T.$$

Hence

$$\mathbf{V}^n = M^n\mathbf{V}^0.$$

By Exercise 7.60, the characteristic polynomial of M is $\rho(z)$, i.e., $p_M(z) = \rho(z)$. Therefore the set of eigenvalues of M is the same as the set of roots of ρ, and these eigenvalues determine how the powers M^n behave asymptotically as $n \to +\infty$. The scalar sequence $\{U^n\}_{n=0}^{+\infty}$ is bounded as $n \to +\infty$ if and only if the vector sequence $\{\mathbf{V}^n\}$ is bounded,

and $\{\mathbf{V}^n\}$ is bounded if and only if all elements of M^n is bounded. Since $\|\mathbf{V}^n\| \le \|M^n\|\|\mathbf{V}^0\|$, the zero-stability is now equivalent to the power-boundedness of M.

By Theorem 7.68, we have

$$M = RJR^{-1} \Rightarrow M^n = RJ^nR^{-1}.$$

Therefore M^n's growth or boundedness is determined by the boundedness of

$$J_i^n = \begin{bmatrix} \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \binom{n}{2}\lambda_i^{n-2} & \cdots & \binom{n}{m_i-1}\lambda_i^{n-m_i+1} \\ & \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \cdots & \binom{n}{m_i-2}\lambda_i^{n-m_i+2} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} \\ & & & & \lambda_i^n \end{bmatrix},$$

which follows from $J_i^n = (\lambda_i I + \eta)^n$ where η is the nilpotent matrix with $\eta^{m_i} = \mathbf{0}$,

$$\eta_{ij} = \begin{cases} 1 & \text{if } j - i = 1; \\ 0 & \text{otherwise.} \end{cases}$$

By Definition 7.62, the dimension of the eigenspace of the companion matrix M is 1 for each eigenvalue of M because the upper-right $(s-1)\times(s-1)$ block of $zI - M$ is nonsingular for any $z \in \mathbb{C}$. Hence the geometric multiplicity $m_g(\lambda)$ is 1 for any eigenvalue λ of M. By Theorem 7.68, there is exactly one Jordan block for each eigenvalue of M.

As $n \to \infty$, the nonzero elements of J_i^n approach 0 if $|\lambda_i| < 1$ and ∞ if $|\lambda_i| > 1$. For $|\lambda_i| = 1$, they are bounded in the case of a 1×1 block, but unbounded if $m_i \ge 2$. □

7.3.4 Linear difference equations

Definition 7.128. A *system of linear difference equations* is a set of equations of the form

$$X_n = A_n X_{n-1} + \phi_n, \tag{7.78}$$

where $n, s \in \mathbb{N}^+$, $X_n \in \mathbb{C}^s$, $\phi_n \in \mathbb{C}^s$, and $A_n \in \mathbb{C}^{s\times s}$. With the initial vector X_0 specified, the system of linear difference equations becomes an initial value problem. The system is *homogeneous* if $\phi_n = \mathbf{0}$.

Example 7.129. A linear difference equation of the form

$$y_n = \alpha_{n1}y_{n-1} + \alpha_{n2}y_{n-2} + \cdots + \alpha_{ns}y_{n-s} + \psi_n$$

can be easily recast in the form (7.78) by writing

$$A_n = \begin{bmatrix} \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{n,s-1} & \alpha_{ns} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, X_{n-1} = \begin{bmatrix} y_{n-1} \\ y_{n-2} \\ \vdots \\ y_{n-s} \end{bmatrix},$$

$$\phi_n = \begin{bmatrix} \psi_n & 0 & 0 & \cdots & 0 \end{bmatrix}^T.$$

Theorem 7.130. The problem (7.78) with initial value X_0 has the unique solution

$$X_n = \left(\prod_{i=1}^n A_i\right) X_0 + \left(\prod_{i=2}^n A_i\right)\phi_1 + \left(\prod_{i=3}^n A_i\right)\phi_2 + \cdots + A_n\phi_{n-1} + \phi_n, \tag{7.79}$$

where

$$\prod_{i=m}^n A_i = \begin{cases} A_nA_{n-1}\cdots A_{m+1}A_m & \text{if } m \le n; \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Proof. For $n = 1$, (7.79) reduces to (7.78). The rest of the proof is a straightforward induction. □

Theorem 7.131. Let θ_n be the solution to the homogeneous linear difference equation

$$\theta_{n+s} + \sum_{i=0}^{s-1}\alpha_i\theta_{n+i} = 0 \tag{7.80}$$

with constant coefficients α_i's and the initial values

$$\begin{bmatrix} \theta_0 \\ \theta_{-1} \\ \vdots \\ \theta_{-s+2} \\ \theta_{-s+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{7.81}$$

Then the inhomogeneous equation

$$y_{n+s} + \sum_{i=0}^{s-1}\alpha_i y_{n+i} = \psi_{n+s} \tag{7.82}$$

with initial values $y_0, y_1, \cdots, y_{s-1}$ is uniquely solved by

$$y_n = \sum_{i=0}^{s-1}\theta_{n-i}\tilde{y}_i + \sum_{i=s}^{n}\theta_{n-i}\psi_i \tag{7.83}$$

where

$$\begin{bmatrix} \tilde{y}_{s-1} \\ \tilde{y}_{s-2} \\ \tilde{y}_{s-3} \\ \vdots \\ \tilde{y}_1 \\ \tilde{y}_0 \end{bmatrix} = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_{s-2} & \theta_{s-1} \\ 0 & 1 & \theta_1 & \cdots & \theta_{s-3} & \theta_{s-2} \\ 0 & 0 & 1 & \cdots & \theta_{s-4} & \theta_{s-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \theta_1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} y_{s-1} \\ y_{s-2} \\ y_{s-3} \\ \vdots \\ y_1 \\ y_0 \end{bmatrix}. \tag{7.84}$$

Exercise 7.132. Prove Theorem 7.131.

7.3.5 Convergence

Definition 7.133. Given initial values

$$\forall i = 0, 1, \ldots, s-1, \quad \mathbf{U}^i = \phi^i(\mathbf{u}(0), k)$$

satisfying

$$\forall i = 0, 1, \ldots, s-1, \quad \lim_{k\to0}\|\phi^i(\mathbf{u}(0), k) - \mathbf{u}(0)\| = 0, \tag{7.85}$$

an LMM is *convergent* if it yields

$$\lim_{\substack{k\to 0\\ Nk=T}} \mathbf{U}^N = \mathbf{u}(T) \tag{7.86}$$

for *any* fixed $T > 0$ and *any* IVP with $\mathbf{f}(\mathbf{u}, t)$ Lipschitz continuous in $\mathbf{u}$ and continuous in t.

Lemma 7.134. A convergent LMM is zero-stable.

Proof. Apply the convergent LMM to the trivial IVP

$$u'(t) = 0; \qquad u(0) = 0 \tag{7.87}$$

and we have (7.77). Suppose, by contradiction, that the characteristic polynomial $\rho(Z)$ has a root ζ_1 with $|\zeta_1| > 1$ or a multiple root ζ_2 with $|\zeta_2| = 1$. In the first case,

$$\sum_{j=0}^{s} \alpha_j \zeta_1^j = 0 \;\Rightarrow\; \zeta_1^m \sum_{j=0}^{s} \alpha_j \zeta_1^j = 0 \;\Rightarrow\; \sum_{j=0}^{s} \alpha_j Z^j \zeta_1^m = 0,$$

i.e., $U^n = \zeta_1^n$ is a solution of (7.77). In the second case, we have $\rho(\zeta_2)\zeta_2^m = 0$ for any $m \in \mathbb{N}^+$. In addition, define

$$\chi(z) := (\rho(z)z^m)' = \left(\sum_{i=0}^{s} \alpha_i z^{m+i}\right)' = \sum_{i=1}^{s} \alpha_i (m+i) z^{m+i-1}$$

and we know from $\chi(\zeta_2) = 0$ that

$$\forall n \in \mathbb{N},\ V^n = n\zeta_2^{n-1}$$

is a solution of (7.77). For (7.87) and any fixed $T > 0$,

$$U_k(T) = \sqrt{k}\zeta_1^{T/k}, \qquad V_k(T) = \frac{T}{\sqrt{k}}\zeta_2^{T/k-1}$$

satisfy condition (7.85) for initial values, but diverge as $k \to 0$, i.e. $n \to \infty$. This contradicts the convergence of the LMM and completes the proof. □

Lemma 7.135. A convergent LMM is preconsistent.

Proof. By (7.86) and the continuity of $\mathbf{u}$ in time, we have

$$\lim_{k\to 0} U^N = \lim_{k\to 0} U^{N-1} = \cdots = \lim_{k\to 0} U^{N-s} = \mathbf{u}(T),$$

where $N = T/k$. Substituting this equation into the limit of the LMM equation (7.56) yields preconsistency as in Definition 7.112. □

Lemma 7.136. A convergent LMM is consistent.

Exercise 7.137. Prove Lemma 7.136.

Lemma 7.138. For an autonomous IVP, the one-step error of a consistent LMM satisfies

$$\|\mathcal{L}\mathbf{u}(t_n)\| \le \sum_{j=0}^{s-1} \Big(\frac{1}{2}(s-j)^2|\alpha_j| + (s-j)|\beta_j|\Big) LMk^2, \tag{7.88}$$

where L is the Lipschitz constant, and M is an upper bound of $\|\mathbf{f}(\mathbf{u}(t))\|$ on $t \in [0, T]$.

Proof. By definition of the one-step error (7.43), we have

$$\begin{aligned}
\mathcal{L}\mathbf{u}(t_n) =& \sum_{j=0}^{s} \alpha_j \mathbf{u}(t_{n+j}) - k\sum_{j=0}^{s} \beta_j \mathbf{u}'(t_{n+j}) \\
=& \sum_{j=0}^{s-1} \alpha_j \mathbf{u}(t_{n+j}) - \sum_{j=0}^{s-1} \alpha_j \mathbf{u}(t_{n+s}) \\
& - k\sum_{j=0}^{s-1} \Big((j-s)\alpha_j - \beta_j\Big)\mathbf{u}'(t_{n+s}) - k\sum_{j=0}^{s-1} \beta_j \mathbf{u}'(t_{n+j}) \\
=& \sum_{j=0}^{s-1} \alpha_j \Big(\mathbf{u}(t_{n+j}) - \mathbf{u}(t_{n+s}) - (j-s)k\mathbf{u}'(t_{n+s})\Big) \\
& + k\sum_{j=0}^{s-1} \beta_j \Big(\mathbf{u}'(t_{n+s}) - \mathbf{u}'(t_{n+j})\Big),
\end{aligned}$$

where the second step follows from the consistency condition (7.71), i.e.,

$$\begin{aligned}
&\alpha_s = -\textstyle\sum_{j=0}^{s-1} \alpha_j, \\
&\beta_s = \textstyle\sum_{j=0}^{s} j\alpha_j - \sum_{j=0}^{s-1} \beta_j = \sum_{j=0}^{s-1} \Big((j-s)\alpha_j - \beta_j\Big).
\end{aligned}$$

Taylor expansions yield the identity

$$\begin{aligned}
&\mathbf{u}(t_{n+j}) - \mathbf{u}(t_{n+s}) - (j-s)k\mathbf{u}'(t_{n+s}) \\
=& k\int_{s-j}^{0} [\mathbf{f}(\mathbf{u}(t_{n+s} - \xi k)) - \mathbf{f}(\mathbf{u}(t_{n+s}))]\, \mathrm{d}\xi,
\end{aligned}$$

which, together with the Lipschitz condition, implies

$$\begin{aligned}
&\|\mathbf{u}(t_{n+j}) - \mathbf{u}(t_{n+s}) - (j-s)k\mathbf{u}'(t_{n+s})\| \\
\le& kL\int_{0}^{s-j} \|\mathbf{u}(t_{n+s} - \xi k) - \mathbf{u}(t_{n+s})\|\, \mathrm{d}\xi \\
\le& \frac{1}{2}(s-j)^2 k^2 LM,
\end{aligned}$$

where the second step follows from the mean value theorem and the condition of M being an upper bound of $\|\mathbf{f}(\mathbf{u}(t))\|$. Similarly, we have

$$\|\mathbf{f}(\mathbf{u}(t_{n+s})) - \mathbf{f}(\mathbf{u}(t_{n+j}))\| \le LM(s-j)k.$$

Take a norm of $\mathcal{L}\mathbf{u}(t_n)$, apply the above two inequalities, and we have (7.88). □

Lemma 7.139. For an autonomous IVP, the solution errors of a consistent LMM with $k < k_0$ and $k_0|\beta_s|L < 1$ satisfy

$$\left\|\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i}\right\| \le Ck \max_{i=0}^{s-1} \|\mathbf{E}^{n+i}\| + Dk^2, \tag{7.89}$$

where both C and D are positive constants.

Proof. By definitions of the LMM, its one-step errors, and its solution errors, we have

$$\begin{aligned}
&\mathbf{E}^{n+s} + \sum_{i=0}^{s-1} \alpha_i \mathbf{E}^{n+i} \\
=& \mathbf{U}^{n+s} - \mathbf{u}(t_{n+s}) + \sum_{i=0}^{s-1} \alpha_i (\mathbf{U}^{n+i} - \mathbf{u}(t_{n+i})) \\
=& k\sum_{i=0}^{s} \beta_i (\mathbf{f}(\mathbf{U}^{n+i}) - \mathbf{f}(\mathbf{u}(t_{n+i}))) - \mathcal{L}\mathbf{u}(t_n),
\end{aligned}$$

which yields

$$\begin{aligned}
&\left\|\mathbf{E}^{n+s}+\sum_{i=0}^{s-1}\alpha_i\mathbf{E}^{n+i}\right\| \\
\le&\|\mathcal{L}\mathbf{u}(t_n)\|+k|\beta_s|\left\|\mathbf{f}(\mathbf{U}^{n+s})-\mathbf{f}(\mathbf{u}(t_{n+s}))\right\| \\
&+k\sum_{i=0}^{s-1}|\beta_i|\left\|\mathbf{f}(\mathbf{U}^{n+i})-\mathbf{f}(\mathbf{u}(t_{n+i}))\right\| \\
\le&\|\mathcal{L}\mathbf{u}(t_n)\|+kL|\beta_s|\left\|\mathbf{E}^{n+s}\right\|+kL\sum_{i=0}^{s-1}|\beta_i|\left\|\mathbf{E}^{n+i}\right\| \\
\le&\|\mathcal{L}\mathbf{u}(t_n)\|+kL|\beta_s|\left\|\mathbf{E}^{n+s}+\sum_{i=0}^{s-1}\alpha_i\mathbf{E}^{n+i}\right\| \\
&+kL\sum_{i=0}^{s-1}|\alpha_i\beta_s|\left\|\mathbf{E}^{n+i}\right\|+kL\sum_{i=0}^{s-1}|\beta_i|\left\|\mathbf{E}^{n+i}\right\|.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&(1-kL|\beta_s|)\left\|\mathbf{E}^{n+s}+\sum_{i=0}^{s-1}\alpha_i\mathbf{E}^{n+i}\right\| \\
\le&kL\sum_{i=0}^{s-1}(|\alpha_i\beta_s|+|\beta_i|)\left\|\mathbf{E}^{n+i}\right\|+\|\mathcal{L}\mathbf{u}(t_n)\|.
\end{aligned}$$

For any $k<k_0<\frac{1}{|\beta_s L|}$, dividing both sides by $(1-kL|\beta_s|)$ and applying Lemma 7.138 yield (7.89). □

Theorem 7.140. An LMM is convergent if and only if it is consistent and stable.

Proof. We only prove the sufficiency since the necessity has been stated in Lemmas 7.134 and 7.136. By Lemma 7.139, we have

$$\mathbf{E}^{n+s}=-\sum_{i=0}^{s-1}\alpha_i\mathbf{E}^{n+i}+\psi_{n+s},$$

where $\|\psi_n\|\le Ck\max_{i=1}^{s}\|\mathbf{E}^{n-i}\|+Dk^2$ for any k sufficiently small. Then the zero-stability of the LMM and Theorem 7.131 imply the existence of bounded constants θ_i's such that

$$\mathbf{E}^n=\sum_{i=0}^{s-1}\theta_{n-i}\widetilde{\mathbf{E}^i}+\sum_{i=s}^{n}\theta_{n-i}\psi_i,$$

where $\widetilde{\mathbf{E}^i}'$s are linear combinations of $\mathbf{E}^j$'s for $i,j=0,1,\ldots,s-1$; see (7.84). Note that, in order to apply Theorem 7.131, we have shifted $\mathbf{E}^{n+i}$ to $\mathbf{E}^{n+i-s}$. It follows that

$$\|\mathbf{E}^n\|\le\theta_m\sum_{i=0}^{s-1}\left\|\widetilde{\mathbf{E}^i}\right\|+\theta_m Cks\sum_{i=s}^{n-1}\|\mathbf{E}^i\|+\theta_m D(n-s+1)k^2,$$

where $\theta_m=\sup_{i=1}^{n}|\theta_i|$ and the factor s of the second summation is introduced to account for the fact that a local maximum value of $\|\mathbf{E}^{n-i}\|$ may appear in at most s adjacent terms. Define a sequence (v_i) as

$$\begin{cases}
v_0=\theta_m\sum_{i=0}^{s-1}\left\|\widetilde{\mathbf{E}^i}\right\|; \\
v_1=\theta_m Dk^2+v_0; \\
\quad\cdots \\
v_n=\theta_m Cks\sum_{i=1}^{n-1}v_i+n\theta_m Dk^2+v_0,
\end{cases}$$

where $\lim_{k\to0}v_0=0$ because Definition 7.133 implies $\lim_{k\to0}\left\|\widetilde{\mathbf{E}^i}\right\|=0$ for each $i=0,1,\ldots,s-1$. It is straightforward to show that, for $n>1$,

$$v_n+\frac{Dk}{Cs}=(1+\theta_m Cks)\left(v_{n-1}+\frac{Dk}{Cs}\right),$$

which implies

$$\begin{aligned}
v_n&=-\tfrac{Dk}{Cs}+(1+\theta_m Cks)^{n-1}\left(v_1+\tfrac{Dk}{Cs}\right) \\
&=(1+\theta_m Cks)^{n-1}v_0+[(1+\theta_m Cks)^n-1]\tfrac{Dk}{Cs} \\
&<\exp(\theta_m Csnk)v_0+[\exp(\theta_m Csnk)-1]\tfrac{Dk}{Cs}.
\end{aligned}$$

For $n=T/k$, we have $\lim_{k\to0}v_n=0$. The proof is completed by the fact of $\|\mathbf{E}^n\|<v_n$ for each n. □

Theorem 7.141. Consider an IVP of which $\mathbf{f}(\mathbf{u},t)$ is p times continuously differentiable with respect to both t and $\mathbf{u}$. For a convergent LMM with consistency of order p and with its initial conditions satisfying

$$\forall i=0,1,\ldots,s-1,\qquad\|\mathbf{U}^i-\mathbf{u}(t_i)\|=O(k^p),$$

its numerical solution of the IVP satisfies

$$\|\mathbf{U}^{t/k}-\mathbf{u}(t)\|=O(k^p)\tag{7.90}$$

for all $t\in[0,T]$ and sufficiently small $k>0$.

Proof. This proof is similar to that of Theorem 7.140. □

7.3.6 Absolute stability

Definition 7.142. The *stability polynomial* of an LMM is

$$\pi_\kappa(\zeta):=\rho(\zeta)-\kappa\sigma(\zeta)=\sum_{j=0}^{s}(\alpha_j-\kappa\beta_j)\zeta^j.\tag{7.91}$$

Definition 7.143. An LMM is *absolutely stable* for some κ if all solutions $\{\mathbf{U}^n\}$ of

$$\pi_\kappa(Z)\mathbf{U}^n=[\rho(Z)-\kappa\sigma(Z)]\,\mathbf{U}^n=\mathbf{0}$$

are bounded as $n\to+\infty$, where Z is the time-shift operator in Notation 5.

Theorem 7.144 (*Root condition* for absolute stability)**.** An LMM is absolutely stable for $\kappa:=k\lambda$ if and only if all the roots of $\pi_\kappa(\zeta)$ satisfy $|\zeta|\le1$, and any root satisfying $|\zeta|=1$ is simple.

Proof. This proof is the same as that of Theorem 7.127. □

Definition 7.145. The *region of absolute stability (RAS)* for an LMM is the set of all $\kappa\in\mathbb{C}$ for which the method is absolutely stable.

Example 7.146. For Euler's method (7.39),

$$\pi_\kappa(\zeta)=(\zeta-1)-\kappa=\zeta-(1+\kappa),\tag{7.92}$$

with the single root $\zeta_1=1+\kappa$. Thus the RAS for Euler's method is the disk:

$$\mathcal{R}=\{\kappa:|1+\kappa|\le1\}.\tag{7.93}$$

Example 7.147. For backward Euler's method (7.40),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \kappa\zeta = (1 - \kappa)\zeta - 1, \tag{7.94}$$

with root $\zeta_1 = (1 - \kappa)^{-1}$. Thus the RAS for backward Euler's method is:

$$\mathcal{R} = \{\kappa : |(1 - \kappa)^{-1}| \le 1\} = \{\kappa : |1 - \kappa| \ge 1\}. \tag{7.95}$$

Example 7.148. For the trapezoidal method (7.41),

$$\pi_\kappa(\zeta) = (\zeta - 1) - \frac{1}{2}\kappa(\zeta + 1) = \left(1 - \frac{1}{2}\kappa\right)\zeta - \left(1 + \frac{1}{2}\kappa\right). \tag{7.96}$$

Thus the RAS for the trapezoidal method is the left half-plane:

$$\begin{aligned}\mathcal{R} &= \left\{\kappa \in \mathbb{C} : \left|\frac{2 + \kappa}{2 - \kappa}\right| \le 1\right\} \\ &= \{\kappa \in \mathbb{C} : \operatorname{Re}\kappa \le 0\}.\end{aligned} \tag{7.97}$$

Example 7.149. For the midpoint method (7.42),

$$\pi_\kappa(\zeta) = \zeta^2 - 2\kappa\zeta - 1. \tag{7.98}$$

$\pi_\kappa(\zeta) = 0$ implies

$$2\kappa = \zeta - \frac{1}{\zeta}.$$

Since $\zeta = ae^{i\theta}$ and $\frac{1}{\zeta} = a^{-1}e^{-i\theta}$, there are always one zero with $|\zeta_1| \le 1$ and another zero with $|\zeta_2| \ge 1$, depending on the sign of κ. The only possibility for both roots to have a modulus no greater than one is $|\zeta_1| = |\zeta_2| = 1 = a$. So the stability region consists only of the open interval from $-i$ to i on the imaginary axis:

$$\mathcal{R} = \{\kappa \in \mathbb{C} : \kappa = i\alpha \text{ with } |\alpha| < 1\}. \tag{7.99}$$

Definition 7.150. The *boundary locus* method finds the RAS of an LMM (ρ, σ) with $\sigma(e^{i\theta}) \ne 0$ by steps as follows.

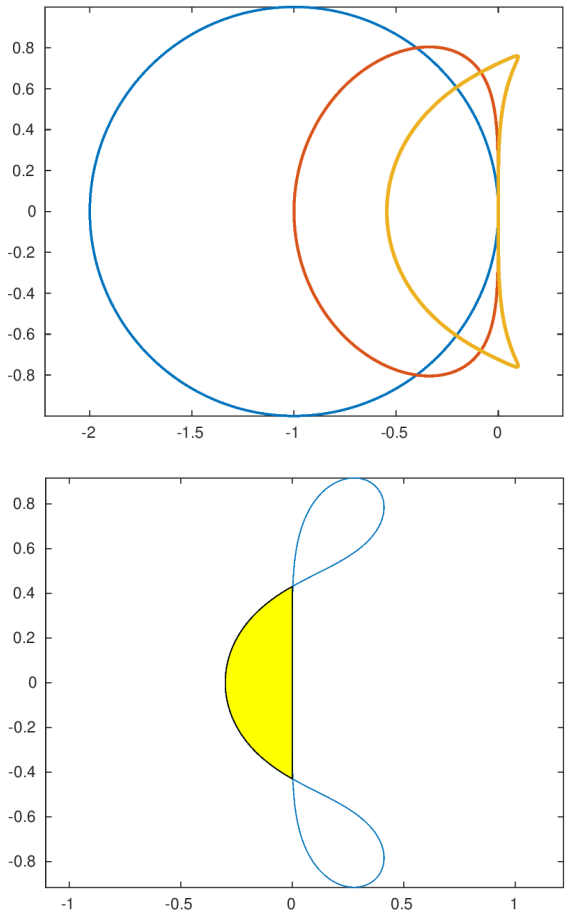
(a) compute the *root locus curve*

$$\gamma(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \qquad \theta \in [0, 2\pi]; \tag{7.100}$$

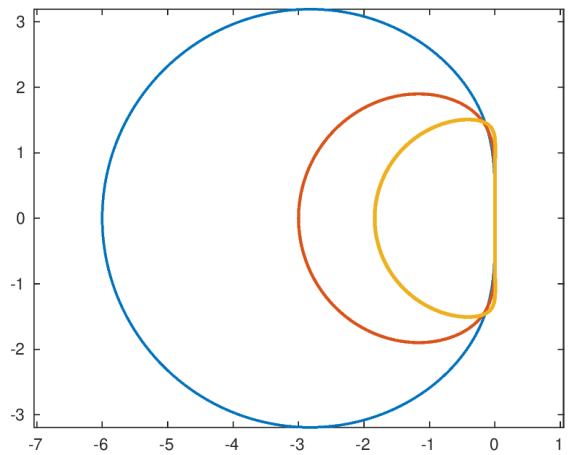
(b) the closed curve γ divides the complex plane $\mathbb{C}$ into a number of connected regions;

(c) for each connected region $S \subset \mathbb{C}$, choose a convenient interior point $\kappa_p \in S$ and solve the equation $\rho(\zeta) - \kappa_p\sigma(\zeta) = 0$: S is part of the RAS if all roots are in the unit disk; otherwise S is not.

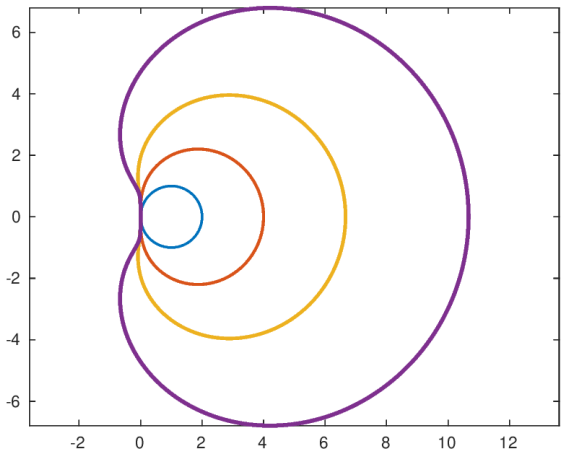
Example 7.151. The RASs of Adams-Bashforth formulas are shown below, with the first plot as those of $p = 1, 2, 3$ and the second as that of $p = 4$. Each RAS is bounded.



Example 7.152. The RASs of Adams-Moulton formulas with $p = 3, 4, 5$ are shown below. Each RAS is bounded.



Example 7.153. The RASs of backward differentiation formulas with $p = 1, 2, 3, 4$ are shown below. Each RAS is unbounded.



Exercise 7.154. Write a program to reproduce the RAS plots in Examples 7.151, 7.152, and 7.153.

7.3.7 The first Dahlquist barrier

Theorem 7.155. The s-step Adams and Nystrom formulas are stable for all $s \geq 1$. The s-step backward differentiation formulas are stable for $s = 1, 2, \ldots, 6$, but unstable for $s \geq 7$.

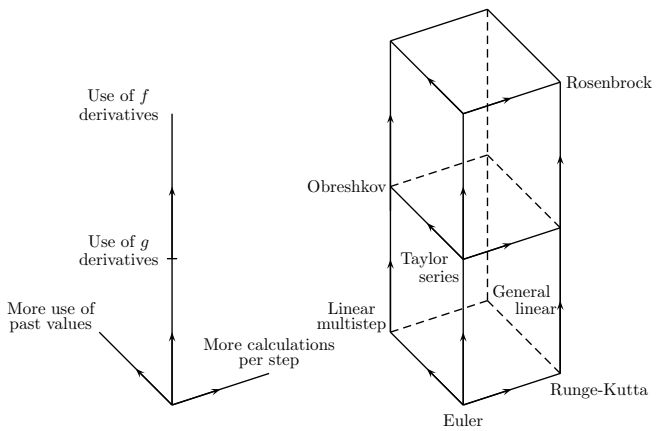
Proof. See Hairer et al. [1993]. □

Theorem 7.156. The order of accuracy p of a stable s-step LMM satisfies

$$p \leq \begin{cases} s & \text{if the LMM is explicit,} \\ s+1 & \text{else if } s \text{ is odd,} \\ s+2 & \text{else if } s \text{ is even.} \end{cases} \tag{7.101}$$

Proof. See Hairer et al. [1993]. □

7.4 Runge-Kutta methods



Definition 7.157. A *one-step method* or *multistage method* constructs numerical solutions of a scalar IVP (7.3) at each time step $n = 0, 1, \ldots$ by a formula of the form

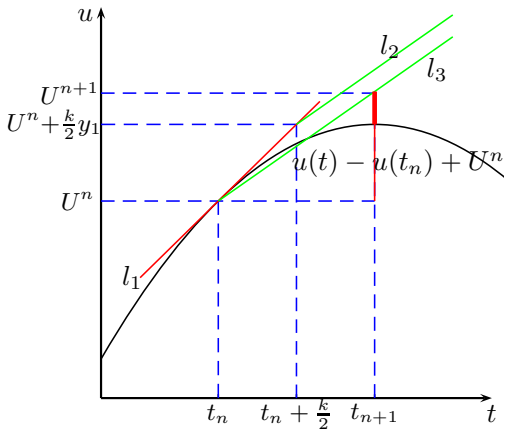
$$U^{n+1} = U^n + k\Phi(U^n, t_n; k), \tag{7.102}$$

where the *increment function* $\Phi : \mathbb{R} \times [0, T] \times (0, +\infty) \to \mathbb{R}$ is given in terms of the function $f : \mathbb{R} \times [0, T] \to \mathbb{R}$ in (7.3).

7.4.1 Classical formulas

Definition 7.158. The *modified Euler method* or the *improved polygon method* is a one-step method of the form

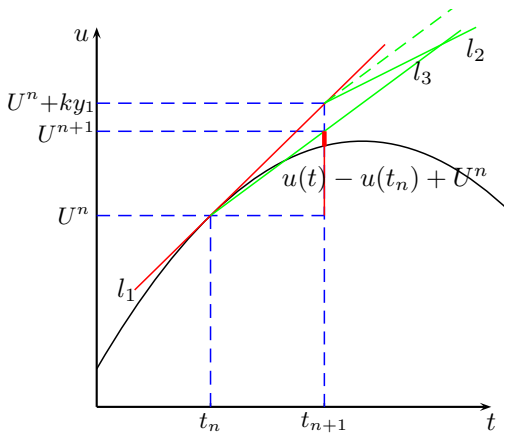
$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + \frac{k}{2}y_1, t_n + \frac{k}{2}), \\ U^{n+1} = U^n + ky_2. \end{cases} \tag{7.103}$$



Exercise 7.159. Does the length of the thick red line segment in the above figure represent the one-step error in Definition 7.172? If so, prove it; otherwise derive an expression of the represented quantity.

Definition 7.160. The *improved Euler method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + ky_1, t_n + k), \\ U^{n+1} = U^n + \frac{k}{2}(y_1 + y_2). \end{cases} \tag{7.104}$$



Definition 7.161. *Heun's third-order formula* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + \frac{k}{3}y_1, t_n + \frac{k}{3}), \\ y_3 = f(U^n + \frac{2k}{3}y_2, t_n + \frac{2k}{3}), \\ U^{n+1} = U^n + \frac{k}{4}(y_1 + 3y_3). \end{cases} \tag{7.105}$$

Definition 7.162. The *classical fourth-order Runge-Kutta method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + \frac{k}{2}y_1, t_n + \frac{k}{2}), \\ y_3 = f(U^n + \frac{k}{2}y_2, t_n + \frac{k}{2}), \\ y_4 = f(U^n + ky_3, t_n + k), \\ U^{n+1} = U^n + \frac{k}{6}(y_1 + 2y_2 + 2y_3 + y_4). \end{cases} \tag{7.106}$$

Definition 7.163. An *s-stage explicit Runge-Kutta (ERK) method* is a one-step method of the form

$$\begin{cases} y_1 = f(U^n, t_n), \\ y_2 = f(U^n + ka_{2,1}y_1, t_n + c_2k), \\ y_3 = f(U^n + k(a_{3,1}y_1 + a_{3,2}y_2), t_n + c_3k), \\ \quad \dots \\ y_s = f(U^n + k(a_{s,1}y_1 + \ldots + a_{s,s-1}y_{s-1}), t_n + c_sk), \\ U^{n+1} = U^n + k(b_1y_1 + b_2y_2 + \ldots + b_sy_s), \end{cases} \tag{7.107}$$

where $a_{i,j}$, b_i, and c_i are real coefficients for $i, j = 1, 2, \dots, s$, $a_{i,j} = 0$ for $i \le j$, and

$$\forall i = 1, 2, \dots, s, \quad c_i = \sum_{j=1}^{s} a_{i,j}. \tag{7.108}$$

Definition 7.164. An *s-stage Runge-Kutta method* is a one-step method of the form

$$\begin{cases} y_i = f(U^n + k\sum_{j=1}^s a_{i,j}y_j, t_n + c_ik), \\ U^{n+1} = U^n + k\sum_{j=1}^s b_jy_j, \end{cases} \tag{7.109}$$

where $i = 1, 2, \dots, s$ and the coefficients $a_{i,j}$, b_j, c_i are real.

Definition 7.165. The *Butcher tableau* is one way to organize the coefficients of a Runge-Kutta method as follows.

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \cdots & a_{1,s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & \cdots & a_{s,s} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{7.110}$$

Definition 7.166. An *implicit Runge-Kutta (IRK) method* is a Runge-Kutta method with at least one $a_{i,j} \neq 0$ for $i \le j$. A *diagonal implicit Runge-Kutta (DIRK) method* is an IRK method with $a_{i,j} = 0$ whenever $i < j$. A *singly diagonal implicit Runge-Kutta (SDIRK) method* is a DIRK method with $a_{1,1} = a_{2,2} = \cdots = a_{s,s} = \gamma \neq 0$.

Example 7.167. The Butcher tableau of an s-stage ERK method is

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a_{2,1} & 0 & & & \\ c_3 & a_{3,1} & a_{3,2} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \tag{7.111}$$

Example 7.168. The Butcher tableau of the classical fourth-order RK method (7.106), is

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{7.112}$$

Exercise 7.169. Write down the Butcher tableaux of the modified Euler method, the improved Euler method, and Heun's third-order method.

Definition 7.170. The *TR-BDF2 method* is a second-order DIRK method of the form

$$\begin{cases} U^* = U^n + \frac{k}{4}\left(f(U^n, t_n) + f(U^*, t_n + \frac{k}{2})\right), \\ U^{n+1} = \frac{1}{3}\left(4U^* - U^n + kf(U^{n+1}, t_{n+1})\right). \end{cases} \tag{7.113}$$

Exercise 7.171. Rewrite the TR-BDF2 method in the standard form of a Runge-Kutta method and derive its Butcher tableau. Give a geometric interpretation of TR-BDF2 by drawing a figure similar to those for the modified Euler method and the improved Euler method.

7.4.2 Consistency and convergence

Definition 7.172. The *one-step error of a multistage method* (7.102) is

$$\mathcal{L}u(t_n) := u(t_{n+1}) - u(t_n) - k\Phi(u(t_n), t_n; k). \tag{7.114}$$

Definition 7.173. A multistage method is said to have *order of accuracy p* if

$$\mathcal{L}u(t_n) = \Theta(k^{p+1}) \text{ as } k \to 0. \tag{7.115}$$

Definition 7.174. A multistage method is *consistent* if

$$\lim_{k\to 0} \frac{1}{k}\mathcal{L}u(t_n) = 0. \tag{7.116}$$

Example 7.175. For the modified Euler method, we have

$$\frac{U^{n+1} - U^n}{k} = f\left(U^n + \frac{k}{2}f(U^n, t_n), t_n + \frac{k}{2}\right) \tag{7.117}$$

and thus the one-step error is

$$\begin{aligned} \mathcal{L}u(t_n) =& u(t_{n+1}) - u(t_n) \\ & - kf\left(u(t_n) + \frac{k}{2}f(u(t_n), t_n), t_n + \frac{k}{2}\right) \\ =& u(t_{n+1}) - u(t_n) - kf\left(u(t_n) + \frac{1}{2}ku'(t_n), t_n + \frac{k}{2}\right) \\ =& ku'\left(t_n + \frac{k}{2}\right) + O(k^3) \\ & - kf\left(u\left(t_n + \frac{k}{2}\right) + O(k^2), t_n + \frac{k}{2}\right) \\ =& ku'\left(t_n + \frac{k}{2}\right) + O(k^3) - kf\left(u\left(t_n + \frac{k}{2}\right), t_n + \frac{k}{2}\right) \\ =& O(k^3), \end{aligned}$$

where the second and last equality hold since u satisfies the IVP and the third and fourth follow from Taylor expansions. Hence the method is at least second-order accurate.

Exercise 7.176. Derive the $O(k^3)$ term in Example 7.175 to verify that it does not vanish.

Theorem 7.177. A multistage method is consistent if and only if

$$\lim_{k\to 0} \Phi(u,t;k) = f(u,t) \tag{7.118}$$

for any (u,t) in the domain of f.

Proof. Definition 7.157 and a Taylor expansion of $u(t_{n+1})$ at t_n yield

$$\frac{\mathcal{L}u(t_n)}{k} = f(u(t_n),t_n) - \Phi(u(t_n),t_n;k) + \Theta(k).$$

The proof is completed by taking limit of the above equation in the asymptotic range of $k \to 0$, c.f. Definition 7.174. □

Corollary 7.178. The Euler method is consistent.

Proof. This follows from Theorem 7.177 and the fact that $\Phi(u,t;0) = f(u,t)$ for Euler's method. □

Definition 7.179. A multistage method is *convergent* if its solution error tends to zero as $k \to 0$ for any $T > 0$ and for any initial condition $u_0 = u(0) + o(1)$, i.e.,

$$\lim_{k\to 0; Nk=T} U^N = u(T). \tag{7.119}$$

Lemma 7.180. Let (ξ_n) be a sequence in $\mathbb{R}$ such that

$$|\xi_{n+1}| \le (1+C)|\xi_n| + D, \quad n \in \mathbb{N} \tag{7.120}$$

for some positive constants C and D. Then we have

$$|\xi_n| \le e^{nC}|\xi_0| + \frac{D}{C}(e^{nC} - 1), \quad n \in \mathbb{N}. \tag{7.121}$$

Proof. The induction basis $n = 0$ clearly holds. Now suppose (7.121) holds for n, then for the inductive step, we have

$$\begin{aligned}|\xi_{n+1}| &\le (1+C)e^{nC}|\xi_0| + (1+C)\frac{D}{C}(e^{nC}-1) + D \\ &\le e^{(n+1)C}|\xi_0| + \frac{D}{C}(e^{(n+1)C} - 1),\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second from $1 + C \le e^C$. Thus the estimate (7.121) holds for $n+1$ as well. □

Theorem 7.181. Suppose the increment function Φ that describes a multistage method is continuous (in u, t, and k) and satisfies a Lipschitz condition

$$|\Phi(u,t;k) - \Phi(v,t;k)| \le M|u-v| \tag{7.122}$$

for all (u,t) and (v,t) in the domain of f and for all sufficiently small k. Also suppose that the initial condition satisfies $|E^0| = O(k)$. Then the multistage method is convergent if and only if it is consistent. Furthermore, if the method has order of accuracy p, i.e., $\mathcal{L}u(t_n) \le Kk^{p+1}$, and the initial condition satisfies $|E^0| = O(k^{p+1})$, then its solution error can be bounded as

$$|E^n| \le \frac{K}{M}\left(e^{MT} - 1\right)k^p. \tag{7.123}$$

Proof. For sufficiency, we assume that the multistage method is consistent and compute

$$\begin{aligned}|E^{n+1} - E^n| &= |(U^{n+1} - U^n) - (u(t_{n+1}) - u(t_n))| \\ &= |k\Phi(U^n,t_n;k) - (u(t_{n+1}) - u(t_n))| \\ &= |k\Phi(U^n,t_n;k) - k\Phi(u(t_n),t_n;k) - \mathcal{L}u(t_n)| \\ &\le kM|U^n - u(t_n)| + kc(k),\end{aligned}$$

where the last step follows from the Lipschitz condition (7.122) and $\lim_{k\to 0} c(k) = \lim_{k\to 0} \frac{1}{k}\max|\mathcal{L}u(t)| = 0$. Hence we have

$$|E^{n+1}| \le (1+kM)|E^n| + kc(k).$$

Applying Lemma 7.180 with $C = kM$ and $D = kc(k)$ yields

$$\begin{aligned}|E^n| &\le |E^0|e^{nkM} + \frac{c(k)}{M}\left(e^{nkM} - 1\right) \\ &= |E^0|e^{MT} + \frac{c(k)}{M}\left(e^{MT} - 1\right),\end{aligned}$$

which establishes the convergence since $|E^0|$ and $c(k)$ both tend to 0 as $k \to 0$. In particular, (7.123) follows from this inequality and the condition of $c(k) \le Kk^p$.

For necessity, we assume that the multistage method is convergent, i.e., the multistage method (7.102) converges to the solution of

$$u'(t) = f(u,t), \quad u(0) = u_0,$$

for all final time $T > 0$. Consider

$$g(u,t) := \Phi(u,t;0)$$

and observe that by Theorem 7.177 the multistage method is consistent with the new IVP

$$u'(t) = g(u,t), \quad u(0) = u_0.$$

Since we have already shown that consistency implies convergence, the multistage method also converges to this new IVP. Hence the solutions of the two IVPs coincide and we have $f(u(\tau),\tau) = g(u(\tau),\tau)$ for all $(u(\tau),\tau)$ in the domain of f. Then the continuity of Φ in k at $k = 0$ implies

$$\begin{aligned}&\forall \epsilon > 0,\ \exists \delta \text{ s.t. } \forall k < \delta,\ \forall t \in [0,T], \\ &\quad |\Phi(u,t;k) - f(u,t)| \\ &\le |\Phi(u,t;0) - f(u,t)| + |\Phi(u,t;k) - \Phi(u,t;0)| \\ &< \epsilon,\end{aligned}$$

which implies uniform convergence of $\Phi(u,t;k)$ to f. Then the proof is completed by Theorem 7.177. □

Corollary 7.182. Both the modified Euler method and the improved Euler method are convergent. If f in the IVP is twice continuously differentiable, then each of them has order of accuracy two.

Proof. For the modified Euler method (7.103), we have

$$\Phi(u,t;k) = f\left(u + \frac{k}{2}f(u,t), t + \frac{k}{2}\right),$$

which clearly satisfies the consistency condition (7.118), and hence by Theorem 7.181, it only remains to verify the Lipschitz condition of Φ. From the Lipschitz condition for f we obtain

$$\begin{aligned}
&|\Phi(u,t;k) - \Phi(v,t;k)| \\
&= \left| f\left(u + \frac{k}{2}f(u,t), t + \frac{k}{2}\right) - f\left(v + \frac{k}{2}f(v,t), t + \frac{k}{2}\right)\right| \\
&\le L\left(|u-v| + \frac{k}{2}|f(u,t) - f(v,t)|\right) \\
&\le L\left(1 + \frac{kL}{2}\right)|u-v|,
\end{aligned}$$

hence Φ also satisfies a Lipschitz condition.

If f is twice continuously differentiable, then by Example 7.175, the one-step error of the modified Euler method satisfies

$$\mathcal{L}u(t_n) \le Kk^3,$$

Therefore the modified Euler method (7.103) has order of accuracy two by Theorem 7.181.

The same result concerning the improved Euler method (7.104) can be proved in a similar manner. □

Lemma 7.183. The one-step error of the classical Runge-Kutta method (7.106) is

$$\mathcal{L}u(t_n) = O(k^5). \tag{7.124}$$

Exercise 7.184. Prove Lemma 7.183.

Corollary 7.185. The classical Runge-Kutta method (7.106) is convergent. If f in the IVP is four-times continuously differentiable, then it is convergent with order of accuracy four.

Proof. The function Φ describing the classical Runge-Kutta method (7.106) is given by

$$\Phi = \frac{1}{6}(\Phi_1 + 2\Phi_2 + 2\Phi_3 + \Phi_4),$$

where

$$\begin{aligned}
\Phi_1(u,t;k) &= f(u,t), \\
\Phi_2(u,t;k) &= f\left(u + \frac{k}{2}\Phi_1(u,t;k), t + \frac{k}{2}\right), \\
\Phi_3(u,t;k) &= f\left(u + \frac{k}{2}\Phi_2(u,t;k), t + \frac{k}{2}\right), \\
\Phi_4(u,t;k) &= f(u + k\Phi_3(u,t;k), t + k).
\end{aligned}$$

From this, consistency follows immediately by Theorem 7.177. Since Φ clearly satisfies a Lipschitz condition, it follows from Theorem 7.181 that the classical Runge-Kutta method (7.106) is convergent.

If f is four-times continuously differentiable, Lemma 7.183 shows that the classical Runge-Kutta method (7.106) has a one-step error of $O(k^5)$, hence it has order of accuracy four by Theorem 7.181. □

7.4.3 Absolute stability

Definition 7.186. The *stability function of a one-step method* is a ratio of two polynomials

$$R(z) = \frac{P(z)}{Q(z)} \tag{7.125}$$

that satisfies

$$U^{n+1} = R(z)U^n \tag{7.126}$$

for the test problem $u'(t) = \lambda u$ where $z := k\lambda$.

Example 7.187. The fourth-order Runge-Kutta method has its stability function as

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4. \tag{7.127}$$

Example 7.188. The trapezoidal rule, when viewed as a one-step method, has its stability function as

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, \tag{7.128}$$

which is also the root of the LMM stability polynomial in Example 7.148.

Exercise 7.189. Show that the TR-BDF2 method (7.113) has

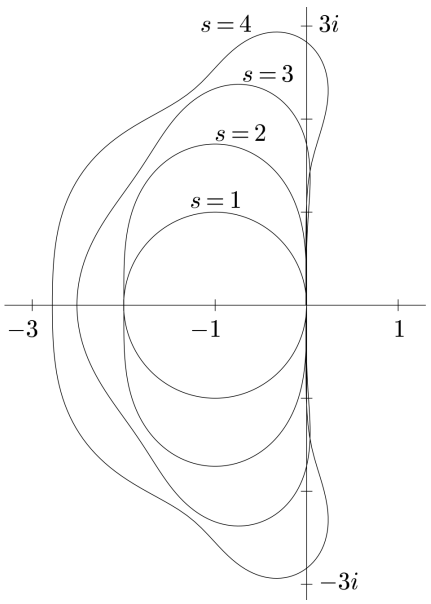
$$R(z) = \frac{1 + \frac{5}{12}z}{1 - \frac{7}{12}z + \frac{1}{12}z^2}, \tag{7.129}$$

and $R(z) - e^z = O(z^3)$ as $z \to 0$.

Definition 7.190. The *region of absolute stability (RAS) of a one-step method* is a subset of the complex plane

$$\mathcal{R} := \{z \in \mathbb{C} : |R(z)| \le 1\}. \tag{7.130}$$

Example 7.191. The boundaries of RASs for ERKs with $s = 1, 2, 3, 4$ are shown below.



7.5 Stiff IVPs

Example 7.192. Consider the IVP

$$u'(t) = \lambda(u - \cos t) - \sin t, \quad u(0) = \eta. \tag{7.131}$$

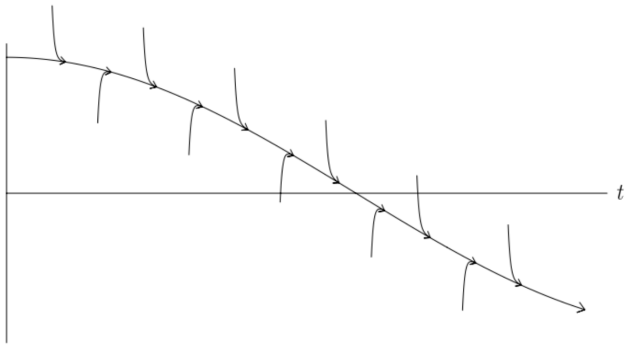
By Duhamel's principle (7.25), the exact solution is

$$\begin{aligned} u_\eta(t) &= e^{\lambda t}\eta - \int_0^t e^{\lambda(t-\tau)}(\lambda\cos\tau + \sin\tau)\mathrm{d}\tau \\ &= e^{\lambda t}\eta - \int_0^t \lambda e^{\lambda(t-\tau)}\cos\tau\mathrm{d}\tau - \int_0^t e^{\lambda(t-\tau)}\sin\tau\mathrm{d}\tau \\ &= e^{\lambda t}(\eta - 1) + \cos t, \end{aligned}$$

where the third equality follows from the integration-by-parts formula.

If $\eta = \cos(0) = 1$, then $u_1(t) = \cos t$ is the unique solution. If $\eta \neq 1$ and $\lambda < 0$, then the solution curve $u_\eta(t)$ decays exponentially to $u_1(t)$.

A negative λ with large magnitude has a dominant effect on nearby solutions of the ODE corresponding to different initial data; the following picture shows some solution curves with $\lambda = -100$.



For six values of k, the following table compares the results at $T = 1$ computed by the second-order Adams-Bashforth and the second-order BDF method.

k	AB2	BDF2
0.2	14.40	0.5404
0.1	-5.70×10^4	0.54033
0.05	-1.91×10^9	0.540309
0.02	-5.77×10^{10}	0.5403034
0.01	0.5403019	0.54030258
0.005	0.54030222	0.54030238
⋮	⋮	⋮
0	0.540302306	0.540302306

The results indicate the curious effect that this property of the ODE has on numerical computations. To achieve a solution error $E(T) \leq \epsilon = 4 \times 10^{-5}$, the BDF2 method may use $k = 0.1$, the AB2 method has to use $k \leq 0.01$ while the time scale of the IVP is 1.

7.5.1 The notion of stiffness

Definition 7.193. An IVP is said to be *stiff in an interval* if for some initial condition any numerical method with a finite RAS is forced to use a time-step size that is excessively smaller than the time scale of the true solution of the IVP.

Definition 7.194. For an IVP

$$\mathbf{u}'(t) = A\mathbf{u} + \mathbf{b}(t), \tag{7.132}$$

where $\mathbf{u}, \mathbf{b} \in \mathbb{R}^n$ and A is a constant, diagonalizable, $n \times n$ matrix with eigenvalues $\lambda_i \in \mathbb{C}, i = 1, 2, \cdots, n$, its *stiffness ratio* is

$$\frac{\max\limits_{\lambda\in\Lambda(A)} |\mathrm{Re}\,\lambda|}{\min\limits_{\lambda\in\Lambda(A)} |\mathrm{Re}\,\lambda|}. \tag{7.133}$$

Example 7.195. Consider the linear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -1000 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \qquad t \in [0, 1] \tag{7.134}$$

with initial value $\mathbf{u}(0) = (1, 1)^T$. Suppose we want

$$\|\mathbf{E}\|_\infty \leq \epsilon,$$

that is

$$|U_1^N - e^{-1000}| \leq \epsilon, \quad |U_2^N - e^{-1}| \leq \epsilon.$$

If (7.134) is solved by a p-th order LMM with time-step size k. For U_2^N to be sufficiently accurate, we need $k = O(\epsilon^{1/p})$. But for U_1^N to be sufficiently accurate, if the formula has a stability region of finite size like the Euler formula, we need k to be on the order 10^{-3}. Most likely this is a much tighter restriction.

Example 7.196. Consider the nonlinear IVP

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} -u_1 u_2 \\ \cos(u_1) - \exp(u_2) \end{pmatrix}. \tag{7.135}$$

The Jacobian matrix is

$$J = -\begin{pmatrix} u_2 & u_1 \\ \sin(u_1) & \exp(u_2) \end{pmatrix}.$$

Near a point t with $u_1(t) = 0$ and $u_2(t) \gg 1$, the matrix is diagonal with widely differing eigenvalues and the behavior will probably be stiff.

Example 7.197. Read Example 8.2 (pp 167) in the book by Leveque.

7.5.2 A-stability and L-stability

Definition 7.198. An ODE method is *A-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \mathrm{Re}\, z \leq 0\} \subseteq \mathcal{R}. \tag{7.136}$$

Example 7.199. The backward Euler's method and trapezoidal method are A-stable.

Theorem 7.200 (Dahlquist's Second Barrier)**.** The order of accuracy of an implicit A-stable LMM satisfies $p \leq 2$. An explicit LMM cannot be A-stable.

Definition 7.201. An ODE method is *A(α)-stable* if its region of absolute stability $\mathcal{R}$ satisfies

$$\{z \in \mathbb{C} : \pi - \alpha \leq \arg(z) \leq \pi + \alpha\} \subseteq \mathcal{R}. \tag{7.137}$$

It is *A(0)-stable* if it is A(α)-stable for some $\alpha > 0$.

Example 7.202. As shown in Example 7.153, the BDFs are A(α)-stable with $\alpha = 90^\circ$ for $p = 1, 2$ and $\alpha \approx 86^\circ$, 73°, 51°, and 17° for $p = 3, 4, 5, 6$ respectively. Note the large drop of α from $p = 5$ to $p = 6$.

Definition 7.203. A one-step method is *L-stable* if it is A-stable and

$$\lim_{z\to\infty} |R(z)| = 0, \tag{7.138}$$

where $U^{n+1} = R(z)U^n$.

Example 7.204. We use the trapezoidal and backward Euler's methods to solve the IVP (7.131) with $\lambda = -10^6$. The following table shows the errors at $T = 3$ with various values of k and the initial data $u(0) = \eta$.

	k	Backward Euler	Trapezoidal
$\eta = 1$	0.2	9.7731e-08	4.7229e-10
	0.1	4.9223e-08	1.1772e-10
	0.05	2.4686e-08	2.9406e-11
$\eta = 1.5$	0.2	9.7731e-08	4.9985e-01
	0.1	4.9223e-08	4.9940e-01
	0.05	2.4686e-08	4.9761e-01

The results are caused by the fact that the backward Euler's method is L-stable while the trapezoidal method is not.

Exercise 7.205. Reproduce the results in Example 7.204 and explain in your own language why the first-order backward Euler method is superior to the second-order trapezoidal method.

Chapter 8

Boundary Value Problems (BVPs)

Definition 8.1. A *partial differential equation* (PDE) is an equation involving an unknown function of two or more variables and some of its partial derivatives.

Definition 8.2. *Laplace equation* is a second-order PDE of the form

$$\Delta u(\mathbf{x}) = 0, \tag{8.1}$$

where the unknown is a function $u : \overline{\Omega} \to \mathbb{R}$, $\overline{\Omega}$ is the closure of an open set $\Omega \subset \mathbb{R}^n$, and the *Laplacian operator* $\Delta : \mathcal{C}^2(\Omega) \to \mathcal{C}(\Omega)$ is

$$\Delta := \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}. \tag{8.2}$$

Example 8.3. The fundamental theorem of vector calculus, (a.k.a. Helmholtz's theorem) states that a sufficiently continuous vector field $\mathbf{v}^*$ on a bounded domain can be uniquely decomposed into a divergence-free part and a curl-free part:

$$\begin{cases} \mathbf{v}^* &= \mathbf{v} + \nabla\phi, \\ \nabla \cdot \mathbf{v} &= 0, \qquad \nabla \times \nabla\phi = \mathbf{0}. \end{cases} \tag{8.3}$$

Potential flow is a special type of flow where the velocity $\mathbf{u}$ can be expressed as the gradient of a scalar function:

$$\mathbf{u} = \nabla\varphi,$$

where φ is called the *velocity potential*. If the fluid is also incompressible, i.e. $\nabla \cdot \mathbf{u} = 0$, the velocity potential satisfies a Laplace equation $\Delta\varphi = 0$.

Definition 8.4. *Poisson's equation* is a second-order PDE of the form

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \tag{8.4}$$

where the unknown is a function $u : \overline{\Omega} \to \mathbb{R}$, $\overline{\Omega}$ is the closure of an open set $\Omega \subset \mathbb{R}^n$, and the RHS function $f : \Omega \to \mathbb{R}$ is given a priori.

Definition 8.5. A *boundary value problem* (BVP) is a differential equation together with a set of additional constraints, called the *boundary conditions*, that hold only on the domain boundary.

Definition 8.6. Common types of boundary conditions for a one-dimensional interval $\Omega = (a, b)$ are

- *Dirichlet conditions*: $u(a) = \alpha$ and $u(b) = \beta$;
- *Mixed conditions*: $u(a) = \alpha$ and $\left.\frac{\partial u}{\partial x}\right|_b = \beta$;
- *Neumann conditions*: $\left.\frac{\partial u}{\partial x}\right|_a = \alpha$ and $\left.\frac{\partial u}{\partial x}\right|_b = \beta$.

Example 8.7. The Helmholtz decomposition in Example 8.3 is realized by solving the Neumann BVP

$$\Delta\phi = \nabla \cdot \mathbf{v}^* \qquad \text{in } \Omega, \tag{8.5a}$$

$$\mathbf{n} \cdot \nabla\phi = \mathbf{n} \cdot (\mathbf{v}^* - \mathbf{v}) \qquad \text{on } \partial\Omega. \tag{8.5b}$$

8.1 Finite difference (FD) methods

Formula 8.8. In solving a linear BVP, the general procedures of an FD method are as follows.

(FD-1) Discretize the problem domain by a grid.

(FD-2) Approximate each spatial derivative in the PDE with some finite difference formula at every grid point to get a system of linear equations $A\mathbf{U} = \mathbf{F}$ where the vector $\mathbf{U}$ approximates the unknown variable on the grid while the vector $\mathbf{F}$ contains given conditions of the BVP such as boundary conditions and derivatives of the unknown function.

(FD-3) Solve the system of algebraic equations.

Example 8.9 (An FD method for Poisson's equation in a unit interval)**.** Consider the one-dimensional BVP

$$-u''(x) = f(x) \text{ in } \Omega := (0, 1) \tag{8.6}$$

with Dirichlet boundary conditions

$$u(0) = \alpha, \quad u(1) = \beta. \tag{8.7}$$

The general procedures of an FD method based on the central difference are as follows.

(a) Discretize Ω by a Cartesian grid with uniform spacing,

$$x_j = jh, \qquad h = \frac{1}{m+1}, \quad j = 0, 1, \ldots, m+1.$$

Set $U_0 = \alpha$, $U_{m+1} = \beta$ and we will compute m values $U_1, \ldots, U_m$ where each U_j approximates $u(x_j)$.

(b) Approximate the second derivative u'' with a centered difference

$$u''(x_j) = \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} + O(h^2) \quad (8.8)$$

and we get the following system of linear equations:

$$\begin{aligned} -\frac{\alpha - 2U_1 + U_2}{h^2} &= f(x_1), \\ -\frac{U_{j-1} - 2U_j + U_{j+1}}{h^2} &= f(x_j), \quad j = 2, \cdots, m-1, \\ -\frac{U_{m-1} - 2U_m + \beta}{h^2} &= f(x_m). \end{aligned}$$

These equations are written in the form

$$A\mathbf{U} = \mathbf{F}, \quad (8.9)$$

where

$$\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{m-1} \\ U_m \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} f(x_1) + \frac{\alpha}{h^2} \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) + \frac{\beta}{h^2} \end{bmatrix}, \quad (8.10)$$

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}. \quad (8.11)$$

(c) Solve the linear system (8.9).

Formula 8.10 (The method of undetermined coefficients)**.** A *general method to derive FD formulas* that approximate $u^{(k)}(\bar{x})$ is based on an arbitrary stencil of $n > k$ distinct points $x_1, x_2, \ldots, x_n$. Taylor expansions of u at each point x_i in the stencil about $u(\bar{x})$ yield

$$u(x_i) = u(\bar{x}) + (x_i - \bar{x})u'(\bar{x}) + \cdots + \frac{1}{k!}(x_i - \bar{x})^k u^{(k)}(\bar{x}) + \cdots$$

for $i = 1, 2, \ldots, n$. This leads to a linear combination of point values that approximates $u^{(k)}(\bar{x})$,

$$u^{(k)}(\bar{x}) = c_1 u(x_1) + c_2 u(x_2) + \cdots + c_n u(x_n) + O(h^p),$$

where the c_j's are chosen to make p as large as possible:

$$\forall i = 0, \ldots, p-1, \qquad \frac{1}{i!}\sum_{j=1}^{n} c_j (x_j - \bar{x})^i = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise.} \end{cases} \quad (8.12)$$

Example 8.11. To approximate $u'(\bar{x})$ with an FD formula

$$D_2 u(\bar{x}) = au(\bar{x}) + bu(\bar{x} - h) + cu(\bar{x} - 2h), \quad (8.13)$$

we determine the coefficients a, b, and c to give the best possible accuracy. Taylor expansions at $\bar{x}$ yield

$$\begin{aligned} D_2 u(\bar{x}) =& (a + b + c)u(\bar{x}) - (b + 2c)hu'(\bar{x}) \\ &+ \frac{1}{2}(b + 4c)h^2 u''(\bar{x}) - \frac{1}{6}(b + 8c)h^3 u'''(\bar{x}) \\ &+ O\left(h^4\right). \end{aligned}$$

Set $a + b + c = 0$, $b + 2c = -\frac{1}{h}$, and $b + 4c = 0$, solve

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{h} \\ 0 \end{bmatrix}, \quad (8.14)$$

and we get

$$a = \frac{3}{2h}, \quad b = -\frac{2}{h}, \quad c = \frac{1}{2h}. \quad (8.15)$$

Therefore the FD formula is determined as

$$D_2 u(\bar{x}) = \frac{1}{2h}[3u(\bar{x}) - 4u(\bar{x} - h) + u(\bar{x} - 2h)]. \quad (8.16)$$

Definition 8.12. In approximating a derivative, an FD formula is *p-th order accurate* if its error E has the form

$$E(h) = \Theta(h^p), \quad (8.17)$$

where h is the maximum distance of adjacent points in the stencil.

Exercise 8.13. Consider approximating $u'(x)$ at a point $\bar{x}$ using the nearby function values $u(\bar{x} \pm h)$. Three commonly used formulas are

$$D_+ u(\bar{x}) := \frac{u(\bar{x} + h) - u(\bar{x})}{h}, \quad (8.18)$$

$$D_- u(\bar{x}) := \frac{u(\bar{x}) - u(\bar{x} - h)}{h}, \quad (8.19)$$

$$D_0 u(\bar{x}) := \frac{u(\bar{x} + h) - u(\bar{x} - h)}{2h} = \frac{1}{2}(D_+ + D_-)u(\bar{x}). \quad (8.20)$$

For $u(x) = \sin(x)$ and $\bar{x} = 1$, calculate the errors of the above three formulas in approximating $u'(1) = \cos(1) \approx 0.5403023$ with $h = 0.01$ and 0.005. Deduce orders of accuracy of these formulas and give a geometric interpretation.

Exercise 8.14. Show that the FD formulas $D_+ u(\bar{x})$ and $D_- u(\bar{x})$ are first-order accurate while $D_0 u(\bar{x})$ is second-order accurate.

Exercise 8.15. Construct a table of divided difference (as in Definition 3.18) to derive a quadratic polynomial that agrees with $u(x)$ at $\bar{x}, \bar{x} - h$, and $\bar{x} - 2h$. Then take derivative of this polynomial to obtain the FD formula (8.16).

8.2 Errors and consistency

Definition 8.16. The *global error* or *solution error* of an FD method in Formula 8.8 is

$$\mathbf{E} = \mathbf{U} - \hat{\mathbf{U}}, \quad (8.21)$$

where $\hat{\mathbf{U}} = [u(x_1), u(x_2), \cdots, u(x_m)]^T$ is the vector of true values and $\mathbf{U}$ the computed solution.

Definition 8.17. A *grid function* is a function $\mathbf{g} : \mathbf{X} \to \mathbb{R}$ on a discrete grid $\mathbf{X}$ that contains a finite number of points.

Definition 8.18. The *q-norm of a grid function* $\mathbf{g}$ on a one-dimensional grid $\mathbf{X} := \{x_1, x_2, \ldots, x_N\}$ is

$$\|\mathbf{g}\|_q = \left(h \sum_{i=1}^{N} |g_i|^q \right)^{\frac{1}{q}}, \tag{8.22}$$

where $\mathbf{g} = (g_1, g_2, \cdots, g_N)$. In particular, the *1-norm* is

$$\|\mathbf{g}\|_1 = h \sum_{i=1}^{N} |g_i| \tag{8.23}$$

and the *max-norm* is

$$\|\mathbf{g}\|_\infty = \max_{1 \le i \le N} |g_i|. \tag{8.24}$$

Exercise 8.19. Suppose a grid function $\mathbf{g} : \mathbf{X} \to \mathbb{R}$ has $\mathbf{X} := \{x_1, x_2, \ldots, x_N\}$, $g_1 = O(h)$, $g_N = O(h)$, and $g_j = O(h^2)$ for all $j = 2, \ldots, N-1$. Show that

$$\|\mathbf{g}\|_\infty = O(h),\ \|\mathbf{g}\|_1 = O(h^2),\ \|\mathbf{g}\|_2 = O(h^{\frac{3}{2}}). \tag{8.25}$$

As the main point of this exercise, the differences in the max-norm, 1-norm, and 2-norm of a grid function often reveal the percentage of components with large magnitude.

Definition 8.20. The *local truncation error* (LTE) of an FD method in Formula 8.8 is the error caused by replacing a continuous derivative with an FD formula.

Example 8.21. When we approximate Δu with

$$D^2 u(x_j) := \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2}, \tag{8.26}$$

the LTE of the FD method in Example 8.9 is

$$\tau_j = -D^2 u(x_j) - \big(-u''(x_j)\big) = -\frac{h^2}{12} u''''(x_j) + O(h^4).$$

Lemma 8.22. The LTE of an FD method in Formula 8.8 is the error of calculating the righthand side function $\mathbf{F}$ in (8.9) by replacing U_j with the exact solution $u(x_j)$,

$$\boldsymbol{\tau} = A\hat{\mathbf{U}} - \mathbf{F}, \tag{8.27}$$

where $\hat{\mathbf{U}}$ is the vector of true solution values.

Proof. By (8.6), any component F_j of $\mathbf{F}$ is

$$F_j = f(x_j) = -\Delta u(x_j).$$

The rest of the proof follows from Definition 8.20. □

Lemma 8.23. The LTE and the global error are related as

$$A\mathbf{E} = -\boldsymbol{\tau}. \tag{8.28}$$

Proof. $A\mathbf{E} = A(\mathbf{U} - \hat{\mathbf{U}}) = \mathbf{F} - (\mathbf{F} + \boldsymbol{\tau}) = -\boldsymbol{\tau}$. □

Definition 8.24. An FD method in Formula 8.8 is said to be *consistent* with the BVP if

$$\lim_{h \to 0} \|\boldsymbol{\tau}^h\| = 0, \tag{8.29}$$

where $\boldsymbol{\tau}^h$ is the LTE.

8.3 Convergence and stability

Definition 8.25. An FD method is *convergent* if

$$\lim_{h \to 0} \left\|\mathbf{E}^h\right\| = 0, \tag{8.30}$$

where $\mathbf{E}^h$ is the solution error in Definition 8.16.

Definition 8.26. An FD method in Formula 8.8 is *stable* if

(a) $\lim_{h \to 0} A$ is invertible,

(b) $\lim_{h \to 0} \left\|A^{-1}\right\| = O(1)$.

Theorem 8.27. A consistent and stable FD method is convergent.

Proof. Lemma 8.23 and Definitions 8.24 and 8.26 yield

$$\lim \left\|\mathbf{E}^h\right\| \le \lim \left\|(A^h)^{-1}\right\| \lim \left\|\boldsymbol{\tau}^h\right\| \le C \lim \left\|\boldsymbol{\tau}^h\right\| = 0. \quad \square$$

8.3.1 Stability in the 2-norm

Definition 8.28 (Matrix norms induced by vector norms)**.** The *norm* of a matrix $A \in \mathbb{R}^{n \times n}$ is defined by

$$\begin{aligned} \|A\| &= \sup \left\{ \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0} \right\} \\ &= \sup \Big\{ \|A\mathbf{x}\| : \|\mathbf{x}\| = 1 \Big\}. \end{aligned}$$

Example 8.29. Commonly used matrix norms include

$$\begin{cases} \|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|, \\ \|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|, \\ \|A\|_2 = \sqrt{\rho(A^T A)}, \end{cases} \tag{8.31}$$

where $\rho(B) := \max_i |\lambda_i(B)|$ is the spectral radius of the matrix B, i.e. the maximum modulus of eigenvalues of B.

Lemma 8.30. The eigenvalues λ_k and eigenvectors $\mathbf{w}_k$ of the matrix A in (8.11) are

$$\lambda_k(A) = \frac{4}{h^2} \sin^2 \frac{k\pi}{2(m+1)}, \tag{8.32}$$

$$w_{k,j} = \sin \frac{jk\pi}{m+1}, \tag{8.33}$$

where $j, k = 1, 2, \ldots, m$.

Proof. It is straightforward to verify the conclusions using the trigonometric identity

$$\sin\alpha + \sin\beta = 2 \sin \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2}. \quad \square$$

Theorem 8.31. The FD method in Example 8.9 is second-order convergent in the 2-norm.

Proof. We have $\|A\|_2 = \rho(A)$ since the matrix A is symmetric. Then Lemma 8.30 yields

$$\lim_{h \to 0} \|A^{-1}\|_2 = \lim_{h \to 0} \frac{1}{\min |\lambda_p|} = \lim_{h \to 0} \frac{h^2}{4 \sin^2 \frac{\pi h}{2}} = \frac{1}{\pi^2} = O(1).$$

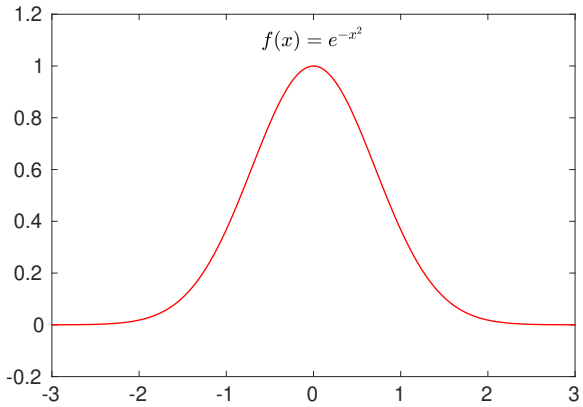
By Definition 8.26, the method is stable. The rest of the proof follows from Example 8.21, Definition 8.24, and Theorem 8.27. □

8.3.2 Gaussian and Dirac delta functions

Definition 8.32. A Gaussian function, often simply referred to as a *Gaussian*, is a function of the form

$$f(x) = a\exp\left(-\frac{(x-b)^2}{2c^2}\right), \tag{8.34}$$

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



Lemma 8.33. The integral of a Gaussian is

$$\int_{-\infty}^{+\infty} ae^{-\frac{(x-b)^2}{2c^2}}\mathrm{d}x = ac\sqrt{2\pi}. \tag{8.35}$$

Proof. By the trick of combining two one-dimensional Gaussians and the Polar coordinate transformation, we have

$$\begin{aligned}\int_{-\infty}^{+\infty} e^{-x^2}\mathrm{d}x &= \sqrt{\left(\int_{-\infty}^{+\infty} e^{-x^2}\mathrm{d}x\right)\left(\int_{-\infty}^{+\infty} e^{-y^2}\mathrm{d}y\right)} \\ &= \sqrt{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)}\,\mathrm{d}x\mathrm{d}y} \\ &= \sqrt{\int_0^{2\pi}\int_0^{+\infty} e^{-r^2} r\,\mathrm{d}r\mathrm{d}\theta} \\ &= \sqrt{2\pi\cdot\left(-\frac{1}{2}e^{-r^2}\right)\Big|_0^{+\infty}} = \sqrt{\pi},\end{aligned}$$

and hence

$$\int_{-\infty}^{+\infty} ae^{-\frac{(x-b)^2}{2c^2}}\mathrm{d}x = \sqrt{2}ac\int_{-\infty}^{+\infty} e^{-y^2}\mathrm{d}y = ac\sqrt{2\pi},$$

where it follows from the transformation of $x = b+\sqrt{2}cy$. □

Definition 8.34. A *normal distribution* or *Gaussian distribution* is a continuous probability distribution of the form

$$f_{\mu,\sigma} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{8.36}$$

where μ is the *mean* or *expectation* and σ is the *standard deviation*.

Definition 8.35. The *Dirac delta function* $\delta(x-\bar{x})$ centered at $\bar{x}$ is

$$\delta(x-\bar{x}) = \lim_{\epsilon\to 0}\phi_\epsilon(x-\bar{x}) \tag{8.37}$$

where $\phi_\epsilon(x-\bar{x}) = f_{\bar{x},\epsilon}$ is a normal distribution with its mean at $\bar{x}$ and its standard deviation as ϵ.

Lemma 8.36. The Dirac delta function satisfies

$$\delta(x-\bar{x}) = \begin{cases}+\infty, & x=\bar{x},\\ 0, & x\neq\bar{x};\end{cases} \tag{8.38a}$$

$$\int_{-\infty}^{+\infty}\delta(x-\bar{x})\mathrm{d}x = 1. \tag{8.38b}$$

Lemma 8.37 (Sifting property of δ)**.** If $f:\mathbb{R}\to\mathbb{R}$ is continuous, then

$$\int_{-\infty}^{+\infty}\delta(x-\bar{x})f(x)\mathrm{d}x = f(\bar{x}). \tag{8.39}$$

Proof. Since $I_\epsilon := [\bar{x}-\epsilon, \bar{x}+\epsilon]$ is a compact interval and $f(x)$ is continuous over I_ϵ, $f(x)$ is bounded over I_ϵ, say, $f(x)\in[m,M]$. The nonnegativeness of ϕ_ϵ and the integral mean value theorem C.64 imply that

$$(*): \qquad \int_{-\infty}^{+\infty}\delta(x-\bar{x})f(x)\mathrm{d}x = \int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon}\delta(x-\bar{x})f(x)\mathrm{d}x$$

is bounded within the interval

$$\left[m\int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon}\delta(x-\bar{x})\mathrm{d}x,\ M\int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon}\delta(x-\bar{x})\mathrm{d}x\right].$$

It follows that

$$\lim_{\epsilon\to 0}\int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon}\delta(x-\bar{x})f(x)\mathrm{d}x = f(\bar{x})\int_{\bar{x}-\epsilon}^{\bar{x}+\epsilon}\delta(x-\bar{x})\mathrm{d}x = f(\bar{x}).$$

Apply $\lim_{\epsilon\to 0}$ to $(*)$ and we have (8.39). □

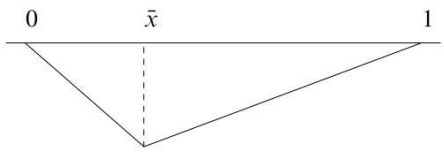
8.3.3 Green's function

Definition 8.38. For any fixed $\bar{x}\in[0,1]$, the *Green's function* $G(x;\bar{x})$ is the function of x that solves the BVP

$$\begin{cases}u''(x) = \delta(x-\bar{x}),\\ u(0) = u(1) = 0.\end{cases} \tag{8.40}$$

Lemma 8.39. The Green's function $G(x;\bar{x})$ that solves (8.40) is

$$G(x;\bar{x}) = \begin{cases}(\bar{x}-1)x, & x\in[0,\bar{x}],\\ \bar{x}(x-1), & x\in[\bar{x},1].\end{cases} \tag{8.41}$$



Proof. For any fixed ϵ, we have from (8.40) and (8.38b) that

$$\int_{x_0-\epsilon}^{x_0+\epsilon} G''(x)\mathrm{d}x = \int_{x_0-\epsilon}^{x_0+\epsilon} \delta(x-\bar{x})\mathrm{d}x = \begin{cases} 0, & \bar{x} \notin (x_0-\epsilon, x_0+\epsilon), \\ 1, & \bar{x} \in (x_0-\epsilon, x_0+\epsilon). \end{cases}$$

Take limit $\epsilon \to 0$ of the above equation and we deduce from the second fundamental theorem of calculus (Theorem C.66) that

$$\lim_{\epsilon\to 0} G'(x_0+\epsilon) - \lim_{\epsilon\to 0} G'(x_0-\epsilon) = \begin{cases} 0 & \text{if } x_0 \in (0,\bar{x}) \cup (\bar{x},1), \\ 1 & \text{if } x_0 = \bar{x}. \end{cases} \tag{8.42}$$

Substitute

$$G(x;\bar{x}) = \begin{cases} ax+b, & x \in [0,\bar{x}], \\ cx+d, & x \in [\bar{x},1] \end{cases}$$

into (8.42) and (8.40) and the continuity of $G(x;\bar{x})$ yields

$$\begin{cases} c = a+1 \\ b = 0 \\ c+d = 0 \\ a\bar{x}+b = c\bar{x}+d \end{cases} \Rightarrow \begin{cases} a = \bar{x}-1 \\ b = 0 \\ c = \bar{x} \\ d = -\bar{x} \end{cases},$$

which completes the proof. □

Corollary 8.40. The solution to the linear BVP

$$\begin{cases} u''(x) = c\delta(x-\bar{x}), \\ u(0) = u(1) = 0. \end{cases}$$

is

$$u(x) = cG(x;\bar{x}).$$

Proof. This follows directly from Lemma 8.39. □

8.3.4 Stability in the max-norm

Lemma 8.41. For the matrix A in (8.11), any element of its inverse $B = A^{-1}$ is

$$b_{ij} = -hG(x_i;x_j) = \begin{cases} -h(x_j-1)x_i, & i \le j, \\ -hx_j(x_i-1), & i \ge j. \end{cases} \tag{8.43}$$

More explicitly, the matrix B is

$$B = -h\begin{bmatrix} x_1(x_1-1) & x_1(x_2-1) & \cdots & x_1(x_m-1) \\ x_1(x_2-1) & x_2(x_2-1) & \cdots & x_2(x_m-1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(x_m-1) & x_2(x_m-1) & \cdots & x_m(x_m-1) \end{bmatrix}.$$

Proof. To verify that B is indeed the inverse of A, it suffices to multiply the ith row of h^2A and the jth column of $-\frac{1}{h}B$,

$$\begin{bmatrix} 0,\cdots,0,-1,2,-1,0,\cdots,0 \end{bmatrix} \begin{bmatrix} x_1(x_j-1) \\ \vdots \\ x_{j-1}(x_j-1) \\ x_j(x_j-1) \\ x_j(x_{j+1}-1) \\ \vdots \\ x_j(x_m-1) \end{bmatrix},$$

the only nonzero case is when $i = j$:

$$2x_j(x_j-1) - x_{j-1}(x_j-1) - x_j(x_{j+1}-1) = -h. \qquad \square$$

Theorem 8.42. The max-norm of $B = A^{-1}$ satisfies

$$\|B\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{m} |b_{ij}| \le 1. \tag{8.44}$$

Proof. Lemma 8.41 yields

$$\begin{aligned} \sum_{j=1}^{m} |b_{ij}| &= \sum_{j=1}^{i} hx_j|x_i-1| + \sum_{j=i+1}^{m} hx_i|x_j-1| \\ &\le \sum_{j=1}^{i} h\left(\frac{m}{m+1}\right)^2 + \sum_{j=i+1}^{m} h\left(\frac{m}{m+1}\right)^2 \\ &= mh\left(\frac{m}{m+1}\right)^2 = \left(\frac{m}{m+1}\right)^3 \le 1. \end{aligned}$$ □

8.4 A solution via Green's function

Lemma 8.43. Suppose $\mathcal{L}$ is an invertible linear differential operator that satisfies

$$\mathcal{L}u(x) = f(x). \tag{8.45}$$

Then we have

$$u(x) = \int G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x}, \tag{8.46}$$

where G is the Green's function satisfying

$$\mathcal{L}G(x;\bar{x}) = \delta(x-\bar{x}). \tag{8.47}$$

Proof. Multiply (8.47) by $f(\bar{x})$, integrate w.r.t. $\bar{x}$, and we have

$$\int \mathcal{L}G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x} = \int \delta(x-\bar{x})f(\bar{x})\mathrm{d}\bar{x} = f(x),$$

where the second equality follows from the sifting property of the delta function (Lemma 8.37). Therefore

$$\mathcal{L}\int G(x;\bar{x})f(\bar{x})\mathrm{d}\bar{x} = f(x),$$

which further implies (8.46) since $\mathcal{L}$ is invertible. □

Theorem 8.44. The Dirichlet BVP

$$\begin{cases} u''(x) = f(x), \\ u(0) = \alpha, u(1) = \beta \end{cases} \tag{8.48}$$

is solved by

$$u(x) = \alpha G_0(x) + \beta G_1(x) + \hat{U}(x), \tag{8.49}$$

where $G_0(x), G_1(x)$ and $\hat{U}(x)$ are defined by BVPs as follows

$$\begin{cases} G_0''(x) = 0, \\ G_0(0) = 1, G_0(1) = 0 \end{cases} \Rightarrow \quad G_0(x) = 1-x, \tag{8.50a}$$

$$\begin{cases} G_1''(x) = 0, \\ G_1(0) = 0, G_1(1) = 1 \end{cases} \Rightarrow \quad G_1(x) = x, \tag{8.50b}$$

$$\begin{cases} \hat{U}''(x) = f(x), \\ \hat{U}(0) = 0, \hat{U}(1) = 0 \end{cases} \Rightarrow \quad \hat{U}(x) = \int_0^1 f(\bar{x})G(x;\bar{x})\mathrm{d}\bar{x}. \tag{8.50c}$$

Proof. This follows from the linearity of the BVP (8.48). □

8.5 Other boundary conditions

Example 8.45. Consider the second order BVP

$$u''(x) = f(x) \quad \text{in } (0,1) \tag{8.51}$$

with mixed boundary conditions

$$u'(0) = \sigma, \quad u(1) = \beta. \tag{8.52}$$

As the crucial difference between this BVP and the BVP with pure Dirichlet conditions in Example 8.9, the value of $u(x)$ at $x = 0$ becomes an unknown to be solved for.

The first approach is to use a one-sided expression

$$\frac{U_1 - U_0}{h} = \sigma \tag{8.53}$$

to arrive at

$$A_E \mathbf{U}_E = \mathbf{F}_E \tag{8.54}$$

where

$$A_E = \frac{1}{h^2}\begin{bmatrix} -h & h & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & 0 & h^2 \end{bmatrix},$$

$$\mathbf{U}_E = \begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_{m-1} \\ U_m \\ U_{m+1} \end{bmatrix}, \qquad \mathbf{F}_E = \begin{bmatrix} \sigma \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) \\ \beta \end{bmatrix}.$$

The LTE at $x_0 = 0$ is

$$\begin{aligned} \tau_0 &= \frac{1}{h^2}(hu(x_1) - hu(x_0)) - \sigma \\ &= u'(x_0) + \frac{1}{2}hu''(x_0) + O(h^2) - \sigma \\ &= \frac{1}{2}hu''(x_0) + O(h^2), \end{aligned}$$

which is only first order accurate.

The second approach is to extend the domain with a *ghost cell* $x_{-1} = -h$ and use a central difference to obtain

$$\frac{U_1 - U_{-1}}{2h} = \sigma \tag{8.55}$$

that is second-order accurate for the LTE. We do not have any information for U_{-1}, so we want to eliminate it by

$$\frac{1}{h^2}(U_{-1} - 2U_0 + U_1) = f(x_0). \tag{8.56}$$

(8.55) and (8.56) yield

$$\frac{1}{h}(-U_0 + U_1) = \sigma + \frac{h}{2}f(x_0). \tag{8.57}$$

The resulting matrix is the same as that of the first approach in (8.54) except that the first component of $\mathbf{F}_E$ has an additional term $\frac{h}{2}f(x_0)$.

The third approach is to use U_0, U_1, and U_2 to approximate $u'(0)$ and we can get a second-order FD formula, c.f. Example 8.11,

$$-\frac{1}{h}\left(\frac{3}{2}U_0 - 2U_1 + \frac{1}{2}U_2\right) = \sigma + O(h^2). \tag{8.58}$$

This results in the linear system

$$A_F \mathbf{U}_E = \mathbf{F}_E \tag{8.59}$$

where

$$A_F = \frac{1}{h^2}\begin{bmatrix} -\frac{3}{2}h & 2h & -\frac{1}{2}h & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & 0 & h^2 \end{bmatrix}.$$

Exercise 8.46. Show that the first column of $B_E = A_E^{-1}$ contains elements that are $O(1)$.

Example 8.47. Consider the second-order BVP

$$u''(x) = f(x) \text{ in } (0,1), \tag{8.60}$$

with pure Neumann conditions

$$u'(0) = \sigma_0, \quad u'(1) = \sigma_1. \tag{8.61}$$

To ensure the existence of a solution, the following compatibility condition on $f(x)$, σ_0, and σ_1 must be satisfied:

$$\int_0^1 f(x)\mathrm{d}x = \int_0^1 u''(x)\mathrm{d}x = u'(1) - u'(0) = \sigma_1 - \sigma_0. \tag{8.62}$$

In fact, if (8.62) holds, there are an infinite number of solutions: if v is a solution of (8.60), $v + \mathbb{R}$ are also solutions.

Using procedures similar to those in Example 8.45, we can discretize (8.60) and (8.61) as

$$A_F \mathbf{U}_E = \mathbf{F}_F, \tag{8.63}$$

where

$$A_F = \frac{1}{h^2}\begin{bmatrix} -h & h & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & h & -h \end{bmatrix}, \tag{8.64}$$

$$\mathbf{F}_F = \begin{bmatrix} \sigma_0 + \frac{h}{2}f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) \\ -\sigma_1 + \frac{h}{2}f(x_{m+1}) \end{bmatrix}.$$

Lemma 8.48. The matrix A_F in (8.64) satisfies

$$\dim \mathcal{N}(A_F) = 1. \tag{8.65}$$

Proof. Clearly, $\mathbf{e} = (1, 1, \cdots, 1)^T$ is in the null space of A_F. The rest follows from the well-posedness of the BVP with mixed conditions. □

Theorem 8.49 (Solvability condition)**.** The linear system (8.63) has a solution if and only if

$$\frac{h}{2} f(x_0) + h \sum_{i=1}^{m} f(x_i) + \frac{h}{2} f(x_{m+1}) = \sigma_1 - \sigma_0. \tag{8.66}$$

Proof. The fundamental theorem of linear algebra (Theorem B.81) implies

$$\mathbb{R}^{m+2} = \mathcal{R}(A_F) \oplus \mathcal{N}(A_F^T) \tag{8.67}$$

and $\dim \mathcal{N}(A_F^T) = \dim \mathcal{N}(A_F)$. Lemma 8.48 further yields $\dim \mathcal{N}(A_F^T) = 1$. Then it is readily verified that

$$\mathcal{N}(A_F^T) = \operatorname{span}\left\{(1, h, h, \ldots, h, 1)^T\right\}.$$

For sufficiency, the above equation and (8.66) imply that $\mathbf{F}_F$ is orthogonal to $\mathcal{N}(A_F^T)$ and thus (8.67) yields $\mathbf{F}_F \in \mathcal{R}(A_F)$. Hence (8.63) must have a solution. As for necessity, the existence of a solution of (8.63) implies $\mathbf{F}_F \in \mathcal{R}(A_F)$, which, together with the above two equations, implies (8.66). □

8.6 BVPs in two dimensions

Example 8.50 (An FD method for Poisson's equation in a unit square)**.** Consider the two-dimensional BVP

$$-\frac{\partial^2}{\partial x^2} u(x, y) - \frac{\partial^2}{\partial y^2} u(x, y) = f(x, y) \tag{8.68}$$

in $\Omega := (0, 1) \times (0, 1)$ with homogeneous Dirichlet conditions

$$u(x, y)|_{\partial\Omega} = 0. \tag{8.69}$$

A uniform Cartesian grid can be generated with

$$x_i = ih, \quad y_j = jh, \quad i, j = 1, 2, \cdots, m, \tag{8.70}$$

where $h = \Delta x = \Delta y = \frac{1}{m+1}$ is the uniform grid size.

Approximate $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ separately and we have, $\forall i, j = 1, 2, \ldots, m$,

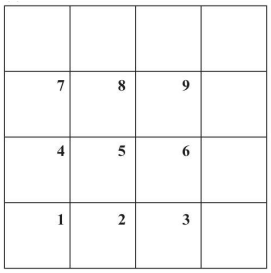
$$-\frac{U_{i-1,j} - 2U_{ij} + U_{i+1,j}}{h^2} - \frac{U_{i,j-1} - 2U_{ij} + U_{i,j+1}}{h^2} = f_{ij}. \tag{8.71}$$

These $m \times m$ equations organize into a single system

$$A_{2D}\mathbf{U} = \mathbf{F}. \tag{8.72}$$

Exercise 8.51. Show that the LTE $\boldsymbol{\tau}$ of the FD method in Example 8.50 is

$$\tau_{i,j} = -\frac{1}{12} h^2 \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right)\Big|_{(x_i, y_j)} + O(h^3). \tag{8.73}$$



Example 8.52. For $m = 3$ with ordering as shown above, we have

$$A_{2D} = \frac{1}{h^2} \begin{bmatrix} +4 & -1 & & -1 & & & & & \\ -1 & +4 & -1 & & -1 & & & & \\ & -1 & +4 & & & -1 & & & \\ -1 & & & +4 & -1 & & -1 & & \\ & -1 & & -1 & +4 & -1 & & -1 & \\ & & -1 & & -1 & +4 & & & -1 \\ & & & -1 & & & +4 & -1 & \\ & & & & -1 & & -1 & +4 & -1 \\ & & & & & -1 & & -1 & +4 \end{bmatrix}$$

$$= \frac{1}{h^2} \begin{bmatrix} T & -I & \\ -I & T & -I \\ & -I & T \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \\ U_{12} \\ U_{22} \\ U_{32} \\ U_{13} \\ U_{23} \\ U_{33} \end{bmatrix},$$

where

$$T = \begin{bmatrix} +4 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & +4 \end{bmatrix}. \tag{8.74}$$

where $\mathbf{U}$ is obtained by stacking the columns on top of each other.

Lemma 8.53. Let $\mathbf{E} = \mathbf{U} - \hat{\mathbf{U}}$ denote the global error of the linear system (8.72). Then the LTE (8.73) satisfies

$$A_{2D}\mathbf{E} = -\boldsymbol{\tau}. \tag{8.75}$$

Proof. The proof is the same as that of Lemma 8.23. □

8.6.1 Kronecker product

Definition 8.54. The *Kronecker product* of two matrices $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{p \times q}$ is another matrix $A \otimes B \in \mathbb{C}^{mp \times nq}$ given by

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}, \tag{8.76}$$

where a_{ij} is the (i, j)th element of A.

Example 8.55.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} 0 & 5 \\ 6 & 7 \end{bmatrix} = \begin{bmatrix} 0 & 5 & 0 & 10 \\ 6 & 7 & 12 & 14 \\ 0 & 15 & 0 & 20 \\ 18 & 21 & 24 & 28 \end{bmatrix}.$$

Definition 8.56. For $X \in \mathbb{C}^{m\times n}$, vec$(X)$ is defined to be a column vector of size mn made of the columns of X stacked on top of one another from left to right.

Lemma 8.57. Any $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$, and $X \in \mathbb{R}^{m\times n}$ satisfy

$$\text{vec}(AX) = (I_n \otimes A)\text{vec}(X), \tag{8.77}$$
$$\text{vec}(XB) = (B^T \otimes I_m)\text{vec}(X). \tag{8.78}$$

Proof. We have

$$\begin{aligned}\text{vec}(AX) &= \text{vec}([A\mathbf{X}_1, A\mathbf{X}_2, \cdots, A\mathbf{X}_n]) \\ &= \begin{bmatrix} A\mathbf{X}_1 \\ A\mathbf{X}_2 \\ \vdots \\ A\mathbf{X}_n \end{bmatrix} = \begin{bmatrix} A & & & \\ & A & & \\ & & \ddots & \\ & & & A \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \\ &= (I_n \otimes A) \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}.\end{aligned}$$

Let $Y = XB$, then

$$\mathbf{Y}_j = X\mathbf{b}_j \Rightarrow y_{kj} = \sum_{i=1}^{n} x_{ki}b_{ij}. \tag{8.79}$$

Let $C = B^T \otimes I_m$, then the (i,j)-th sub-block of C is

$$C_{ij} = b_{ji}I_m. \tag{8.80}$$

Let $D = C\text{vec}(X)$, then the j-th block of D is

$$\mathbf{D}_j = \sum_{i=1}^{n} C_{ji}\mathbf{X}_i = \sum_{i=1}^{n} b_{ij}I_m\mathbf{X}_i = \sum_{i=1}^{n} b_{ij}\mathbf{X}_i, \tag{8.81}$$

and the (k,j)-th entry of D is (Here we also use (k,j) to denote the scalar index corresponding to the multi-index (k,j).)

$$d_{(k,j)} = \sum_{i=1}^{n} b_{ij}x_{ki} = \sum_{i=1}^{n} x_{ki}b_{ij}, \tag{8.82}$$

Combining (8.79) and (8.82) yields (8.78). □

8.6.2 Convergence in the 2-norm

Lemma 8.58. The linear system (8.72) is equivalent to

$$AU_{m\times m} + U_{m\times m}A = F_{m\times m}, \tag{8.83}$$

where the (i,j)th element of $U_{m\times m}$ is the computed solution at the (i,j)th grid point, the (i,j)th element of $F_{m\times m}$ is

$$(F_{m\times m})_{ij} = f(ih, jh),$$

and A is the 1D discrete Laplacian in (8.11).

Proof. A direct computation gives

$$\begin{cases} (AU_{m\times m})_{ij} = \dfrac{1}{h^2}(-U_{i-1,j} + 2U_{ij} - U_{i+1,j}), \\ (U_{m\times m}A)_{ij} = \dfrac{1}{h^2}(-U_{i,j-1} + 2U_{ij} - U_{i,j+1}), \end{cases} \tag{8.84}$$

and the *homogeneous* Dirichlet condition yields

$$AU_{m\times m} + U_{m\times m}A = F_{m\times m}.$$ □

Lemma 8.59. The 1D discrete Laplacian A in (8.11) satisfies

$$\text{vec}(AU_{m\times m} + U_{m\times m}A) = (I_m \otimes A + A \otimes I_m)\text{vec}(U_{m\times m}).$$

Proof. By Lemma 8.57, we have

$$\text{vec}(AU_{m\times m}) = (I_m \otimes A)\text{vec}(U_{m\times m}),$$

and

$$\text{vec}(U_{m\times m}A) = (A^T \otimes I_m)\text{vec}(U_{m\times m}) = (A \otimes I_m)\text{vec}(U_{m\times m}),$$

where the second equality follows from the symmetry of A. Adding these two equations gives the desired result. □

Theorem 8.60. With matrix ordering, the linear system (8.72) can be written as

$$A_{2D} = I_m \otimes A + A \otimes I_m,\ \mathbf{U} = \text{vec}(U_{m\times m}),\ \mathbf{F} = \text{vec}(F_{m\times m}).$$

Proof. This follows from Lemma 8.58 and Lemma 8.59. □

Definition 8.61. The *discrete Laplacian* in n-dimensional space analogous to the 1D discrete Laplacian (8.11) is

$$A_{nD} = \sum_{j=0}^{n-1} \underbrace{I_m \otimes \cdots \otimes I_m}_{\#I_m=j} \otimes A \otimes \underbrace{I_m \otimes \cdots \otimes I_m}_{\#I_m=n-j-1}. \tag{8.85}$$

Example 8.62. For $n = 3$, we have

$$A_{3D} = A \otimes I_m \otimes I_m + I_m \otimes A \otimes I_m + I_m \otimes I_m \otimes A.$$

Theorem 8.63. The eigen-pairs of A_{2D} are

$$\lambda_{ij} = \lambda_i + \lambda_j, \quad \mathbf{W}_{ij} = \text{vec}(\mathbf{w}^i(\mathbf{w}^j)^T), \tag{8.86}$$

where $i, j = 1, 2, \ldots, m$ and $(\lambda_i, \mathbf{w}_i)$ is an eigen-pair of A in Lemma 8.30.

Proof. By Lemma 8.30, we have

$$\begin{aligned} A\mathbf{w}^i(\mathbf{w}^j)^T + \mathbf{w}^i(\mathbf{w}^j)^T A &= \lambda_i\mathbf{w}^i(\mathbf{w}^j)^T + \mathbf{w}^i(\mathbf{w}^j)^T A^T \\ &= \lambda_i\mathbf{w}^i(\mathbf{w}^j)^T + \mathbf{w}^i(A\mathbf{w}^j)^T \\ &= \lambda_i\mathbf{w}^i(\mathbf{w}^j)^T + \lambda_j\mathbf{w}^i(\mathbf{w}^j)^T \\ &= (\lambda_i + \lambda_j)\mathbf{w}^i(\mathbf{w}^j)^T. \end{aligned}$$

Then Theorem 8.60 and Lemma 8.59 yield

$$\begin{aligned} A_{2D}\text{vec}(\mathbf{w}^i(\mathbf{w}^j)^T) &= (I_m \otimes A + A \otimes I_m)\text{vec}(\mathbf{w}^i(\mathbf{w}^j)^T) \\ &= \text{vec}(A(\mathbf{w}^i(\mathbf{w}^j)^T) + (\mathbf{w}^i(\mathbf{w}^j)^T)A) \\ &= (\lambda_i + \lambda_j)\text{vec}(\mathbf{w}^i(\mathbf{w}^j)^T), \end{aligned}$$

and hence $\lambda_i + \lambda_j$ is an eigenvalue of A with corresponding eigenvector $\mathbf{W}_{ij}$. □

Theorem 8.64. The FD method in Example 8.50 is second-order convergent in the 2-norm.

Proof. We have $\|A_{2D}\|_2 = \rho(A_{2D})$ since A_{2D} is symmetric. Then Theorem 8.63 yields

$$\lim_{h\to 0} \|A_{2D}^{-1}\|_2 = \lim_{h\to 0} \frac{1}{\min|\lambda_{ij}|} = \lim_{h\to 0} \frac{h^2}{8\sin^2\frac{\pi h}{2}} = \frac{1}{2\pi^2} = O(1).$$

By Definition 8.26, the method is stable. The proof is completed by (8.73), Definition 8.24, Theorem 8.27, and Lemma 8.53. □

8.6.3 Convergence in the max-norm via a discrete maximum principle

Theorem 8.65. The FD method in Example 8.50 is second-order convergent in the max-norm.

Proof. Define a *comparison function* ϕ as

$$\phi(x,y) := \left(x-\frac{1}{2}\right)^2 + \left(y-\frac{1}{2}\right)^2 \tag{8.87}$$

and write $\phi_{i,j} = \phi(ih, jh)$. By Example 8.50, A_{2D} can be expressed componentwise as

$$\begin{aligned} A_{2D}U_{i,j} &:= (A_{2D}\mathbf{U})_{i,j} \\ &= \frac{1}{h^2}(4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1}). \end{aligned} \tag{8.88}$$

(8.87) and (8.88) yield

$$\forall i,j = 1,2,\ldots,m, \quad A_{2D}\phi_{i,j} = -4. \tag{8.89}$$

Let $E_{i,j}$ denote the solution error at the grid points (ih, jh) where $i,j = 0,1,\cdots,m+1$. Write $\tau_m := \|\boldsymbol{\tau}\|_\infty$. For

$$\psi_{i,j} := E_{i,j} + \frac{1}{4}\tau_m\phi_{i,j}, \tag{8.90}$$

we have, at each grid point (x_i, y_j) in (8.70),

$$A_{2D}\psi_{i,j} = A_{2D}E_{i,j} + \frac{1}{4}\tau_m A_{2D}\phi_{i,j} = -\tau_{i,j} - \tau_m \le 0.$$

By (8.88), $A_{2D}\psi_{i,j} \le 0$ dictates that $\psi_{i,j}$ be smaller than at least one of its neighbors $U_{i+1,j}$, $U_{i-1,j}$, $U_{i,j+1}$, and $U_{i,j-1}$. Therefore, the maximum value of ψ must occur at a boundary point, i.e. a point with $i = 0, m+1$ or $j = 0, m+1$. Consequently, there exists some constant $C > 0$ such that

$$E_{i,j} \le \psi_{i,j} \le \frac{1}{8}\tau_m < Ch^2,$$

where the first step follows from (8.90) and $\tau_m\phi_{i,j} \ge 0$, the second step from $E_{i,j}$ being zero at all boundary points and the fact that the maximum of ϕ is $\frac{1}{2}$ at the domain corners, and the last step from Exercise 8.51.

By similar arguments, the function

$$\chi_{i,j} := -E_{i,j} + \frac{1}{4}\tau_m\phi_{i,j} \tag{8.91}$$

satisfies $A_{2D}\chi_{i,j} \le 0$ for all grid points and thus

$$-E_{i,j} \le \chi_{i,j} \le \frac{1}{8}\tau_m < Ch^2.$$

To sum up, we have $|E_{i,j}| = O(h^2)$ for all grid points. □

Notation 6. Consider discretizing Poisson's equation in Example 8.4 on domain Ω. Denote by $\mathbf{X}_{\partial\Omega}$ the set of *boundary points* where values of the unknown function u are prescribed by Dirichlet conditions. Let $\mathbf{X}$ denote the set of grid points such that the grid function $\mathbf{X} \to \mathbb{R}$ is the numerical approximation of $u : \Omega \to \mathbb{R}$. The set of *interior points* is defined as

$$\mathbf{X}_\Omega = \mathbf{X} \setminus \mathbf{X}_{\partial\Omega}. \tag{8.92}$$

Lemma 8.66 (Discrete maximum principle)**.** Suppose that an FD discretization of (8.4) yields

$$\forall P \in \mathbf{X}_\Omega, \quad L_hU_P - f_P + g_P = 0, \tag{8.93}$$

where f_P corresponds to the RHS of (8.4), g_P corresponds to all boundary data other than Dirichlet conditions, and L_h and $\mathbf{X}_\Omega$ satisfy

(DMP-1) for each interior point $P \in \mathbf{X}_\Omega$, L_h is of the form

$$L_hU_P = c_PU_P - \sum_Q c_QU_Q, \tag{8.94}$$

where c_P and all c_Q's are positive and the sum is taken over all neighboring grid points of P;

(DMP-2) $\forall P \in \mathbf{X}_\Omega$, $c_P \ge \sum_Q c_Q$;

(DMP-3) $\mathbf{X}_\Omega$ is *connected*, i.e., $\forall P_1, P_{m+1} \in \mathbf{X}_\Omega$, there exists a sequence of points P_1, P_2, …, P_{m+1} such that, $\forall r = 1,2,\ldots,m$, both U_{P_r} and $U_{P_{r+1}}$ appear in some equation (8.93);

(DMP-4) at least one equation (8.94) involves a boundary value U_Q given by a Dirichlet condition.

Then any grid function $\psi : \mathbf{X} \to \mathbb{R}$ satisfying

$$\forall P \in \mathbf{X}_\Omega, \quad L_h\psi_P \le 0 \tag{8.95}$$

cannot attain its nonnegative maximum at an interior point, i.e.,

$$\max_{P\in\mathbf{X}} \psi_P \ge 0 \;\Rightarrow\; \max_{P\in\mathbf{X}_\Omega} \psi_P \le \max_{Q\in\mathbf{X}_{\partial\Omega}} \psi_Q. \tag{8.96}$$

Proof. Suppose $M_\Omega > M_{\partial\Omega}$, i.e., the interior maximum at some interior point P is greater than the boundary maximum. Then we have

$$M_\Omega = U_P \le \frac{1}{c_P}\sum_Q c_QU_Q \le \frac{1}{c_P}\sum_Q c_QM_\Omega \le M_\Omega,$$

where the first inequality follows from (8.95) and (8.94), the second from the definition of M_Ω, and the third from (DMP-2) and $M_\Omega \ge 0$. For the above equation to hold, we must have $U_Q = U_P$ for each Q. By (DMP-3), U takes the same value M_Ω on all interior points. Then (DMP-4) implies $M_\Omega = M_{\partial\Omega}$, which contradicts the starting point $M_\Omega > M_{\partial\Omega}$. □

Theorem 8.67. Suppose the discretization (8.93) of Poisson's equation (8.4) satisfies the conditions (DMP-1,2,3,4) in Lemma 8.66. Then the solution error $E_P := U_P - u(P)$ of the FD method (8.93) is bounded by

$$\forall P \in \mathbf{X}, \quad |E_P| \le T_{\max}\left(\max_{Q\in\mathbf{X}_{\partial\Omega}} \phi(Q)\right), \tag{8.97}$$

where $T_{\max} = \max_{P\in\mathbf{X}_\Omega}|T_P|$, T_P is the LTE at P satisfying $L_hE_P = -T_P$, and $\phi : \mathbf{X} \to \mathbb{R}$ is a nonnegative grid function satisfying

$$\forall P \in \mathbf{X}_\Omega, \quad L_h\phi_P \le -1. \tag{8.98}$$

Proof. $L_h E_P = -T_P$ and (8.98) yield

$$\psi_P := E_P + T_{\max}\phi_P \;\Rightarrow\; L_h\psi_P \le -T_P - T_{\max} \le 0.$$

Then we have

$$\begin{aligned} E_P &\le \max_{P\in\mathbf{X}_\Omega} (E_P + T_{\max}\phi_P) \\ &\le \max_{Q\in\mathbf{X}_{\partial\Omega}} (E_Q + T_{\max}\phi_Q) = T_{\max} \max_{Q\in\mathbf{X}_{\partial\Omega}} (\phi_Q), \end{aligned}$$

where the first step follows from $T_{\max}\phi_P \ge 0$, the second from Lemma 8.66, and the third from the fact that $E_Q = 0$ at each boundary point, c.f. Notation 6. Repeat the above arguments with $\psi_P = -E_P + T_{\max}\phi_P$ and we have

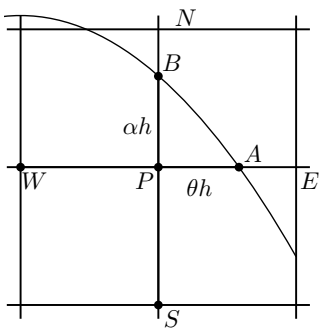
$$-E_P \le T_{\max} \max_{Q\in\mathbf{X}_{\partial\Omega}} (\phi_Q). \qquad \square$$

8.6.4 Convergence on irregular domains

Example 8.68 (An FD method for Poisson's equation in 2D irregular domains)**.** Consider the BVP

$$-\frac{\partial^2}{\partial x^2}u(x,y) - \frac{\partial^2}{\partial y^2}u(x,y) = f(x,y) \tag{8.99}$$

in a 2D irregular domain Ω with Dirichlet conditions.



An interior point is said to be *regular* if the standard 5-point stencil is applicable; otherwise it is *irregular*. For an irregular point, we modify the FD discretization in Example 8.50 to incorporate the info of local geometry and Dirichlet conditions. For example, in the above plot, the discrete operator becomes

$$L_h U_P := \frac{(1+\theta)U_P - U_A - \theta U_W}{\frac{1}{2}\theta(1+\theta)h^2} + \frac{(1+\alpha)U_P - U_B - \alpha U_S}{\frac{1}{2}\alpha(1+\alpha)h^2}. \tag{8.100}$$

Together with the equations of the form (8.71) at regular interior points, the equations of the form $L_h U_P - f_P = 0$ form a linear system. However, a global analysis of this linear system is difficult.

Exercise 8.69. Show that, in Example 8.68, the LTE at an irregular interior point is $O(h)$ while the LTE at a regular interior point is $O(h^2)$.

Theorem 8.70. Suppose that, in the notation of Theorem 8.67, the set $\mathbf{X}_\Omega$ of interior points is partitioned as

$$\mathbf{X}_\Omega = \mathbf{X}_1 \cup \mathbf{X}_2, \quad \mathbf{X}_1 \cap \mathbf{X}_2 = \emptyset,$$

the nonnegative function $\phi : \mathbf{X} \to \mathbb{R}$ satisfies

$$\forall P \in \mathbf{X}_1, \quad L_h\phi_P \le -C_1 < 0; \tag{8.101a}$$
$$\forall P \in \mathbf{X}_2, \quad L_h\phi_P \le -C_2 < 0, \tag{8.101b}$$

and the LTE of (8.93) satisfy

$$\forall P \in \mathbf{X}_1, \quad |T_P| < T_1; \tag{8.102a}$$
$$\forall P \in \mathbf{X}_2, \quad |T_P| < T_2. \tag{8.102b}$$

Then the solution error $E_P := U_P - u(P)$ of the FD method (8.93) is bounded by

$$\forall P \in \mathbf{X}, \quad |E_P| \le \left(\max_{Q\in\mathbf{X}_{\partial\Omega}} \phi(Q)\right) \max_{P\in\mathbf{X}_\Omega} \left\{\frac{T_1}{C_1}, \frac{T_2}{C_2}\right\}.$$

Exercise 8.71. Prove Theorem 8.70 by choosing a function ψ to which Lemma 8.66 applies.

Theorem 8.72. The FD method in Example 8.68 is second-order convergent in the max-norm.

Proof. Define a comparison function ϕ as

$$\phi(x,y) := \begin{cases} F_1\left[(x-p)^2 + (y-q)^2\right] & \text{if } (x,y) \in \mathbf{X}_\Omega; \\ F_1\left[(x-p)^2 + (y-q)^2\right] + F_2 & \text{if } (x,y) \in \mathbf{X}_{\partial\Omega}, \end{cases}$$

where (p,q) is the geometric center of Ω and $F_1, F_2 > 0$ are constants to be chosen later. For a regular point Q, we have

$$L_h\phi_Q = -4F_1.$$

As for an irregular point P shown in Example 8.68, the coefficient of U_A is

$$-\frac{2}{\theta(1+\theta)h^2} < -\frac{1}{h^2}$$

because $\theta \in (0,1)$. Hence we have

$$L_h\phi_P < -4F_1 - \frac{1}{h^2}F_2 < -\frac{1}{h^2}F_2.$$

By Exercise 8.69, we write the maximum LTEs on regular and irregular points as $T_1 = K_1h^2$ and $T_2 = K_2h$, respectively. Then Theorem 8.70 implies

$$|E_P| \le (F_1R^2 + F_2)\max\left\{\frac{K_1h^2}{4F_1}, \frac{K_2h^3}{F_2}\right\},$$

where R is the maximum distance of a point in Ω to the geometric center of Ω. The RHS of the above equation is minimized when we choose $\frac{F_1}{F_2} = \frac{K_1}{4K_2h}$ so that the two terms in max{} equal. It follows that

$$\forall P \in \mathbf{X}, \quad |E_P| \le \frac{1}{4}K_1R^2h^2 + K_2h^3. \qquad \square$$

Chapter 9

Multigrid Methods

9.1 The model problem

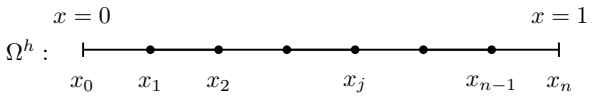
Definition 9.1. The *model problem* for our exposition of multigrid methods is the one-dimensional Poisson equation with homogeneous boundary condition

$$\begin{cases} -\Delta u = f & \text{in } \Omega := (0,1); \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{9.1}$$

Example 9.2. As a special case of Example 8.9 with $\alpha = 0$, $\beta = 0$, and $m = n - 1$, our discretization of (9.1) yields a linear system

$$A\mathbf{u} = \mathbf{f}, \tag{9.2}$$

where the unit interval Ω is discretized by uniform grid size $h = \frac{1}{n}$ into n cells with cell boundaries at $x_j = jh = \frac{j}{n}$ for $j = 0, 1, \ldots, n$, the knowns $f_j = f(x_j)$ and the unknowns u_j are located at the internal nodes x_j with $j = 1, 2, \ldots, n-1$.



The matrix $A \in \mathbb{R}^{(n-1)\times(n-1)}$ is the same as that in (8.11), i.e., a Toeplitz matrix given by

$$a_{ij} = \begin{cases} \frac{2}{h^2} & \text{if } i = j; \\ -\frac{1}{h^2} & \text{if } i - j = \pm 1; \\ 0, & \text{otherwise.} \end{cases} \tag{9.3}$$

By Lemma 8.30, the eigenvalues and eigenvectors of A are

$$\lambda_k(A) = \frac{4}{h^2}\sin^2\frac{k\pi}{2n} = \frac{4}{h^2}\sin^2\frac{kh\pi}{2}, \tag{9.4}$$

$$w_{k,j} = \sin\frac{jk\pi}{n} = \sin(x_j k\pi), \tag{9.5}$$

where $j, k = 1, 2, \ldots, n-1$.

9.1.1 The residual equation

Definition 9.3. The *error of a multigrid method* is

$$\mathbf{e} = \mathbf{u} - \tilde{\mathbf{u}}, \tag{9.6}$$

where $\mathbf{u}$ is the vector of exact solutions and $\tilde{\mathbf{u}}$ that of computed solutions.

Definition 9.4. The *residual of an approximate solution* $\tilde{\mathbf{u}} \approx \mathbf{u}$ to (9.2) is $\mathbf{r} = \mathbf{f} - A\tilde{\mathbf{u}}$.

Lemma 9.5. The error and the residual of an approximate solution $\tilde{u}_j$ satisfy the *residual equation*

$$A\mathbf{e} = \mathbf{r}. \tag{9.7}$$

Proof. This follows from Definition 9.4, in the same way that Lemma 8.23 follows from Lemma 8.22. □

Definition 9.6. The *condition number of a matrix* B is

$$\text{cond}(B) := \|B\|_2\|B^{-1}\|_2. \tag{9.8}$$

Theorem 9.7. The relative error of an approximate solution is bounded by its relative residual.

$$\frac{1}{\text{cond}(A)}\frac{\|\mathbf{r}\|_2}{\|\mathbf{f}\|_2} \le \frac{\|\mathbf{e}\|_2}{\|\mathbf{u}\|_2} \le \text{cond}(A)\frac{\|\mathbf{r}\|_2}{\|\mathbf{f}\|_2}. \tag{9.9}$$

Exercise 9.8. Prove Theorem 9.7.

Exercise 9.9. Compute by hand the values of $\text{cond}(A)$ for $n = 8$ and $n = 1024$.

9.1.2 Fourier modes on Ω^h

Notation 7. Ω^h denotes the uniform grid of n intervals that discretizes the problem domain Ω. Occasionally we also abuse the notation to mean the corresponding vector space of continuous grid functions $\{\Omega^h \to \mathbb{R}\}$.

Definition 9.10. The *wavelength of a sinusoidal function* is the distance of one sinusoidal period. The *wavenumber of a sinusoidal function* k is the number of half sinusoidal waves in unit distance.

Lemma 9.11. The kth Fourier mode with its jth component as $w_{k,j} = \sin(x_j k\pi)$ has wavelength $L = \frac{2}{k}$.

Proof. By Definition 9.10, $\sin(x_j k\pi) = -\sin(x_j + \frac{L}{2})k\pi$ implies $x_j k\pi = (x_j + \frac{L}{2})k\pi - \pi$. Hence $k = \frac{2}{L}$. □

Exercise 9.12. For $\Omega = (0,1)$, plot to show that the maximum wavenumber that is representable on Ω^h is $n_{\max} = \frac{1}{h}$. What if we require that the Fourier mode be 0 at the boundary points?

Proof. Alternate from local maximum and local minimum at the $n+1$ grid points and we have $n_{\max} = \frac{1}{h}$. When homogeneous Dirichlet conditions are imposed on the boundary points, the maximum number of alternation between local extrema is reduced by one. Hence we have $n_{\max} = \frac{1}{h} - 1$. □

Lemma 9.13 (Aliasing)**.** For $k \in (n, 2n)$ on Ω^h, the Fourier mode $\mathbf{w}_k$ of which the jth component is $w_{k,j} = \sin(x_j k\pi)$ is actually represented as the additive inverse of the mode $\mathbf{w}_{k'}$ where $k' = 2n - k$.

Proof. It is readily verified that

$$\begin{aligned}\sin(x_j k\pi) &= -\sin(2j\pi - x_j k\pi) = -\sin(x_j(2n-k)\pi)\\ &= -\sin(x_j k'\pi) = -w_{k',j}.\end{aligned}$$ □

Example 9.14. According to Lemma 9.13, the mode with $k = \frac{3}{2}n$ is represented by $k = \frac{1}{2}n$.

Exercise 9.15. Plot the case of $n = 6$ for Example 9.14.

Definition 9.16. On Ω^h, the Fourier modes with wavenumbers $k \in [1, \frac{n}{2})$ are called the *low-frequency* (LF) or *smooth* modes, those with $k \in [\frac{n}{2}, n)$ the *high-frequency* (HF) or *oscillatory* modes.

9.2 Classical iterative methods

Definition 9.17. The *fixed point iteration* for solving a linear system such as $A\mathbf{u} = \mathbf{f}$ in (9.2) is an iteration of the form

$$\mathbf{u}^{(\ell+1)} = T\mathbf{u}^{(\ell)} + \mathbf{c}, \tag{9.10}$$

where T and $\mathbf{c}$ are functions of A and $\mathbf{f}$ such that $\mathbf{u} = T\mathbf{u}+\mathbf{c}$.

Lemma 9.18. ℓ times of fixed point iteration yield

$$\mathbf{e}^{(\ell)} = T^\ell \mathbf{e}^{(0)}. \tag{9.11}$$

The iteration converges to 0 iff the spectral radius $\rho(T) < 1$.

Exercise 9.19. Prove Lemma 9.18.

Definition 9.20. The *Jacobi method* is a fixed point iteration in which the matrix A is split as $A = D + L + U$ with D, L, U being the diagonal, lower triangular, and upper triangular part of A, respectively, and in which the iteration (9.10) is given by

$$T = -D^{-1}(L+U), \qquad \mathbf{c} = D^{-1}\mathbf{f}. \tag{9.12}$$

Example 9.21. The iteration matrix of the Jacobi method $T = -D^{-1}(L+U)$ is given by

$$t_{ij} = \begin{cases} \frac{1}{2} & \text{if } i-j = \pm 1;\\ 0 & \text{otherwise.}\end{cases} \tag{9.13}$$

Definition 9.22. The *weighted Jacobi method* is a fixed point iteration of the form

$$\mathbf{u}^* = T\mathbf{u}^{(\ell)} + \mathbf{c} \tag{9.14a}$$

$$\mathbf{u}^{(\ell+1)} = (1-\omega)\mathbf{u}^{(\ell)} + \omega\mathbf{u}^*, \tag{9.14b}$$

where T and $\mathbf{c}$ are given in (9.12).

Lemma 9.23. The weighted Jacobi has the iteration matrix

$$T_\omega = (1-\omega)I - \omega D^{-1}(L+U) = I - \frac{\omega h^2}{2}A, \tag{9.15}$$

whose eigenvectors are the same as those of A, with the corresponding eigenvalues as

$$\lambda_k(T_\omega) = 1 - 2\omega \sin^2 \frac{k\pi}{2n}, \tag{9.16}$$

where $k = 1, 2, \ldots, n-1$.

Exercise 9.24. Prove Lemma 9.23.

Exercise 9.25. Write a Matlab program to reproduce Fig. 2.7 in the book by Briggs et al. [2000]. For $n = 64$, $\omega \in [0, 1]$, verify $\rho(T_\omega) \geq 0.9986$ and hence slow convergence.

Definition 9.26. The *smoothing factor* μ is the maximal factor of damping for HF modes. An iterative method is said to have the *smoothing property* if μ is small and independent of the grid size.

Example 9.27. The smoothing factor of the weighted Jacobi is determined by the optimization problem

$$\mu = \min_{\omega\in(0,1]} \max_{k\in[\frac{n}{2},n)} |\lambda_k(T_\omega)|. \tag{9.17}$$

Since $\lambda_k(T_\omega)$ is a monotonically decreasing function, the minimum is obtained by setting $\lambda_{\frac{n}{2}}(T_\omega) = -\lambda_n(T_\omega)$, which implies $\omega = \frac{2}{3}$. Consequently we have $|\lambda_k| \leq \mu = \frac{1}{3}$.

Exercise 9.28. Write a Matlab program to reproduce Figure 2.8 in the book by Briggs et al. [2000], verifying that regular Jacobi is only good for damping modes $16 \leq k \leq 48$. In contrast, for $\omega = \frac{2}{3}$, the modes $16 \leq k < 64$ are all damped out quickly.

9.3 Key elements of multigrid

9.3.1 Restriction and prolongation

Lemma 9.29. The kth LF mode on Ω^h becomes the kth mode (LF or HF) on Ω^{2h}:

$$w^h_{k,2j} = w^{2h}_{k,j}. \tag{9.18}$$

LF modes $k \in [\frac{n}{4}, \frac{n}{2})$ of Ω^h will become HF modes on Ω^{2h}.

Proof. It is readily verified that

$$w^h_{k,2j} = \sin\frac{2jk\pi}{n} = \sin\frac{jk\pi}{\frac{n}{2}} = w^{2h}_{k,j}, \tag{9.19}$$

where $k \in [1, \frac{n}{2})$. Because of the smaller range of k on Ω^{2h}, the modes with $k \in [\frac{n}{4}, \frac{n}{2})$ are HF by definition since the highest wavenumber is $\frac{n}{2}$ on Ω^{2h}. □

Definition 9.30. The *restriction* operator

$$I^{2h}_h : \mathbb{R}^{n-1} \to \mathbb{R}^{\frac{n}{2}-1}$$

maps a vector on the fine grid Ω^h to its counterpart on the coarse grid Ω^{2h}:

$$I^{2h}_h v^h = v^{2h}. \tag{9.20}$$

Definition 9.31. The *full-weighting* operator is a restriction operator given by

$$v_j^{2h} = \frac{1}{4}\left(v_{2j-1}^h + 2v_{2j}^h + v_{2j+1}^h\right), \tag{9.21}$$

where $j = 1, 2, \ldots, \frac{n}{2} - 1$.

Example 9.32. For $n = 8$, the full-weighting operator is

$$I_h^{2h} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 & & & & \\ & & 1 & 2 & 1 & & \\ & & & & 1 & 2 & 1 \end{bmatrix}. \tag{9.22}$$

Definition 9.33. The *prolongation or interpolation* operator

$$I_{2h}^h : \mathbb{R}^{\frac{n}{2}-1} \to \mathbb{R}^{n-1}$$

maps a vector on the coarse grid Ω^{2h} to its counterpart on the fine grid Ω^h:

$$I_{2h}^h v^{2h} = v^h. \tag{9.23}$$

Definition 9.34. The *linear interpolation* operator is a prolongation operator given by

$$\begin{array}{rcl} v_{2j}^h & = & v_j^{2h}, \\ v_{2j+1}^h & = & \frac{1}{2}(v_j^{2h} + v_{j+1}^{2h}). \end{array} \tag{9.24}$$

Example 9.35. For $n = 8$, the linear interpolation operator is

$$I_{2h}^h = \frac{1}{2}\begin{bmatrix} 1 & & \\ 2 & & \\ 1 & 1 & \\ & 2 & \\ & 1 & 1 \\ & & 2 \\ & & 1 \end{bmatrix}. \tag{9.25}$$

9.3.2 Two-grid correction

Definition 9.36. The *two-grid correction scheme*

$$v^h \leftarrow \mathtt{TG}(v^h, f^h, \nu_1, \nu_2) \tag{9.26}$$

solves $A\mathbf{u} = \mathbf{f}$ in (9.2) via steps as follows.

(TG-1) Relax $A^h\mathbf{u}^h = f^h$ for ν_1 times on Ω^h with initial guess v^h: $v^h \leftarrow T_\omega^{\nu_1} v^h + \mathbf{c}'(f)$,

(TG-2) Compute the fine-grid residual $r^h = f^h - A^h v^h$ and restrict it to the coarse grid by $r^{2h} = I_h^{2h} r^h$: $r^{2h} \leftarrow I_h^{2h}(f^h - A^h v^h)$,

(TG-3) Solve $A^{2h}e^{2h} = r^{2h}$ on Ω^{2h}: $e^{2h} \leftarrow (A^{2h})^{-1} r^{2h}$,

(TG-4) Interpolate the coarse-grid error to the fine grid by $e^h = I_{2h}^h e^{2h}$ and correct the fine-grid approximation: $v^h \leftarrow v^h + I_{2h}^h e^{2h}$,

(TG-5) Relax $A^h\mathbf{u}^h = f^h$ for ν_2 times on Ω^h with initial guess v^h: $v^h \leftarrow T_\omega^{\nu_2} v^h + \mathbf{c}'(f)$.

Lemma 9.37. Acting on the error vector, the iteration matrix of the two-grid correction scheme (9.26) is

$$TG = T_\omega^{\nu_2}\left[I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right] T_\omega^{\nu_1}. \tag{9.27}$$

Proof. By Definition 9.36, the residual on the fine grid is

$$r^h(v^h) = f^h - A^h\left(T_\omega^{\nu_1} v^h + \mathbf{c}'(f)\right).$$

The two-grid correction scheme with $\nu_2 = 0$ replaces the initial guess with

$$v^h \leftarrow T_\omega^{\nu_1} v^h + \mathbf{c}'(f) + I_{2h}^h (A^{2h})^{-1} I_h^{2h} r^h(v^h)$$

which also holds for the exact solution u^h

$$u^h \leftarrow T_\omega^{\nu_1} u^h + \mathbf{c}'(f) + I_{2h}^h (A^{2h})^{-1} I_h^{2h} r^h(u^h).$$

Subtracting the two equations yields

$$e^h \leftarrow T_\omega^{\nu_1} e^h - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h T_\omega^{\nu_1} e^h.$$

Similar arguments applied to step (TG-5) yield (9.27). □

9.3.3 Multigrid cycles

Definition 9.38. The *V-cycle scheme*

$$\mathbf{v}^h \leftarrow \mathtt{VC}^h(\mathbf{v}^h, \mathbf{f}^h, \nu_1, \nu_2) \tag{9.28}$$

solves $A\mathbf{u} = \mathbf{f}$ in (9.2) via steps as follows.

(VC-1) Relax ν_1 times on $A^h\mathbf{u}^h = \mathbf{f}^h$ with a given initial guess $\mathbf{v}^h$,

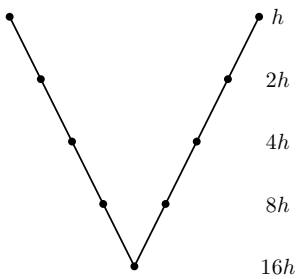
(VC-2) If Ω^h is the coarsest grid, go to (VC-4), otherwise

$$\begin{array}{rl} \mathbf{f}^{2h} & \leftarrow I_h^{2h}(\mathbf{f}^h - A^h\mathbf{v}^h), \\ \mathbf{v}^{2h} & \leftarrow \mathbf{0}, \\ \mathbf{v}^{2h} & \leftarrow \mathtt{VC}^{2h}(\mathbf{v}^{2h}, \mathbf{f}^{2h}, \nu_1, \nu_2). \end{array}$$

(VC-3) Interpolate error back and correct the solution:

$$\mathbf{v}^h \leftarrow \mathbf{v}^h + I_{2h}^h \mathbf{v}^{2h}.$$

(VC-4) Relax ν_2 times on $A^h\mathbf{u}^h = \mathbf{f}^h$ with the initial guess as $\mathbf{v}^h$.



Lemma 9.39. In a D-dimensional domain with $n = 2^m$ cells ($m \in \mathbb{N}^+$) along each dimension, the storage cost of V-cycles is

$$2n^{\mathrm{D}}\left(1 + 2^{-\mathrm{D}} + 2^{-2\mathrm{D}} + \cdots + 2^{-m\mathrm{D}}\right) < \frac{2n^{\mathrm{D}}}{1 - 2^{-\mathrm{D}}}. \tag{9.29}$$

Let WU denote the computational cost of performing one relaxation sweep on the finest grid. After neglecting the intergrid transfer, the computational cost of a single V-cycle with $\nu_1 = \nu_2 = 1$ is

$$2\mathrm{WU}\left(1 + 2^{-\mathrm{D}} + 2^{-2\mathrm{D}} + \cdots + 2^{-m\mathrm{D}}\right) < \frac{2}{1 - 2^{-\mathrm{D}}}\mathrm{WU}. \tag{9.30}$$

Proof. On each grid, both vectors of errors and residuals must be stored, and this justifies the factor of 2 in (9.29); the rest of (9.29) follows from Definition 9.38. A similar argument yields (9.30). □

Definition 9.40. The *full multigrid V-cycle*

$$\mathbf{v}^h \leftarrow \texttt{FMG}^h(\mathbf{f}^h, \nu_1, \nu_2) \tag{9.31}$$

solves $A\mathbf{u} = \mathbf{f}$ in (9.2) via steps as follows.

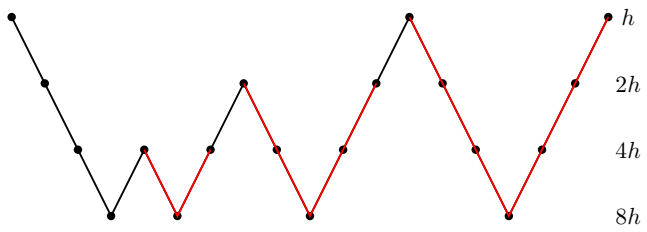
(FMG-1) If Ω^h is the coarsest grid, set $\mathbf{v}^h \leftarrow \mathbf{0}$ and go to (FMG-3), otherwise

$$\begin{aligned} \mathbf{f}^{2h} &\leftarrow I_h^{2h}\mathbf{f^h}, \\ \mathbf{v}^{2h} &\leftarrow \texttt{FMG}^{2h}(\mathbf{f}^{2h}, \nu_1, \nu_2). \end{aligned}$$

(FMG-2) Correct $\mathbf{v}^h \leftarrow I_{2h}^h \mathbf{v}^{2h}$.

(FMG-3) Perform a V-cycle with the initial guess as $\mathbf{v}^h$:

$$\mathbf{v}^h \leftarrow \texttt{VC}^h(\mathbf{v}^h, \mathbf{f}^h, \nu_1, \nu_2).$$



Exercise 9.41. Show that, for $\nu_0 = \nu_1 = \nu_2 = 1$, the computational cost of an FMG cycle is less than $\frac{2}{(1-2^{-\mathrm{D}})^2}$ WU. Give upper bounds as tight as possible for computational costs of an FMG cycle for $\mathrm{D} = 1, 2, 3$.

9.4 Why multigrid methods work?

9.4.1 The spectral picture

Definition 9.42. A Fourier mode $\mathbf{w}_k^h$ with $k \in [1, \frac{n}{2})$ and the mode $\mathbf{w}_{k'}^h$ with $k' = n - k$ are called *complementary modes* on Ω^h.

Lemma 9.43. A pair of complementary modes satisfy

$$w_{k',j}^h = (-1)^{j+1} w_{k,j}^h. \tag{9.32}$$

Proof. This follows from

$$w_{k',j}^h = \sin\frac{(n-k)j\pi}{n} = \sin\left(j\pi - \frac{kj\pi}{n}\right) = (-1)^{j+1} w_{k,j}^h.$$

□

Lemma 9.44. The action of the full-weighting operator on a pair of complementary modes on Ω^h is

$$I_h^{2h}\mathbf{w}_k^h = c_k \mathbf{w}_k^{2h} := \cos^2\frac{k\pi}{2n}\mathbf{w}_k^{2h}, \tag{9.33a}$$

$$I_h^{2h}\mathbf{w}_{k'}^h = -s_k \mathbf{w}_k^{2h} := -\sin^2\frac{k\pi}{2n}\mathbf{w}_k^{2h}, \tag{9.33b}$$

where $k \in [1, \frac{n}{2})$, $k' = n - k$. In addition, $I_h^{2h}\mathbf{w}_{\frac{n}{2}}^h = \mathbf{0}$.

Proof. Recall that $\frac{4}{h^2}s_k$ is the eigenvalue of A^h. For a smooth mode, we have

$$\begin{aligned} \left(I_h^{2h}\mathbf{w}_k^h\right)_j &= \frac{1}{4}\sin\frac{(2j-1)k\pi}{n} + \frac{1}{2}\sin\frac{2jk\pi}{n} + \frac{1}{4}\sin\frac{(2j+1)k\pi}{n} \\ &= \frac{1}{2}\left(1+\cos\frac{k\pi}{n}\right)\sin\frac{2jk\pi}{n} = \cos^2\frac{k\pi}{2n} w_{k,j}^{2h}, \end{aligned}$$

where the last step follows from Lemma 9.29. (9.33b) can be proved by similar steps with k replaced with $n-k$. □

Lemma 9.45. The action of the interpolation operator on Ω^{2h} is

$$I_{2h}^h \mathbf{w}_k^{2h} = c_k \mathbf{w}_k^h - s_k \mathbf{w}_{k'}^h, \tag{9.34}$$

where $k' = n - k$.

Proof. Lemma 9.43 and trigonometric identities yield

$$\begin{aligned} c_k w_{k,j}^h - s_k w_{k',j}^h &= \left(\cos^2\frac{k\pi}{2n} + (-1)^j \sin^2\frac{k\pi}{2n}\right) w_{k,j}^h \\ &= \begin{cases} w_{k,j}^h & \text{if } j \text{ is even;} \\ \cos\frac{k\pi}{n} w_{k,j}^h & \text{if } j \text{ is odd.} \end{cases} \end{aligned}$$

On the other hand, by Definition 9.34, we have

$$\left(I_{2h}^h \mathbf{w}_k^{2h}\right)_j = \begin{cases} w_{k,j}^h, & \text{if } j \text{ is even,} \\ \frac{1}{2}\sin\frac{k\pi(j-1)}{n} + \frac{1}{2}\sin\frac{k\pi(j+1)}{n} & \text{if } j \text{ is odd,} \end{cases}$$

where last expression simplifies to $\cos\frac{k\pi}{n} w_{k,j}^h$. □

Theorem 9.46. The two-grid correction operator is invariant on the subspace $W_k^h = \mathrm{span}\{\mathbf{w}_k^h, \mathbf{w}_{k'}^h\}$.

$$TG\mathbf{w}_k = \lambda_k^{\nu_1+\nu_2} s_k \mathbf{w}_k + \lambda_k^{\nu_1}\lambda_{k'}^{\nu_2} s_k \mathbf{w}_{k'} \tag{9.35a}$$

$$TG\mathbf{w}_{k'} = \lambda_{k'}^{\nu_1}\lambda_k^{\nu_2} c_k \mathbf{w}_k + \lambda_{k'}^{\nu_1+\nu_2} c_k \mathbf{w}_{k'}, \tag{9.35b}$$

where λ_k is the eigenvalue of T_ω.

Proof. Consider first the case of $\nu_1 = \nu_2 = 0$.

$$A^h\mathbf{w}_k^h = \frac{4s_k}{h^2}\mathbf{w}_k^h \tag{9.36a}$$

$$\Rightarrow I_h^{2h}A^h\mathbf{w}_k^h = \frac{4c_k s_k}{h^2}\mathbf{w}_k^{2h} \tag{9.36b}$$

$$\Rightarrow (A^{2h})^{-1} I_h^{2h} A^h \mathbf{w}_k^h = \frac{4c_k s_k}{h^2}\frac{(2h)^2}{4\sin^2\frac{k\pi}{n}}\mathbf{w}_k^{2h} = \mathbf{w}_k^{2h} \tag{9.36c}$$

$$\Rightarrow -I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h \mathbf{w}_k^h = -c_k\mathbf{w}_k^h + s_k\mathbf{w}_{k'}^h \tag{9.36d}$$

$$\Rightarrow \left[I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right]\mathbf{w}_k^h = s_k\mathbf{w}_k^h + s_k\mathbf{w}_{k'}^h. \tag{9.36e}$$

Similarly, we have

$$A^h\mathbf{w}_{k'}^h = \frac{4s_{k'}}{h^2}\mathbf{w}_{k'}^h = \frac{4c_k}{h^2}\mathbf{w}_{k'}^h \tag{9.37a}$$

$$\Rightarrow I_h^{2h}A^h\mathbf{w}_{k'}^h = -\frac{4c_k s_k}{h^2}\mathbf{w}_k^{2h} \tag{9.37b}$$

$$\Rightarrow (A^{2h})^{-1} I_h^{2h} A^h \mathbf{w}_{k'}^h = -\frac{4c_k s_k}{h^2}\frac{(2h)^2}{4\sin^2\frac{k\pi}{n}}\mathbf{w}_k^{2h} = -\mathbf{w}_k^{2h} \tag{9.37c}$$

$$\Rightarrow -I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h \mathbf{w}_{k'}^h = c_k\mathbf{w}_k^h - s_k\mathbf{w}_{k'}^h \tag{9.37d}$$

$$\Rightarrow \left(I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right)\mathbf{w}_{k'}^h = c_k\mathbf{w}_k^h + c_k\mathbf{w}_{k'}^h, \tag{9.37e}$$

where $c_k = s_{k'}$ is applied in (9.37a).

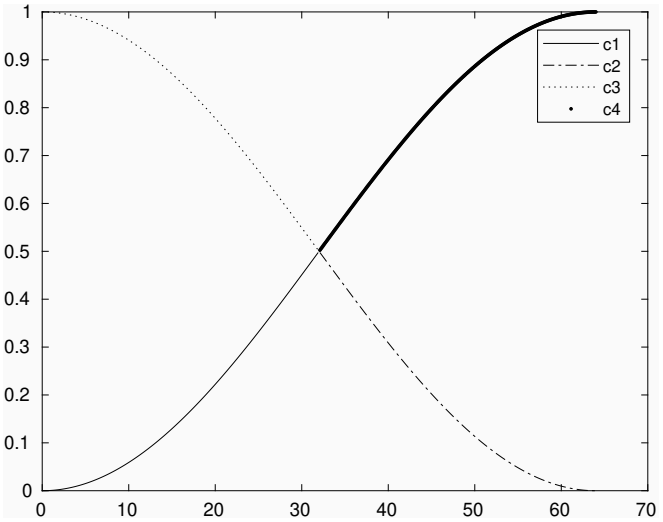
Adding pre-smoothing incurs a scaling of $\lambda_k^{\nu_1}$ for (9.36e) and $\lambda_{k'}^{\nu_1}$ for (9.37e). In contrast, adding post-smoothing incurs a scaling of $\lambda_k^{\nu_2}$ for $\mathbf{w}_k^h$ and a scaling of $\lambda_{k'}^{\nu_2}$ for $\mathbf{w}_{k'}^h$ in both (9.36e) and (9.37e). Hence (9.35) holds. □

Exercise 9.47. Rewrite (9.35) as

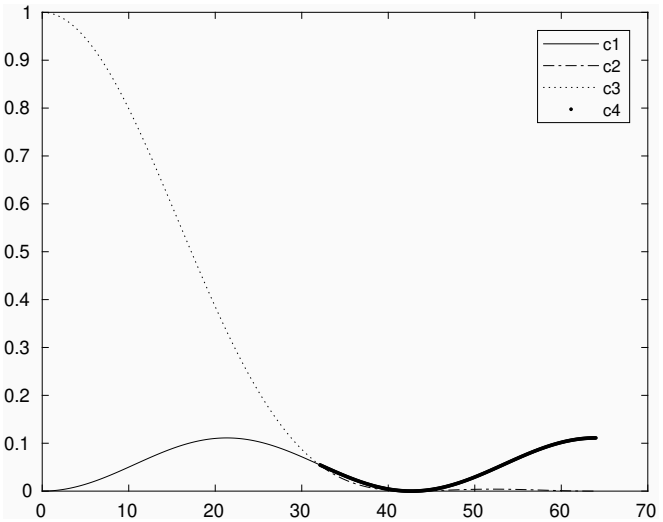
$$TG\begin{bmatrix}\mathbf{w}_k\\ \mathbf{w}_{k'}\end{bmatrix} = \begin{bmatrix}\lambda_k^{\nu_1+\nu_2}s_k & \lambda_k^{\nu_1}\lambda_{k'}^{\nu_2}s_k\\ \lambda_{k'}^{\nu_1}\lambda_k^{\nu_2}c_k & \lambda_{k'}^{\nu_1+\nu_2}c_k\end{bmatrix}\begin{bmatrix}\mathbf{w}_k\\ \mathbf{w}_{k'}\end{bmatrix} = \begin{bmatrix}c_1 & c_2\\ c_3 & c_4\end{bmatrix}\begin{bmatrix}\mathbf{w}_k\\ \mathbf{w}_{k'}\end{bmatrix}. \tag{9.38}$$

Explain why the magnitude of all four c_i's are small. Reproduce the following plots of the damping coefficients of two-grid correction with weighted Jacobi for $n = 64$ and $\omega = \frac{2}{3}$. The x-axis represents the wave number k.

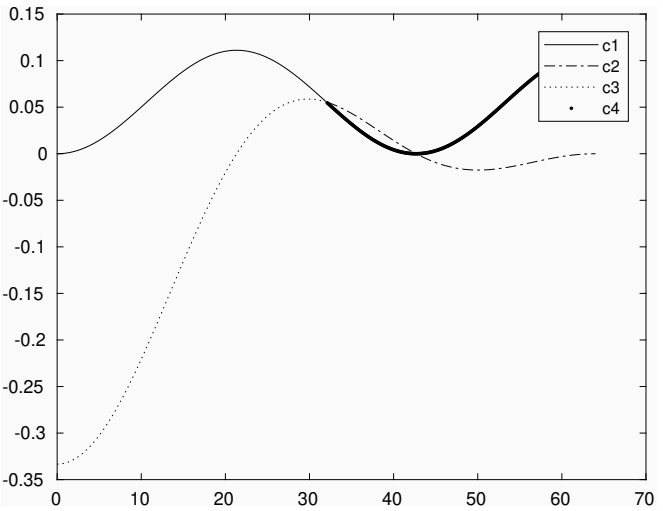
$\nu_1 = 0$, $\nu_2 = 0$.



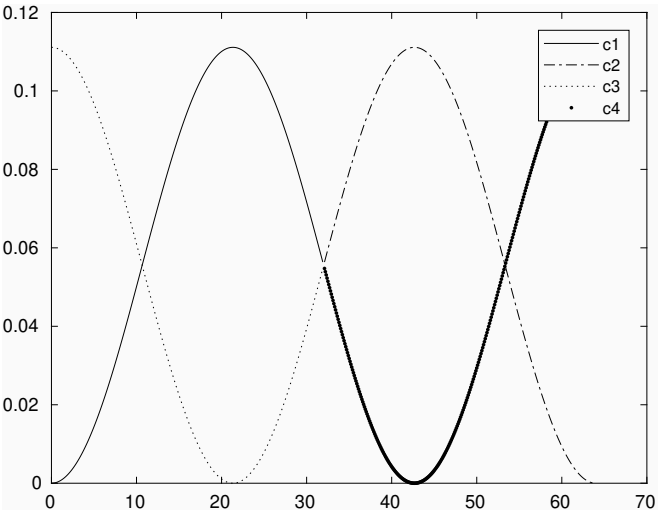
$\nu_1 = 0$, $\nu_2 = 2$.



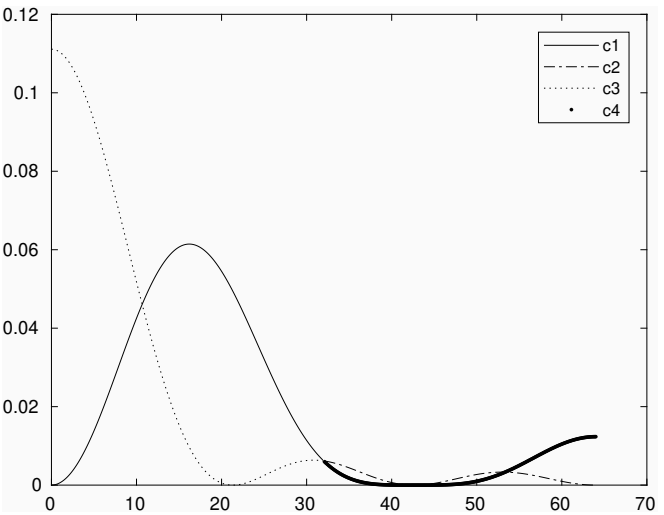
$\nu_1 = 1$, $\nu_2 = 1$.



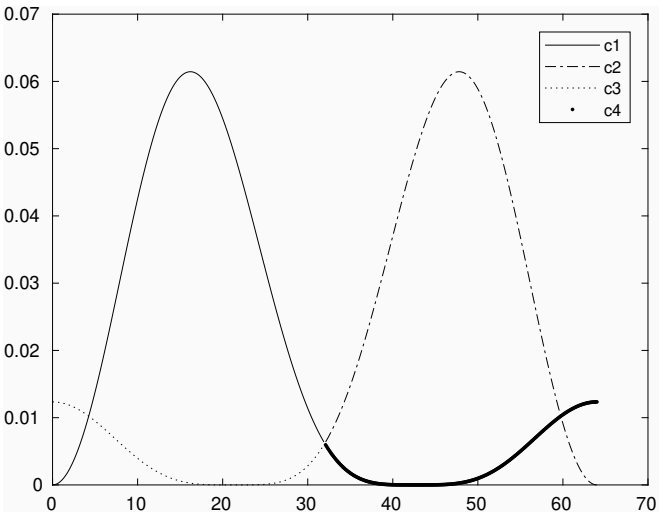
$\nu_1 = 2$, $\nu_2 = 0$.



$\nu_1 = 2$, $\nu_2 = 2$.



$\nu_1 = 4$, $\nu_2 = 0$.



Hint: It is tricky to plot the coefficients defined in (9.38), especially in Matlab. Since c_2, c_4 act on HF modes, one has to ensure that the components in the vectors s_k and c_k indeed correspond to those in $\mathbf{w}_{k'}$. If s_k and c_k are computed from an increasing order of the frequencies, then their components will have to be reversed for plotting. Physical intuition helps in this case: c_1 and c_4 should form one curve while c_2 and c_3 should form another.

9.4.2 The algebraic picture

Lemma 9.48. The full-weighting operator and the linear-interpolation operator satisfy the *variational properties*

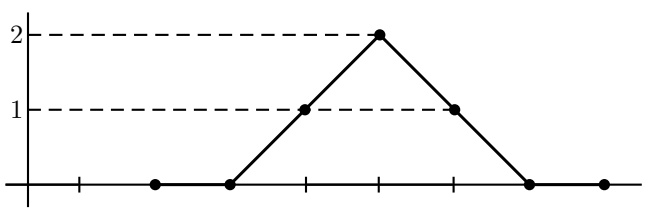
$$I_{2h}^h = c(I_h^{2h})^T,\ c \in \mathbb{R}^+. \tag{9.39a}$$

$$I_h^{2h} A^h I_{2h}^h = A^{2h}. \tag{9.39b}$$

(9.39b) is also called the *Galerkin condition*.

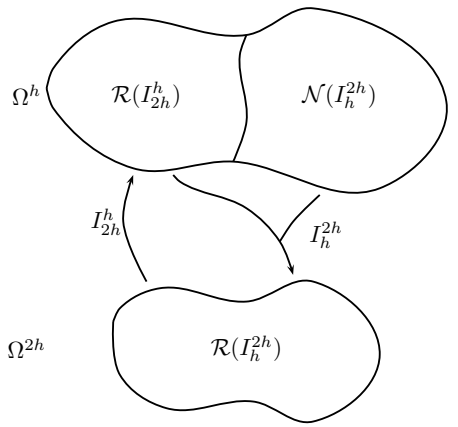
Lemma 9.49. A basis for the range of the interpolation operator $\mathcal{R}(I_{2h}^h)$ is given by its columns, hence dim $\mathcal{R}(I_{2h}^h) = \frac{n}{2} - 1$. Its null space $\mathcal{N}(I_{2h}^h) = \{\mathbf{0}\}$.

Proof. $\mathcal{R}(I_{2h}^h) = \{I_{2h}^h v^{2h} : v^{2h} \in \Omega^{2h}\}$. The maximum dimension of $\mathcal{R}(I_{2h}^h)$ is thus $\frac{n}{2} - 1$. Any v^{2h} can be expressed as $v^{2h} = \sum v_j^{2h} \mathbf{e}_j^{2h}$. It is obvious that the columns of I_{2h}^h are linearly independent. □



Lemma 9.50. The full-weighting operator satisfies

$$\dim \mathcal{R}(I_h^{2h}) = \frac{n}{2} - 1, \qquad \dim \mathcal{N}(I_h^{2h}) = \frac{n}{2}. \tag{9.40}$$



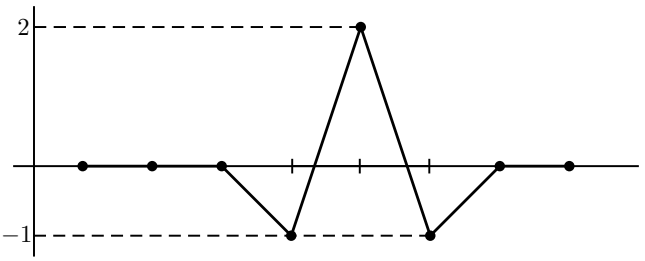
Exercise 9.51. Prove Lemma 9.50.

Lemma 9.52. A basis for the null space of the full-weighting operator is given by

$$\mathcal{N}(I_h^{2h}) = \text{span}\{A^h \mathbf{e}_j^h : j \text{ is odd}\}, \tag{9.41}$$

where $\mathbf{e}_j^h$ is the jth unit vector on Ω^h.

Proof. Consider $I_h^{2h} A^h$. The jth row of I_h^{2h} has $2(j-1)$ leading zeros and the next three nonzero entries are $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. Since the bandwidth of A^h is 3, it suffices to consider only five columns of A^h for potentially non-zero dot-product $\sum_i (I_h^{2h})_{ji} (A^h)_{ik}$. For $2j \pm 1$, these dot products are zero; for $2j$, the dot product is $\frac{1}{2}$; for $2j \pm 2$, the dot product is $-\frac{1}{4}$. Hence for any odd j, we have $I_h^{2h} A^h \mathbf{e}_j^h = \mathbf{0}$. □



Theorem 9.53. The null space of the two-grid correction operator (without relaxation) is the range of interpolation:

$$\mathcal{N}(TG) = \mathcal{R}(I_{2h}^h). \tag{9.42}$$

Proof. If $\mathbf{s}^h \in \mathcal{R}(I_{2h}^h)$, then $\mathbf{s}^h = I_{2h}^h \mathbf{q}^{2h}$.

$$TG\mathbf{s}^h = \left[I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right] I_{2h}^h \mathbf{q}^{2h} = \mathbf{0},$$

where the last step comes from (9.39b). Hence we have $\mathcal{R}(I_{2h}^h) \subseteq \mathcal{N}(TG)$. By Lemma 9.52, $\mathbf{t}^h \in \mathcal{N}(I_h^{2h} A^h)$ implies $\mathbf{t}^h = \sum_{j \text{ is odd}} t_j \mathbf{e}_j$. Consequently, we have

$$TG\mathbf{t}^h = \left[I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h\right] \mathbf{t}^h = \mathbf{t}^h,$$

i.e., TG is the identity operator when acting on $\mathcal{N}(I_h^{2h} A^h)$. Hence the dimension of $\mathcal{N}(TG)$ is no greater than the dimension of $\mathcal{R}(I_h^{2h} A^h)$, which is the same as dim $\mathcal{R}(I_{2h}^h)$ since A^h is a bijection with full rank on $\mathbb{R}^{n-1}$. This implies that dim$\mathcal{N}(TG) \le$ dim$\mathcal{R}(I_{2h}^h)$, which completes the proof. □

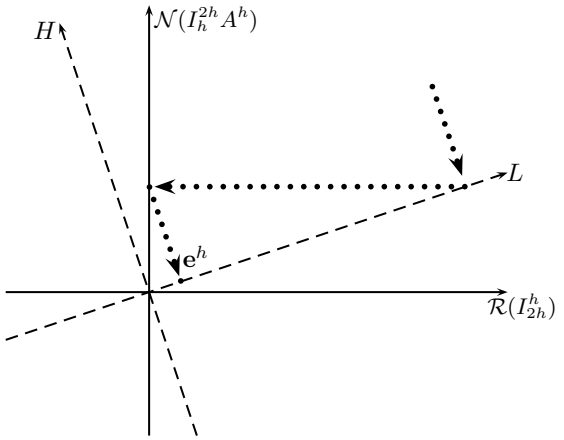
Definition 9.54. Let A be an $n \times n$ symmetric positive definite matrix. The *A-inner product* or *energy inner product* of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is defined as

$$(\mathbf{u}, \mathbf{v})_A := (A\mathbf{u}, \mathbf{v}), \tag{9.43}$$

where $(\cdot, \cdot)$ is the Euclidean inner product on $\mathbb{R}^n$. Naturally, the *A-norm* or *energy norm* is defined as

$$\|\mathbf{u}\|_A := \sqrt{(\mathbf{u}, \mathbf{u})_A}. \tag{9.44}$$

Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are *A-orthogonal* iff $(\mathbf{u}, \mathbf{v})_A = 0$.



9.4.3 The optimal complexity of FMG

Definition 9.55. Denote the errors of computed results from exact solutions are

$$E_i^h = v_i^h - u(x_i) = v_i^h - u_i^h + u_i^h - u(x_i).$$

The *discretization error* is the error $u_i^h - u(x_i)$ incurred by truncating the Taylor series of exact values and the *algebraic error* is the error $v_i^h - u_i^h$ incurred by inexact solution of the linear system.

Lemma 9.56. When interpolating errors from a coarse grid to the fine grid, we have

$$\left\|\mathbf{v}^{2h} - \mathbf{u}^{2h}\right\|_{A^{2h}} = c\left\|I_{2h}^h \mathbf{v}^{2h} - I_{2h}^h \mathbf{u}^{2h}\right\|_{A^h}. \tag{9.45}$$

where $c \in \mathbb{R}^+$.

Proof. Definition 9.54 and Lemma 9.48 yield

$$\begin{aligned}
&\|\mathbf{v}^{2h} - \mathbf{u}^{2h}\|_{A^{2h}} \\
=&\left(A^{2h}(\mathbf{v}^{2h} - \mathbf{u}^{2h}), \mathbf{v}^{2h} - \mathbf{u}^{2h}\right) \\
=&\left(I_h^{2h} A^h I_{2h}^h (\mathbf{v}^{2h} - \mathbf{u}^{2h}), \mathbf{v}^{2h} - \mathbf{u}^{2h}\right) \\
=&\left(A^h I_{2h}^h (\mathbf{v}^{2h} - \mathbf{u}^{2h}), c I_{2h}^h (\mathbf{v}^{2h} - \mathbf{u}^{2h})\right) \\
=&c\|I_{2h}^h \mathbf{v}^{2h} - I_{2h}^h \mathbf{u}^{2h}\|_{A^h}.
\end{aligned}$$

□

Lemma 9.57. An FMG cycle reduces the algebraic error from $O(1)$ to $O(h^p)$, i.e.,

$$\|\mathbf{e}^h\|_{A^h} \leq Kh^p, \tag{9.46}$$

where p is the order of accuracy of the discrete Laplacian.

Theorem 9.58. With a pth-order FD discretization on Ω^h, the FMG solves the model problem in Definition 9.1 in $O(\frac{1}{h})$ time.

Chapter 10

Parabolic Problems

10.1 Parabolic equations

Definition 10.1. A second-order, constant-coefficient, linear partial differential equation (PDE) of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \tag{10.1}$$

is called a *parabolic PDE* if its coefficients satisfy

$$B^2 - 4AC = 0. \tag{10.2}$$

Definition 10.2. The *one-dimensional heat equation* is a parabolic PDE of the form

$$u_t = \nu u_{xx} \text{ in } \Omega := (0,1) \times (0,T), \tag{10.3}$$

where $x \in (0,1)$ is the spatial location, $t \in (0,T)$ the time and $\nu > 0$ the dynamic viscosity; the equation has to be supplemented with an *initial condition*

$$u(x,0) = \eta(x), \text{ on } (0,1) \times \{0\} \tag{10.4}$$

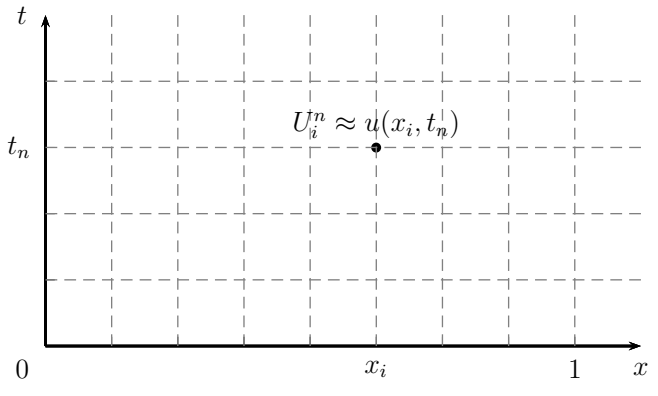
and appropriate boundary conditions at $\{0,1\} \times (0,T)$.

10.2 The method of lines (MOL)

Notation 8. The space-time domain of the PDE (10.3) can be discretized by the rectangular grids

$$x_i = ih, \quad t_n = nk, \tag{10.5}$$

$h = \frac{1}{m+1}$ is the uniform mesh spacing and $k = \Delta t$ is the uniform time-step size. The unknowns U_i^n are located at nodes (x_i, t_n).



Definition 10.3. The *method of lines* (MOL) is a technique for solving PDEs via

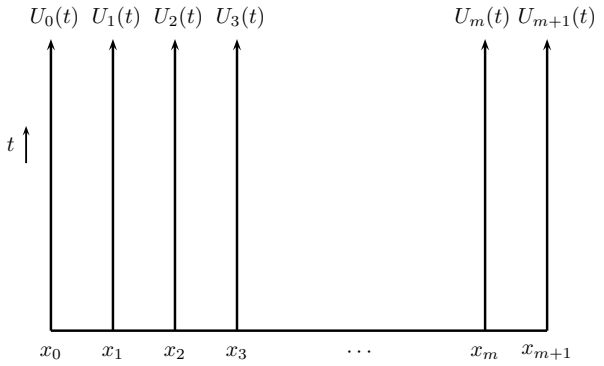
(a) discretizing the spatial derivatives while leaving the time variable continuous;

(b) solving the resulting ODEs with a numerical method designed for IVPs.

Example 10.4. Discretize the heat equation (10.3) in space at grid point x_i by

$$U_i'(t) = \frac{\nu}{h^2}\Big(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)\Big), \tag{10.6}$$

where $U_i(t) \approx u(x_i, t)$ for $i = 1, 2, \cdots, m$.



For Dirichlet conditions

$$\begin{cases} u(0,t) = g_0(t), & \text{on } \{0\} \times (0,T); \\ u(1,t) = g_1(t), & \text{on } \{1\} \times (0,T), \end{cases} \tag{10.7}$$

this semi-discrete system (10.6) can be written as

$$\mathbf{U}'(t) = A\mathbf{U}(t) + g(t), \tag{10.8}$$

where

$$A = \frac{\nu}{h^2}\begin{bmatrix} -2 & +1 & & & & \\ +1 & -2 & +1 & & & \\ & +1 & -2 & +1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & +1 & -2 & +1 \\ & & & & +1 & -2 \end{bmatrix}, \tag{10.9}$$

$$\mathbf{U}(t) := \begin{bmatrix} U_1(t) \\ U_2(t) \\ U_3(t) \\ \vdots \\ U_{m-1}(t) \\ U_m(t) \end{bmatrix}, \quad g(t) = \frac{\nu}{h^2} \begin{bmatrix} g_0(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}. \tag{10.10}$$

Definition 10.5. The *FTCS (forward in time, centered in space) method* solves the heat equation (10.3) by

$$\frac{U_i^{n+1} - U_i^n}{k} = \frac{\nu}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n), \tag{10.11}$$

or, equivalently

$$U_i^{n+1} = U_i^n + 2r(U_{i-1}^n - 2U_i^n + U_{i+1}^n), \tag{10.12}$$

where $r := \frac{k\nu}{2h^2}$.

Example 10.6. For homogeneous Dirichlet boundary conditions, the FTCS method can be written as

$$\mathbf{U}^{n+1} = (I + kA)\mathbf{U}^n, \tag{10.13}$$

where A is the matrix in (10.9) and

$$\mathbf{U}^n := \begin{bmatrix} U_1^n \\ U_2^n \\ \vdots \\ U_m^n \end{bmatrix}. \tag{10.14}$$

Definition 10.7. The *Crank-Nicolson method* solves the heat equation (10.3) by

$$\begin{aligned} \frac{U_i^{n+1} - U_i^n}{k} &= \frac{1}{2}\Big(f(U_i^n, t_n) + f(U_i^{n+1}, t_{n+1})\Big) \\ &= \frac{\nu}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}), \end{aligned} \tag{10.15}$$

or, equivalently

$$\begin{aligned} &- rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} \\ &= rU_{i-1}^n + (1-2r)U_i^n + rU_{i+1}^n. \end{aligned} \tag{10.16}$$

Exercise 10.8. Show that the matrix form of the Crank-Nicolson method for solving the heat equation (10.3) with Dirichlet conditions is

$$\left(I - \frac{k}{2}A\right)\mathbf{U}^{n+1} = \left(I + \frac{k}{2}A\right)\mathbf{U}^n + \mathbf{b}^n, \tag{10.17}$$

where $r = \frac{k\nu}{2h^2}$ and

$$\mathbf{b}^n = r \begin{bmatrix} g_0(t_n) + g_0(t_{n+1}) \\ 0 \\ \vdots \\ 0 \\ g_1(t_n) + g_1(t_{n+1}) \end{bmatrix}.$$

10.3 Accuracy and consistency

Definition 10.9. The *local truncation error (LTE) of an MOL* for solving a PDE is the error caused by replacing continuous derivatives with finite difference formulas.

Example 10.10. The LTE of the FTCS method in Definition 10.5 is

$$\begin{aligned} \tau(x,t) =& \frac{u(x,t+k) - u(x,t)}{k} \\ &- \frac{\nu}{h^2}\Big(u(x-h,t) - 2u(x,t) + u(x+h,t)\Big) \\ =& \Big(u_t + \frac{1}{2}ku_{tt} + \frac{1}{6}k^2u_{ttt} + \cdots\Big) \\ &- \nu\Big(u_{xx} + \frac{1}{12}h^2u_{xxxx} + \cdots\Big) \\ =& \Big(\frac{1}{2}k\nu^2 - \frac{\nu}{12}h^2\Big)u_{xxxx} + O(k^2 + h^4), \end{aligned}$$

where the first step follows from the Definition 7.74, the second from Taylor expansions and the last from $u_t = \nu u_{xx}$ and $u_{tt} = \nu u_{xxt} = \nu u_{txx} = \nu^2 u_{xxxx}$. Due to $\tau(x,t) = O(k + h^2)$, this method is said to be second order accurate in space and first order accurate in time.

Exercise 10.11. Show that the Crank-Nicolson method in Definition 10.7 is second order accurate in both space and time by calculating its LTE as

$$\tau(x,t) = O(k^2 + h^2).$$

Definition 10.12. An MOL is said to be *consistent* if

$$\lim_{k,h\to 0} \tau(x,t) = 0. \tag{10.18}$$

Definition 10.13. The *solution error* of an MOL is

$$E_i^n = U_i^n - u(x_i, t_n), \tag{10.19}$$

where $u(x_i, t_n)$ is the exact solution of the PDE at the grid point (x_i, t_n).

10.4 Stability

Lemma 10.14. The eigenvalues λ_p and eigenvectors $\mathbf{w}^p$ of A in (10.9) are

$$\lambda_p = -\frac{4\nu}{h^2}\sin^2\Big(\frac{p\pi h}{2}\Big), \tag{10.20}$$

$$w_j^p = \sin(p\pi jh), \tag{10.21}$$

where $p, j = 1, 2, \cdots, m$ and $h = \frac{1}{m+1}$.

Proof. This follows directly from Lemma 8.30. □

Example 10.15. For the FTCS method (10.11) to be absolutely stable, we must have $|1 + k\lambda| \le 1$ for each eigenvalue in (10.20), which implies $-2 \le -4\nu k/h^2 \le 0$ and thus limits the time-step size to

$$k \le \frac{h^2}{2\nu}. \tag{10.22}$$

Definition 10.16. An MOL is said to be *unconditionally stable* for a PDE if in solving the semi-discrete system of the PDE its ODE solver is absolutely stable for any $k > 0$.

Lemma 10.17. Suppose the ODE solver of the MOL is $A(\alpha)$-stable for the semi-discrete system that results from spatially discretizing the heat equation. Then the MOL is unconditionally stable for the heat equation.

Proof. The RAS of an $A(\alpha)$-stable method contains the negative real axis. All eigenvalues of the heat equations are negative real numbers, hence $k\lambda$ is in the RAS for any $k > 0$. □

Corollary 10.18. The Crank-Nicolson method (10.16) is unconditionally stable for the heat equation.

Proof. The ODE solver of the Crank-Nicolson method (10.16) is the trapezoidal rule, which is A-stable and hence $A(\alpha)$-stable. The proof is completed by Lemma 10.17. □

Definition 10.19. A linear MOL of the form

$$\mathbf{U}^{n+1} = B(k)\mathbf{U}^n + \mathbf{b}^n(k) \tag{10.23}$$

is *Lax-Richtmyer stable* if

$$\forall T > 0,\ \exists C_T > 0, \quad \forall k > 0, \forall n \in \mathbb{N}^+ \text{ satisfying } nk \le T, \ \|B(k)^n\| \le C_T. \tag{10.24}$$

Definition 10.20. A linear MOL (10.23) is said to have *strong stability* if

$$\|B\|_2 \le 1. \tag{10.25}$$

Corollary 10.21. The Crank-Nicolson method has strong stability with

$$B = \Big(I - \frac{k}{2}A\Big)^{-1}\Big(I + \frac{k}{2}A\Big). \tag{10.26}$$

Proof. (10.26) follows directly from Exercise 10.8. The symmetry of A implies the symmetry of B and thus the spectral radius of B satisfies

$$\rho(B) = \max\left|\frac{1 + k\lambda_p/2}{1 - k\lambda_p/2}\right| \le 1.$$

Then the proof is completed by Definition 10.20. □

10.5 Convergence

Theorem 10.22 (Lax Equivalence Theorem)**.** A consistent linear MOL (10.23) is convergent if and only if it is Lax-Richtmyer stable.

Proof. We only prove the sufficiency. Apply the numerical method (10.23) to the exact solution $\hat{\mathbf{U}}^n$ and we obtain

$$(*) \qquad \hat{\mathbf{U}}^{n+1} = B\hat{\mathbf{U}}^n + \mathbf{b}^n + k\tau^n,$$

where the dependence on k has been suppressed for clarity and where

$$\hat{\mathbf{U}}^n := \begin{bmatrix} u(x_1, t_n) \\ u(x_2, t_n) \\ \vdots \\ u(x_m, t_n) \end{bmatrix}, \quad \tau^n := \begin{bmatrix} \tau(x_1, t_n) \\ \tau(x_2, t_n) \\ \vdots \\ \tau(x_m, t_n) \end{bmatrix}.$$

Subtracting (*) from (10.23) gives the difference equation for the global error $E^n = \mathbf{U}^n - \hat{\mathbf{U}}^n$:

$$E^{n+1} = BE^n - k\tau^n,$$

and hence, by induction,

$$E^N = B^N E^0 - k\sum_{n=1}^{N} B^{N-n}\tau^{n-1},$$

from which we have

$$\left\|E^N\right\| \le \left\|B^N\right\| \left\|E^0\right\| + k\sum_{n=1}^{N} \left\|B^{N-n}\right\| \left\|\tau^{n-1}\right\|.$$

If the method is Lax-Richtmyer stable, then for $Nk < T$, we have

$$\left\|E^N\right\| \le C_T \left\|E^0\right\| + kN \cdot C_T \max_{1\le n\le N} \left\|\tau^{n-1}\right\|,$$

the RHS goes to 0 as $k \to 0$ and $h \to 0$. □

Corollary 10.23. The Crank-Nicolson method is convergent for any $k > 0$.

Proof. This follows from Theorem 10.22 and Corollary 10.21. □

Example 10.24. For the FTCS method, (10.13) implies

$$B = I + kA \tag{10.27}$$

and thus the convergence depends on

$$\rho(B) \le 1 + O(k),$$

which is a form of Lax-Richtmyer stability.

Exercise 10.25. Prove the necessity part of Theorem 10.22.

10.6 Von Neumann analysis

Theorem 10.26. The exact solution to the heat equation (10.3) with Dirichlet conditions $g_0(t) = g_1(t) = 0$ is

$$u(x,t) = \sum_{j=0}^{\infty} \hat{u}_j(t)\sin(\pi j x), \tag{10.28}$$

where

$$\hat{u}_j(t) = \exp(-j^2\pi^2\nu t)\hat{u}_j(0), \tag{10.29}$$

and $\hat{u}_j(0)$ is determined as the Fourier coefficients of the initial data $\eta(x)$.

Proof. It is straightforward to verify that (10.28) is indeed the solution of (10.3). □

Example 10.27. Consider the FTCS method. To apply von Neumann analysis we consider how this method works on a single wave number ξ, i.e., we set

$$U_j^n = [g(\xi)]^n e^{ix_j\xi}. \tag{10.30}$$

Then we expect that

$$U_j^{n+1} = g(\xi)U_j^n, \tag{10.31}$$

where $g(\xi)$ is the amplification factor for this wave number. Inserting these expressions into (10.12) gives

$$g(\xi)U_j^n = \Big[1 + \frac{\nu k}{h^2}\Big(e^{-i\xi h} - 2 + e^{i\xi h}\Big)\Big]U_j^n,$$

i.e.,

$$g(\xi) = 1 - \frac{4\nu k}{h^2}\sin^2\Big(\frac{\xi h}{2}\Big).$$

To guarantee $|g(\xi)| \le 1$, we take

$$1 - \frac{4\nu k}{h^2} \ge -1,$$

which implies (10.22), i.e. $k \le \frac{h^2}{2\nu}$.

Exercise 10.28. For the Crank-Nicolson method, show that the modulus of its amplification factor is never greater than 1 for any choice of $k, h > 0$.

10.7 Green's function of the heat equation in $(-\infty, +\infty)$

Lemma 10.29. The Fourier transform of a Gaussian centered at the origin is another such Gaussian.

Proof. First we consider the case $f(x) = e^{-x^2}$, then

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}e^{-i\xi x}\,\mathrm{d}x,$$

Differentiating with respect to ξ yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\xi}\hat{f}(\xi) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}(-ix)e^{-i\xi x}\,\mathrm{d}x\\ &= \frac{i}{2\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{\mathrm{d}e^{-x^2}}{\mathrm{d}x}e^{-i\xi x}\,\mathrm{d}x\\ &= -\frac{\xi}{2\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}e^{-i\xi x}\,\mathrm{d}x\\ &= -\frac{\xi}{2}\hat{f}(\xi),\end{aligned}$$

where the third line follows from the integration by parts formula. The unique solution to this ODE is given by

$$\hat{f}(\xi) = c\cdot e^{-\frac{\xi^2}{4}},$$

where $c = \hat{f}(0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x = \frac{\sqrt{2}}{2}$. The proof is completed by the dilation property (D.10) of Fourier transform. In particular, the Fourier transform of a Gaussian with $a = 1$, $b = 0$, and c is another Gaussian with $a' = c$, $b' = 0$, and $c' = \frac{1}{c}$. □

Lemma 10.30. For any $u \in L^2$ satisfying

$$\forall n \in \mathbb{N}, \qquad \lim_{x\to\pm\infty} u^{(n)}(x) = 0, \tag{10.32}$$

we have

$$\widehat{\frac{\mathrm{d}^2u}{\mathrm{d}x^2}} = -\xi^2\hat{u}. \tag{10.33}$$

Proof. Repeated application of (10.32) yields

$$\begin{aligned}\sqrt{2\pi}\cdot\widehat{\frac{\mathrm{d}^2u}{\mathrm{d}x^2}} &= \int_{-\infty}^{+\infty} e^{-i\xi x}\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\mathrm{d}x\\ &= e^{-i\xi x}\frac{\mathrm{d}u}{\mathrm{d}x}\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty}\frac{\mathrm{d}u}{\mathrm{d}x}(-i\xi)e^{-i\xi x}\mathrm{d}x\\ &= i\xi\int_{-\infty}^{+\infty}\frac{\mathrm{d}u}{\mathrm{d}x}e^{-i\xi x}\mathrm{d}x\\ &= i\xi(e^{-i\xi x}u)\Big|_{-\infty}^{+\infty} + (i\xi)^2\int_{-\infty}^{+\infty} ue^{-i\xi x}\mathrm{d}x\\ &= -\xi^2\int_{-\infty}^{+\infty} ue^{-i\xi x}\mathrm{d}x = -\xi^2\sqrt{2\pi}\hat{u},\end{aligned}$$

where the first and last lines follow from Definition D.2, the second and fourth lines from the integration by parts formula, and the third line from (10.32). □

Theorem 10.31. The solution to the heat equation

$$u_t = \nu u_{xx} \text{ on } (-\infty, +\infty) \tag{10.34}$$

with the initial condition $\eta(x) = e^{-\beta x^2}$ is

$$u(x,t) = \frac{1}{\sqrt{4\beta\nu t + 1}}e^{-\frac{x^2}{4\nu t + 1/\beta}}. \tag{10.35}$$

Proof. By Lemma 10.30, the Fourier transform of (10.34) leads to the ODE

$$\hat{u}_t(\xi,t) = -\nu\xi^2\hat{u}(\xi,t),$$

the solution of which with the initial data $\hat{u}(\xi,0) = \hat{\eta}(\xi)$ yields

$$\hat{u}(\xi,t) = e^{-\nu\xi^2 t}\hat{\eta}(\xi).$$

Then

$$\begin{aligned}u(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{u}(\xi,t)e^{i\xi x}\mathrm{d}\xi\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\nu\xi^2 t}\hat{\eta}(\xi)e^{i\xi x}\mathrm{d}\xi\\ &= \frac{1}{2\sqrt{\pi\beta}}\int_{-\infty}^{+\infty} e^{-\xi^2(\nu t + \frac{1}{4\beta})}e^{i\xi x}\mathrm{d}\xi.\end{aligned}$$

Define $C = \frac{1}{4\nu t + 1/\beta}$, then

$$\begin{aligned}u(x,t) &= \frac{1}{2\sqrt{\pi\beta}}\int_{-\infty}^{+\infty} e^{-\frac{\xi^2}{4C}}e^{i\xi x}\mathrm{d}\xi\\ &= \frac{1}{2\sqrt{\pi\beta}}\sqrt{4\pi C}\cdot e^{-x^2C}\\ &= \frac{1}{\sqrt{4\beta\nu t + 1}}e^{-\frac{x^2}{4\nu t + 1/\beta}}.\end{aligned}$$

As t increases this Gaussian becomes more spread out and the magnitude decreases. □

Corollary 10.32. A translation of the initial condition

$$\eta(x) = e^{-\beta(x-\bar{x})^2} \tag{10.36}$$

of the heat equation (10.34) leads to a translation of the solution, i.e.,

$$u(x,t) = \frac{1}{\sqrt{4\beta\nu t+1}} e^{-\frac{(x-\bar{x})^2}{4\nu t+1/\beta}}. \tag{10.37}$$

Corollary 10.33. For the heat equation (10.34) with the initial condition as

$$\omega_\beta(x,0;\bar{x}) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(x-\bar{x})^2}, \tag{10.38}$$

its solution is

$$\omega_\beta(x,t;\bar{x}) = \frac{1}{\sqrt{4\pi\nu t+\pi/\beta}} e^{-\frac{(x-\bar{x})^2}{4\nu t+1/\beta}}. \tag{10.39}$$

Definition 10.34. The *Green's function*

$$G(x,t;\bar{x}) := \lim_{\beta\to+\infty} \omega_\beta(x,t;\bar{x}) = \frac{1}{\sqrt{4\pi\nu t}} e^{-\frac{(x-\bar{x})^2}{4\nu t}} \tag{10.40}$$

is the solution of the heat equation (10.34) with its initial condition as the Dirac delta function in Definition 8.35.

Chapter 11

Hyperbolic Problems

Definition 11.1. A second-order, constant-coefficient, linear partial differential equation (PDE) of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \tag{11.1}$$

is called a *hyperbolic PDE* if its coefficients satisfy

$$B^2 - 4AC > 0. \tag{11.2}$$

Definition 11.2. The *one-dimensional wave equation* is a hyperbolic PDE of the form

$$u_{tt} = a^2 u_{xx}, \tag{11.3}$$

where $a > 0$ is the *wave speed*.

Definition 11.3. The *one-dimensional advection equation* is

$$u_t = -au_x \text{ in } \Omega := (0,1) \times (0,T), \tag{11.4}$$

where $x \in (0,1)$ is the spatial location and $t \in (0,T)$ the time; the equation has to be supplemented with an *initial condition*

$$u(x,0) = \eta(x), \text{ on } (0,1) \times \{0\} \tag{11.5}$$

and appropriate boundary conditions at either $\{0\} \times (0,T)$ or $\{1\} \times (0,T)$, depending on the sign of a.

Theorem 11.4. The exact solution of the Cauchy problem (11.4) is

$$u(x,t) = \eta(x - at). \tag{11.6}$$

Proof. It is straightforward to verify that

$$u_t + au_x = -a\eta'(x-at) + a\eta'(x-at) = 0. \qquad \square$$

Definition 11.5. A system of PDEs of the form

$$\mathbf{u}_t + A\mathbf{u}_x = \mathbf{0} \tag{11.7}$$

is *hyperbolic* if A is diagonalizable and its eigenvalues are all real.

Example 11.6. The Euler equations are

$$\frac{\partial}{\partial t}\begin{bmatrix} p \\ u \end{bmatrix} + \begin{bmatrix} 0 & \kappa_0 \\ \frac{1}{\rho_0} & 0 \end{bmatrix} \frac{\partial}{\partial x}\begin{bmatrix} p \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{11.8}$$

The equation for the pressure p can be further written as

$$p_{tt} = a^2 p_{xx} \text{ with } a = \pm\sqrt{\kappa_0/\rho_0}.$$

11.1 Classical MOLs

Example 11.7. Discretize the advection equation (11.4) in space at grid point x_j by

$$U_j'(t) = -\frac{a}{2h}\left(U_{j+1}(t) - U_{j-1}(t)\right), \quad 2 \le j \le m, \tag{11.9}$$

where $U_j(t) \approx u(x_j,t)$ for $j = 1, 2, \cdots, m+1$. For periodic boundary conditions

$$u(0,t) = u(1,t) = g_0(t), \tag{11.10}$$

the discretizations of (11.4) at $j = 1$ and $j = m+1$ are

$$U_1'(t) = -\frac{a}{2h}\left(U_2(t) - U_{m+1}(t)\right), \tag{11.11}$$

$$U_{m+1}'(t) = -\frac{a}{2h}\left(U_1(t) - U_m(t)\right). \tag{11.12}$$

Then the semi-discrete system can be written as

$$\mathbf{U}'(t) = A\mathbf{U}(t), \tag{11.13}$$

where

$$A = -\frac{a}{2h}\begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{bmatrix}, \tag{11.14}$$

and $\mathbf{U}(t) = [U_1(t), U_2(t), \cdots, U_{m+1}(t)]^T$.

Lemma 11.8. The eigenvalues of A in (11.13) are

$$\lambda_p = -\frac{ia}{h}\sin(2\pi ph) \text{ for } p = 1, 2, \ldots, m+1. \tag{11.15}$$

The corresponding eigenvector $\mathbf{w}^p$ has components

$$w_j^p = e^{2\pi ipjh} \text{ for } j = 1, 2, \ldots, m+1. \tag{11.16}$$

Proof. For $j = 2, 3, \ldots, m$, we have

$$\begin{aligned}(A\mathbf{w}^p)_j &= -\frac{a}{2h}\left(w_{j+1}^p - w_{j-1}^p\right) \\ &= -\frac{a}{2h}e^{2\pi ipjh}\left(e^{2\pi iph} - e^{-2\pi iph}\right) \\ &= -\frac{ia}{h}\sin(2\pi ph)e^{2\pi ipjh} \\ &= \lambda_p w_j^p.\end{aligned}$$

Similarly for $j = 1$ and $j = m+1$. $\square$

Notation 9. Hereafter we define the Courant number as

$$\mu := \frac{ak}{h}, \tag{11.17}$$

where k is the uniform time-step size.

11.1.1 The FTCS method

Definition 11.9. The FTCS method for the advection equation (11.4) is

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right), \tag{11.18}$$

or in matrix form

$$\mathbf{U}^{n+1} = (I + kA)\mathbf{U}^n. \tag{11.19}$$

Corollary 11.10. The FTCS method for the advection equation (11.4) is unconditionally unstable for $k = O(h)$.

Proof. The RAS of the forward Euler's method is

$$\mathcal{R} = \{z \in \mathbb{C} : |1 + z| \le 1\}.$$

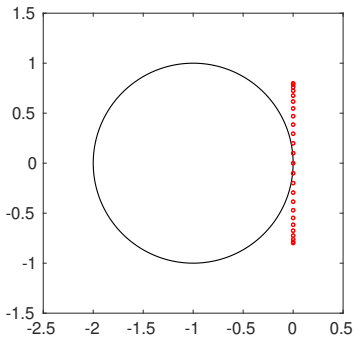
For (11.19), we have

$$z_p = k\lambda_p = -i\mu \sin(2\pi ph),$$

which lies on the imaginary axis between $-i\mu$ and $i\mu$, and thus if $k = O(h)$, then

$$\forall p = 1, 2, \ldots, m+1, \quad z_p \notin \mathcal{R},$$

which implies the instability, as shown below. □



Lemma 11.11. The FTCS method for the advection equation has Lax-Richtmyer stability for $k = O(h^2)$.

Proof. Since λ_p is purely imaginary, we have

$$|1 + k\lambda_p|^2 = 1 + k\frac{k}{h^2}a^2 \sin^2(2\pi ph) \le 1 + k\alpha,$$

for some $\alpha = O(1)$, hence the skew-symmetry of A implies

$$\|(I + kA)^n\|_2 \le \left(\|I + kA\|_2^2\right)^{\frac{n}{2}} \le (1 + k\alpha)^{n/2} \le e^{\alpha T/2},$$

which shows the uniform boundedness of the iteration matrix needed for Lax-Richtmyer stability. □

11.1.2 The leapfrog method

Definition 11.12. The *leapfrog method* for the advection equation (11.4) is

$$\frac{U_j^{n+1} - U_j^{n-1}}{2k} = -\frac{a}{2h}\left(U_{j+1}^n - U_{j-1}^n\right),$$

or, equivalently

$$U_j^{n+1} = U_j^{n-1} - \mu\left(U_{j+1}^n - U_{j-1}^n\right). \tag{11.20}$$

11.1.3 Lax-Friedrichs

Definition 11.13. The *Lax-Friedrichs method* for the advection equation (11.4) is

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right). \tag{11.21}$$

Lemma 11.14. Consider the IVP system

$$\mathbf{U}'(t) = A_\epsilon \mathbf{U}(t), \tag{11.22}$$

where

$$A_\epsilon = A + \frac{\epsilon}{h^2}\begin{bmatrix} -2 & 1 & & & & 1 \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ 1 & & & & 1 & -2 \end{bmatrix} \tag{11.23}$$

with A defined in (11.14). The eigenvalues of A_ϵ are

$$\lambda_p = -\frac{ia}{h}\sin(2\pi ph) - \frac{2\epsilon}{h^2}\left[1 - \cos(2\pi ph)\right] \tag{11.24}$$

for $p = 1, 2, \ldots, m+1$. The corresponding eigenvector $\mathbf{w}^p$ has components

$$w_j^p = e^{2\pi ipjh} \text{ for } j = 1, 2, \ldots, m+1. \tag{11.25}$$

Proof. This follows from Lemma 11.8 and the result on the eigenpair of the second-order discrete Laplacian. □

Lemma 11.15. The Lax-Friedrichs method can be considered as the MOL obtained by applying the forward Euler to the semidiscrete system (11.22) with $\epsilon = \frac{h^2}{2k}$.

Proof. The Lax-Friedrichs method can be rewritten as

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{1}{2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n\right),$$

which is equivalent to

$$\frac{U_j^{n+1} - U_j^n}{k} + a\left(\frac{U_{j+1}^n - U_{j-1}^n}{2h}\right) = \epsilon\frac{U_{j-1}^n - 2U_j^n + U_{j+1}^n}{h^2};$$

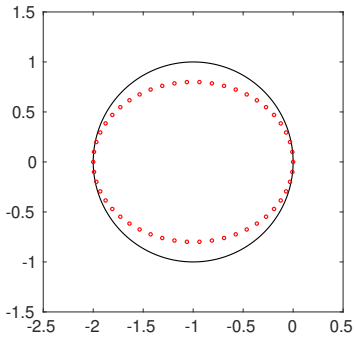
and this shows the standard discretization from the advection-diffusion equation. □

Theorem 11.16. The Lax-Friedrichs method (11.21) is convergent provided that $|\mu| \le 1$.

Proof. By Lemma 11.15, we have

$$z_p = k\lambda_p = -i\mu \sin(2\pi ph) - \frac{2k\epsilon}{h^2}\left[1 - \cos(2\pi ph)\right],$$

thus z_p's lie on an ellipse centered at $\frac{-2k\epsilon}{h^2} = -1$ with semi-axes $\left(\frac{2k\epsilon}{h^2}, \mu\right) = (1, \mu)$. If $|\mu| \le 1$, then this ellipse lies entirely inside the absolute region of stability of the forward Euler's method. Hence the Lax-Friedrichs method is convergent provided that $|\mu| \le 1$. □



11.1.4 Lax-Wendroff

Definition 11.17. The *Lax-Wendroff method* for the advection equation (11.4) is

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right). \end{aligned} \tag{11.26}$$

Lemma 11.18. The Lax-Wendroff method (11.26) is second-order accurate both in space and in time.

Proof. We calculate the LTE as

$$\begin{aligned} \tau(x,t) =& \frac{u(x,t+k) - u(x,t)}{k} + a\frac{u(x+h,t) - u(x-h,t)}{2h} \\ &- \frac{ka^2}{2}\frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} \\ =& u_t(x,t) + \frac{k}{2}u_{tt}(x,t) + au_x(x,t) - \frac{ka^2}{2}u_{xx}(x,t) \\ &+ O(k^2 + h^2) \\ =& O(k^2 + h^2), \end{aligned}$$

where the first step follows from the definition of LTE, the second from Taylor expansions and the last from $u_t = -au_x$ and $u_{tt} = -au_{tx} = a^2u_{xx}$. □

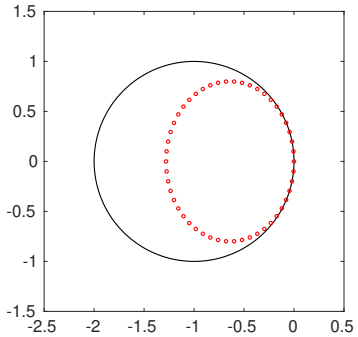
Lemma 11.19. The Lax-Wendroff method (11.26) can be considered as the MOL obtained by applying the forward Euler to the semidiscrete system (11.22) with $\epsilon = \frac{1}{2}ka^2$.

Theorem 11.20. The Lax-Wendroff method (11.26) is convergent provided $|\mu| \le 1$.

Proof. By Lemma 11.19, we have

$$z_p = k\lambda_p = -i\mu \sin(2\pi ph) + \mu^2\left[\cos(2\pi ph) - 1\right].$$

These values all lie on an ellipse centered at $-\mu^2$ with semi-axes of length μ^2 and $|\mu|$. If $|\mu| \le 1$, then all of these values lie inside the stability region of the forward Euler's method, thus ensuring the stability of the Lax-Wendroff method. □



11.1.5 The Upwind method

Definition 11.21. The *upwind method* for the advection equation (11.4) is

$$U_j^{n+1} = \begin{cases} U_j^n - \mu\left(U_j^n - U_{j-1}^n\right) & \text{if } a \ge 0; \\ U_j^n - \mu\left(U_{j+1}^n - U_j^n\right) & \text{if } a < 0. \end{cases} \tag{11.27}$$

Lemma 11.22. The upwind method (11.27) can be considered as the MOL obtained by applying the forward Euler to the semidiscrete system (11.22) with $\epsilon = \frac{h}{2}|a|$.

Proof. We only prove the case of $a > 0$. Then the upwind method can be rewritten as

$$U_j^{n+1} = U_j^n - \frac{\mu}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{\mu}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right),$$

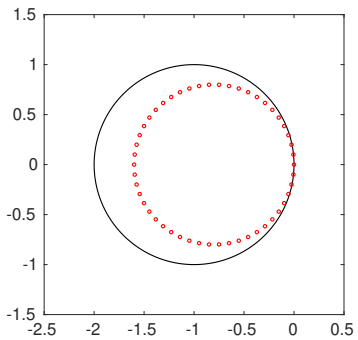
which is the forward Euler's method applied to (11.22) with $\epsilon = \frac{ah}{2}$. □

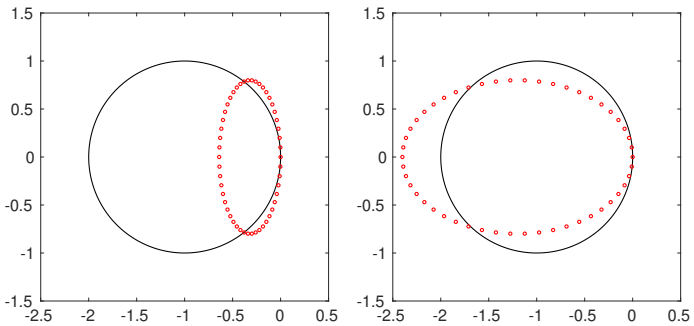
Theorem 11.23. For $a > 0$, the upwind method is convergent if and only if $\mu \le 1$; for $a < 0$, the upwind method is convergent if and only if $\mu \ge -1$.

Proof. We only prove the case of $a > 0$. By Lemma 11.22, we have

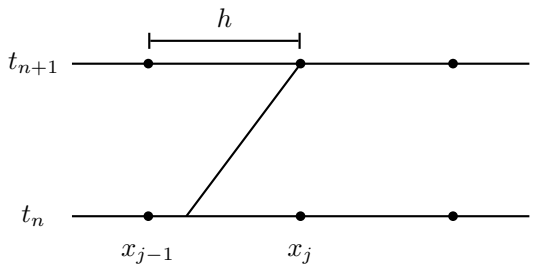
$$z_p = k\lambda_p = -i\mu \sin(2\pi ph) + \mu[\cos(2\pi ph) - 1].$$

These values all lie on a circle centered at $(-\mu, 0)$ with radius μ. If $\mu \le 1$, then all of these values lie inside the RAS of the forward Euler's method, thus ensuring the stability of the upwind method. For any choice of k, h satisfying $\mu > 1$, z_p would lie outside of the RAS and hence be unstable. □





Corollary 11.24. The upwind method is equivalent to characteristic tracing followed by a linear interpolation.



Proof. If we take $\mu = 1$, then the upwind method (11.27) reduces to

$$U_j^{n+1} = U_j^n - U_j^n + U_{j-1}^n = U_{j-1}^n.$$

Therefore for exact initial conditions, this method yields the exact solution by simply shifting the solution.

For $\mu < 1$, using characteristic tracing, we know

$$u(x_j, t + k) = u(x_j - ak, t).$$

Linear interpolating $u(x_j - ak, t)$ yields

$$u(x_j - ak, t) = \mu U_{j-1}^n + (1 - \mu) U_j^n + O(h^2),$$

which leads to the upwind method

$$U_j^{n+1} = \mu U_{j-1}^n + (1 - \mu) U_j^n = U_j^n - \mu \left(U_j^n - U_{j-1}^n\right). \quad \square$$

11.1.6 The Beam-Warming method

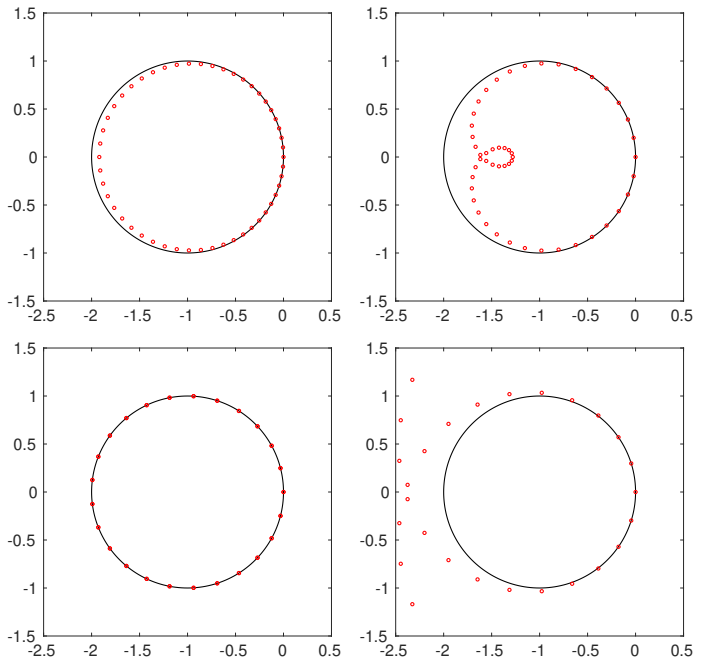
Definition 11.25. The *Beam-Warming method* solves the advection equation (11.4) by

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(3U_j^n - 4U_{j-1}^n + U_{j-2}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_j^n - 2U_{j-1}^n + U_{j-2}^n\right) \quad \text{if } a \geq 0; \end{aligned} \tag{11.28}$$

$$\begin{aligned} U_j^{n+1} =& U_j^n - \frac{\mu}{2}\left(-3U_j^n + 4U_{j+1}^n - U_{j+2}^n\right) \\ &+ \frac{\mu^2}{2}\left(U_j^n - 2U_{j+1}^n + U_{j+2}^n\right) \quad \text{if } a < 0. \end{aligned} \tag{11.29}$$

Exercise 11.26. Show that the Beam-Warming method is second-order accurate both in time and in space.

Exercise 11.27. Show that the Beam-Warming methods (11.28) and (11.29) are stable for $\mu \in [0, 2]$ and $\mu \in [-2, 0]$, respectively. Reproduce the following plots for $\mu = 0.8$, 1.6, 2, and 2.4.



11.2 The CFL condition

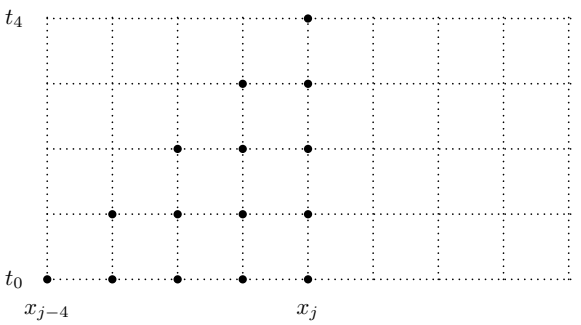
Definition 11.28. For the advection equation (11.4), the *domain of dependence* of a point $(X, T) \in \Omega$ is

$$\mathcal{D}_{\text{ADV}}(X, T) = \{X - aT\}. \tag{11.30}$$

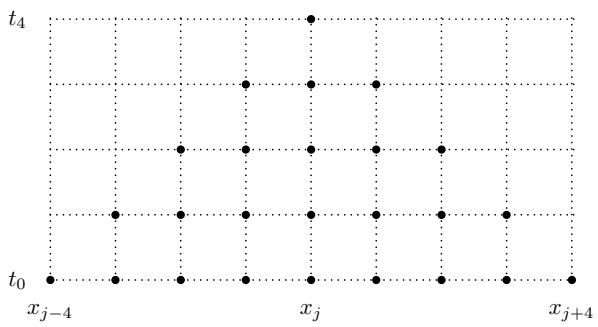
Definition 11.29. The *numerical domain of dependence* of a grid point (x_j, t_n) is the set of all grid points x_i such that U_i^0 at x_i has an effect on U_j^n.

$$\mathcal{D}_N(x_j, t_n) = \{x_i : U_i^0 \text{ affects } U_j^n\}. \tag{11.31}$$

Example 11.30. Numerical domain of dependence of a grid point using the upwind method.



Example 11.31. Numerical domain of dependence of a grid point using a 3-point explicit method.



Theorem 11.32 (Courant-Friedrichs-Lewy)**.** A numerical method can be convergent only if its numerical domain of dependence contains the domain of dependence of the PDE, at least in the limit of $k, h \to 0$.

Proof. It suffices to say that if some point p in the domain of dependence is not contained in the numerical domain of dependence, then we have no control over the value of p in the numerical method. Consequently, the numerical method cannot converge. □

Example 11.33. In solving the advection equation with $a = 1$, any choice of $k > h$ will result in instability. Take an extreme example of $k = 3h$, the numerical domain of dependence for U_j^3 would only contain x_j and $x_{j-1} = x_j - h$ while the domain of dependence is the singleton set $\{x_j - 3h\}$. The former does not contain the latter and thus this choice of k will result in divergence.

Example 11.34. The heat equation

$$\begin{cases} u_t = \nu u_{xx} \\ u(x,0) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(x-\bar{x})^2}, \end{cases} \tag{11.32}$$

has its exact solution as

$$u(x,t) = \frac{1}{\sqrt{4\pi\nu t + \pi/\beta}} e^{-(x-\bar{x})^2/(4\nu t + 1/\beta)}. \tag{11.33}$$

The domain of dependence is the whole line, i.e.,

$$\mathcal{D}_{\text{DIFF}}(X,T) = (-\infty, +\infty), \tag{11.34}$$

because an initial point source

$$\lim_{\beta\to\infty} u(x,0) = \delta(x - \bar{x})$$

instantaneously affects each point on the real line:

$$\lim_{\beta\to\infty} u(x,t) = \frac{1}{\sqrt{4\pi\nu t}} e^{-\frac{(x-\bar{x})^2}{4\nu t}}.$$

This is very much an artifact of the mathematical model rather than the true physics.

11.3 Modified equations

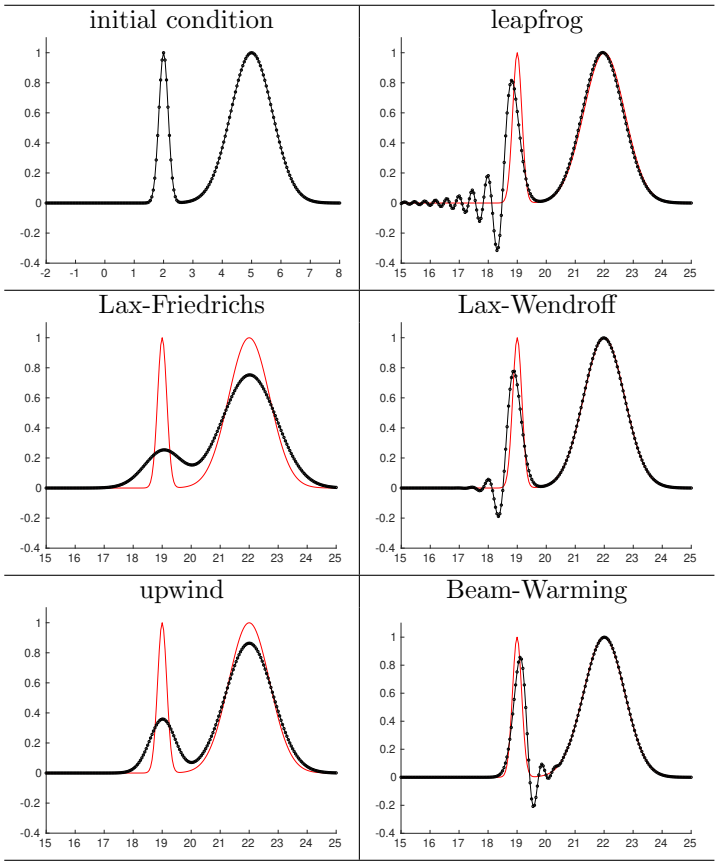
Example 11.35. For the advection equation

$$u_t + u_x = 0$$

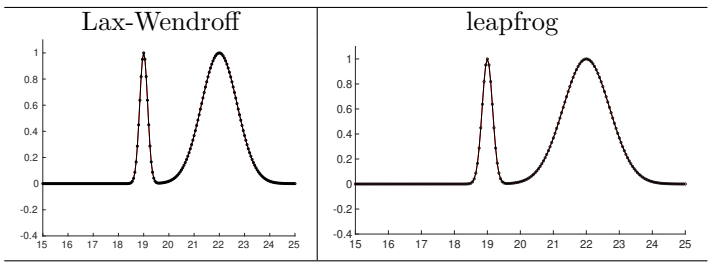
with initial condition

$$u(x,0) = \eta(x) = \exp(-20(x-2)^2) + \exp(-(x-5)^2), \tag{11.35}$$

the exact solution at $t = T$ is simply the initial data shifted by T. We solve this problem with $h = 0.05$ to $T = 17$ using the leapfrog method, the Lax-Friedrichs method, the Lax-Wendroff method, the upwind method, and the Beam-Warming method. The final results with $k = 0.8h$ are shown below.



If we keep all parameters the same except the change $k = h$, we have the following results.



These results invite a number of questions as follows.

(a) Why are the solutions of Lax-Friedrichs and upwind much smoother than those of the other three methods?

(b) What caused the ripples in the solutions of the three methods in the right column?

(c) Why does the numerical solution of the leapfrog method contain more oscillations than that of the Lax-Wendroff method?

(d) For the Lax-Wendroff method, why do the ripples of numerical solutions lag behind the true crest?

(e) For the Beam-Warming method, why do the ripples of numerical solutions move ahead of the true crest?

(f) Why are numerical results with $k = h$ much better than those with $k = 0.8h$?

These questions concern the physics behind the different features of the results of different methods; they can be answered by the modified equations.

Exercise 11.36. Reproduce all results in Example 11.35.

Definition 11.37. The *modified equation of an MOL for solving a PDE* (the original equation) is a PDE obtained from the formula of the MOL by

(1) replacing U_j^n with a smooth grid function $v(x_j, t_n)$ in the MOL formula,

(2) expanding all terms in Taylor series at (x_j, t_n),

(3) neglecting potentially high-order terms.

Example 11.38. Consider the upwind method for solving the advection equation

$$U_j^{n+1} = U_j^n - \mu\left(U_j^n - U_{j-1}^n\right).$$

The modified equation can be derived as follows.

(1) Replace U_j^n with $v(x_j, t_n)$ and we have

$$v(x, t+k) = v(x,t) - \mu\left(v(x,t) - v(x-h,t)\right).$$

(2) Expand all terms in Taylor series at (x, t) in a way similar to the calculation of the LTE.

$$\begin{aligned} 0 =& \frac{v(x,t+k) - v(x,t)}{k} + \frac{a}{h}\left(v(x,t) - v(x-h,t)\right) \\ =& \left(v_t + \frac{1}{2}kv_{tt} + \frac{1}{6}k^2 v_{ttt} + \cdots\right) \\ &+ a\left(v_x - \frac{1}{2}hv_{xx} + \frac{1}{6}h^2 v_{xxx} + \cdots\right), \end{aligned}$$

and thus

$$v_t + av_x = \frac{1}{2}\left(ahv_{xx} - kv_{tt}\right) - \frac{1}{6}\left(ah^2 v_{xxx} + k^2 v_{ttt}\right) + \cdots,$$

differentiating with respect to t and x gives

$$\begin{aligned} v_{tt} &= -av_{xt} + \frac{1}{2}\left(ahv_{xxt} - kv_{ttt}\right) + \cdots, \\ v_{tx} &= -av_{xx} + \frac{1}{2}\left(ahv_{xxx} - kv_{ttx}\right) + \cdots. \end{aligned}$$

Combining these gives

$$v_{tt} = a^2 v_{xx} + O(h+k).$$

Therefore we have

$$v_t + av_x = \frac{1}{2}ah\left(1-\mu\right)v_{xx} + O(h^2 + k^2),$$

(3) Neglect the high-order terms and we have the modified equation as

$$v_t + av_x = \frac{1}{2}ah\left(1-\mu\right)v_{xx} := \beta v_{xx}, \tag{11.36}$$

which is satisfied better by the grid function than the advection equation $v_t + av_x = 0$.

Exercise 11.39. Derive the modified equation of the Lax-Wendroff method for the advection equation as

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = 0. \tag{11.37}$$

Example 11.40. By Lemma D.16, the solution to the modified equation (11.37) is

$$v(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i\xi(x - C_p t)} \mathrm{d}\xi.$$

For Lax-Wendroff, (11.37) and Example D.27 yield

$$\begin{aligned} C_p(\xi) &= a - \frac{ah^2}{6}\left(1-\mu^2\right)\xi^2, \\ C_g(\xi) &= a - \frac{ah^2}{2}\left(1-\mu^2\right)\xi^2. \end{aligned}$$

First, the phase velocity $C_p \neq a$ for $\mu \neq 1$, and its value depends on ξ; this answers Question (b) of Example 11.35. For $\mu \neq 1$, both C_p and C_g have a magnitude smaller than $|a|$, hence numerical oscillations lag behind the true wave crest; this answers Question (d) of Example 11.35.

Exercise 11.41. Show that the modified equation of the leapfrog method is also (11.37). However, if one more term of higher-order derivative had been retained, the modified equation of the leapfrog method would have been

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = \epsilon_f v_{xxxxx} \tag{11.38}$$

while that of the Lax-Wendroff method would have been

$$v_t + av_x + \frac{ah^2}{6}\left(1-\mu^2\right)v_{xxx} = \epsilon_w v_{xxxx}. \tag{11.39}$$

Exercise 11.42. Show that the modified equation of the Beam-Warming method applied to the advection equation (11.4) with $a \geq 0$ is

$$v_t + av_x + \frac{ah^2}{6}\left(-2 + 3\mu - \mu^2\right)v_{xxx} = 0. \tag{11.40}$$

Thus we have

$$\begin{aligned} C_p(\xi) &= a + \frac{ah^2}{6}\left(\mu-1\right)\left(\mu-2\right)\xi^2, \\ C_g(\xi) &= a + \frac{ah^2}{2}\left(\mu-1\right)\left(\mu-2\right)\xi^2. \end{aligned}$$

How do these facts answer Question (e) of Example 11.35?

Exercise 11.43. What if $\mu = 1$? Discuss this case for both Lax-Wendroff and leapfrog methods to answer Question (f) of Example 11.35.

11.4 Von Neumann analysis

Exercise 11.44. For the advection equation (11.4) with $a \geq 0$, apply the von Neumann analysis to the upwind method to derive its amplification factor as

$$g(\xi) = (1 - \mu) + \mu e^{-i\xi h}. \tag{11.41}$$

For which values of μ would the method be stable?

Exercise 11.45. Apply the von Neumann analysis to the Lax-Friedrichs method to derive its amplification factor as

$$g(\xi) = \cos(\xi h) - \mu i \sin(\xi h). \tag{11.42}$$

For which values of μ would the method be stable?

Exercise 11.46. Apply the von Neumann analysis to the Lax-Wendroff method to derive its amplification factor as

$$g(\xi) = 1 - 2\mu^2 \sin^2 \frac{\xi h}{2} - i\mu \sin(\xi h). \tag{11.43}$$

For which values of μ would the method be stable?

Example 11.47. When performing the analysis of modified equations, we typically neglect some higher-order terms of ξh in deriving the group velocity and the phase velocity. For ξh sufficiently small, the modified equation would be a reasonable model. However, for large ξh the terms we have neglected may play an equally important role. In this case it might be better to use an approach similar to von Neumann analysis by setting

$$v(x_j, t_n) := e^{i(\xi x_j - \omega t_n)}. \tag{11.44}$$

For the leapfrog method, this form yields

$$\sin(\omega k) = \mu \sin(\xi h), \tag{11.45}$$

which yields the group velocity as

$$\frac{\mathrm{d}\omega}{\mathrm{d}\xi} = \pm \frac{a \cos(\xi h)}{\sqrt{1 - \mu^2 \sin^2(\xi h)}}, \tag{11.46}$$

where the $\pm$ follows from the multivalued dispersion relation (11.45). For high-frequency modes satisfying $\xi h \approx \pi$, the group velocity may have a sign different from that of a.

Chapter 12

Fourth-order Finite Volume (FV) Methods

Notation 10. We discretize a rectangular problem domain Ω into a collection of rectangular grid cells. Each cell is denoted by a multi-index $\mathbf{i} \in \mathbb{Z}^{\mathrm{D}}$, its region by

$$\mathcal{C}_{\mathbf{i}} = \left[\mathbf{x}_O + \mathbf{i}h,\ \mathbf{x}_O + (\mathbf{i} + \mathbb{1})\, h\right], \tag{12.1}$$

and the region of its higher face in dimension d by

$$\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \left[\mathbf{x}_O + \left(\mathbf{i} + \mathbf{e}^d\right) h,\ \mathbf{x}_O + (\mathbf{i} + \mathbb{1})\, h\right], \tag{12.2}$$

where $\mathbf{x}_O \in \mathbb{R}^{\mathrm{D}}$ is some fixed origin of the coordinates, h the uniform mesh spacing, $\mathbb{1} \in \mathbb{Z}^{\mathrm{D}}$ the multi-index with all its components equal to one, and $\mathbf{e}^d \in \mathbb{Z}^{\mathrm{D}}$ a multi-index with 1 as its d-th component and 0 otherwise.

Exercise 12.1. Denote the unit interval of cell $\mathbf{i}$ along the dth dimension by

$$I^d_{\mathbf{i}} := x_{O,d} + [i_d h, (i_d + 1)h], \tag{12.3}$$

where $x_{O,d}$ is the dth component of $\mathbf{x}_O$ in Notation 10. Use (12.1) and (12.2) to show that the region of cell $\mathbf{i}$ is the Cartesian product of its lower face in the dth dimension and the interval $I^d_{\mathbf{i}}$, i.e.,

$$\mathcal{C}_{\mathbf{i}} = \mathcal{F}_{\mathbf{i}-\frac{1}{2}\mathbf{e}^d} \times I^d_{\mathbf{i}}. \tag{12.4}$$

Proof. By (12.2), we have

$$\mathcal{F}_{\mathbf{i}-\frac{1}{2}\mathbf{e}^d} = \mathcal{F}_{\mathbf{i}-\mathbf{e}^d+\frac{1}{2}\mathbf{e}^d} = \left[\mathbf{x}_O + \mathbf{i}h,\ \mathbf{x}_O + \left(\mathbf{i} + \mathbb{1} - \mathbf{e}^d\right) h\right].$$

The proof is completed by (12.1) and (12.3). □

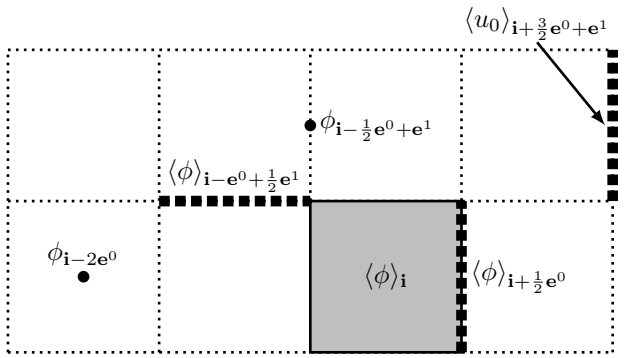
12.1 The FV formulation

Definition 12.2. The *cell average* of a function $\phi : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ over the cell $\mathbf{i}$ is a function $\mathbb{Z}^{\mathrm{D}} \to \mathbb{R}$ given by

$$\langle \phi \rangle_{\mathbf{i}} = \frac{1}{h^{\mathrm{D}}} \int_{\mathcal{C}_{\mathbf{i}}} \phi(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{12.5}$$

Definition 12.3. The *face average* of a function $\phi : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ over the face $\mathbf{i} + \frac{1}{2}\mathbf{e}^d$ is a function $\mathbb{Z}^{\mathrm{D}} \to \mathbb{R}$ given by

$$\langle \phi \rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{h^{\mathrm{D}-1}} \int_{\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}} \phi(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{12.6}$$



Notation 11. We strictly distinguishes three different types of quantities, viz. point values, cell averages, and face averages. A symbol without the averaging operator $\langle\cdot\rangle$ denotes a point value; otherwise it denotes either a cell-averaged value if the subscript is an integer multi-index, or a face-averaged value if the subscript is a fractional multi-index. In the above plot, $\phi_{\mathbf{i}-2\mathbf{e}^0}$, and $\phi_{\mathbf{i}-\frac{1}{2}\mathbf{e}^0+\mathbf{e}^1}$ are point values; $\langle \phi \rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^0}$, $\langle \phi \rangle_{\mathbf{i}-\mathbf{e}^0+\frac{1}{2}\mathbf{e}^1}$, and $\langle u_0 \rangle_{\mathbf{i}+\frac{3}{2}\mathbf{e}^0+\mathbf{e}^1}$ are face-averaged values; $\langle \phi \rangle_{\mathbf{i}}$ is a cell-averaged. Horizontal and vertical bold dotted lines represent the averaging processes over a vertical cell face and a horizontal cell face, respectively. Light gray area represents averaging over a cell.

Lemma 12.4. Point values can be converted to face averages and cell averages to the fourth-order accuracy via

$$\langle \phi \rangle_{\mathbf{i}} = \phi_{\mathbf{i}} + \frac{h^2}{24} \sum_{d=1}^{\mathrm{D}} \left.\frac{\partial^2 \phi(\mathbf{x})}{\partial x_d^2}\right|_{\mathbf{i}} + O(h^4), \tag{12.7}$$

$$\langle \phi \rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \phi_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + \frac{h^2}{24} \sum_{d' \neq d} \left.\frac{\partial^2 \phi(\mathbf{x})}{\partial x_{d'}^2}\right|_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + O(h^4). \tag{12.8}$$

Proof. The above identities follow from Taylor expansions of the integrands in (12.5) and (12.6). □

Theorem 12.5. Cell averages can be converted to face averages to the fourth-order accuracy via

$$\begin{aligned} \langle \phi \rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} =& \frac{7}{12}\left(\langle \phi \rangle_{\mathbf{i}} + \langle \phi \rangle_{\mathbf{i}+\mathbf{e}^d}\right) \\ & - \frac{1}{12}\left(\langle \phi \rangle_{\mathbf{i}-\mathbf{e}^d} + \langle \phi \rangle_{\mathbf{i}+2\mathbf{e}^d}\right) + O(h^4), \end{aligned} \tag{12.9}$$

$$\left\langle \frac{\partial \phi}{\partial x_d} \right\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{15}{12h}\Big(\langle \phi \rangle_{\mathbf{i}+\mathbf{e}^d} - \langle \phi \rangle_{\mathbf{i}} \Big) - \frac{1}{12h}\Big(\langle \phi \rangle_{\mathbf{i}+2\mathbf{e}^d} - \langle \phi \rangle_{\mathbf{i}-\mathbf{e}^d} \Big) + O(h^4). \tag{12.10}$$

Proof. We prove (12.9) and (12.10) via the first fundamental theorem of calculus (Theorem C.65). Define an indefinite integral $\Phi^d : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ of ϕ along the dth axis as

$$\Phi^d(\mathbf{x}) := \int_{\xi}^{x_d} \phi(x_1, \ldots, x_{d-1}, \eta, x_{d+1}, \ldots, x_{\mathrm{D}}) \, \mathrm{d}\eta, \tag{12.11}$$

where the lower limit $\xi \in \mathbb{R}$ is fixed. Theorem C.65 implies

$$\phi(\mathbf{x}) = \frac{\partial \Phi^d}{\partial x_d}. \tag{12.12}$$

For the first dimension, define a function $\delta_{\mathbf{i}} : \mathbb{R}^{\mathrm{D}-1} \to \mathbb{R}$ to represent the integral of ϕ over the interval $I^1_{\mathbf{i}}$ in (12.3):

$$\begin{aligned} \varphi_j &:= \Phi^1 \left(x_{O,1} + jh, x_2, \ldots, x_{\mathrm{D}} \right); \\ \delta_{\mathbf{i}} \left(x_2, \ldots, x_{\mathrm{D}} \right) &:= \varphi_{i_1+1} - \varphi_{i_1}. \end{aligned} \tag{12.13}$$

Taylor expansions of φ_{i_1+2}, φ_{i_1}, φ_{i_1+3}, φ_{i_1-1} in the variable x_1 at $\bar{x} := x_{O,1} + (i_1+1)h$ yield

$$\begin{aligned} &\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 2 & 4 & 8 & 16 \\ -2 & 4 & -8 & 16 \end{bmatrix} \begin{bmatrix} h & & & \\ & \frac{h^2}{2} & & \\ & & \frac{h^3}{6} & \\ & & & \frac{h^4}{24} \end{bmatrix} \begin{bmatrix} \frac{\partial \Phi^1}{\partial x_1} \\ \frac{\partial^2 \Phi^1}{\partial x_1^2} \\ \frac{\partial^3 \Phi^1}{\partial x_1^3} \\ \frac{\partial^4 \Phi^1}{\partial x_1^4} \end{bmatrix}_{x_1=\bar{x}} \\ &= \begin{bmatrix} \delta_{\mathbf{i}+\mathbf{e}^1} \\ -\delta_{\mathbf{i}} \\ \delta_{\mathbf{i}+\mathbf{e}^1} + \delta_{\mathbf{i}+2\mathbf{e}^1} \\ -\delta_{\mathbf{i}-\mathbf{e}^1} - \delta_{\mathbf{i}} \end{bmatrix} + O(h^5). \end{aligned} \tag{12.14}$$

Then (12.12) yields

$$\begin{bmatrix} h\phi \\ h^2 \frac{\partial \phi}{\partial x_1} \\ h^3 \frac{\partial^2 \phi}{\partial x_1^2} \\ h^4 \frac{\partial^3 \phi}{\partial x_1^3} \end{bmatrix}_{x_1=\bar{x}} = \mathbf{C} \begin{bmatrix} \delta_{\mathbf{i}+\mathbf{e}^1} \\ -\delta_{\mathbf{i}} \\ \delta_{\mathbf{i}+\mathbf{e}^1} + \delta_{\mathbf{i}+2\mathbf{e}^1} \\ -\delta_{\mathbf{i}-\mathbf{e}^1} - \delta_{\mathbf{i}} \end{bmatrix} + O(h^5), \tag{12.15}$$

where

$$\mathbf{C} = \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & -\frac{1}{12} & \frac{1}{12} \\ \frac{4}{3} & \frac{4}{3} & -\frac{1}{12} & -\frac{1}{12} \\ -1 & 1 & \frac{1}{2} & -\frac{1}{2} \\ -4 & -4 & 1 & 1 \end{bmatrix}.$$

Construct an auxiliary matrix

$$\mathbf{M} = \begin{bmatrix} 1 & & & \\ & -1 & & \\ -1 & & 1 & \\ & 1 & & -1 \end{bmatrix}, \tag{12.16}$$

add $\mathbf{M}^{-1}\mathbf{M}$ into the middle of the right-hand side of (12.15), and we have

$$\begin{bmatrix} \phi \\ h \frac{\partial \phi}{\partial x_1} \\ h^2 \frac{\partial^2 \phi}{\partial x_1^2} \\ h^3 \frac{\partial^3 \phi}{\partial x_1^3} \end{bmatrix}_{x_1=\bar{x}} = \frac{1}{h} \mathbf{T}^{(4)} \begin{bmatrix} \delta_{\mathbf{i}+\mathbf{e}^1} \\ \delta_{\mathbf{i}} \\ \delta_{\mathbf{i}+2\mathbf{e}^1} \\ \delta_{\mathbf{i}-\mathbf{e}^1} \end{bmatrix} + O(h^4), \tag{12.17}$$

where the fourth-order interpolation matrix $\mathbf{T}^{(4)}$ is

$$\mathbf{T}^{(4)} = \begin{bmatrix} \frac{7}{12} & \frac{7}{12} & -\frac{1}{12} & -\frac{1}{12} \\ \frac{5}{4} & -\frac{5}{4} & -\frac{1}{12} & \frac{1}{12} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -3 & 3 & 1 & -1 \end{bmatrix}. \tag{12.18}$$

Integrating the first row of (12.17) over $\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}$ yields (12.9) with $d = 1$, because (12.4) implies that, for any $j \in \mathbb{Z}$,

$$\begin{aligned} &\frac{1}{h^{\mathrm{D}}} \int_{\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^1}} \delta_{\mathbf{i}+j\mathbf{e}^1} \\ =&\frac{1}{h^{\mathrm{D}}} \int_{\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^1}} \int_{(i_1+j)h}^{(i_1+j+1)h} \phi(\eta, x_2, \ldots, x_{\mathrm{D}}) \, \mathrm{d}\eta \, \mathrm{d}x_2 \cdots \mathrm{d}x_{\mathrm{D}} \\ =&\frac{1}{h^{\mathrm{D}}} \int_{\mathcal{C}_{\mathbf{i}+j\mathbf{e}^1}} \phi = \langle \phi \rangle_{\mathbf{i}+j\mathbf{e}^1}. \end{aligned}$$

The proof of (12.9) is completed by a repetition of the above arguments on the first dimension to all other dimensions. (12.10) can be proved by repeating the above procedures to obtain a fifth-order counterpart of (12.17) and then integrating its second equation; see Exercise 12.6. □

Exercise 12.6. Repeat the procedures in the above proof to generate the fifth-order interpolation matrix

$$\mathbf{T}^{(5)} = \begin{bmatrix} \frac{47}{60} & \frac{9}{20} & -\frac{13}{60} & -\frac{1}{20} & \frac{1}{30} \\ \frac{5}{4} & -\frac{5}{4} & -\frac{1}{12} & \frac{1}{12} & 0 \\ -2 & \frac{1}{2} & \frac{3}{2} & \frac{1}{4} & -\frac{1}{4} \\ -3 & 3 & 1 & -1 & 0 \\ 6 & -4 & -4 & 1 & 1 \end{bmatrix}, \tag{12.19}$$

where the additional column is associated with $\langle \phi \rangle_{i+3}$. Explain why the formulas of the fourth order and the fifth order coincide for $\frac{\partial \phi}{\partial x}$.

12.2 Discrete operators

Definition 12.7. The *discrete gradient*, the *discrete divergence*, and the *discrete Laplacian* are defined as

$$\mathbf{G}_d \langle \phi \rangle_{\mathbf{i}} = \frac{1}{12h}\Big(- \langle \phi \rangle_{\mathbf{i}+2\mathbf{e}^d} + 8 \langle \phi \rangle_{\mathbf{i}+\mathbf{e}^d} - 8 \langle \phi \rangle_{\mathbf{i}-\mathbf{e}^d} + \langle \phi \rangle_{\mathbf{i}-2\mathbf{e}^d} \Big), \tag{12.20}$$

$$\mathbf{D} \langle \mathbf{u} \rangle_{\mathbf{i}} = \frac{1}{12h} \sum_d \Big(- \langle u_d \rangle_{\mathbf{i}+2\mathbf{e}^d} + 8 \langle u_d \rangle_{\mathbf{i}+\mathbf{e}^d} - 8 \langle u_d \rangle_{\mathbf{i}-\mathbf{e}^d} + \langle u_d \rangle_{\mathbf{i}-2\mathbf{e}^d} \Big), \tag{12.21}$$

$$\mathbf{L}\langle\phi\rangle_{\mathbf{i}} = \frac{1}{12h^2}\sum_d \Big(-\langle\phi\rangle_{\mathbf{i}+2\mathbf{e}^d} + 16\langle\phi\rangle_{\mathbf{i}+\mathbf{e}^d} - 30\langle\phi\rangle_{\mathbf{i}} + 16\langle\phi\rangle_{\mathbf{i}-\mathbf{e}^d} - \langle\phi\rangle_{\mathbf{i}-2\mathbf{e}^d}\Big). \tag{12.22}$$

The discrete divergence operator also acts on tensor averages,

$$\mathbf{D}\langle\mathbf{uu}\rangle_{\mathbf{i}} = \frac{1}{h}\sum_d \left(\mathbf{F}\langle u_d,\mathbf{u}\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} - \mathbf{F}\langle u_d,\mathbf{u}\rangle_{\mathbf{i}-\frac{1}{2}\mathbf{e}^d}\right), \tag{12.23}$$

where the discrete face average of the product of two scalar functions is

$$\mathbf{F}\langle\phi,\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\langle\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + \frac{h^2}{12}\sum_{d'\neq d}\left(\mathbf{G}^{\perp}_{d'}\phi\right)_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\left(\mathbf{G}^{\perp}_{d'}\psi\right)_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}, \tag{12.24}$$

and $\mathbf{G}^{\perp}_{d'}$ is the discrete gradient operator in the transverse directions,

$$\left(\mathbf{G}^{\perp}_{d'}\phi\right)_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{2h}\left(\langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d+\mathbf{e}^{d'}} - \langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d-\mathbf{e}^{d'}}\right). \tag{12.25}$$

Lemma 12.8. The operators in Definition 12.7 are fourth-order accurate, i.e.,

$$\mathbf{G}_d\langle\phi\rangle_{\mathbf{i}} = \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}\frac{\partial\phi}{\partial x_d} + O(h^4), \tag{12.26a}$$

$$\mathbf{D}\langle\mathbf{u}\rangle_{\mathbf{i}} = \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}\nabla\cdot\mathbf{u} + O(h^4), \tag{12.26b}$$

$$\mathbf{L}\langle\phi\rangle_{\mathbf{i}} = \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}\nabla^2\phi + O(h^4), \tag{12.26c}$$

$$\mathbf{D}\langle\mathbf{u}\rangle_{\mathbf{i}} = O(h^4) \;\Rightarrow\; \mathbf{D}\langle\mathbf{uu}\rangle_{\mathbf{i}} = \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}(\mathbf{u}\cdot\nabla)\mathbf{u} + O(h^4), \tag{12.26d}$$

$$\mathbf{F}\langle\phi,\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{h^{\mathrm{D}-1}}\int_{\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}}\phi\psi + O(h^4). \tag{12.26e}$$

Proof. (12.26a) follows from (12.9), the second fundamental theorem of calculus, and the fact that $\frac{\partial}{\partial x_d}$ commutes with $\langle\cdot\rangle$. (12.26b) follows from the divergence theorem and (12.9). (12.26c) follows from the divergence theorem and (12.10). The rest of the proof concerns (12.26d) and (12.26e).

Denote the cell center of $\mathcal{C}_{\mathbf{i}}$ by $\mathbf{x}_{\mathbf{i}} = \left(\mathbf{i}+\frac{1}{2}\mathbb{1}\right)h$, and the face centers by $\mathbf{x}_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}^d} = \mathbf{x}_{\mathbf{i}} \pm \frac{h}{2}\mathbf{e}^d$. Let $\mathbf{x}_c = \mathbf{x}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}$ be the center of $\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}$. Then the Taylor series of a function ϕ about $\mathbf{x}_c$ can be expressed in multi-index notation as

$$\phi(\mathbf{x}) = \sum_{|\mathbf{j}|\le 3}\frac{1}{\mathbf{j}!}(\mathbf{x}-\mathbf{x}_c)^{\mathbf{j}}\phi^{(\mathbf{j})}(\mathbf{x}_c) + O(h^4) = \sum_{|\mathbf{j}|\le 3}\frac{1}{\mathbf{j}!}\boldsymbol{\eta}^{\mathbf{j}}\phi^{(\mathbf{j})}(\mathbf{x}_c) + O(h^4), \tag{12.27}$$

where $\boldsymbol{\eta} = \mathbf{x}-\mathbf{x}_c$, so that $\eta_d = 0$ and $|\boldsymbol{\eta}| \approx O(h)$ on $\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}$.

Then the product of two functions $\phi,\psi:\mathbb{R}^{\mathrm{D}}\to\mathbb{R}$ is

$$\begin{aligned}\phi(\mathbf{x})\psi(\mathbf{x}) &= \left(\sum_{|\mathbf{j}|\le 3}\frac{1}{\mathbf{j}!}\boldsymbol{\eta}^{\mathbf{j}}\phi^{(\mathbf{j})}(\mathbf{x}_c)\right)\left(\sum_{|\mathbf{k}|\le 3}\frac{1}{\mathbf{k}!}\boldsymbol{\eta}^{\mathbf{k}}\psi^{(\mathbf{k})}(\mathbf{x}_c)\right) + O(h^4)\\ &= \sum_{\mathbf{k}:|\mathbf{k}|\le 3}\frac{1}{\mathbf{k}!}\boldsymbol{\eta}^{\mathbf{k}}\sum_{\mathbf{j}:\mathbf{j}\le\mathbf{k}}\binom{\mathbf{k}}{\mathbf{j}}\phi^{(\mathbf{j})}(\mathbf{x}_c)\psi^{(\mathbf{k}-\mathbf{j})}(\mathbf{x}_c) + O(h^4),\end{aligned}$$

where the second step follows from a variable substitution. Then the average over $\mathcal{F}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}$ (dropping indices on $\mathcal{F}$ and evaluation at $\mathbf{x}_c$) is

$$\begin{aligned}&\frac{1}{h^{\mathrm{D}-1}}\int_{\mathcal{F}}\phi\psi\,\mathrm{d}\mathbf{x}\\ =&\frac{1}{h^{\mathrm{D}-1}}\int_{\mathcal{F}}\sum_{\mathbf{k}:|\mathbf{k}|\le 3}\frac{1}{\mathbf{k}!}\boldsymbol{\eta}^{\mathbf{k}}\sum_{\mathbf{j}:\mathbf{j}\le\mathbf{k}}\binom{\mathbf{k}}{\mathbf{j}}\phi^{(\mathbf{j})}\psi^{(\mathbf{k}-\mathbf{j})}\,\mathrm{d}\mathbf{x} + O(h^4)\\ =&\sum_{\mathbf{k}:|\mathbf{k}|\le 3}\frac{1}{\mathbf{k}!}\left(\frac{1}{h^{\mathrm{D}-1}}\int_{\mathcal{F}}\boldsymbol{\eta}^{\mathbf{k}}\,\mathrm{d}\mathbf{x}\right)\sum_{\mathbf{j}:\mathbf{j}\le\mathbf{k}}\binom{\mathbf{k}}{\mathbf{j}}\phi^{(\mathbf{j})}\psi^{(\mathbf{k}-\mathbf{j})} + O(h^4).\end{aligned}$$

If $k_d \neq 0$ or $\mathbf{k}$ is odd in any component, then $\int_{\mathcal{F}}\boldsymbol{\eta}^{\mathbf{k}}\,\mathrm{d}\mathbf{x} = 0$. Hence, the only nonzero terms come from the two cases $\mathbf{k}=\mathbf{0}$, $\mathbf{j}=\mathbf{0}$ and $\mathbf{k}=2\mathbf{e}^{d'}$, $\mathbf{j}=\mathbf{0},\mathbf{e}^{d'},2\mathbf{e}^{d'}$ with $d'\neq d$. Thus,

$$\begin{aligned}&\frac{1}{h^{\mathrm{D}-1}}\int_{\mathcal{F}}\phi\psi\,\mathrm{d}\mathbf{x}\\ =&\phi\psi + \frac{h^2}{24}\sum_{d'\neq d}\left(\phi^{\left(2\mathbf{e}^{d'}\right)}\psi + \psi^{\left(2\mathbf{e}^{d'}\right)}\phi\right)\\ &+\frac{h^2}{12}\sum_{d'\neq d}\left(\phi^{\left(\mathbf{e}^{d'}\right)}\psi^{\left(\mathbf{e}^{d'}\right)}\right) + O(h^4)\\ =&\left(\phi + \frac{h^2}{24}\sum_{d'\neq d}\phi^{\left(2\mathbf{e}^{d'}\right)}\right)\left(\psi + \frac{h^2}{24}\sum_{d'\neq d}\psi^{\left(2\mathbf{e}^{d'}\right)}\right)\\ &+\frac{h^2}{12}\sum_{d'\neq d}\left(\phi^{\left(\mathbf{e}^{d'}\right)}\psi^{\left(\mathbf{e}^{d'}\right)}\right) + O(h^4)\\ =&\langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\langle\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + \frac{h^2}{12}\sum_{d'\neq d}\left(\phi^{\left(\mathbf{e}^{d'}\right)}\psi^{\left(\mathbf{e}^{d'}\right)}\right) + O(h^4),\end{aligned}$$

where we have used (12.8) to convert the first two terms in parentheses to face averages. The last term representing the product of transverse gradients can be approximated with

$$\mathbf{G}^{\perp}_{d'}\phi\big|_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{2h}\left(\langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d+\mathbf{e}^{d'}} - \langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d-\mathbf{e}^{d'}}\right) = \left.\frac{\partial\phi}{\partial x_{d'}}\right|_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + O(h^2),$$

leading to $O(h^4)$ overall for the average flux formula:

$$\langle\phi\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \langle\phi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\langle\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + \frac{h^2}{12}\sum_{d'\neq d}\left(\mathbf{G}^{\perp}_{d'}\phi\big|_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\,\mathbf{G}^{\perp}_{d'}\psi\big|_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\right) + C_4(\mathbf{x}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d})h^4 + O(h^5). \tag{12.28}$$

Furthermore,

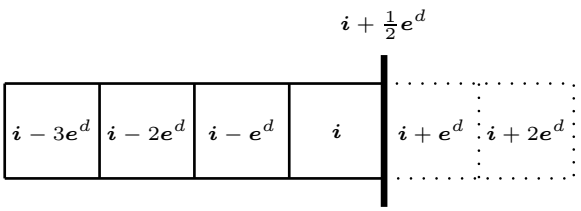
$$\frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}(\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}\nabla\cdot(\mathbf{uu}) - \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}(\nabla\cdot\mathbf{u})\mathbf{u}$$
$$= \frac{1}{h}\sum_d \mathbf{e}^d\cdot\left(\langle\mathbf{uu}\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} - \langle\mathbf{uu}\rangle_{\mathbf{i}-\frac{1}{2}\mathbf{e}^d}\right) + O(h^4),$$

where we have applied the chain rule, the divergence theorem, (12.26b), and the given condition $\mathbf{D}\langle\mathbf{u}\rangle_{\mathbf{i}} = O(h^4)$. Then (12.26d) follows from (12.23) and (12.26e). Another necessary condition is the cancellation from the symmetry of the difference stencils, i.e.,

$$C_4(\mathbf{x}_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}) - C_4(\mathbf{x}_{\mathbf{i}-\frac{1}{2}\mathbf{e}^d}) = O(h). \qquad \square$$

12.3 Ghost cells

Definition 12.9. *Ghost cells* are convenience devices for evaluating FD or FV discrete operators at cells near non-periodic domain boundaries.



Example 12.10. Two layers of ghost cells are used to enforce boundary conditions for fourth-order FV methods. We set the values of ghost cells by extrapolating those of the interior cells, with the boundary conditions incorporated in the extrapolation formulas. Referring to the above plot, different boundary conditions entail different cell-averaged values for the cells $\mathbf{i}+\mathbf{e}^d$ and $\mathbf{i}+2\mathbf{e}^d$. In particular, Dirichlet boundary conditions g are fulfilled to fourth-order accuracy by filling ghost cells with the following formulas:

$$\langle\phi\rangle_{\mathbf{i}+\mathbf{e}^d} = \frac{1}{3}\left(-13\langle\phi\rangle_{\mathbf{i}} + 5\langle\phi\rangle_{\mathbf{i}-\mathbf{e}^d} - \langle\phi\rangle_{\mathbf{i}-2\mathbf{e}^d}\right) + 4\langle g\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + O(h^4); \tag{12.29a}$$

$$\langle\phi\rangle_{\mathbf{i}+2\mathbf{e}^d} = \frac{1}{3}\left(-70\langle\phi\rangle_{\mathbf{i}} + 32\langle\phi\rangle_{\mathbf{i}-\mathbf{e}^d} - 7\langle\phi\rangle_{\mathbf{i}-2\mathbf{e}^d}\right) + 12\langle g\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + O(h^4). \tag{12.29b}$$

Neumann boundary conditions are fulfilled to the fifth order by

$$\langle\psi\rangle_{\mathbf{i}+\mathbf{e}^d} = \frac{1}{10}\left(5\langle\psi\rangle_{\mathbf{i}} + 9\langle\psi\rangle_{\mathbf{i}-\mathbf{e}^d} - 5\langle\psi\rangle_{\mathbf{i}-2\mathbf{e}^d} + \langle\psi\rangle_{\mathbf{i}-3\mathbf{e}^d}\right) + \frac{6}{5}h\left\langle\frac{\partial\psi}{\partial n}\right\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + O(h^5), \tag{12.30a}$$

$$\langle\psi\rangle_{\mathbf{i}+2\mathbf{e}^d} = \frac{1}{10}\left(-75\langle\psi\rangle_{\mathbf{i}} + 145\langle\psi\rangle_{\mathbf{i}-\mathbf{e}^d} - 75\langle\psi\rangle_{\mathbf{i}-2\mathbf{e}^d} + 15\langle\psi\rangle_{\mathbf{i}-3\mathbf{e}^d}\right) + 6h\left\langle\frac{\partial\psi}{\partial n}\right\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} + O(h^5), \tag{12.30b}$$

where $\left\langle\frac{\partial\psi}{\partial n}\right\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}$ is the Neumann condition for ψ.

Exercise 12.11. When no boundary conditions are known, a scalar ψ can be smoothly extended to fill a ghost cell abutting the boundary. Derive the following formulas

$$\langle\psi\rangle_{\mathbf{i}+\mathbf{e}^d} = 5\langle\psi\rangle_{\mathbf{i}} - 10\langle\psi\rangle_{\mathbf{i}-\mathbf{e}^d} + 10\langle\psi\rangle_{\mathbf{i}-2\mathbf{e}^d} - 5\langle\psi\rangle_{\mathbf{i}-3\mathbf{e}^d} + \langle\psi\rangle_{\mathbf{i}-4\mathbf{e}^d} + O(h^5); \tag{12.31}$$

and those for its faced-averaged value at the boundary,

$$\langle\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{60}\left(137\langle\psi\rangle_{\mathbf{i}} - 163\langle\psi\rangle_{\mathbf{i}-\mathbf{e}^d} + 137\langle\psi\rangle_{\mathbf{i}-2\mathbf{e}^d} - 63\langle\psi\rangle_{\mathbf{i}-3\mathbf{e}^d} + 12\langle\psi\rangle_{\mathbf{i}-4\mathbf{e}^d}\right) + O(h^5). \tag{12.32}$$

Lemma 12.12. Face-averaged derivatives can be calculated from known boundary conditions and interior cell averages:

$$\left\langle\frac{\partial\psi}{\partial n}\right\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{72h}\left(-415\langle\psi\rangle_{\mathbf{i}} + 161\langle\psi\rangle_{\mathbf{i}-\mathbf{e}^d} - 55\langle\psi\rangle_{\mathbf{i}-2\mathbf{e}^d} + 9\langle\psi\rangle_{\mathbf{i}-3\mathbf{e}^d} + 300\langle\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\right) + O(h^4), \tag{12.33a}$$

$$\left\langle\frac{\partial^2\psi}{\partial n^2}\right\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d} = \frac{1}{48h^2}\left(-755\langle\psi\rangle_{\mathbf{i}} + 493\langle\psi\rangle_{\mathbf{i}-\mathbf{e}^d} - 191\langle\psi\rangle_{\mathbf{i}-2\mathbf{e}^d} + 33\langle\psi\rangle_{\mathbf{i}-3\mathbf{e}^d} + 420\langle\psi\rangle_{\mathbf{i}+\frac{1}{2}\mathbf{e}^d}\right) + O(h^3). \tag{12.33b}$$

Algorithm 12.13. The discrete convection $\mathbf{D}\langle\mathbf{uu}\rangle$ are evaluated as follows:

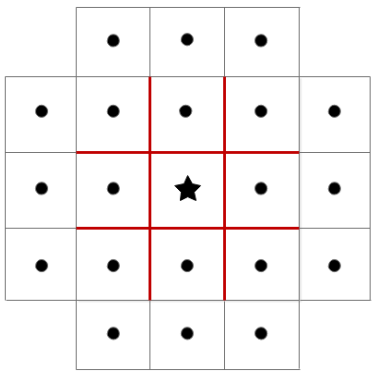
(Cnv-1) fill the ghost cells of cell-averaged velocity $\langle\mathbf{u}\rangle_{\mathbf{i}}$ using (12.29),

(Cnv-2) convert $\langle\mathbf{u}\rangle_{\mathbf{i}}$ to face-averaged *normal* velocity $\langle u_d\rangle_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}^d}$ for each dimension d using (12.9),

(Cnv-3) convert $\langle\mathbf{u}\rangle_{\mathbf{i}}$ to face-averaged velocity $\langle\mathbf{u}\rangle_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}^d}$ for *each* component of the velocity and each dimension d using (12.9).

(Cnv-4) calculate the discrete velocity product $\mathbf{F}\langle u_d, \mathbf{u}\rangle$ using (12.24) and (12.25).

(Cnv-5) compute $\mathbf{D}\langle\mathbf{uu}\rangle$ using (12.23).



12.4 FV methods for BVPs

Algorithm 12.14. To solve the Neumann BVP of the form

$$\Delta\phi = \nabla\cdot\mathbf{u}, \tag{12.34a}$$
$$\mathbf{n}\cdot\nabla\phi = \mathbf{n}\cdot\mathbf{u}, \tag{12.34b}$$

a fourth-order FV approximation is

$$\mathbf{L}_H\langle\phi\rangle = \mathbf{D}_H\langle\mathbf{u}\rangle, \tag{12.35}$$

where $\mathbf{L}_H$, $\mathbf{D}_H$ are the same as $\mathbf{L}$, $\mathbf{D}$ defined in (12.22) and (12.21), except that the fluxes at the wall boundaries are set to zero. More precisely, $\mathbf{D}_H\langle\mathbf{u}\rangle$ is calculated as follows,

(DvH-1) smoothly extend $\langle u_n\rangle$ to the ghost cells abutting the domain by (12.31),

(DvH-2) convert $\langle u_n\rangle$ to face averages using (12.9),

(DvH-3) zero out the face averages on the wall boundaries,

(DvH-4) sum up the face-averaged normal velocities for each interior cell.

$\mathbf{L}_H\langle\phi\rangle$ can also be calculated by first converting cell averages of $\nabla\phi$ to face averages and then apply the above steps, but this is not amenable to a multigrid solver. To facilitate the smoothing operations, $\mathbf{L}_H\langle\phi\rangle$ is evaluated by (12.22), with ghost cells that immediately abut the wall filled by (12.31) and those away from the wall by

$$\begin{aligned}\langle\phi\rangle_{\mathbf{i}+2\mathbf{e}^d} =& 60\langle\phi\rangle_{\mathbf{i}} - 149\langle\phi\rangle_{\mathbf{i}-\mathbf{e}^d} + 150\langle\phi\rangle_{\mathbf{i}-2\mathbf{e}^d} - 75\langle\phi\rangle_{\mathbf{i}-3\mathbf{e}^d}\\ &+ 15\langle\phi\rangle_{\mathbf{i}-4\mathbf{e}^d} + O(h^5)\end{aligned} \tag{12.36}$$

so that the resulting flux $\left\langle\frac{\partial\phi}{\partial x_d}\right\rangle_{\mathbf{i}+\frac12\mathbf{e}^d}$ given by (12.10) vanishes.

If $\mathbf{G}\langle\phi\rangle$ is needed, the ghost cells of $\langle\phi\rangle$ are filled by the formulas in (12.30), where the Neumann boundary value $\left\langle\frac{\partial\phi}{\partial x_d}\right\rangle_{\mathbf{i}+\frac12\mathbf{e}^d}$ is obtained by extrapolating cell-averaged normal velocity $\langle u_n\rangle$'s to the wall boundary using (12.32).

12.5 FV-MOL algorithms for the advection-diffusion equation

Definition 12.15. The *advection-diffusion equation* is a PDE of the form

$$\frac{\partial\phi}{\partial t} = -\nabla\cdot(\mathbf{u}\phi) + \nu\Delta\phi + f, \tag{12.37}$$

where the constant diffusivity ν, the velocity field $\mathbf{u}:\mathbb{R}^{\mathrm{D}}\times\mathbb{R}\to\mathbb{R}^{\mathrm{D}}$, and the forcing term $f:\mathbb{R}^{\mathrm{D}}\times\mathbb{R}\to\mathbb{R}$ are known *a priori*.

Example 12.16. To generate a semi-discrete system of the advection-diffusion equation in the FV formulation, we average (12.37) over a control volume $\mathcal{C}_{\mathbf{i}}$ and apply the divergence theorem to obtain a system of ODEs:

$$\frac{\mathrm{d}\langle\phi\rangle_{\mathbf{i}}}{\mathrm{d}t} = L_{\mathrm{adv}}(\langle\phi\rangle,t)_{\mathbf{i}} + L_{\mathrm{diff}}(\langle\phi\rangle)_{\mathbf{i}} + \langle f\rangle_{\mathbf{i}}, \tag{12.38}$$

where

$$\begin{aligned}L_{\mathrm{adv}}(\langle\phi\rangle,t)_{\mathbf{i}} &= -\frac1h\sum_{d=1}^{D}\left(\langle u_d\phi\rangle_{\mathbf{i}+\frac12\mathbf{e}^d} - \langle u_d\phi\rangle_{\mathbf{i}-\frac12\mathbf{e}^d}\right),\\ L_{\mathrm{diff}}(\langle\phi\rangle)_{\mathbf{i}} &= \nu\langle\Delta\phi\rangle_{\mathbf{i}}\\ &= \frac{\nu}{h}\sum_{d=1}^{D}\left(\left\langle\frac{\partial\phi}{\partial x_d}\right\rangle_{\mathbf{i}+\frac12\mathbf{e}^d} - \left\langle\frac{\partial\phi}{\partial x_d}\right\rangle_{\mathbf{i}-\frac12\mathbf{e}^d}\right).\end{aligned}$$

Definition 12.17. An *ERK-ESDIRK IMplicit-EXplicit (IMEX) Runge-Kutta scheme* for solving an ODE

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \mathbf{X}^{[\mathrm{E}]}(\phi,t) + \mathbf{X}^{[\mathrm{I}]}(\phi) \tag{12.39}$$

consists of steps as follows:

$$\phi^{(1)} = \phi^n \approx \phi(t^n), \tag{12.40a}$$
$$\forall s = 2,3,\dots,n_{\mathbf{s}},$$
$$\begin{aligned}\left(I - k\gamma\mathbf{X}^{[\mathrm{I}]}\right)\phi^{(s)} =& \phi^n + k\sum_{j=1}^{s-1}a_{s,j}^{[E]}\mathbf{X}^{[\mathrm{E}]}\left(\phi^{(j)},t^{(j)}\right)\\ &+ k\sum_{j=1}^{s-1}a_{s,j}^{[I]}\mathbf{X}^{[\mathrm{I}]}\phi^{(j)},\end{aligned} \tag{12.40b}$$
$$\begin{aligned}\phi^{n+1} =& \phi^n + k\sum_{s=1}^{n_{\mathbf{s}}}b_s^{[E]}\mathbf{X}^{[\mathrm{E}]}\left(\phi^{(s)},t^{(s)}\right)\\ &+ k\sum_{s=1}^{n_{\mathbf{s}}}b_s^{[I]}\mathbf{X}^{[\mathrm{I}]}\phi^{(s)},\end{aligned} \tag{12.40c}$$

where the superscript $^{(s)}$ denotes an intermediate stage, $t^{(s)} = t^n + c_s k$ the time of that stage, $n_{\mathbf{s}}$ the number of stages, and $A, \mathbf{b}, \mathbf{c}$ standard coefficients of the Butcher tableau.

Algorithm 12.18. A *fourth-order FV method for solving the advection-diffusion equation on periodic domains* is obtained by directly applying the ERK-ESDIRK IMEX algorithm (12.40) to the ODE system (12.38) with

$$\mathbf{X}^{[\mathrm{E}]} = L_{\mathrm{adv}} + \langle f\rangle, \qquad \mathbf{X}^{[\mathrm{I}]} = L_{\mathrm{diff}}. \tag{12.41}$$

Example 12.19. Kennedy and Carpenter [2003] studied a group of implicit-explicit Runge-Kutta schemes from third- to fifth-order accurate with the following form:

$$
\begin{array}{c|c}
\mathbf{c}^{[E]} & A^{[E]} \\
\hline
 & \left(\mathbf{b}^{[E]}\right)^T \\
\hline
 & \left(\hat{\mathbf{b}}^{[E]}\right)^T
\end{array}
=
\begin{array}{c|cccccc}
0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
2\gamma & 2\gamma & 0 & 0 & \cdots & 0 & 0 \\
c_3 & a_{31}^{[E]} & a_{32}^{[E]} & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
c_{s-1} & a_{s-1,1}^{[E]} & a_{s-1,2}^{[E]} & a_{s-1,3}^{[E]} & \cdots & 0 & 0 \\
1 & a_{s,1}^{[E]} & a_{s,2}^{[E]} & a_{s,3}^{[E]} & \cdots & a_{s,s-1}^{[E]} & 0 \\
\hline
 & b_1 & b_2 & b_3 & \cdots & b_{s-1} & \gamma \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & \cdots & \hat{b}_{s-1} & \hat{b}_s
\end{array}, \tag{12.42}
$$

$$
\begin{array}{c|c}
\mathbf{c}^{[I]} & A^{[I]} \\
\hline
 & \left(\mathbf{b}^{[I]}\right)^T \\
\hline
 & \left(\hat{\mathbf{b}}^{[I]}\right)^T
\end{array}
=
\begin{array}{c|cccccc}
0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
2\gamma & \gamma & \gamma & 0 & \cdots & 0 & 0 \\
c_3 & a_{31}^{[I]} & a_{32}^{[I]} & \gamma & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
c_{s-1} & a_{s-1,1}^{[I]} & a_{s-1,2}^{[I]} & a_{s-1,3}^{[I]} & \cdots & \gamma & 0 \\
1 & b_1 & b_2 & b_3 & \cdots & b_{s-1} & \gamma \\
\hline
 & b_1 & b_2 & b_3 & \cdots & b_{s-1} & \gamma \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & \cdots & \hat{b}_{s-1} & \hat{b}_s
\end{array}, \tag{12.43}
$$

where coefficients in $\hat{\mathbf{b}}$ are useful for error estimation.

The coefficients of `ARK4(3)6L[2]SA`, a particular IMEX scheme, are, in decimal form, $\gamma = 0.25$, $\mathbf{c}^{[E]} = \mathbf{c}^{[I]} = \mathbf{c}$, $\mathbf{b}^{[E]} = \mathbf{b}^{[I]} = \mathbf{b}$ where

$$\mathbf{c} = (0.0,\ 0.5,\ 0.332,\ 0.62,\ 0.85,\ 1.0)^T$$

$$
\begin{aligned}
b_1 &= 0.15791629516167136,\\
b_2 &= 0.,\\
b_3 &= 0.18675894052400077,\\
b_4 &= 0.6805652953093346,\\
b_5 &= -0.27524053099500667,\\
a_{31}^{[E]} &= 0.221776,\\
a_{32}^{[E]} &= 0.110224,\\
a_{41}^{[E]} &= -0.04884659515311857,\\
a_{42}^{[E]} &= -0.17772065232640102,\\
a_{43}^{[E]} &= 0.8465672474795197,\\
a_{51}^{[E]} &= -0.15541685842491548,\\
a_{52}^{[E]} &= -0.3567050098221991,\\
a_{53}^{[E]} &= 1.0587258798684427,\\
a_{54}^{[E]} &= 0.30339598837867193,\\
a_{61}^{[E]} &= 0.2014243506726763,\\
a_{62}^{[E]} &= 0.008742057842904185,\\
a_{63}^{[E]} &= 0.15993995707168115,\\
a_{64}^{[E]} &= 0.4038290605220775,\\
a_{65}^{[E]} &= 0.22606457389066084
\end{aligned}
$$

$$
\begin{aligned}
a_{31}^{[I]} &= 0.137776,\\
a_{32}^{[I]} &= -0.055776,\\
a_{41}^{[I]} &= 0.14463686602698217,\\
a_{42}^{[I]} &= -0.22393190761334475,\\
a_{43}^{[I]} &= 0.4492950415863626,\\
a_{51}^{[I]} &= 0.09825878328356477,\\
a_{52}^{[I]} &= -0.5915442428196704,\\
a_{53}^{[I]} &= 0.8101210538282996,\\
a_{54}^{[I]} &= 0.283164405707806,
\end{aligned}
$$

Lemma 12.20. The *stability function* of the IMEX Runge-Kutta methods in Definition 12.17 is

$$R\left(\bar{\lambda}^d + \mathtt{i}\bar{\lambda}^a\right) = \frac{\det\left(I - \bar{\lambda}^d A^{[I]} - \mathtt{i}\bar{\lambda}^a A^{[E]} + \left(\bar{\lambda}^d + \mathtt{i}\bar{\lambda}^a\right)\mathbb{1}\otimes\mathbf{b}^T\right)}{\det\left(I - \bar{\lambda}^d A^{[I]} - \mathtt{i}\bar{\lambda}^a A^{[E]}\right)}, \tag{12.44}$$

where $\bar{\lambda}^d = \lambda^d k$, $\bar{\lambda}^a = \lambda^a k$. The vector $\mathbf{b}$ and the matrices $A^{[E]}$ and $A^{[I]}$ are the coefficients in (12.42) and (12.43).

Proof. See Calvo et al. [2001]. □

Lemma 12.21. For a periodic domain, the ODE system (12.38) with constant velocity $\mathbf{u}$ and $\langle f\rangle = 0$ can be converted to a system of decoupled ODEs of the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y = \left(\lambda^d + \mathtt{i}\lambda^a\right) y, \tag{12.45}$$

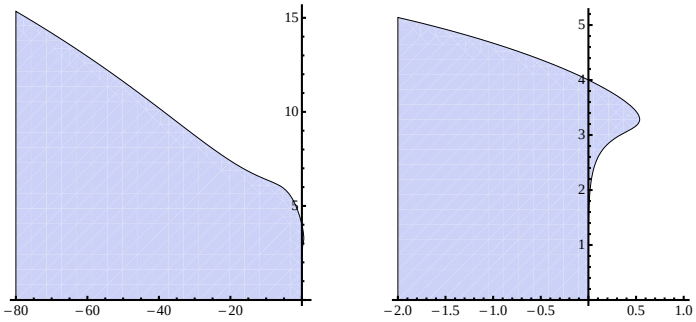
where $\lambda^d, \mathtt{i}\lambda^a$ are the eigenvalues of the diffusion and advection operators:

$$
\begin{aligned}
\lambda^d &= -4\tfrac{\nu}{h^2}\textstyle\sum_{d=1}^{\mathrm{D}} \sin^2\tfrac{\theta_d}{2}\left(1 + \tfrac{1}{3}\sin^2\tfrac{\theta_d}{2}\right),\\
\lambda^a &= -\tfrac{1}{h}\textstyle\sum_{d=1}^{\mathrm{D}} u_d \sin\theta_d\left(1 + \tfrac{2}{3}\sin^2\tfrac{\theta_d}{2}\right),
\end{aligned} \tag{12.46}
$$

with $\theta_d = \xi_d h_d \in (0, 2\pi)$. Here ξ_d and h_d are the wave number and grid size in the dth dimension; we assume that a single Fourier mode of the cell average $\langle\phi\rangle_{\mathbf{i}}$ is of the form $y(t)e^{i\boldsymbol{\xi}\cdot\mathbf{x}_{\mathbf{i}}}$.

Exercise 12.22. Prove Lemma 12.21.

Exercise 12.23. Reproduce the following stability region $|R(z)| < 1$ in the complex plane for the fourth-order advection-diffusion solver that discretizes (12.37) with operators in Definition 12.7 and adopts `ARK4(3)6L[2]SA` to solve the resulting ODE system (12.38).



The first plot shows the stability region in the range $(\bar{\lambda}_d, \bar{\lambda}_a) \in [-80, 0] \times [0, 15]$ and the second is a zoom-in of the first near the origin. It is clear from these plots that

- the maximum stable Courant number increases as diffusion becomes stronger,
- in the absence of diffusion the scaled advection eigenvalue should be less than 4.

Deduce from the second plot and (12.46) that the range of stable Courant numbers for `ARK4(3)6L[2]SA` is

$$\mu \leq \frac{2.91}{\mathrm{D}}. \tag{12.47}$$

Theorem 12.24. The FV method in Algorithm 12.18 is convergent with fourth-order accuracy.

Proof. This follows from Lemma 12.8, Lemma 12.21, and Theorem 10.22. □

Chapter 13

FV Methods for the Incompressible Navier-Stokes Equations (INSE)

Definition 13.1. A *domain* is a connected and bounded regular open subset $\Omega \subset \mathbb{R}^{\mathrm{D}}$ with $\mathrm{D} = 2, 3$.

Definition 13.2. The incompressible Navier-Stokes equations (INSE) is

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = \mathbf{g} - \nabla p + \nu \Delta \mathbf{u}, \tag{13.1a}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{13.1b}$$

where $t \in [0, +\infty)$ is time, $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}\,(\mathrm{D} = 2, 3)$ the spatial location, $\mathbf{g}$ the external forcing term, p the pressure, $\mathbf{u}$ the velocity, ν the dynamic viscosity.

Definition 13.3. The *Eulerian accelerations* of the INSE are vectors of time derivatives

$$\mathbf{a} := \frac{\partial \mathbf{u}}{\partial t}, \qquad \mathbf{a}^* := -\mathbf{u} \cdot \nabla \mathbf{u} + \mathbf{g} + \nu \Delta \mathbf{u}. \tag{13.2}$$

Definition 13.4. The *pressure Poisson equation* (PPE) is an elliptic equation that describes the relation between the pressure and the velocity in the INSE,

$$\Delta p = \nabla \cdot (\mathbf{g} - \mathbf{u} \cdot \nabla \mathbf{u}) \qquad \text{in } \Omega, \tag{13.3a}$$

$$\mathbf{n} \cdot \nabla p = \mathbf{n} \cdot (\mathbf{a}^* - \mathbf{a}) \qquad \text{on } \partial\Omega. \tag{13.3b}$$

13.1 Leray-Helmholtz Projection

Definition 13.5. A *projection* $\overline{\mathbf{P}}$ is a linear transformation from a vector space to itself such that the idempotent condition holds

$$\overline{\mathbf{P}}^2 = \overline{\mathbf{P}}. \tag{13.4}$$

Definition 13.6. The *Leray-Helmholtz projection* on a domain $\Omega \subset \mathbb{R}^{\mathrm{D}}$ is a projection $\mathscr{P} : \mathcal{C}^1(\Omega) \to \mathcal{C}^1(\Omega)$

$$\mathscr{P}\mathbf{v}^* := \mathbf{v} = \mathbf{v}^* - \nabla \phi \tag{13.5}$$

such that $\mathbf{v}, \mathbf{v}^* : \Omega \to \mathbb{R}^{\mathrm{D}}$ are vector fields, $\mathbf{v}$ satisfies $\nabla \cdot \mathbf{v} = 0$, and $\phi : \Omega \to \mathbb{R}$ is a scalar function.

Definition 13.7. The *no-penetration condition* for a vector field $\mathbf{v}$ on a domain Ω is the boundary condition

$$\mathbf{v} \cdot \mathbf{n} = 0, \tag{13.6}$$

where $\mathbf{n}$ is the outward normal of the domain boundary $\partial\Omega$.

Lemma 13.8. On domains with periodic or no-penetration conditions, the Leray-Helmholtz projection is well defined and can be expressed as

$$\mathscr{P} = 1 - \nabla (\Delta_n)^{-1} \nabla \cdot, \tag{13.7}$$

where $(\Delta_n)^{-1}$ denotes solving Poisson's equation with pure Neumann conditions.

Proof. By Helmholtz's theorem, a sufficiently continuous vector field $\mathbf{v}^*$ on regular compact domain $\overline{\Omega}$ in $\mathbb{R}^{\mathrm{D}}$ can be uniquely decomposed into a divergence-free part $\mathbf{v}$ and a curl-free part $\nabla\phi$:

$$\mathbf{v}^* = \mathbf{v} + \nabla \phi,$$
$$\nabla \cdot \mathbf{v} = 0, \qquad \nabla \times \nabla \phi = \mathbf{0}.$$

This is achieved by solving the Poisson's equation with pure Neumann boundary conditions:

$$\Delta \phi = \nabla \cdot \mathbf{v}^* \qquad \text{in } \Omega, \tag{13.8a}$$

$$\mathbf{n} \cdot \nabla \phi = \mathbf{n} \cdot (\mathbf{v}^* - \mathbf{v}) \qquad \text{on } \partial\Omega, \tag{13.8b}$$

where $\mathbf{n}$ denotes the outward normal of the domain boundary $\partial\Omega$. The above arguments justifies (13.7) and it remains to show that the BVP (13.8) admits a unique solution.

Periodic conditions imply $\oint_{\partial\Omega} \mathbf{n} \cdot \mathbf{v} = 0$. As for no-penetration conditions, Gauss theorem and $\nabla \cdot \mathbf{v} = 0$ yield

$$0 = \int_\Omega \nabla \cdot \mathbf{v} = \oint_{\partial\Omega} \mathbf{n} \cdot \mathbf{v},$$

thus the solvability of (13.8) holds. Since ϕ in (13.8) is determined up to an additive constant, $\nabla\phi$ is unique, which further implies the uniqueness of $\mathbf{v}$ in (13.5). □

Lemma 13.9. The Leray-Helmholtz projection $\mathscr{P}$ satisfies

$$\mathscr{P}^2 = \mathscr{P}, \quad \nabla \cdot \mathscr{P}\mathbf{v}^* = 0, \quad \mathscr{P}\nabla\phi = \mathbf{0}. \tag{13.9}$$

Proof. These identities follow from Definition 13.6 and Lemma 13.8. □

13.2 The approximate projection

Definition 13.10. The fourth-order *approximate projection* associated with the Leray-Helmholtz projection is the discrete operator

$$\mathbf{P} = \mathbf{I} - \mathbf{G}\mathbf{L}^{-1}\mathbf{D}, \tag{13.10}$$

where $\mathbf{I}$ is the identity operator and the other operators are the same as those in Definition 12.7.

Exercise 13.11. For periodic domains, express $\mathbf{DG}$ as a linear combination of cell averages to verify that $\mathbf{DG} \neq \mathbf{L}$. What is the one-dimensional stencil of this operator? Can you say anything about this stencil?

Lemma 13.12. The approximate projection $\mathbf{P}$ is not a projection.

Proof. Exercise 13.11 implies $\mathbf{P}^2 \neq \mathbf{P}$, and the conclusion follows from Definition 13.5. □

Exercise 13.13. Show that the discrete operator

$$\mathbf{P}_E = \mathbf{I} - \mathbf{G}(\mathbf{D}\mathbf{G})^{-1}\mathbf{D} \tag{13.11}$$

is indeed a projection.

Definition 13.14. The *FV scalar inner product* and the *FV vector inner product* on a domain Ω are respectively

$$\begin{aligned} \langle \phi, \psi \rangle_S &= h^{\mathrm{D}} \textstyle\sum_{\mathbf{i}} \langle \phi \rangle_{\mathbf{i}} \langle \psi \rangle_{\mathbf{i}}, \\ \langle \mathbf{v}, \mathbf{w} \rangle_V &= h^{\mathrm{D}} \textstyle\sum_{\mathbf{i}} \langle \mathbf{v} \rangle_{\mathbf{i}} \cdot \langle \mathbf{w} \rangle_{\mathbf{i}}, \end{aligned} \tag{13.12}$$

where $\phi, \psi : \Omega \to \mathbb{R}$ are scalar functions and $\mathbf{v}, \mathbf{w} : \Omega \to \mathbb{R}^{\mathrm{D}}$ are vector functions.

Lemma 13.15. On periodic domains, the linear maps $\mathbf{G}$ and $-\mathbf{D}$ in (12.20) and (12.21) are adjoint in the sense that

$$\langle \mathbf{D}\mathbf{u}, \phi \rangle_S = -\langle \mathbf{u}, \mathbf{G}\phi \rangle_V. \tag{13.13}$$

The corresponding matrices satisfy $\mathbf{G} = -\mathbf{D}^T$.

Proof. It suffices to show

$$\begin{aligned} &\langle \mathbf{D}\mathbf{u}, \phi \rangle_S + \langle \mathbf{u}, \mathbf{G}\phi \rangle_V \\ =& h^{\mathrm{D}} \sum_{\mathbf{i}} \big(\langle \phi \rangle_{\mathbf{i}} \mathbf{D} \langle \mathbf{u} \rangle_{\mathbf{i}} + \langle \mathbf{u} \rangle_{\mathbf{i}} \cdot \mathbf{G} \langle \phi \rangle_{\mathbf{i}} \big) = 0. \end{aligned}$$

Consider all possible terms containing $\langle \phi \rangle_{\mathbf{i}}$. For dimension d, $\sum_{\mathbf{j}} \big(\langle \phi \rangle_{\mathbf{j}} \mathbf{D} \langle \mathbf{u} \rangle_{\mathbf{j}} \big)$ expands to

$$\langle \phi \rangle_{\mathbf{i}} \big(8 \langle u_d \rangle_{\mathbf{i}+\mathbf{e}^d} - \langle u_d \rangle_{\mathbf{i}+2\mathbf{e}^d} - 8 \langle u_d \rangle_{\mathbf{i}-\mathbf{e}^d} + \langle u_d \rangle_{\mathbf{i}-2\mathbf{e}^d} \big).$$

Similarly, $\sum_{\mathbf{j}} \big(\langle \mathbf{u} \rangle_{\mathbf{j}} \cdot \mathbf{G} \langle \phi \rangle_{\mathbf{j}} \big)$ expands to

$$\langle \phi \rangle_{\mathbf{i}} \big(-8 \langle u_d \rangle_{\mathbf{i}+\mathbf{e}^d} + \langle u_d \rangle_{\mathbf{i}+2\mathbf{e}^d} + 8 \langle u_d \rangle_{\mathbf{i}-\mathbf{e}^d} - \langle u_d \rangle_{\mathbf{i}-2\mathbf{e}^d} \big).$$

In the former, all the terms are contributed by $\langle \phi \rangle_{\mathbf{i}} \mathbf{D} \langle \mathbf{u} \rangle_{\mathbf{i}}$; in the latter, no terms come from $\langle \mathbf{u} \rangle_{\mathbf{i}} \cdot \mathbf{G} \langle \phi \rangle_{\mathbf{i}}$, e.g., $-8 \langle \phi \rangle_{\mathbf{i}} \langle u_d \rangle_{\mathbf{i}+\mathbf{e}^d}$ is contributed by $\langle \mathbf{u} \rangle_{\mathbf{i}+\mathbf{e}^d} \cdot \mathbf{G} \langle \phi \rangle_{\mathbf{i}+\mathbf{e}^d}$. Because of periodicity, all five multi-indices are well defined for the cells to remain inside the domain, hence these above terms cancel. The same argument also applies to the terms containing $\langle u_d \rangle_{\mathbf{i}}$ for all d. □

Corollary 13.16. On periodic domains, the corresponding matrix of the approximate projection operator $\mathbf{P}$ defined in (13.10) is symmetric, i.e. $\mathbf{P}^T = \mathbf{P}$.

Proof. The symmetry of $\mathbf{L}$ and Lemma 13.15 yield

$$\mathbf{P}^T = (\mathbf{I} - \mathbf{G}\mathbf{L}^{-1}\mathbf{D})^T = \mathbf{I} - \mathbf{D}^T\mathbf{L}^{-1}\mathbf{G}^T = \mathbf{I} - \mathbf{G}\mathbf{L}^{-1}\mathbf{D} = \mathbf{P}.$$

□

Lemma 13.17. On periodic domains, the discrete gradient $\mathbf{G}$ in (12.20) and discrete Laplacian $\mathbf{L}$ in (12.22) commute,

$$\mathbf{G}\mathbf{L} = \mathbf{L}\mathbf{G}. \tag{13.14}$$

Consequently, the discrete Laplacian and the approximate projection commute,

$$\mathbf{P}\mathbf{L} = \mathbf{L}\mathbf{P}. \tag{13.15}$$

Proof. For a 1D periodic domain, let $\bar{\mathbf{G}}$ and $\bar{\mathbf{L}}$ denote the matrices of the discrete gradient operator and Laplacian operator scaled by $12h$ and $12h^2$, respectively. From (12.20) and (12.22), we have

$$\bar{\mathbf{G}}_{i,j} = \begin{cases} \pm 8, & j = \mathrm{mod}(i \pm 1, m) \\ \mp 1, & j = \mathrm{mod}(i \pm 2, m) \\ 0, & \text{otherwise} \end{cases}, \tag{13.16}$$

$$\bar{\mathbf{L}}_{i,j} = \begin{cases} -30, & j = i \\ 16, & j = \mathrm{mod}(i \pm 1, m) \\ -1, & j = \mathrm{mod}(i \pm 2, m) \\ 0, & \text{otherwise} \end{cases}, \tag{13.17}$$

where m is the number of cells. To avoid clustering of notation, I drop "mod" in the indices of matrix entries to use the cyclic shorthands "$i \pm \cdot$" for "$\mathrm{mod}(i \pm \cdot, m)$," "$k \pm \cdot$" for "$\mathrm{mod}(k \pm \cdot, m)$," and so on. It follows that

$$\begin{aligned} (\bar{\mathbf{G}}\bar{\mathbf{L}})_{k,\ell} &= \textstyle\sum_{j=k-2}^{k+2} \bar{\mathbf{G}}_{k,j} \bar{\mathbf{L}}_{j,\ell} \\ &= -\bar{\mathbf{L}}_{k+2,\ell} + 8\bar{\mathbf{L}}_{k+1,\ell} - 8\bar{\mathbf{L}}_{k-1,\ell} + \bar{\mathbf{L}}_{k-2,\ell}, \\ (\bar{\mathbf{L}}\bar{\mathbf{G}})_{k,\ell} &= \textstyle\sum_{j=k-2}^{k+2} \bar{\mathbf{L}}_{k,j} \bar{\mathbf{G}}_{j,\ell} \\ &= -\bar{\mathbf{L}}_{k,\ell-2} + 8\bar{\mathbf{L}}_{k,\ell-1} - 8\bar{\mathbf{L}}_{k,\ell+1} + \bar{\mathbf{L}}_{k,\ell+2}. \end{aligned}$$

Since $\bar{\mathbf{L}}$ is a Toeplitz matrix, we have $\bar{\mathbf{L}}_{k+2,\ell} = \bar{\mathbf{L}}_{k,\ell-2}$, $\bar{\mathbf{L}}_{k+1,\ell} = \bar{\mathbf{L}}_{k,\ell-1}$, and so on. It follows that

$$\bar{\mathbf{G}}\bar{\mathbf{L}} = \bar{\mathbf{L}}\bar{\mathbf{G}}. \tag{13.18}$$

For a 2D domain, the matrices of the discrete operators can be expressed as Kronecker products of their 1D counterparts and identity matrices

$$12h^2\mathbf{L} = \bar{\mathbf{L}} \otimes \mathbf{I} + \mathbf{I} \otimes \bar{\mathbf{L}}, \quad 12h\mathbf{G} = \big(\bar{\mathbf{G}} \otimes \mathbf{I},\ \mathbf{I} \otimes \bar{\mathbf{G}} \big)^T. \tag{13.19}$$

In order to prove (13.14), it suffices to show that $\mathbf{L}$ commutes with each subblock of $\mathbf{G}$, which follows from (13.18) and the mixed-product property of Kronecker products, i.e. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. For three and higher dimensions, (13.14) can be proved by a straightforward induction based on the 1D and 2D arguments.

Finally, (13.15) follows directly from (13.10), (13.14), and Lemma 13.15. □

Theorem 13.18. On periodic domains, both the spectral radius and the Euclidean 2-norm of the approximate projection operator are one,

$$\rho(\mathbf{P}) = \|\mathbf{P}\|_2 = 1. \tag{13.20}$$

Furthermore, $\mathbf{P}$ is a fourth-order approximation to the Leray-Helmholtz projection,

$$\mathbf{P}\langle\mathbf{u}\rangle_{\mathbf{i}} - \frac{1}{h^{\mathrm{D}}}\int_{\mathcal{C}_{\mathbf{i}}}\left(\mathbf{I} - \nabla(\Delta_n^{-1})\nabla\cdot\right)\mathbf{u} = O(h^4). \tag{13.21}$$

Proof. It follows from Corollary 13.16 that $\rho(\mathbf{P}) = 1$ implies $\|\mathbf{P}\|_2 = 1$, hence we only need to show the former. For the projection $\mathbf{P}_E$ in (13.11), we have $\lambda(\mathbf{P}_E) \in \{0, 1\}$ since its *minimal polynomial* is $\mathbf{P}_E^2 - \mathbf{P}_E = 0$. Let $\mathbf{P}_E = \mathbf{I} - \mathbf{Q}_E$, $\mathbf{P} = \mathbf{I} - \mathbf{Q}$. Clearly, $\lambda(\mathbf{Q}_E) \in \{0, 1\}$, and we only need to show $0 \le \lambda(\mathbf{Q}) \le 1$. By Lemma 13.17, $\mathbf{G}$ commutes with both $\mathbf{L}^{-1}$ and $\mathbf{DG}$. Thus we have

$$\begin{aligned}\mathbf{Q} =& \mathbf{G}\mathbf{L}^{-1}(\mathbf{DG})(\mathbf{DG})^{-1}\mathbf{D}\\ =& \mathbf{L}^{-1}(\mathbf{DG})\mathbf{G}(\mathbf{DG})^{-1}\mathbf{D}\\ =& \mathbf{L}^{-1}\mathbf{DG}\mathbf{Q}_E,\end{aligned}$$

and it suffices to shows

$$0 \le \lambda\left(\mathbf{L}^{-1}\mathbf{DG}\right) \le 1.$$

Using discrete Fourier analysis, we can define the shift operator as

$$s_d\langle\phi\rangle_{\mathbf{i}} = \langle\phi\rangle_{\mathbf{i}+\mathbf{e}^d}, \tag{13.22}$$

whose eigenvectors are the single Fourier modes with the eigenvalues $e^{\mathrm{i}\beta_d}$, where $\beta_d = \kappa_d\frac{\pi}{N}$, $\kappa_d = 1, \cdots, N-1$, and N the total number of points in dimension d. It follows from (12.20), (12.21), and (12.22) that for a given Fourier component,

$$\begin{aligned}\lambda(\mathbf{DG}) =& -\frac{4}{h^2}\sum_{d=1}^{\mathrm{D}}\sin^2\frac{\beta_d}{2}\left(1-\sin^2\frac{\beta_d}{2}\right)\left(1+\frac{2}{3}\sin^2\frac{\beta_d}{2}\right)^2;\\ \mathbf{L} =& \frac{1}{12h^2}\sum_{d=1}^{\mathrm{D}}\left(16s_d + \frac{16}{s_d} - 30 - s_d^2 - \frac{1}{s_d^2}\right)\\ \Rightarrow \lambda(\mathbf{L}) =& -\frac{4}{h^2}\sum_{d=1}^{\mathrm{D}}\sin^2\frac{\beta_d}{2}\left(1+\frac{1}{3}\sin^2\frac{\beta_d}{2}\right).\end{aligned}$$

$\lambda\left(\mathbf{L}^{-1}\mathbf{DG}\right) \ge 0$ follows from the negative definiteness of $\lambda(\mathbf{L})$ and $\lambda(\mathbf{DG})$. $\lambda\left(\mathbf{L}^{-1}\mathbf{DG}\right) \le 1$ holds because

$$\eta(1-\eta)\left(1+\tfrac{2}{3}\eta\right)^2 - \eta\left(1+\tfrac{1}{3}\eta\right) = -\tfrac{4}{9}\eta^2(2\eta+\eta^2) \le 0,$$

for $\eta = \sin^2\frac{\beta_d}{2} \in [0, 1]$.

Finally, (13.21) follows directly from (12.26a), (12.26b), and (12.26c) in Lemma 12.8; it can also be proved via considering the Taylor expansions of the symbols of individual operators for any *fixed* Fourier mode. □

13.3 The INSE on periodic domains

Theorem 13.19. The following ODE is a fourth-order approximation to the INSE (13.1) on periodic domains:

$$\frac{\mathrm{d}\langle\mathbf{u}\rangle}{\mathrm{d}t} = \mathbf{P}\Big(-\mathbf{D}\langle\mathbf{uu}\rangle + \langle\mathbf{g}\rangle + \nu\mathbf{L}\langle\mathbf{u}\rangle\Big). \tag{13.23}$$

Proof. This follows from applying the Leray-Helmholtz projection to the INSE (13.1), taking average of the resulting equation, using (13.21) and Lemma 12.8. □

Algorithm 13.20. A *fourth-order FV method for solving the INSE on periodic domains* is obtained by directly applying the ERK-ESDIRK IMEX algorithm (12.40) to the ODE system (13.23) with

$$\mathbf{X}^{[\mathrm{E}]}\langle\mathbf{u}\rangle = -\mathbf{D}\langle\mathbf{uu}\rangle + \langle\mathbf{g}\rangle, \qquad \mathbf{X}^{[\mathrm{I}]}\langle\mathbf{u}\rangle = \nu\mathbf{L}\langle\mathbf{u}\rangle. \tag{13.24}$$

More precisely, the algorithmic steps are

$$\langle\mathbf{u}\rangle^{(1)} = \langle\mathbf{u}\rangle^n \approx \langle\mathbf{u}(t^n)\rangle, \tag{13.25a}$$

for $s = 2, 3, \ldots, n_{\mathbf{s}}$,

$$\begin{aligned}(\mathbf{I}-k\nu\gamma\mathbf{L})\langle\mathbf{u}\rangle^{(s)} =& \langle\mathbf{u}\rangle^n + k\sum_{j=1}^{s-1}a_{s,j}^{[\mathrm{E}]}\mathbf{P}\mathbf{X}^{[\mathrm{E}]}\langle\mathbf{u}\rangle^{(j)}\\ &+ k\nu\sum_{j=1}^{s-1}a_{s,j}^{[\mathrm{I}]}\mathbf{L}\langle\mathbf{u}\rangle^{(j)}, \qquad (13.25b)\\ \langle\mathbf{u}^*\rangle^{n+1} =& \langle\mathbf{u}\rangle^n + k\sum_{s=1}^{n_{\mathbf{s}}}b_s^{[\mathrm{E}]}\mathbf{X}^{[\mathrm{E}]}\langle\mathbf{u}\rangle^{(s)}\\ &+ k\nu\sum_{s=1}^{n_{\mathbf{s}}}b_s^{[\mathrm{I}]}\mathbf{L}\langle\mathbf{u}\rangle^{(s)}, \qquad (13.25c)\\ \langle\mathbf{u}\rangle^{n+1} =& \mathbf{P}\langle\mathbf{u}^*\rangle^{n+1}. \qquad (13.25d)\end{aligned}$$

At any time instant, the pressure can be extracted from the velocity by solving the discrete pressure Poisson equation

$$\mathbf{L}\langle p\rangle = \mathbf{D}\big(\langle\mathbf{g}\rangle - \mathbf{D}\langle\mathbf{uu}\rangle + \nu\mathbf{L}\langle\mathbf{u}\rangle\big) \tag{13.26}$$

with periodic boundary conditions.

Theorem 13.21. The solution $\langle\mathbf{u}\rangle$ produced by Algorithm 13.20 evolves in a vector space that is solenoidal with fourth-order accuracy, i.e. $\mathbf{D}\langle\mathbf{u}\rangle = O(h^4)$.

Proof. It suffices to point out that the velocity at each intermediate stage satisfies $\mathbf{D}\langle\mathbf{u}\rangle^{(s)} = O(h^4)$ because

- the initial velocity $\langle\mathbf{u}\rangle^n$ satisfies $\mathbf{D}\langle\mathbf{u}\rangle^n = O(h^4)$,
- $\mathbf{DPX}^{[\mathrm{E}]}\langle\mathbf{u}\rangle = O(h^4)$,
- $\mathbf{L}$ is linear and commutes with $\mathbf{P}$. □

Appendix A

Sets, Logic, and Functions

A.1 First-order logic

Definition A.1. A *set* $\mathcal{S}$ is a collection of *distinct* objects that share a common quality; it is often denoted with the following notation

$$\mathcal{S} = \{x \mid \text{ the conditions that } x \text{ satisfies. } \}. \tag{A.1}$$

Notation 12. $\mathbb{R}, \mathbb{Z}, \mathbb{N}, \mathbb{Q}, \mathbb{C}$ denote the sets of real numbers, integers, natural numbers, rational numbers and complex numbers, respectively. $\mathbb{R}^+, \mathbb{Z}^+, \mathbb{N}^+, \mathbb{Q}^+$ the sets of positive such numbers. In particular, $\mathbb{N}$ contains the number zero while $\mathbb{N}^+$ does not.

Definition A.2. $\mathcal{S}$ is a *subset* of $\mathcal{U}$, written $\mathcal{S} \subseteq \mathcal{U}$, if and only if (iff) $x \in \mathcal{S} \Rightarrow x \in \mathcal{U}$. $\mathcal{S}$ is a *proper subset* of $\mathcal{U}$, written $\mathcal{S} \subset \mathcal{U}$, if $\mathcal{S} \subseteq \mathcal{U}$ and $\exists x \in \mathcal{U}$ s.t. $x \notin \mathcal{S}$.

Definition A.3 (Statements of first-order logic)**.** A *universal statement* is a logical statement of the form

$$\mathsf{U} = (\forall x \in \mathcal{S},\ \mathsf{A}(x)). \tag{A.2}$$

An *existential statement* has the form

$$\mathsf{E} = (\exists x \in \mathcal{S},\ \text{s.t. } \mathsf{A}(x)), \tag{A.3}$$

where $\forall$ ("for each") and $\exists$ ("there exists") are the *quantifiers*, $\mathcal{S}$ is a set, "s.t." means "such that," and $\mathsf{A}(x)$ is the *formula*.
A statement of *implication/conditional* has the form

$$\mathsf{A} \Rightarrow \mathsf{B}. \tag{A.4}$$

Example A.4. Universal and existential statements:
$\forall x \in [2, +\infty),\ x > 1$;
$\forall x \in \mathbb{R}^+,\ x > 1$;
$\exists p, q \in \mathbb{Z},\ \text{s.t. } p/q = \sqrt{2}$;
$\exists p, q \in \mathbb{Z},\ \text{s.t. } \sqrt{p} = \sqrt{q} + 1$.

Definition A.5. *Uniqueness quantification* or *unique existential quantification*, written $\exists!$ or $\exists_{=1}$, indicates that exactly one object with a certain property exists.

Exercise A.6. Express the logical statement $\exists! x,\ \text{s.t. } \mathsf{A}(x)$ with $\exists$, $\forall$, and $\Leftrightarrow$.

Definition A.7. A *universal-existential statement* is a logical statement of the form

$$\mathsf{U}_E = (\forall x \in \mathcal{S},\ \exists y \in \mathcal{T} \text{ s.t. } \mathsf{A}(x, y)). \tag{A.5}$$

An *existential-universal statement* has the form

$$\mathsf{E}_U = (\exists y \in \mathcal{T},\ \text{s.t. } \forall x \in \mathcal{S},\ \mathsf{A}(x, y)). \tag{A.6}$$

Example A.8. True or false:
$\forall x \in [2, +\infty),\ \exists y \in \mathbb{Z}^+ \text{ s.t. } x^y < 10^5$;
$\exists y \in \mathbb{R} \text{ s.t. } \forall x \in [2, +\infty),\ x > y$;
$\exists y \in \mathbb{R} \text{ s.t. } \forall x \in [2, +\infty),\ x < y$.

Example A.9 (Translating an English statement into a logical statement)**.** Goldbach's conjecture states *every even natural number greater than 2 is the sum of two primes.* Let $\mathbb{P} \subset \mathbb{N}^+$ denote the set of prime numbers. Then Goldbach's conjecture is $\forall a \in 2\mathbb{N}^+ + 2, \exists p, q \in \mathbb{P}, \text{ s.t. } a = p + q$.

Theorem A.10. The existential-universal statement implies the corresponding universal-existential statement, but not vice versa.

Example A.11 (Translating a logical statement to an English statement)**.** Let $\mathcal{S}$ be the set of all human beings.
U_E =($\forall p \in \mathcal{S}, \exists q \in \mathcal{S}$ s.t. q is p's mom.)
E_U=($\exists q \in \mathcal{S}$ s.t. $\forall p \in \mathcal{S}$, q is p's mom.)
U_E is probably true, but E_U is certainly false.
If E_U were true, then U_E would be true. Why?

Axiom A.12 (First-order negation of logical statements)**.** The negations of the statements in Definition A.3 are

$$\neg\mathsf{U} = (\exists x \in \mathcal{S},\ \text{s.t. } \neg\mathsf{A}(x)). \tag{A.7}$$
$$\neg\mathsf{E} = (\forall x \in \mathcal{S},\ \neg\mathsf{A}(x)). \tag{A.8}$$

Rule A.13. The negation of a more complicated logical statement abides by the following rules:

- switch the type of each quantifier until you reach the last formula without quantifiers;
- negate the last formula.

One might need to group quantifiers of like type.

Example A.14 (The negation of Goldbach's conjecture)**.** $\exists a \in 2\mathbb{N}^+ + 2 \text{ s.t. } \forall p, q \in \mathbb{P},\ a \neq p + q$.

Exercise A.15. Negate the logical statement in Definition C.57.

Axiom A.16 (Contraposition)**.** A conditional statement is logically equivalent to its contrapositive.

$$(\mathsf{A} \Rightarrow \mathsf{B}) \Leftrightarrow (\neg\mathsf{B} \Rightarrow \neg\mathsf{A}) \tag{A.9}$$

Example A.17. "If Jack is a man, then Jack is a human being." is equivalent to "If Jack is not a human being, then Jack is not a man."

Exercise A.18. Draw an Euler diagram of subsets to illustrate Example A.17.

Exercise A.19. Rewrite each of the following statements and its *negation* into *logical statements* using symbols, quantifiers, and formulas.

(a) The only even prime is 2.

(b) Multiplication of integers is associative.

(c) Goldbach's conjecture has at most a finite number of counterexamples.

A.2 Ordered sets

Definition A.20. The *Cartesian product* $\mathcal{X} \times \mathcal{Y}$ between two sets $\mathcal{X}$ and $\mathcal{Y}$ is the set of all possible ordered pairs with first element from $\mathcal{X}$ and second element from $\mathcal{Y}$:

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) \mid x \in \mathcal{X},\ y \in \mathcal{Y}\}. \tag{A.10}$$

Axiom A.21 (Fundamental principle of counting)**.** 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

$$\prod_{i=1}^{k} n_i = n_1 n_2 \cdots n_k. \tag{A.11}$$

Example A.22. Let A, E, D be the set of appetizers, main entrees, desserts in a restaurant. $A \times E \times D$ is the set of possible dinner combos. If $\#A = 10$, $\#E = 5$, $\#D = 6$, $\#(A \times E \times D) = 300$.

Definition A.23 (Maximum and minimum)**.** Consider $\mathcal{S} \subseteq \mathbb{R}$, $\mathcal{S} \neq \emptyset$. If $\exists s_m \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S}$, $x \leq s_m$, then s_m is the *maximum* of $\mathcal{S}$ and denoted by $\max \mathcal{S}$. If $\exists s_m \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S}$, $x \geq s_m$, then s_m is the *minimum* of $\mathcal{S}$ and denoted by $\min \mathcal{S}$.

Definition A.24 (Upper and lower bounds)**.** Consider $\mathcal{S} \subseteq \mathbb{R}$, $\mathcal{S} \neq \emptyset$. a is an *upper bound* of $\mathcal{S} \subseteq \mathbb{R}$ if $\forall x \in \mathcal{S}$, $x \leq a$; then the set $\mathcal{S}$ is said to be *bounded above*. a is a *lower bound* of $\mathcal{S}$ if $\forall x \in \mathcal{S}$, $x \geq a$; then the set $\mathcal{S}$ is said to be *bounded below*. $\mathcal{S}$ is *bounded* if it is bounded above and bounded below.

Definition A.25 (Supremum and infimum)**.** Consider a nonempty set $\mathcal{S} \subseteq \mathbb{R}$. If $\mathcal{S}$ is bounded above and $\mathcal{S}$ has a least upper bound then we call it the *supremum* of $\mathcal{S}$ and denote it by $\sup \mathcal{S}$. If $\mathcal{S}$ is bounded below and $\mathcal{S}$ has a greatest lower bound, then we call it the *infimum* of $\mathcal{S}$ and denote it by $\inf \mathcal{S}$.

Example A.26. If a set $\mathcal{S} \subset \mathbb{R}$ has a maximum, we have $\max \mathcal{S} = \sup \mathcal{S}$.

Example A.27. $\sup[a, b] = \sup[a, b) = \sup(a, b] = \sup(a, b)$.

Axiom A.28 (Completeness of $\mathbb{R}$)**.** Every nonempty subset of $\mathbb{R}$ that is bounded above has a least upper bound.

Corollary A.29. Every nonempty subset of $\mathbb{R}$ that is bounded below has a greatest lower bound.

Definition A.30. A *binary relation between two sets* $\mathcal{X}$ and $\mathcal{Y}$ is an ordered triple $(\mathcal{X}, \mathcal{Y}, \mathcal{G})$ where $\mathcal{G} \subseteq \mathcal{X} \times \mathcal{Y}$.
A *binary relation on* $\mathcal{X}$ is the relation between $\mathcal{X}$ and $\mathcal{X}$.
The statement $(x, y) \in R$ is read "x is R-related to y," and denoted by xRy or $R(x, y)$.

Definition A.31. An *equivalence relation* "$\sim$" on $\mathcal{A}$ is a binary relation on $\mathcal{A}$ that satisfies $\forall a, b, c \in \mathcal{A}$,

- $a \sim a$ (reflexivity);
- $a \sim b$ implies $b \sim a$ (symmetry);
- $a \sim b$ and $b \sim c$ imply $a \sim c$ (transitivity).

Definition A.32. A binary relation "$\leq$" on some set $\mathcal{S}$ is a *total order* or *linear order* on $\mathcal{S}$ iff, $\forall a, b, c \in \mathcal{S}$,

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

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

Example A.33. The real numbers with less or equal.

Example A.34. The English letters of the alphabet with dictionary order.

Example A.35. The Cartesian product of a set of totally ordered sets with the *lexicographical order*.

Example A.36. Sort your book in lexicographical order and save a lot of time. $\log_{26} N \ll N!$

Definition A.37. A binary relation "$\leq$" on some set $\mathcal{S}$ is a *partial order* on $\mathcal{S}$ iff, $\forall a, b, c \in \mathcal{S}$, antisymmetry, transitivity, and reflexivity ($a \leq a$) hold.
A set equipped with a partial order is a called a *poset*.

Example A.38. The set of subsets of a set $\mathcal{S}$ ordered by inclusion "$\subseteq$."

Example A.39. The natural numbers equipped with the relation of divisibility.

Example A.40. The set of stuff you will put on your body every morning with the time ordered: undershorts, pants, belt, shirt, tie, jacket, socks, shoes, watch.

Example A.41. Inheritance ("is-a" relation) is a partial order. $A \to B$ reads "B is a special type of A".

Example A.42. Composition ("has-a" relation) is also a partial order. $A \rightsquigarrow B$ reads "B *has an* instance/object of A."

Example A.43. Implication "$\Rightarrow$" is a partial order on the set of logical statements.

Example A.44. The set of definitions, axioms, propositions, theorems, lemmas, etc., is a poset with inheritance, composition, and implication. It is helpful to relate them with these partial orderings.

A.3 Functions

Definition A.45. A *function/map/mapping* f from $\mathcal{X}$ to $\mathcal{Y}$, written $f : \mathcal{X} \to \mathcal{Y}$ or $\mathcal{X} \mapsto \mathcal{Y}$, is a subset of the Cartesian product $\mathcal{X} \times \mathcal{Y}$ satisfying that $\forall x \in \mathcal{X}$, there is exactly one $y \in \mathcal{Y}$ s.t. $(x, y) \in \mathcal{X} \times \mathcal{Y}$. $\mathcal{X}$ and $\mathcal{Y}$ are the *domain* and *range* of f, respectively.

Definition A.46. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be *injective* or *one-to-one* iff

$$\forall x_1 \in \mathcal{X}, \forall x_2 \in \mathcal{X}, \quad x_1 \neq x_2 \;\Rightarrow\; f(x_1) \neq f(x_2). \tag{A.12}$$

It is *surjective* or *onto* iff

$$\forall y \in \mathcal{Y}, \; \exists x \in \mathcal{X}, \text{ s.t. } y = f(x). \tag{A.13}$$

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

Definition A.47. A set $\mathcal{S}$ is *countably infinite* iff there exists a bijective function $f : \mathcal{S} \to \mathbb{N}^+$ that maps $\mathcal{S}$ to $\mathbb{N}^+$. A set is *countable* if it is either finite or countably infinite.

Example A.48. Are the integers countable? Are the rationals countable? Are the real numbers countable?

Definition A.49. A *binary function* or a *binary operation* on a set $\mathcal{S}$ is a map $\mathcal{S} \times \mathcal{S} \to \mathcal{S}$.

Appendix B

Linear Algebra

B.1 Vector spaces

Definition B.1. A *field* $\mathbb{F}$ is a set together with two binary operations, usually called "addition" and "multiplication" and denoted by "$+$" and "$*$", such that $\forall a, b, c \in \mathbb{F}$, the following axioms hold,

- commutativity: $a + b = b + a$, $ab = ba$;
- associativity: $a + (b + c) = (a + b) + c$, $a(bc) = (ab)c$;
- identity: $a + 0 = a$, $a1 = a$;
- invertibility: $a + (-a) = 0$, $aa^{-1} = 1$ $(a \neq 0)$;
- distributivity: $a(b + c) = ab + ac$.

Definition B.2. A *vector space* or *linear space* over a field $\mathbb{F}$ is a set $\mathcal{V}$ together with two binary operations "$+$" and "$\times$" respectively called vector addition and scalar multiplication that satisfy the following axioms:

(VSA-1) commutativity
$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$;

(VSA-2) associativity
$\forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$;

(VSA-3) compatibility
$\forall \mathbf{u} \in \mathcal{V}$, $\forall a, b \in \mathbb{F}$, $(ab)\mathbf{u} = a(b\mathbf{u})$;

(VSA-4) additive identity
$\forall \mathbf{u} \in \mathcal{V}$, $\exists \mathbf{0} \in \mathcal{V}$, 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

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

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

Definition B.3. A 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 B.4. The simplest vector space is $\{\mathbf{0}\}$. Another simple example of a vector space over a field $\mathbb{F}$ is $\mathbb{F}$ itself, equipped with its standard addition and multiplication.

B.1.1 Subspaces

Definition B.5. A subset $\mathcal{U}$ of $\mathcal{V}$ is called a *subspace* of $\mathcal{V}$ if $\mathcal{U}$ is also a vector space.

Definition B.6. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subsets of $\mathcal{V}$. The *sum* of $\mathcal{U}_1, \ldots, \mathcal{U}_m$ is the set of all possible sums of elements of $\mathcal{U}_1, \ldots, \mathcal{U}_m$:

$$\mathcal{U}_1 + \ldots + \mathcal{U}_m := \left\{ \sum_{j=1}^{m} \mathbf{u}_j : \mathbf{u}_j \in \mathcal{U}_j \right\}. \tag{B.1}$$

Example B.7. For $\mathcal{U} = \{(x, x, y, y) \in \mathbb{F}^4 : x, y \in \mathbb{F}\}$ and $\mathcal{W} = \{(x, x, x, y) \in \mathbb{F}^4 : x, y \in \mathbb{F}\}$, we have

$$\mathcal{U} + \mathcal{W} = \{(x, x, z, y) \in \mathbb{F}^4 : x, y, z \in \mathbb{F}\}.$$

Lemma B.8. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subspaces of $\mathcal{V}$. Then $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ is the smallest subspace of $\mathcal{V}$ that contains $\mathcal{U}_1, \ldots, \mathcal{U}_m$.

Definition B.9. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subspaces of $\mathcal{V}$. The sum $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ is called a *direct sum* if each element in $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ can be written in only one way as a sum $\sum_{j=1}^{m} \mathbf{u}_j$ with $\mathbf{u}_j \in \mathcal{U}_j$ for each $j = 1, \ldots, m$. In this case we write the direct sum as $\mathcal{U}_1 \oplus \ldots \oplus \mathcal{U}_m$.

Exercise B.10. Show that $\mathcal{U}_1 + \mathcal{U}_2 + \mathcal{U}_3$ is not a direct sum:

$$\begin{aligned} \mathcal{U}_1 &= \{(x, y, 0) \in \mathbb{F}^3 : x, y \in \mathbb{F}\}, \\ \mathcal{U}_2 &= \{(0, 0, z) \in \mathbb{F}^3 : z \in \mathbb{F}\}, \\ \mathcal{U}_3 &= \{(0, y, y) \in \mathbb{F}^3 : y \in \mathbb{F}\}. \end{aligned}$$

Lemma B.11. Suppose $\mathcal{U}_1, \ldots, \mathcal{U}_m$ are subspaces of $\mathcal{V}$. Then $\mathcal{U}_1 + \ldots + \mathcal{U}_m$ is a direct sum if and only if the only way to write $\mathbf{0}$ as a sum $\sum_{j=1}^{m} \mathbf{u}_j$, where $\mathbf{u}_j \in \mathcal{U}_j$ for each $j = 1, \ldots, m$, is by taking each $\mathbf{u}_j$ equal to $\mathbf{0}$.

Theorem B.12. Suppose $\mathcal{U}$ and $\mathcal{W}$ are subspaces of $\mathcal{V}$. Then $\mathcal{U} + \mathcal{W}$ is a direct sum if and only if $\mathcal{U} \cap \mathcal{W} = \{\mathbf{0}\}$.

B.1.2 Span and linear independence

Definition B.13. A *list of length* n or n*-tuple* is an ordered collection of n elements (which might be numbers, other lists, or more abstract entities) separated by commas and surrounded by parentheses: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Definition B.14. A vector space composed of all the n-tuples of a field $\mathbb{F}$ is known as a *coordinate space*, denoted by $\mathbb{F}^n$ ($n \in \mathbb{N}^+$).

Example B.15. The properties of forces or velocities in the real world can be captured by a coordinate space $\mathbb{R}^2$ or $\mathbb{R}^3$.

Example B.16. The set of continuous real-valued functions on the interval $[a, b]$ forms a real vector space.

Notation 13. 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 B.17. A *linear combination* of a list of vectors $\{\mathbf{v}_i\}$ is a vector of the form $\sum_i a_i \mathbf{v}_i$ where $a_i \in \mathbb{F}$.

Example B.18. $(17, -4, 2)$ is a linear combination of $(2, 1, -3), (1, -2, 4)$ because

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

Example B.19. $(17, -4, 5)$ is not a linear combination of $(2, 1, -3), (1, -2, 4)$ because there do not exist numbers a_1, a_2 such that

$$(17, -4, 5) = a_1(2, 1, -3) + a_2(1, -2, 4).$$

Solving from the first two equations yields $a_1 = 6$, $a_2 = 5$, but $5 \neq -3 \times 6 + 4 \times 5$.

Definition B.20. The *span* of a list of vectors $(\mathbf{v}_i)$ is the set of all linear combinations of $(\mathbf{v}_i)$,

$$\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) = \left\{ \sum_{i=1}^{m} a_i \mathbf{v}_i : \ a_i \in \mathbb{F} \right\}. \tag{B.2}$$

In particular, the span of the empty set is $\{\mathbf{0}\}$. We say that $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ *spans* $\mathcal{V}$ if $\mathcal{V} = \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$.

Example B.21.

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

Definition B.22. A vector space $\mathcal{V}$ is called *finite dimensional* if some list of vectors span $\mathcal{V}$; otherwise it is *infinite dimensional*.

Example B.23. Let $\mathbb{P}_m(\mathbb{F})$ denote the set of all polynomials with coefficients in $\mathbb{F}$ and degree at most m,

$$\mathbb{P}_m(\mathbb{F}) = \left\{ p : \mathbb{F} \to \mathbb{F};\ p(z) = \sum_{i=0}^{m} a_i z^i, a_i \in \mathbb{F} \right\}. \tag{B.3}$$

Then $\mathbb{P}_m(\mathbb{F})$ is a finite-dimensional vector space for each non-negative integer m. The set of all polynomials with coefficients in $\mathbb{F}$, denoted by $\mathbb{P}(\mathbb{F}) := \mathbb{P}_{+\infty}(\mathbb{F})$, is infinite-dimensional. Both are subspaces of $\mathbb{F}^{\mathbb{F}}$ for $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Definition B.24. A list of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ in $\mathcal{V}$ is called *linearly independent* iff

$$a_1 \mathbf{v}_1 + \ldots + a_m \mathbf{v}_m = \mathbf{0} \ \Rightarrow \ a_1 = \cdots = a_m = 0. \tag{B.4}$$

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

Example B.25. The empty list is declared to be linearly independent. A list of one vector $(\mathbf{v})$ is linearly independent iff $\mathbf{v} \neq \mathbf{0}$. A list of two vectors is linearly independent iff neither vector is a scalar multiple of the other.

Example B.26. The list $(1, z, \ldots, z^m)$ is linearly independent in $\mathbb{P}_m(\mathbb{F})$ for each $m \in \mathbb{N}$.

Example B.27. $(2, 3, 1)$, $(1, -1, 2)$, and $(7, 3, 8)$ is linearly dependent in $\mathbb{R}^3$ because

$$2(2, 3, 1) + 3(1, -1, 2) + (-1)(7, 3, 8) = (0, 0, 0).$$

Example B.28. Every list of vectors containing the $\mathbf{0}$ vector is linearly dependent.

Lemma B.29 (Linear dependence lemma)**.** Suppose $V = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m)$ is a linearly dependent list in $\mathcal{V}$. Then there exists $j \in \{1, 2, \ldots, m\}$ such that

- $\mathbf{v}_j \in \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1})$;
- if the jth term is removed from V, the span of the remaining list equals $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$.

Lemma B.30. In a finite-dimensional vector space, the length of every linearly independent list of vectors is less than or equal to the length of every spanning list of vectors.

B.1.3 Bases

Definition B.31. A *basis* of a vector space $\mathcal{V}$ is a list of vectors in $\mathcal{V}$ that is linearly independent and spans $\mathcal{V}$.

Definition B.32. The *standard basis* of $\mathbb{F}^n$ is the list of vectors

$$(1, 0, \cdots, 0)^T,\ (0, 1, 0, \cdots, 0)^T,\ \ldots,\ (0, \cdots, 0, 1)^T. \tag{B.5}$$

Example B.33. $(z^0, z^1, \ldots, z^m)$ is a basis of $\mathbb{P}_m(\mathbb{F})$ in (B.3).

Lemma B.34. A list of vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is a basis of $\mathcal{V}$ iff every vector $\mathbf{u} \in \mathcal{V}$ can be written uniquely as

$$\mathbf{u} = \sum_{i=1}^{n} a_i \mathbf{v}_i, \tag{B.6}$$

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

Lemma B.35. Every spanning list in a vector space $\mathcal{V}$ can be reduced to a basis of $\mathcal{V}$.

Lemma B.36. Every linearly independent list of vectors in a finite-dimensional vector space can be extended to a basis of that vector space.

B.1.4 Dimension

Definition B.37. The *dimension* of a finite-dimensional vector space $\mathcal{V}$, denoted $\dim \mathcal{V}$, is the length of any basis of the vector space.

Lemma B.38. If $\mathcal{V}$ is finite-dimensional, then every spanning list of vectors in $\mathcal{V}$ with length $\dim \mathcal{V}$ is a basis of $\mathcal{V}$.

Lemma B.39. If $\mathcal{V}$ is finite-dimensional, then every linearly independent list of vectors in $\mathcal{V}$ with length $\dim \mathcal{V}$ is a basis of $\mathcal{V}$.

B.2 Linear maps

Definition B.40. A *linear map* or *linear transformation* between two vector spaces $\mathcal{V}$ and $\mathcal{W}$ is a function $T : \mathcal{V} \to \mathcal{W}$ that satisfies

(LNM-1) additivity
$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $T(\mathbf{u} + \mathbf{v}) = T\mathbf{u} + T\mathbf{v}$;

(LNM-2) homogeneity
$\forall a \in \mathbb{F}$, $\forall \mathbf{v} \in \mathcal{V}$, $T(a\mathbf{v}) = a(T\mathbf{v})$,

where $\mathbb{F}$ is the underlying field of $\mathcal{V}$ and $\mathcal{W}$. In particular, a linear map is called a *linear operator* if $\mathcal{W} = \mathcal{V}$.

Notation 14. The set of all linear maps from $\mathcal{V}$ to $\mathcal{W}$ is denoted by $\mathcal{L}(\mathcal{V}, \mathcal{W})$. The set of all linear operators from $\mathcal{V}$ to itself is denoted by $\mathcal{L}(\mathcal{V})$.

Example B.41. The differentiation operator on $\mathbb{R}[x]$ is a linear map $T \in \mathcal{L}(\mathbb{R}[x], \mathbb{R}[x])$

Example B.42. $\mathbb{F}^{m\times n} = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a vector space with the additive identity as the zero map $\mathbf{0}$.

Lemma B.43. The set $\mathcal{L}(\mathcal{V}, \mathcal{W})$, equipped with scalar multiplication $(aT)\mathbf{v} = a(T\mathbf{v})$ and vector addition $(S + T)\mathbf{v} = S\mathbf{v} + T\mathbf{v}$, is a vector space.

Proof. The scalar field $\mathbb{F}$ of $\mathcal{L}(\mathcal{V}, \mathcal{W})$ is the same as that of $\mathcal{V}$ and $\mathcal{W}$. So multiplicative identity is still 1, the same as that of $\mathbb{F}$. However, the additive identity is the zero map $\mathbf{0} \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. □

Definition B.44. The *identity map*, denoted I, is the function on a vector space that assigns to each element to the same element:

$$I\mathbf{v} = \mathbf{v}. \tag{B.7}$$

B.2.1 Null spaces and ranges

Definition B.45. The *null space* of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ is the subset of $\mathcal{V}$ consisting of those vectors that T maps to the additive identity $\mathbf{0}$:

$$\text{null } T = \{\mathbf{v} \in \mathcal{V} : T\mathbf{v} = \mathbf{0}\}. \tag{B.8}$$

Example B.46. The null space of the differentiation map in Example B.41 is $\mathbb{R}$.

Definition B.47. The *range* of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ is the subset of $\mathcal{W}$ consisting of those vectors that are of the form $T\mathbf{v}$ for some $\mathbf{v} \in \mathcal{V}$:

$$\text{range } T = \{T\mathbf{v} : \mathbf{v} \in \mathcal{V}\}. \tag{B.9}$$

Example B.48. The range of $A \in \mathbb{C}^{m\times n}$ is the span of its column vectors.

Theorem B.49 (The counting theorem or the fundamental theorem of linear maps)**.** If $\mathcal{V}$ is a finite-dimensional vector space and $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, then range T is a finite-dimensional subspace of $\mathcal{W}$ and

$$\dim \mathcal{V} = \dim \text{null } T + \dim \text{range } T. \tag{B.10}$$

B.2.2 The matrix of a linear map

Definition B.50. The *matrix of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ with respect to the bases* $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ of $\mathcal{V}$ and $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m)$ of $\mathcal{W}$, denoted by

$$M_T := M(T, (\mathbf{v}_1, \dots, \mathbf{v}_n), (\mathbf{w}_1, \dots, \mathbf{w}_m)), \tag{B.11}$$

is the $m \times n$ matrix $A(T)$ whose entries $a_{i,j} \in \mathbb{F}$ satisfy the linear system

$$\forall j = 1, 2, \dots, n, \qquad T\mathbf{v}_j = \sum_{i=1}^{m} a_{i,j}\mathbf{w}_i. \tag{B.12}$$

Corollary B.51. The matrix M_T in (B.11) of a linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ satisfies

$$T[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m]M_T. \tag{B.13}$$

Proof. This follows directly from (B.11). □

B.2.3 Duality

Dual vector spaces

Definition B.52. The *dual space* of a vector space V is the vector space of all linear functionals on V,

$$V' = \mathcal{L}(V, \mathbb{F}). \tag{B.14}$$

Definition B.53. For a basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ of a vector space V, its *dual basis* is the list $\varphi_1, \dots, \varphi_n$ where each $\varphi_j \in V'$ is

$$\varphi_j(\mathbf{v}_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } k \neq j. \end{cases} \tag{B.15}$$

Exercise B.54. Show that the dual basis is a basis of the dual space.

Lemma B.55. A finite-dimensional vector space V satisfies

$$\dim V' = \dim V. \tag{B.16}$$

Proof. This follows from Definition B.52 and the identity $\dim \mathcal{L}(V, W) = \dim(V) \dim(W)$. □

Definition B.56. The *double dual space* of a vector space V, denoted by V'', is the dual space of V'.

Lemma B.57. The function $\Lambda : V \to V''$ defined as

$$\forall v \in V, \forall \varphi \in V', \qquad (\Lambda v)(\varphi) = \varphi(v) \tag{B.17}$$

is a linear bijection.

Proof. It is easily verified that Λ is a linear map. The rest follows from Definitions B.52, B.56, and Lemma B.55. □

Dual linear maps

Definition B.58. The *dual map* of a linear map $T : V \to W$ is the linear map $T' : W' \to V'$ defined as

$$\forall \varphi \in W', \qquad T'(\varphi) = \varphi \circ T. \tag{B.18}$$

Exercise B.59. Denote by D the linear map of differentiation $Dp = p'$ on the vector space $\mathcal{P}(\mathbb{R})$ of polynomials with real coefficients. Under the dual map of D, what is the image of the linear functional $\varphi(p) = \int_0^1 p$ on $\mathcal{P}(\mathbb{R})$?

Theorem B.60. The matrix of T' is the transpose of the matrix of T.

Proof. Let $(\mathbf{v}_1, \cdots, \mathbf{v}_n)$, $(\varphi_1, \cdots, \varphi_n)$, $(\mathbf{w}_1, \cdots, \mathbf{w}_n)$, $(\psi_1, \cdots, \psi_n)$, be bases of V, V', W, W', respectively. Denote by A and C the matrices of $T : V \to W$ and $T' : W' \to V'$, respectively. We have

$$\psi_j \circ T = T'(\psi_j) = \sum_{r=1}^{n} c_{r,j}\varphi_r.$$

By Corollary B.51, applying this equation to $\mathbf{v}_k$ yields

$$(\psi_j \circ T)(\mathbf{v}_k) = \sum_{r=1}^{n} c_{r,j}\varphi_r(\mathbf{v}_k) = c_{k,j}.$$

On the other hand, we have

$$\begin{aligned}(\psi_j \circ T)(\mathbf{v}_k) = \psi_j(T\mathbf{v}_k) &= \psi_j\left(\sum_{r=1}^{n} a_{r,k}w_r\right)\\ &= \sum_{r=1}^{n} a_{r,k}\psi_j(w_r) = a_{j,k}.\end{aligned}$$

□

Definition B.61. The *double dual map* of a linear map $T : V \to W$ is the linear map $T'' : V'' \to W''$ defined as $T'' = (T')'$.

Theorem B.62. For $T \in \mathcal{L}(V)$ and Λ in (B.17), we have

$$T'' \circ \Lambda = \Lambda \circ T. \tag{B.19}$$

Proof. Definition B.61 and equation (B.17) yields

$$\begin{aligned}&\forall v \in V,\ \forall \varphi \in V',\\ (T'' \circ \Lambda)v\varphi &= ((T')'\Lambda v)\varphi = (\Lambda v \circ T')\varphi = \Lambda v(T'\varphi)\\ &= (T'\varphi)(v) = \varphi(Tv) = \Lambda(Tv)(\varphi)\\ &= (\Lambda \circ T)v\varphi,\end{aligned}$$

where the third step is natural since T' send V' to V'. □

Corollary B.63. For $T \in \mathcal{L}(V)$ where V is finite-dimensional, the double dual map is

$$T'' = \Lambda \circ T \circ \Lambda^{-1}. \tag{B.20}$$

Proof. This follows directly from Theorem B.62 and Lemma B.57. □

The null space and range of the dual of a linear map

Definition B.64. For $U \subset V$, the *annihilator* of U, denoted U^0, is defined by

$$U^0 := \{\varphi \in V' : \ \forall \mathbf{u} \in U,\ \varphi(\mathbf{u}) = 0\}. \tag{B.21}$$

Exercise B.65. Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5$ denote the standard basis of $V = \mathbb{R}^5$, and $\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5$ its dual basis of V'. Suppose

$$U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2) = \{(x_1, x_2, 0, 0, 0) \in \mathbb{R}^5 : x_1, x_2 \in \mathbb{R}\}.$$

Show that $U^0 = \operatorname{span}(\varphi_3, \varphi_4, \varphi_5)$.

Exercise B.66. Let $i : U \hookrightarrow V$ be an inclusion. Show that $\operatorname{null} i' = U^0$.

Lemma B.67. Suppose $U \subset V$. Then U^0 is a subspace of V'.

Exercise B.68. Suppose V is finite-dimensional. Prove every linear map on a subspace of V can be extended to a linear map on V.

Lemma B.69. Suppose V is finite-dimensional and U is a subspace of V. Then

$$\dim U + \dim U^0 = \dim V. \tag{B.22}$$

Proof. Apply Theorem B.49 to the dual of an inclusion $i' : V' \to U'$ and we have

$$\begin{aligned}&\dim \operatorname{range} i' + \dim \operatorname{null} i' = \dim V'\\ \Rightarrow &\dim \operatorname{range} i' + \dim U^0 = \dim V,\end{aligned}$$

where the second line follows from Example B.66 and Lemma B.55. For any $\varphi \in U'$, Exercise B.68 states that $\varphi \in U'$ can be extended to $\psi \in V'$ such that $i'(\psi) = \varphi$. Hence i' is surjective and we have $U' = \operatorname{range} i'$. The proof is then completed by Lemma B.55. □

Lemma B.70. Any linear map $T \in \mathcal{L}(V, W)$ satisfies

$$\operatorname{null} T' = (\operatorname{range} T)^0. \tag{B.23}$$

Proof. Definitions B.45, B.47, B.58, and B.64 yield

$$\begin{aligned}\varphi \in \operatorname{null} T' &\Leftrightarrow 0 = T'(\varphi) = \varphi \circ T\\ &\Leftrightarrow \forall v \in V, \varphi(Tv) = 0\\ &\Leftrightarrow \varphi(\operatorname{range} T) = 0\\ &\Leftrightarrow \varphi \in (\operatorname{range} T)^0.\end{aligned}$$

□

Lemma B.71. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V, W)$ satisfies

$$\dim \operatorname{null} T' = \dim \operatorname{null} T + \dim W - \dim V. \tag{B.24}$$

Proof. Lemma B.70 and Theorem B.49 yield

$$\begin{aligned}\dim \text{null}T' &= \dim(\text{range}T)^0 = \dim W - \dim(\text{range}T)\\ &= \dim W - \dim V + \dim(\text{null}T)\\ &= \dim \text{null}T + \dim W - \dim V.\end{aligned}$$ □

Corollary B.72. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V,W)$ is surjective if and only if T' is injective.

Proof. T is surjective $\Leftrightarrow W = \text{range}T \Leftrightarrow (\text{range}T)^0 = \{0\}$ $\Leftrightarrow \text{null}T' = \{0\} \Leftrightarrow T'$ is injective. The second step follows from Lemma B.69 applied to W:

$$\dim W = \dim(\text{range}T) + \dim(\text{range}T)^0.$$ □

Lemma B.73. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V,W)$ satisfies

$$\dim \text{range}T' = \dim \text{range}T. \tag{B.25}$$

Proof. Theorem B.49, Lemma B.70, and Lemma B.69 yield

$$\begin{aligned}\dim \text{range}T' &= \dim W - \dim \text{null}T'\\ &= \dim W - \dim(\text{range}T)^0\\ &= \dim(\text{range}T).\end{aligned}$$ □

Lemma B.74. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V,W)$ satisfies

$$\text{range}T' = (\text{null}T)^0. \tag{B.26}$$

Proof. Theorem B.49, Lemma B.70, and Lemma B.69 yield

$$\begin{aligned}\varphi \in \text{range}T' &\Rightarrow \exists \psi \in W' \text{ s.t } T'(\psi) = \varphi\\ &\Rightarrow \forall v \in \text{null}T, \varphi(v) = \psi(Tv) = 0\\ &\Rightarrow \varphi \in (\text{null}T)^0.\end{aligned}$$

The proof is completed by

$$\begin{aligned}\dim \text{range}T' &= \dim(\text{range}T)\\ &= \dim V - \dim \text{null}T\\ &= \dim(\text{null}T)^0.\end{aligned}$$ □

Corollary B.75. For finite-dimensional vector spaces V and W, any linear map $T \in \mathcal{L}(V,W)$ is injective if and only if T' is surjective.

Proof. T is injective $\Leftrightarrow \text{null}T = \{0\} \Leftrightarrow (\text{null}T)^0 = V' \Leftrightarrow \text{range}T' = V' \Leftrightarrow T'$ is surjective. The second step follows from Lemmas B.69 and B.55, and the third step follows from Lemma B.74. □

Matrix ranks

Definition B.76. For a matrix $A \in \mathbb{F}^{m\times n} : \mathbb{F}^n \to \mathbb{F}^m$, its *column space* (or range or image) consists of all linear combinations of its columns, its *row space* (or coimage) is the column space of A^T, its *null space* (or kernel) is the null space of A as a linear operator, and the *left null space* (or cokernnel) is the null space of A^T.

Definition B.77. The *column rank* and *row rank* of a matrix $A \in \mathbb{F}^{m\times n}$ is the dimension of its column space and row space, respectively.

Lemma B.78. Let A_T denote the matrix of a linear operator $T \in \mathcal{L}(V,W)$. Then the column rank of A_T is the dimension of range$T$.

Proof. For $\mathbf{u} = \sum_i c_i \mathbf{v}_i$, Corollary B.51 yields

$$T\mathbf{u} = \sum_i c_i T\mathbf{v}_i = T[\mathbf{v}_1,\dots,\mathbf{v}_n]\mathbf{c} = [\mathbf{w}_1,\dots,\mathbf{w}_m]A_T\mathbf{c}.$$

Hence we have

$$\{T\mathbf{u} : \mathbf{c} \in \mathbb{F}^n\} = \{[\mathbf{w}_1,\dots,\mathbf{w}_m]A_T\mathbf{c} : \mathbf{c} \in \mathbb{F}^n\}.$$

The LHS is range$T$ while $\{A_T\mathbf{c} : \mathbf{c} \in \mathbb{F}^n\}$ is the column space of A_T. Since $(\mathbf{w}_1,\dots,\mathbf{w}_m)$ is a basis, by Definition B.77 the column rank of the matrix $[\mathbf{w}_1,\dots,\mathbf{w}_m]$ is m. Taking dim to both side to the above equation yields the conclusion. Note that the RHS is a subspace of $\mathbb{F}^m$ (why?) and the dimension of it does not depend on the special choice of its basis, hence we can choose $(\mathbf{w}_1,\dots,\mathbf{w}_m)$ to be the standard basis and then $[\mathbf{w}_1,\dots,\mathbf{w}_m]$ is simply the identity matrix. □

Theorem B.79. For any $A \in \mathbb{F}^{m\times n}$, its row rank equals its column rank.

Proof. Define a linear map $T : \mathbb{F}^n \to \mathbb{F}^m$ as $T\mathbf{x} = A\mathbf{x}$. Clearly, A is the matrix of T for the standard bases of $\mathbb{F}^n$ and $\mathbb{F}^m$. Then we have,

$$\begin{aligned}\text{column rank of } A &= \dim \text{range}T\\ &= \dim \text{range}T'\\ &= \text{column rank of the matrix of } T'\\ &= \text{column rank of } A^T\\ &= \text{row rank of } A,\end{aligned}$$

where the first step follows from Lemma B.78, the second from Lemma B.73, the third from Lemma B.78, the fourth from Theorem B.60, and the last from the definition of matrix transpose and matrix products. □

Definition B.80. The *rank* of a matrix is its column rank.

Theorem B.81 (Fundamental theorem of linear algebra)**.** For a matrix $A \in \mathbb{F}^{m\times n} : \mathbb{F}^n \to \mathbb{F}^m$, its column space and row space both have dimension $r \le \min(m,n)$; its null space and left null space have dimensions $n-r$ and $m-r$, respectively. In addition, we have

$$\mathbb{F}^m = \text{range}A \oplus \text{null}A^T, \tag{B.27a}$$

$$\mathbb{F}^n = \text{range}A^T \oplus \text{null}A, \tag{B.27b}$$

where $\text{range}A \perp \text{null}A^T$ and $\text{range}A^T \perp \text{null}A$.

Proof. The first sentence is a rephrase of Theorem B.79 and follows from Theorem B.49. For the second sentence, we only prove (B.27b). $\mathbf{x} \in \text{null}A$ implies $\mathbf{x} \in \mathbb{F}^n$ and $A\mathbf{x} = \mathbf{0}$. The latter expands to

$$\begin{bmatrix}\mathbf{a}_1^T\\ \mathbf{a}_2^T\\ \cdots\\ \mathbf{a}_m^T\end{bmatrix}\mathbf{x} = \begin{bmatrix}0\\0\\\cdots\\0\end{bmatrix},$$

which implies that $\forall j = 1, 2, \ldots, m$, $\mathbf{a}_j \perp \mathbf{x}$. Hence $\mathbf{x}$ is orthogonal to each basis vector of rangeA^T. The rest of the proof follows from Lemma B.73, Theorem B.60, Theorem B.49. □

B.3 Eigenvalues, eigenvectors, and invariant subspaces

B.3.1 Invariant subspaces

Definition B.82. Under a linear operator $T \in \mathcal{L}(\mathcal{V})$, a subspace $\mathcal{U}$ of $\mathcal{V}$ is *invariant* if $\mathbf{u} \in \mathcal{U}$ implies $T\mathbf{u} \in \mathcal{U}$.

Example B.83. Under $T \in \mathcal{L}(\mathcal{V})$, each of the following subspaces of $\mathcal{V}$ is invariant: $\{\mathbf{0}\}$, $\mathcal{V}$, null T, and range T.

Definition B.84. A number $\lambda \in \mathbb{F}$ is called an *eigenvalue of an operator* $T \in \mathcal{L}(\mathcal{V})$ if there exists $\mathbf{v} \in \mathcal{V}$ such that $T\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{v} \neq \mathbf{0}$. Then the vector $\mathbf{v}$ is called an *eigenvector* of T corresponding to λ.

Example B.85. For each eigenvector $\mathbf{v}$ of $T \in \mathcal{L}(\mathcal{V})$, the subspace span$(\mathbf{v})$ is a one-dimensional invariant subspace of $\mathcal{V}$.

Lemma B.86. Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of $T \in \mathcal{L}(\mathcal{V})$ with corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$. Then $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is linearly independent.

Lemma B.87. Suppose $\mathcal{V}$ is finite-dimensional. Then each operator on $\mathcal{V}$ has at most dim $\mathcal{V}$ distinct eigenvalues.

B.3.2 Upper-triangular matrices

Notation 15. Suppose $T \in \mathcal{L}(\mathcal{V})$ and $p \in \mathbb{P}(\mathbb{F})$ is a polynomial given by

$$p(z) = a_0 + a_1 z + \cdots + a_m z^m$$

for $z \in \mathbb{F}$. Then $p(T)$ is the operator given by

$$p(T) = a_0 I + a_1 T + \cdots + a_m T^m,$$

where $I = T^0$ is the identity operator.

Example B.88. Suppose $D \in \mathcal{L}(\mathbb{P}(\mathbb{R}))$ is the differentiation operator defined by $Dq = q'$ and p is the polynomial defined by $p(x) = 7 - 3x + 5x^2$. Then we have

$$p(D) = 7 - 3D + 5D^2, \qquad (p(D))q = 7q - 3q' + 5q''.$$

Definition B.89. The *product polynomial* of two polynomials $p, q \in \mathbb{P}(\mathbb{F})$ is the polynomial defined by

$$\forall z \in \mathbb{F}, \qquad (pq)(z) := p(z)q(z). \tag{B.28}$$

Lemma B.90. Any $T \in \mathcal{L}(\mathcal{V})$ and $p, q \in \mathbb{P}(\mathbb{F})$ satisfy

$$(pq)(T) = p(T)q(T) = q(T)p(T). \tag{B.29}$$

Theorem B.91. Every linear operator on a finite-dimensional, nonzero, complex vector space has an eigenvalue.

Definition B.92. The *matrix of a linear operator* $T \in \mathcal{L}(\mathcal{V})$ is the matrix of the linear map $T \in \mathcal{L}(\mathcal{V}, \mathcal{V})$, c.f. Definition B.50.

Theorem B.93. Suppose $T \in \mathcal{L}(\mathcal{V})$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis of $\mathcal{V}$. Then the following are equivalent:

(a) the matrix of T with respect to $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is upper triangular;

(b) $T\mathbf{v}_j \in \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_j)$ for each $j = 1, \ldots, n$;

(c) $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_j)$ is invariant under T for each $j = 1, \ldots, n$.

Theorem B.94. Every linear operator $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional complex vector space $\mathcal{V}$ has an upper-triangular matrix with respect to some basis of $\mathcal{V}$.

Theorem B.95. Suppose $T \in \mathcal{L}(\mathcal{V})$ has an upper-triangular matrix with respect to some basis of $\mathcal{V}$. Then T is invertible if and only if all the entries on the diagonal of that upper-triangular matrix are nonzero.

Theorem B.96. Suppose $T \in \mathcal{L}(\mathcal{V})$ has an upper-triangular matrix with respect to some basis of $\mathcal{V}$. Then the eigenvalues of T are precisely the entries on the diagonal of that upper-triangular matrix.

B.3.3 Eigenspaces and diagonal matrices

Definition B.97. A *diagonal matrix* is a square matrix that is zero everywhere except possibly along the diagonal.

Definition B.98. The *eigenspace* of $T \in \mathcal{L}(\mathcal{V})$ corresponding to $\lambda \in \mathbb{F}$ is

$$E(\lambda, T) := \text{null}(T - \lambda I). \tag{B.30}$$

Lemma B.99. Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional space $\mathcal{V}$. Then

$$E(\lambda_1, T) + \cdots + E(\lambda_m, T)$$

is a direct sum and

$$\dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) \leq \dim \mathcal{V}. \tag{B.31}$$

Definition B.100. An operator $T \in \mathcal{L}(\mathcal{V})$ is *diagonalizable* if it has a diagonal matrix with respect to some basis of $\mathcal{V}$.

Theorem B.101 (Conditions of diagonalizability)**.** Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional space $\mathcal{V}$. Then the following are equivalent:

(a) T is diagonalizable;

(b) $\mathcal{V}$ has a basis consisting of eigenvectors of T;

(c) there exist one-dimensional subspaces $U_1, \ldots, U_n$ of $\mathcal{V}$, each invariant under T, such that $\mathcal{V} = U_1 \oplus \cdots \oplus U_n$;

(d) $\mathcal{V} = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T)$;

(e) $\dim \mathcal{V} = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T)$.

Corollary B.102. An operator $T \in \mathcal{L}(\mathcal{V})$ is diagonalizable if T has dim $\mathcal{V}$ distinct eigenvalues.

B.4 Inner product spaces

B.4.1 Inner products

Definition B.103. Denote by $\mathbb{F}$ the underlying field of a vector space $\mathcal{V}$. The *inner product* $\langle \mathbf{u}, \mathbf{v} \rangle$ on $\mathcal{V}$ is a function $\mathcal{V} \times \mathcal{V} \to \mathbb{F}$ that satisfies

(IP-1) real positivity: $\forall \mathbf{v} \in \mathcal{V}$, $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$;

(IP-2) definiteness: $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ iff $\mathbf{v} = \mathbf{0}$;

(IP-3) additivity in the first slot:
$\forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$;

(IP-4) homogeneity in the first slot:
$\forall a \in \mathbb{F}$, $\forall \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $\langle a\mathbf{v}, \mathbf{w} \rangle = a \langle \mathbf{v}, \mathbf{w} \rangle$;

(IP-5) conjugate symmetry: $\forall \mathbf{v}, \mathbf{w} \in \mathcal{V}$, $\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}$.

An *inner product space* is a vector space $\mathcal{V}$ equipped with an inner product on $\mathcal{V}$.

Corollary B.104. An inner product has additivity in the second slot, i.e. $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$.

Corollary B.105. An inner product has conjugate homogeneity in the second slot, i.e.

$$\forall a \in \mathbb{F},\ \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \qquad \langle \mathbf{v}, a\mathbf{w} \rangle = \bar{a} \langle \mathbf{v}, \mathbf{w} \rangle . \tag{B.32}$$

Exercise B.106. Prove Corollaries B.104 and B.105 from Definition B.103.

Definition B.107. The *Euclidean inner product* on $\mathbb{F}^n$ is

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{n} v_i \overline{w_i}. \tag{B.33}$$

B.4.2 Norms induced from inner products

Definition B.108. Let $\mathbb{F}$ be the underlying field of an inner product space $\mathcal{V}$. The *norm induced by an inner product* on $\mathcal{V}$ is a function $\mathcal{V} \to \mathbb{F}$:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \tag{B.34}$$

Definition B.109. The *Euclidean ℓ_p norm* of a vector $\mathbf{v} \in \mathbb{F}^n$ is

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \tag{B.35}$$

and the *Euclidean ℓ_∞ norm* is

$$\|\mathbf{v}\|_\infty = \max_i |v_i|. \tag{B.36}$$

Theorem B.110 (Equivalence of norms)**.** Any two norms $\|\cdot\|_N$ and $\|\cdot\|_M$ on a finite dimensional vector space $\mathcal{V} = \mathbb{C}^n$ satisfy

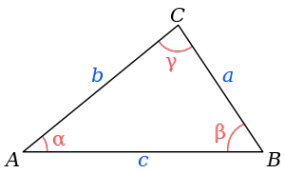
$$\exists c_1, c_2 \in \mathbb{R}^+,\ \text{s.t. } \forall \mathbf{x} \in \mathcal{V},\ c_1 \|\mathbf{x}\|_M \leq \|\mathbf{x}\|_N \leq c_2 \|\mathbf{x}\|_M. \tag{B.37}$$

Definition B.111. The angle between two vectors $\mathbf{v}, \mathbf{w}$ in an inner product space with $\mathbb{F} = \mathbb{R}$ is the number $\theta \in [0, \pi]$,

$$\theta = \arccos \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\| \|\mathbf{w}\|}. \tag{B.38}$$

Theorem B.112 (The law of cosines)**.** Any triangle satisfies

$$c^2 = a^2 + b^2 - 2ab \cos\gamma. \tag{B.39}$$



Proof. The dot product of AB to $AB = CB - CA$ yields

$$c^2 = \langle AB, CB \rangle - \langle AB, CA \rangle .$$

The dot products of CB and CA to $AB = CB - CA$ yield

$$\begin{aligned} \langle CB, AB \rangle &= a^2 - \langle CB, CA \rangle ; \\ - \langle CA, AB \rangle &= - \langle CA, CB \rangle + b^2. \end{aligned}$$

The proof is completed by adding up all three equations and applying (B.38). □

Theorem B.113 (The law of cosines: abstract version)**.** Any induced norm on a real vector space satisfies

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2 \langle \mathbf{u}, \mathbf{v} \rangle . \tag{B.40}$$

Proof. Definitions B.108 and B.103 and $\mathbb{F} = \mathbb{R}$ yield

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

□

B.4.3 Norms and induced inner-products

Definition B.114. A function $\|\cdot\| : \mathcal{V} \to \mathbb{F}$ is a *norm* for a vector space $\mathcal{V}$ iff it satisfies

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

(NRM-2) point separation: $\|\mathbf{v}\| = 0 \Rightarrow \mathbf{v} = \mathbf{0}$.

(NRM-3) absolute homogeneity:
$\forall a \in \mathbb{F}$, $\forall \mathbf{v} \in \mathcal{V}$, $\|a\mathbf{v}\| = |a| \|\mathbf{v}\|$;

(NRM-4) triangle inequality:
$\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$, $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

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

Exercise B.115. Explain how (NRM-1,2,3,4) relate to the geometric meaning of the norm of vectors in $\mathbb{R}^3$.

Lemma B.116. The norm induced by an inner product is a norm as in Definition B.114.

Proof. The induced norm as in (B.34) satisfies (NRM-1,2) trivially. For (NRM-3),

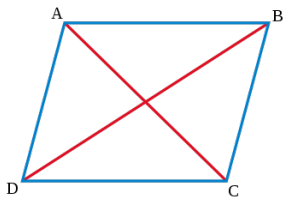
$$\|a\mathbf{v}\|^2 = \langle a\mathbf{v}, a\mathbf{v}\rangle = a\langle \mathbf{v}, a\mathbf{v}\rangle = a\bar{a}\langle \mathbf{v}, \mathbf{v}\rangle = |a|^2\|\mathbf{v}\|^2.$$

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 B.117 (The parallelogram law)**.** The sum of squares of the lengths of the four sides of a parallelogram equals the sum of squares of the two diagonals.



More precisely, we have in the above plot

$$(AB)^2+(BC)^2+(CD)^2+(DA)^2 = (AC)^2+(BD)^2. \quad \text{(B.41)}$$

Proof. Apply the law of cosines to the two diagonals, add the two equations, and we obtain (B.41). □

Theorem B.118 (The parallelogram law: abstract version)**.** Any induced norm (B.34) satisfies

$$2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 = \|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2. \quad \text{(B.42)}$$

Proof. Replace $\mathbf{v}$ in (B.40) with $-\mathbf{v}$ and we have

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

(B.42) follows from adding the above equation to (B.40). □

Exercise B.119. In the case of Euclidean ℓ_p norms, show that the parallelogram law (B.42) holds if and only if $p = 2$.

Theorem B.120. The induced norm (B.34) holds for some inner product $\langle\cdot,\cdot\rangle$ if and only if the parallelogram law (B.42) holds for every pair of $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

Exercise B.121. Prove Theorem B.120.

Example B.122. By Theorem B.120 and Exercise B.119, the ℓ^1 and ℓ^∞ spaces do not have a corresponding inner product for the Euclidean ℓ_1 and ℓ_∞ norms.

B.4.4 Orthonormal bases

Definition B.123. Two vectors $\mathbf{u}, \mathbf{v}$ are called *orthogonal* if $\langle \mathbf{u}, \mathbf{v}\rangle = 0$, i.e., their inner product is the additive identity of the underlying field.

Example B.124. An inner product on the vector space of continuous real-valued functions on the interval $[-1, 1]$ is

$$\langle f, g\rangle = \int_{-1}^{+1} f(x)g(x)\mathrm{d}x.$$

f and g are said to be orthogonal if the integral is zero.

Theorem B.125 (Pythagorean)**.** If $\mathbf{u}, \mathbf{v}$ are orthogonal, then $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

Proof. This follows from (B.40) and Definition B.123. □

Theorem B.126 (Cauchy-Schwarz inequality)**.**

$$|\langle \mathbf{u}, \mathbf{v}\rangle| \le \|\mathbf{u}\|\|\mathbf{v}\|, \quad \text{(B.43)}$$

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 $\mathbf{v} = \mathbf{0}$, (B.43) clearly holds. Otherwise (B.43) follows from substituting $\lambda = -\frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{v},\mathbf{v}\rangle}$ into the above equation. □

Exercise B.127. To explain the choice of λ in the proof of Theorem B.126, what is the geometric meaning of (B.43) in the plane? When will the equality hold?

Example B.128. 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 B.129. If $f, g : [a, b] \to \mathbb{R}$ are continuous, then

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

Definition B.130. A list of vectors $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m)$ is called *orthonormal* if the vectors in it are pairwise orthogonal and each vector has norm 1, i.e.

$$\begin{cases}\forall i = 1, 2, \ldots, m, & \|\mathbf{e}_i\| = 1; \\ \forall i \ne j, & \langle \mathbf{e}_i, \mathbf{e}_j\rangle = 0.\end{cases} \quad \text{(B.44)}$$

Definition B.131. An *orthonormal basis* of an inner-product space $\mathcal{V}$ is an orthonormal list of vectors in $\mathcal{V}$ that is also a basis of $\mathcal{V}$.

Theorem B.132. If $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ is an orthonormal basis of $\mathcal{V}$, then

$$\forall \mathbf{v} \in \mathcal{V}, \quad \mathbf{v} = \sum_{i=1}^n \langle \mathbf{v}, \mathbf{e}_i\rangle \mathbf{e}_i, \quad \text{(B.45a)}$$

$$\|\mathbf{v}\|^2 = \sum_{i=1}^n |\langle \mathbf{v}, \mathbf{e}_i\rangle|^2. \quad \text{(B.45b)}$$

Lemma B.133. Every finite-dimensional inner-product space has an orthonormal basis.

Theorem B.134 (Schur)**.** Every linear operator $T \in \mathcal{L}(\mathcal{V})$ on a finite-dimensional complex vector space $\mathcal{V}$ has an upper-triangular matrix with respect to some orthonormal basis of $\mathcal{V}$.

Proof. This follows from Theorem B.94, Lemma B.133 and the Gram-Schmidt process; see Section 5.1. □

Definition B.135. A *linear functional* on $\mathcal{V}$ is a linear map from $\mathcal{V}$ to $\mathbb{F}$, or, it is an element of $\mathcal{L}(\mathcal{V}, \mathbb{F})$.

Theorem B.136 (Riesz representation theorem)**.** If φ is a linear functional on a finite-dimensional vector space $\mathcal{V}$, then

$$\exists \mathbf{u} \in \mathcal{V} \text{ s.t. } \forall \mathbf{v} \in \mathcal{V}, \quad \varphi(\mathbf{v}) = \langle \mathbf{v}, \mathbf{u} \rangle . \tag{B.46}$$

Furthermore, $\mathbf{u}$ is unique.

Proof. Let $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be an orthonormal basis of $\mathcal{V}$.

$$\begin{aligned}\varphi(\mathbf{v}) &= \varphi\left(\sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{e}_i \rangle \mathbf{e}_i\right) = \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{e}_i \rangle \varphi(\mathbf{e}_i) \\ &= \sum_{i=1}^{n} \left\langle \mathbf{v}, \overline{\varphi(\mathbf{e}_i)} \mathbf{e}_i \right\rangle = \left\langle \mathbf{v}, \sum_{i=1}^{n} \overline{\varphi(\mathbf{e}_i)} \mathbf{e}_i \right\rangle,\end{aligned}$$

where the last two steps follow from Corollaries B.104 and B.105.

As for the uniqueness, suppose that $\exists \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{V}$ s.t. $\varphi(\mathbf{v}) = \langle \mathbf{v}, \mathbf{u}_1 \rangle = \langle \mathbf{v}, \mathbf{u}_2 \rangle$. Then for each $\mathbf{v} \in \mathcal{V}$,

$$0 = \langle \mathbf{v}, \mathbf{u}_1 \rangle - \langle \mathbf{v}, \mathbf{u}_2 \rangle = \langle \mathbf{v}, \mathbf{u}_1 - \mathbf{u}_2 \rangle .$$

Taking $\mathbf{v} = \mathbf{u}_1 - \mathbf{u}_2$ shows that $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{0}$. □

B.5 Operators on inner-product spaces

B.5.1 Adjoint and self-adjoint operators

Definition B.137. The *adjoint of a linear map* $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ between inner-product spaces is a function $T^* : \mathcal{W} \to \mathcal{V}$ that satisfies

$$\forall \mathbf{v} \in \mathcal{V}, \ \forall \mathbf{w} \in \mathcal{W}, \qquad \langle T\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, T^* \mathbf{w} \rangle . \tag{B.47}$$

Example B.138. Define a linear operator $T : \mathbb{R}^3 \to \mathbb{R}^2$,

$$T(x_1, x_2, x_3) = (x_2 + 3x_3, 2x_1).$$

Then $T^*(y_1, y_2) = (2y_2, y_1, 3y_1)$ because

$$\begin{aligned}\langle (x_1, x_2, x_3), T^*(y_1, y_2) \rangle &= \langle T(x_1, x_2, x_3), (y_1, y_2) \rangle \\ &= \langle (x_2 + 3x_3, 2x_1), (y_1, y_2) \rangle \\ &= x_2 y_1 + 3x_3 y_1 + 2x_1 y_2 \\ &= \langle (x_1, x_2, x_3), (2y_2, y_1, 3y_1) \rangle .\end{aligned}$$

Lemma B.139. If $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, then $T^* \in \mathcal{L}(\mathcal{W}, \mathcal{V})$.

Proof. Use Definition B.40. □

Theorem B.140. The adjoint of a linear map has the following properties.

(ADJ-1) additivity:
$\forall S, T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \qquad (S + T)^* = S^* + T^*;$

(ADJ-2) conjugate homogeneity:
$\forall T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \forall a \in \mathbb{F}, \qquad (aT)^* = \bar{a} T^*;$

(ADJ-3) adjoint of adjoint:
$\forall T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), (T^*)^* = T;$

(ADJ-4) identity: $I^* = I;$

(ADJ-5) products: let $\mathcal{U}$ be an inner-product space,
$\forall T \in \mathcal{L}(\mathcal{V}, \mathcal{W}), \forall S \in \mathcal{L}(\mathcal{W}, \mathcal{U}), \ (ST)^* = T^* S^*.$

Proof. Use Definitions B.137 and B.103. □

Definition B.141. The *conjugate transpose*, or *Hermitian transpose*, or *Hermitian conjugate*, or *adjoint matrix*, of a matrix $A \in \mathbb{C}^{m \times n}$ is the matrix $A^* \in \mathbb{C}^{n \times m}$ defined by

$$(A^*)_{ij} = \overline{a_{ji}}, \tag{B.48}$$

where $\overline{a_{ji}}$ denotes the complex conjugate of the entry a_{ji}.

Definition B.142. A matrix $U \in \mathbb{C}^{n \times n}$ is *unitary* iff $U^* U = I$. A matrix $U \in \mathbb{R}^{n \times n}$ is *orthogonal* iff $U^T U = I$.

Theorem B.143. A matrix $U \in \mathbb{C}^{n \times n}$ is unitary if and only if its columns form an orthonormal basis for $\mathbb{C}^n$.

Proof. This follows from considering the (i, j)th element of $U^* U$ and applying $U^* U = I$ in Definition B.142. □

Corollary B.144. A unitary matrix U preserves norms and inner products. More precisely, we have

$$\forall \mathbf{v}, \mathbf{w} \in \mathbb{C}^n, \qquad \langle U\mathbf{v}, U\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle .$$

Proof. This follows from Definitions B.137 and B.142. □

Theorem B.145. Every unitary matrix $U \in \mathbb{C}^{2 \times 2}$ with $\det U = 1$ is of the form

$$U = \begin{bmatrix} a & b \\ -\bar{b} & \bar{a} \end{bmatrix}, \tag{B.49}$$

where $|a|^2 + |b|^2 = 1$.

Proof. Let

$$U = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Then Theorem B.143 and the condition $\det U = 1$ yield

$$\begin{aligned} a\bar{b} + c\bar{d} &= 0, \\ ad - cb &= 1. \end{aligned}$$

In other words, the linear system

$$\begin{bmatrix} \bar{b} & \bar{d} \\ d & -b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

has solution $x = a, y = c$. Furthermore, Theorem B.143 and the form of U yield $|b|^2 + |d|^2 = 1$. Hence the solution $x = a, y = c$ is unique and we have $a = \bar{d}$ and $c = -\bar{b}$, which completes the proof. □

Theorem B.146. Let $T \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. Suppose $e_1, \dots, e_n$ is an orthonormal basis of $\mathcal{V}$ and $f_1, \dots, f_m$ is an orthonormal basis of $\mathcal{W}$. Then

$$M(T^*, (f_1, \dots, f_m), (e_1, \dots, e_n))$$

is the conjugate transpose of

$$M(T, (e_1, \dots, e_n), (f_1, \dots, f_m)).$$

Proof. By Corollary B.51, we have

$$T[e_1, \dots, e_n] = [f_1, \dots, f_m] M_T.$$

The two bases being orthonormal further implies

$$M_T = \begin{bmatrix} \langle f_1, Te_1 \rangle & \langle f_1, Te_2 \rangle & \cdots & \langle f_1, Te_n \rangle \\ \langle f_2, Te_1 \rangle & \langle f_2, Te_2 \rangle & \cdots & \langle f_2, Te_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_m, Te_1 \rangle & \langle f_m, Te_2 \rangle & \cdots & \langle f_m, Te_n \rangle \end{bmatrix}.$$

The proof is completed by repeating the above derivation for T^* and then applying Definitions B.103 and B.137. □

Definition B.147. An operator $T \in \mathcal{L}(\mathcal{V})$ is *self-adjoint* iff $T = T^*$, i.e.

$$\forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \qquad \langle T\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, T\mathbf{w} \rangle . \tag{B.50}$$

Lemma B.148. Suppose $\mathcal{V}$ is a complex inner product space and $T \in \mathcal{L}(\mathcal{V})$. If

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \langle T\mathbf{v}, \mathbf{v} \rangle = 0, \tag{B.51}$$

then $T = \mathbf{0}$.

Proof. By Definition B.103 and (B.51), we have, $\forall \mathbf{u}, \mathbf{w} \in \mathcal{V}$,

$$\begin{aligned} \langle T\mathbf{u}, \mathbf{w} \rangle =& \frac{\langle T(\mathbf{u}+\mathbf{w}), \mathbf{u}+\mathbf{w} \rangle - \langle T(\mathbf{u}-\mathbf{w}), \mathbf{u}-\mathbf{w} \rangle}{4} \\ &+ \mathbf{i} \frac{\langle T(\mathbf{u}+\mathbf{i}\mathbf{w}), \mathbf{u}+\mathbf{i}\mathbf{w} \rangle - \langle T(\mathbf{u}-\mathbf{i}\mathbf{w}), \mathbf{u}-\mathbf{i}\mathbf{w} \rangle}{4} \\ =& 0. \end{aligned}$$

Setting $\mathbf{w} = T\mathbf{u}$ completes the proof. □

Theorem B.149. Suppose $\mathcal{V}$ is a complex inner product space and $T \in \mathcal{L}(\mathcal{V})$. Then T is self-adjoint if and only if

$$\forall \mathbf{v} \in \mathcal{V}, \quad \langle T\mathbf{v}, \mathbf{v} \rangle \in \mathbb{R}. \tag{B.52}$$

Proof. By Definitions B.103, B.137, and B.147, we have

$$\begin{aligned} \langle T\mathbf{v}, \mathbf{v} \rangle - \overline{\langle T\mathbf{v}, \mathbf{v} \rangle} &= \langle T\mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{v}, T\mathbf{v} \rangle \\ &= \langle T\mathbf{v}, \mathbf{v} \rangle - \langle T^*\mathbf{v}, \mathbf{v} \rangle = \langle (T - T^*)\mathbf{v}, \mathbf{v} \rangle . \end{aligned}$$

Then Lemma B.148 completes the proof. □

Lemma B.150. Suppose $\mathcal{V}$ is a real inner product space and $T \in \mathcal{L}(\mathcal{V})$. If T is self-adjoint and satisfies

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \langle T\mathbf{v}, \mathbf{v} \rangle = 0, \tag{B.53}$$

then $T = \mathbf{0}$.

Proof. In Lemma B.148, we have already proved the case of complex inner product spaces. Here we deal with the real case. By the self-adjointness and the underlying field being real, we have

$$\langle T\mathbf{w}, \mathbf{u} \rangle = \langle \mathbf{w}, T\mathbf{u} \rangle = \langle T\mathbf{u}, \mathbf{w} \rangle ,$$

which, together with Definition B.103, implies

$$\langle T\mathbf{u}, \mathbf{w} \rangle = \frac{\langle T(\mathbf{u}+\mathbf{w}), \mathbf{u}+\mathbf{w} \rangle - \langle T(\mathbf{u}-\mathbf{w}), \mathbf{u}-\mathbf{w} \rangle}{4}.$$

Setting $\mathbf{w} = T\mathbf{u}$ completes the proof. □

B.5.2 Normal operators

Definition B.151. An operator $T \in \mathcal{L}(\mathcal{V})$ is *normal* iff $TT^* = T^*T$.

Corollary B.152. Every self-adjoint operator is normal.

Lemma B.153. An operator $T \in \mathcal{L}(\mathcal{V})$ is normal if and only if

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \|T\mathbf{v}\| = \|T^*\mathbf{v}\|. \tag{B.54}$$

Proof. By Lemma B.150 and Definition B.137, we have

$$\begin{aligned} T^*T - TT^* = \mathbf{0} &\Leftrightarrow \forall \mathbf{v} \in \mathcal{V}, \ \langle (T^*T - TT^*)\mathbf{v}, \mathbf{v} \rangle = 0 \\ &\Leftrightarrow \forall \mathbf{v} \in \mathcal{V}, \ \langle T^*T\mathbf{v}, \mathbf{v} \rangle = \langle TT^*\mathbf{v}, \mathbf{v} \rangle \\ &\Leftrightarrow \forall \mathbf{v} \in \mathcal{V}, \ \|T\mathbf{v}\|^2 = \|T^*\mathbf{v}\|^2. \end{aligned}$$

The positivity of a norm completes the proof. □

Theorem B.154. For a linear operator $T \in \mathcal{L}(\mathcal{V})$ on a two-dimensional real inner product space $\mathcal{V}$, the following are equivalent:

(a) T is normal but not self-adjoint.

(b) The matrix of T with respect to every orthonormal basis of $\mathcal{V}$ has the form

$$M(T) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \tag{B.55}$$

where $b \neq 0$.

Proof. (b) ⇒ (a) trivially holds, so we only prove (a) ⇒ (b). Let (e_1, e_2) be an orthonormal basis of $\mathcal{V}$ and set

$$M(T, (e_1, e_2)) = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

By Definition B.50, Theorem B.125, and Definition B.130, we have $\|Te_1\|^2 = a^2 + b^2$. In addition, Theorem B.146 yields $\|T^*e_1\|^2 = a^2 + c^2$. Then Lemma B.153 implies $b^2 = c^2$ and the condition of T being not self-adjoint further yields $c = -b \neq 0$. Considering $\|Te_2\|^2$ and $\|T^*e_2\|^2$ yields $a = d$. □

B.5.3 The spectral theorem

Theorem B.155 (Complex spectral)**.** For a linear operator $T \in \mathcal{L}(\mathcal{V})$ with $\mathbb{F} = \mathbb{C}$, the following are equivalent:

(a) T is normal;

(b) $\mathcal{V}$ has an orthonormal basis consisting of eigenvectors of T;

(c) T has a diagonal matrix with respect to some orthonormal basis of $\mathcal{V}$.

Theorem B.156 (Real spectral)**.** For a linear operator $T \in \mathcal{L}(\mathcal{V})$ with $\mathbb{F} = \mathbb{R}$, the following are equivalent:

(a) T is self-adjoint;

(b) $\mathcal{V}$ has an orthonormal basis consisting of eigenvectors of T;

(c) T has a diagonal matrix with respect to some orthonormal basis of $\mathcal{V}$.

B.5.4 Isometries

Definition B.157. An operator $S \in \mathcal{L}(\mathcal{V})$ is called a (linear) *isometry* iff

$$\forall \mathbf{v} \in \mathcal{V}, \qquad \|S\mathbf{v}\| = \|\mathbf{v}\|. \tag{B.56}$$

Theorem B.158. An operator $S \in \mathcal{L}(\mathcal{V})$ on a real inner product space is an isometry if and only if there exists an orthonormal basis of $\mathcal{V}$ with respect to which $\mathcal{S}$ has a block diagonal matrix such that each block on the diagonal is a 1-by-1 matrix containing 1 or -1, or, is a 2-by-2 matrix of the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{B.57}$$

where $\theta \in (0, \pi)$.

Corollary B.159. For an operator $\mathcal{S} \in \mathcal{L}(\mathcal{V})$ on a two-dimensional real inner product space, the following are equivalent:

(a) $\mathcal{S}$ is an isometry;

(b) $\mathcal{S}$ is either an identity or a reflection or a rotation.

B.5.5 The singular value decomposition

Definition B.160. A self-adjoint linear operator whose eigenvalues are non-negative is called *positive semidefinite* or *positive*, and called *positive definite* if it is also invertible.

Corollary B.161. For any linear operator $f \in \mathcal{L}(\mathcal{V})$, both $f^* \circ f$ and $f \circ f^*$ are self-adjoint and positive semidefinite.

Proof. By Definition B.147, $f^* \circ f$ is self-adjoint since

$$\langle (f^* \circ f)\mathbf{u}, \mathbf{v} \rangle = \langle f\mathbf{u}, f\mathbf{v} \rangle = \langle \mathbf{u}, (f^* \circ f)\mathbf{v} \rangle .$$

Suppose $(\lambda, \mathbf{u})$ is an eigen-pair of $(f^* \circ f)$. Then we have

$$\begin{aligned} \lambda \langle \mathbf{u}, \mathbf{u} \rangle &= \langle (f^* \circ f)\mathbf{u}, \mathbf{u} \rangle = \langle f\mathbf{u}, f\mathbf{u} \rangle \\ \Rightarrow \lambda &= \frac{\langle f\mathbf{u}, f\mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \geq 0. \end{aligned}$$

Similar arguments apply to $f \circ f^*$. □

Definition B.162. The *singular values* of a linear map f are the square roots of the positive eigenvalues of $f^* \circ f$.

Definition B.163. For a rectangular matrix $A \in \mathbb{F}^{m \times n}$, the factorization $A = P\Sigma Q^*$ is a *singular value decomposition* (SVD) iff $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal, and $P \in \mathbb{F}^{m \times m}$ and $Q \in \mathbb{F}^{n \times n}$ are unitary matrices or orthogonal matrices for $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$, respectively. The diagonal entries of Σ, written $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, are the singular values of A. The column vectors of P and Q are the *left singular vectors* and the *right singular vectors* of A, respectively.

Theorem B.164. Any matrix $A \in \mathbb{C}^{m \times n}$ has an SVD.

Definition B.165. Two matrices $A, B \in \mathbb{R}^{n \times n}$ are called *similar* iff there exists an invertible matrix P such that $B = P^{-1}AP$. The map $A \mapsto P^{-1}AP$ is called a *similarity transformation* or *conjugation of the matrix* A.

B.6 Trace and determinant

Definition B.166. The *trace of a matrix* A, denoted by Trace A, is the sum of the diagonal entries of A.

Lemma B.167. The trace of a matrix is the sum of its eigenvalues, each of which is repeated according to its multiplicity.

Definition B.168. A *permutation of a set* A is a bijective function $\sigma : A \to A$.

Definition B.169. Let σ be a permutation of $A = \{1, 2, \ldots, n\}$ and let s denote the number of pairs of integers (j, k) with $1 \leq j < k \leq n$ such that j appears after k in the list $(m_1, \ldots, m_n)$ given by $m_i = \sigma(i)$. The *sign of the permutation* σ is 1 if s is even and -1 if s is odd.

Definition B.170. The *signed volume of a parallelotope* spanned by n vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ is a function $\delta : \mathbb{R}^{n \times n} \to \mathbb{R}$ that satisfies

(SVP-1) $\delta(I) = 1$;

(SVP-2) $\delta(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = 0$ if $\mathbf{v}_i = \mathbf{v}_j$ for some $i \neq j$;

(SVP-3) δ is linear, i.e., $\forall j = 1, \ldots, n,\ \forall c \in \mathbb{R}$,

$$\begin{aligned} &\delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v} + c\mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n) \\ &= \delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n) \\ &\quad + c\delta(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n). \end{aligned} \tag{B.58}$$

Lemma B.171. Adding a multiple of one vector to another does not change the signed volume.

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

Lemma B.172. If the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent, then $\delta(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = 0$.

Proof. WLOG, we assume $\mathbf{v}_1 = \sum_{i=2}^n c_i \mathbf{v}_i$. Then the result follows from (SVP-2,3). □

Lemma B.173. The signed volume δ is alternating, i.e.,

$$\delta(\mathbf{v}_1,\dots,\mathbf{v}_i,\dots,\mathbf{v}_j,\dots,\mathbf{v}_n) = -\delta(\mathbf{v}_1,\dots,\mathbf{v}_j,\dots,\mathbf{v}_i,\dots,\mathbf{v}_n). \tag{B.59}$$

Exercise B.174. Prove Lemma B.173 using (SVP-2,3).

Lemma B.175. Let M_σ denote the matrix of a permutation $\sigma : E \to E$ where E is the set of standard basis vectors in (B.5). Then we have $\delta(M_\sigma) = \operatorname{sgn}(\sigma)$.

Proof. There is a one-to-one correspondence between the vectors in the matrix

$$M_\sigma = [e_{\sigma(1)}, e_{\sigma(2)}, \dots, e_{\sigma(n)}]$$

and the scalars in the one-line notation

$$(\sigma(1)\ \sigma(2)\ \dots\ \sigma(n)).$$

A sequence of transpositions taking σ to the identity map also takes M_σ to the identity matrix. By Lemma B.173, each transposition yields a multiplication factor -1. Definition B.169 and (SVP-1) give $\delta(M_\sigma) = \operatorname{sgn}(\sigma)\delta(I) = \operatorname{sgn}(\sigma)$. □

Definition B.176 (Leibniz formula of determinants)**.** The *determinant of a square matrix* $A \in \mathbb{R}^{n\times n}$ is

$$\det A = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i}, \tag{B.60}$$

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

Lemma B.177. The determinant of a matrix is the product of its eigenvalues, each of which is repeated according to its multiplicity.

Exercise B.178. Show that the determinant formula in (B.60) reduces to

$$\det \begin{bmatrix} a & c \\ b & d \end{bmatrix} = ad - bc \tag{B.61}$$

for $n = 2$. Give a geometric proof that $ad - bc$ is the signed volume δ of the parallelogram determined by two vectors $\mathbf{v}_1 = (a, b)^T$ and $\mathbf{v}_2 = (c, d)^T$ on the plane.

Theorem B.179. The signed volume function satisfying (SVP-1,2,3) in Definition B.170 is unique and is the same as the determinant in (B.60).

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

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

where the first four steps follow from (SVP-3), the sixth step from Lemma B.175, 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 $(i_1, i_2, \dots, i_n)$ is a permutation of $(1, 2, \dots, n)$. □

Exercise B.180. Use the formula in (B.60) to show that $\det A = \det A^T$.

Definition B.181. The i, j *cofactor* of $A \in \mathbb{R}^{n\times n}$ is

$$C_{ij} = (-1)^{i+j} M_{ij}, \tag{B.62}$$

where M_{ij} is the i, j *minor of a matrix* A, i.e. the determinant of the $(n-1)\times(n-1)$ matrix that results from deleting the i-th row and the j-th column of A.

Theorem B.182 (Laplace formula of determinants)**.** Given fixed indices $i, j \in 1, 2, \dots, n$, the determinant of an n-by-n matrix $A = [a_{ij}]$ is given by

$$\det A = \sum_{j'=1}^{n} a_{ij'} C_{ij'} = \sum_{i'=1}^{n} a_{i'j} C_{i'j}. \tag{B.63}$$

Exercise B.183. Prove Theorem B.182 by induction.

Appendix C

Basic Analysis

C.1 Sequences

Definition C.1. A *sequence* is a function on $\mathbb{N}$.

Definition C.2. The *extended real number system* is the real line $\mathbb{R}$ with two additional elements $-\infty$ and $+\infty$:

$$\mathbb{R}^* := \mathbb{R} \cup \{-\infty, +\infty\}. \tag{C.1}$$

An extended real number $x \in \mathbb{R}^*$ is *finite* if $x \in \mathbb{R}$ and it is *infinite* otherwise.

Definition C.3. The *supremum of a sequence* $(a_n)_{n=m}^{\infty}$ is

$$\sup(a_n)_{n=m}^{\infty} := \sup\{a_n : n \geq m\}, \tag{C.2}$$

and the *infimum of a sequence* $(a_n)_{n=m}^{\infty}$ is

$$\inf(a_n)_{n=m}^{\infty} := \inf\{a_n : n \geq m\}. \tag{C.3}$$

C.1.1 Convergence

Definition C.4 (Limit of a sequence)**.** A sequence $\{a_n\}$ has the *limit* L, written $\lim_{n\to\infty} a_n = L$, or $a_n \to L$ as $n \to \infty$, iff

$$\forall \epsilon > 0,\ \exists N,\ \text{s.t.}\ \ \forall n > N,\ |a_n - L| < \epsilon. \tag{C.4}$$

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

Example C.5 (A story of π)**.** A famous estimation of π in ancient China is given by Zu, ChongZhi 1500 years ago,

$$\pi \approx \frac{355}{113} \approx 3.14159292.$$

In modern mathematics, we approximate π with a sequence for increasing accuracy, e.g.

$$\pi \approx 3.141592653589793\ldots \tag{C.5}$$

As of March 2019, we human beings have more than 31 trillion digits of π. However, real world applications never use even a small fraction of the 31 trillion digits:

- If you want to build a fence over your backyard swimming pool, several digits of π is probably enough;
- in NASA, calculations involving π use 15 digits for Guidance Navigation and Control;
- if you want to compute the circumference of the entire universe to the accuracy of less than the diameter of a hydrogen atom, you need only 39 decimal places of π.

On one hand, computational mathematics is judged by a metric that is different from that of pure mathematics; this may cause a huge gap between what needs to be done and what has been done. On the other hand, a computational mathematician cannot assume that a fixed accuracy is good enough for all applications. In the approximation a number or a function, she must develop theory and algorithms to provide the user the choice of an ever-increasing amount of accuracy, so long as the user is willing to invest an increasing amount of computational resources. This is one of the main motivations of infinite sequence and series.

Definition C.6. A sequence $\{a_n\}$ is *Cauchy* if

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } m, n > N \ \Rightarrow\ |a_n - a_m| < \epsilon. \tag{C.6}$$

Exercise C.7. Show that every convergent sequence in $\mathbb{R}$ is Cauchy.

Theorem C.8 (Cauchy criterion)**.** Every Cauchy sequence in $\mathbb{R}$ converges to a limit in $\mathbb{R}$.

Theorem C.9 (Bolzano-Weierstrass)**.** Every bounded sequence has a convergent subsequence.

C.1.2 Limit points

Definition C.10. Let $\epsilon > 0$ be a real number. Two real numbers x, y are said to be *ϵ-close* iff $|x - y| \leq \epsilon$.

Definition C.11. A real number x is said to be *ϵ-adherent* to a sequence $(a_n)_{n=m}^{\infty}$ of real numbers iff there exists an $n \geq m$ such that a_n is ϵ-close to x. x is *continually ϵ-adherent* to $(a_n)_{n=m}^{\infty}$ iff it is ϵ-adherent to $(a_n)_{n=N}^{\infty}$ for every $N \geq m$.

Definition C.12. A real number x is a *limit point* or *adherent point* of a sequence $(a_n)_{n=m}^{\infty}$ of real numbers if it is continually ϵ-adherent to $(a_n)_{n=m}^{\infty}$ for every $\epsilon \geq 0$.

Definition C.13. The *limit superior* of a sequence $(a_n)_{n=m}^{\infty}$ of real numbers is

$$\limsup_{n\to\infty} a_n := \inf(a_N^+)_{N=m}^{\infty}, \tag{C.7}$$

where $a_N^+ = \sup(a_n)_{n=N}^\infty$. The *limit inferior* of $(a_n)_{n=m}^\infty$ is

$$\liminf_{n\to\infty} a_n := \sup(a_N^-)_{N=m}^\infty, \tag{C.8}$$

where $a_N^- = \inf(a_n)_{n=N}^\infty$.

Example C.14. Let $(a_n)_{n=m}^\infty$ be the sequence

$$1.1, -1.01, 1.001, -1.0001, 1.00001, \ldots$$

Then $(a_n^+)_{n=N}^\infty$ is the sequence

$$1.1, 1.001, 1.001, 1.00001, 1.00001, \ldots$$

and $(a_n^-)_{n=N}^\infty$ is the sequence

$$-1.01, -1.01, -1.0001, -1.0001, -1.000001, -1.000001, \ldots$$

Hence we have

$$\limsup_{n\to\infty} a_n = 1, \qquad \liminf_{n\to\infty} a_n = -1.$$

Lemma C.15. Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers. For $L^+ = \limsup a_n$ and $L^- = \liminf a_n$, we have

(a) For every $x > L^+$, elements of the sequence are eventually less than x:

$$\forall x > L^+, \exists N \geq m \text{ s.t. } \forall n \geq N, a_n < x.$$

Similarly, for every $x < L^-$, elements of the sequence are eventually greater than x:

$$\forall x < L^-, \exists N \geq m \text{ s.t. } \forall n \geq N, a_n > x.$$

(b) For every $x < L^+$, there are an infinite number of elements in the sequence that are greater than x:

$$\forall x < L^+, \forall N \geq m, \exists n \geq N \text{ s.t. } a_n > x.$$

Similarly, for every $x > L^-$, there are an infinite number of elements in the sequence that are less than x:

$$\forall x > L^-, \forall N \geq m, \exists n \geq N \text{ s.t. } a_n < x.$$

(c) $\inf(a_n)_{n=m}^\infty \leq L^- \leq L^+ \leq \sup(a_n)_{n=m}^\infty$.

(d) Any limit point c of $(a_n)_{n=m}^\infty$ satisfies $L^- \leq c \leq L^+$.

(e) If L^+ (or L^+) is finite, then it is a limit point of $(a_n)_{n=m}^\infty$.

(f) $\lim_{n\to\infty} a_n = c$ if and only if $L^+ = L^- = c$.

Theorem C.16 (Completeness of $\mathbb{R}$)**.** A sequence of real numbers is Cauchy if and only if it is convergent.

Notation 16 (Asymptotic notation)**.** For $g : \mathbb{R}^+ \to \mathbb{R}^+$, $f : \mathbb{R} \to \mathbb{R}$, and $a \in [0, +\infty]$, we write

$$f(x) = O(g(x)) \text{ as } x \to a$$

iff

$$\limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty.$$

In particular, we have

$$f(x) = o(g(x)) \text{ as } x \to a \;\Leftrightarrow\; \lim_{x\to a} \frac{|f(x)|}{g(x)} = 0.$$

We also write

$$f(x) = \Theta(g(x)) \text{ as } x \to a$$

iff

$$0 < \limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty.$$

C.2 Series

Definition C.17 (Finite series)**.** Let m, n be integers and let $(a_i)_{i=m}^n$ be a finite sequence of real numbers. The *finite series* or *finite sum* associated with the sequence $(a_i)_{i=m}^n$ is the number $\sum_{i=m}^n a_i$ given by the recursive formula

$$\sum_{i=m}^n a_i := \begin{cases} 0 & \text{if } n < m; \\ a_n + \sum_{i=m}^{n-1} a_i & \text{otherwise.} \end{cases} \tag{C.9}$$

Definition C.18 (Formal infinite series)**.** A (formal) *infinite series* associated with an infinite sequence $\{a_n\}$ is the expression $\sum_{n=0}^\infty a_n$.

Definition C.19. The *sequence of partial sums* $(S_n)_{n=0}^\infty$ associated with a formal infinite series $\sum_{i=0}^\infty a_i$ is defined for each n as the sum of the sequence $\{a_i\}$ from a_0 to a_n

$$S_n = \sum_{i=0}^n a_i. \tag{C.10}$$

Definition C.20. A formal infinite series is said to be *convergent* and *converge* to L if its sequence of partial sums converges to some limit L. In this case we write $L = \sum_{n=0}^\infty a_n$ and call L the *sum of the infinite series*.

Definition C.21. A formal infinite series is said to be *divergent* if its sequence of partial sums diverges. In this case we do not assign any real number value to this series.

Lemma C.22. An infinite series $\sum_{n=0}^\infty a_n$ of real numbers is convergent if and only if

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} \text{ s.t. } \forall p, q \geq N,\ \left|\sum_{n=p}^q a_n\right| \leq \epsilon. \tag{C.11}$$

Definition C.23. An infinite series $\sum_{n=0}^\infty a_n$ is *absolutely convergent* iff the series $\sum_{n=0}^\infty |a_n|$ is convergent.

Lemma C.24. An infinite series that is absolutely convergent is convergent.

Theorem C.25 (Root test)**.** For an infinite series $\sum_{n=0}^\infty a_n$, define

$$\alpha := \limsup_{n\to\infty} |a_n|^{\frac{1}{n}}. \tag{C.12}$$

The series is convergent if $\alpha < 1$ and divergent if $\alpha > 1$.

Theorem C.26 (Ratio test)**.** An infinite series $\sum_{n=0}^\infty a_n$ of nonzero real numbers is

- absolutely convergent if $\limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} < 1$;
- divergent if $\liminf_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} > 1$.

C.3 Continuous functions on $\mathbb{R}$

Definition C.27. A *scalar function* is a function whose range is a subset of $\mathbb{R}$.

Definition C.28 (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

$$\forall \epsilon > 0, \exists \delta > 0, \text{ s.t. } \forall x \in I(c,\delta),\ |f(x) - L| < \epsilon. \quad \text{(C.13)}$$

Example C.29. Show that $\lim_{x\to 2} \frac{1}{x} = \frac{1}{2}$.

Proof. If $\epsilon \geq \frac{1}{2}$, choose $\delta = 1$. Then $x \in (1,3)$ implies $\left|\frac{1}{x} - \frac{1}{2}\right| < \frac{1}{2}$ since $\frac{1}{x} - \frac{1}{2}$ is a monotonically decreasing function with its supremum at $x = 1$.

If $\epsilon \in (0, \frac{1}{2})$, choose $\delta = \epsilon$. Then $x \in (2-\epsilon, 2+\epsilon) \subset (\frac{3}{2}, \frac{5}{2})$. Hence $\left|\frac{1}{x} - \frac{1}{2}\right| = \frac{|2-x|}{|2x|} < |2-x| < \epsilon$. The proof is completed by Definition C.28. □

Definition C.30. $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at c iff

$$\lim_{x\to c} f(x) = f(c). \quad \text{(C.14)}$$

Definition C.31. A scalar function f is *continuous on* (a,b), written $f \in \mathcal{C}(a,b)$, if (C.14) holds $\forall x \in (a,b)$.

Theorem C.32 (Intermediate value)**.** A scalar function $f \in \mathcal{C}[a,b]$ satisfies

$$\forall y \in [m,M],\ \exists \xi \in [a,b], \text{ s.t. } y = f(\xi) \quad \text{(C.15)}$$

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

Definition C.33. Let $I = (a,b)$. A function $f : I \to \mathbb{R}$ is *uniformly continuous* on I iff

$$\begin{aligned} &\forall \epsilon > 0, \exists \delta > 0, \text{ s.t.} \\ &\forall x, y \in I,\ |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon. \end{aligned} \quad \text{(C.16)}$$

Example C.34. Show that, on (a,∞), $f(x) = \frac{1}{x}$ is uniformly continuous if $a > 0$ and is not so if $a = 0$.

Proof. If $a > 0$, then $|f(x) - f(y)| = \frac{|x-y|}{xy} < \frac{|x-y|}{a^2}$.
Hence $\forall \epsilon > 0, \exists \delta = a^2\epsilon$, s.t.
$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \frac{|x-y|}{a^2} < \frac{a^2\epsilon}{a^2} = \epsilon$.

If $a = 0$, negating the condition of uniform continuity, i.e. eq. (C.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 C.35. On (a,∞), $f(x) = \frac{1}{x^2}$ is uniformly continuous if $a > 0$ and is not so if $a = 0$.

Theorem C.36. Uniform continuity implies continuity but the converse is not true.

Proof. exercise. □

Theorem C.37. $f : \mathbb{R} \to \mathbb{R}$ is uniformly continuous on (a,b) iff it can be extended to a continuous function $\tilde{f}$ on $[a,b]$.

C.4 Differentiation of functions

Definition C.38. The *derivative* of a function $f : \mathbb{R} \to \mathbb{R}$ at a is the limit

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h}. \quad \text{(C.17)}$$

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

Example C.39. For the power function $f(x) = x^\alpha$, we have $f' = \alpha x^{\alpha-1}$ due to Newton's generalized binomial theorem,

$$(a+h)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} a^{\alpha-n} h^n.$$

Definition C.40. A function $f(x)$ is k times *continuously differentiable* on (a,b) iff $f^{(k)}(x)$ exists on (a,b) and is itself continuous. The set or space of all such functions on (a,b) is denoted by $\mathcal{C}^k(a,b)$. In comparison, $\mathcal{C}^k[a,b]$ is the space of functions f for which $f^{(k)}(x)$ is bounded and uniformly continuous on (a,b).

Theorem C.41. A scalar function f is bounded on $[a,b]$ if $f \in \mathcal{C}[a,b]$.

Theorem C.42. If $f : (a,b) \to \mathbb{R}$ assumes its maximum or minimum at $x_0 \in (a,b)$ and f is differentiable at x_0, then $f'(x_0) = 0$.

Proof. Suppose $f'(x_0) > 0$. Then we have

$$f'(x_0) = \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} > 0.$$

The definition of a limit implies

$$\exists \delta > 0 \text{ s.t. } a < x_0 - \delta < x_0 + \delta < b,$$

which, together with $|x - x_0| < \delta$, implies $\frac{f(x)-f(x_0)}{x-x_0} > 0$. This is a contradiction to $f(x_0)$ being a maximum when we choose $x \in (x_0, x_0 + \delta)$. □

Theorem C.43 (Rolle's)**.** If a function $f : \mathbb{R} \to \mathbb{R}$ satisfies

(i) $f \in \mathcal{C}[a,b]$ and f' exists on (a,b),

(ii) $f(a) = f(b)$,

then $\exists x \in (a,b)$ s.t. $f'(x) = 0$.

Proof. By Theorem C.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 C.42 completes the proof. □

Theorem C.44 (Mean value)**.** If $f \in \mathcal{C}[a,b]$ and if f' exists on (a,b), then $\exists \xi \in (a,b)$ s.t. $f(b) - f(a) = f'(\xi)(b-a)$.

Proof. Construct a linear function $L : [a,b] \to \mathbb{R}$ such that $L(a) = f(a)$, $L(b) = f(b)$, then $\forall x \in (a,b)$, we have $L'(x) = \frac{f(b)-f(a)}{b-a}$. Consider $g(x) = f(x) - L(x)$ on $[a,b]$. $g(a) = 0$, $g(b) = 0$. By Theorem C.43, $\exists \xi \in [a,b]$ such that $g'(\xi) = 0$, which completes the proof. □

C.5 Taylor series

Lemma C.45. A series converges to L iff the associated sequence of partial sums converges to L.

Definition C.46. A *power series* centered at c is a series of the form

$$p(x) = \sum_{n=0}^{\infty} a_n (x-c)^n, \tag{C.18}$$

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

$$I_c(p) = \{x \mid p(x) \text{ converges}\}. \tag{C.19}$$

Definition C.47. If the derivatives $f^{(i)}(x)$ with $i = 1, 2, \ldots, n$ exist for a function $f : \mathbb{R} \to \mathbb{R}$ at $x = c$, then

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k \tag{C.20}$$

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

$$T_1(x) = f(c) + f'(c)(x-c). \tag{C.21}$$

Example C.48. 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 C.49. The *Taylor series* (or Taylor expansion) for $f(x)$ at c is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!}(x-c)^k. \tag{C.22}$$

Definition C.50. The *remainder* of the nth *Taylor polynomial* in approximating $f(x)$ is

$$E_n(x) = f(x) - T_n(x). \tag{C.23}$$

Theorem C.51. Let T_n be the nth Taylor polynomial for $f(x)$ at c.

$$\lim_{n\to\infty} E_n(x) = 0 \;\Leftrightarrow\; \lim_{n\to\infty} T_n(x) = f(x). \tag{C.24}$$

Lemma C.52. $\forall m = 0, 1, 2, \ldots, n$, $E_n^{(m)}(c) = 0$.

Proof. This follows from Definition C.47 and Example C.48. □

Theorem C.53 (Taylor's theorem with Lagrangian form)**.** Consider a function $f : \mathbb{R} \to \mathbb{R}$. If $f \in \mathcal{C}^n[c-d, c+d]$ and $f^{(n+1)}(x)$ exists on $(c-d, c+d)$, then $\forall x \in [c-d, c+d]$, there exists some ξ between c and x such that

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}. \tag{C.25}$$

Proof. Fix $x \neq c$, let M be the unique solution of

$$E_n(x) = f(x) - T_n(x) = \frac{M(x-c)^{n+1}}{(n+1)!}.$$

Consider the function

$$g(t) := E_n(t) - \frac{M(t-c)^{n+1}}{(n+1)!}. \tag{C.26}$$

Clearly $g(x) = 0$. By Lemma C.52, $g^{(k)}(c) = 0$ for each $k = 0, 1, \ldots, n$. Then Rolle's theorem implies that

$$\exists x_1 \in (c, x) \text{ s.t. } g'(x_1) = 0.$$

If $x < c$, change (c, x) above to (x, c). Apply Rolle's theorem to $g'(t)$ on (c, x_1) and we have

$$\exists x_2 \in (c, x_1) \text{ s.t. } g^{(2)}(x_2) = 0.$$

Repeatedly using Rolle's theorem,

$$\exists x_{n+1} \in (c, x_n) \text{ s.t. } g^{(n+1)}(x_{n+1}) = 0. \tag{C.27}$$

Since T_n is a polynomial of degree n, we have $T_n^{(n+1)}(t) = 0$, which, together with (C.27) and (C.26), yields

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

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

Example C.54. How many terms are needed to compute e^2 correctly to four decimal places?

The requirement of four decimal places means an accuracy of at least $\epsilon = 10^{-5}$. By Definition C.49, the Taylor series of e^x at $c = 0$ is

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

By Theorem C.53, 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 $e^2 2^{n+1}/(n+1)! \le \epsilon$ yields $n \ge 12$, i.e., 13 terms.

C.6 Riemann integral

Definition C.55. A *partition of an interval* $I = [a, b]$ is a finite ordered subset $T_n \subseteq I$ of the form

$$T_n(a, b) = \{a = x_0 < x_1 < \cdots < x_n = b\}. \tag{C.28}$$

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,

$$h_n = h(T_n) = \max(x_i - x_{i-1}), \qquad i = 1, 2, \ldots, n. \tag{C.29}$$

Definition C.56. The *Riemann sum* of $f : \mathbb{R} \to \mathbb{R}$ over a partition T_n is

$$S_n(f) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}), \tag{C.30}$$

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

Definition C.57. $f : \mathbb{R} \to \mathbb{R}$ is *integrable* (or more precisely *Riemann integrable*) on $[a, b]$ iff

$$\begin{aligned} &\exists L \in \mathbb{R}, \text{ s.t. } \forall \epsilon > 0, \ \exists \delta > 0 \text{ s.t.} \\ &\forall T_n(a, b) \text{ with } h(T_n) < \delta, \ |S_n(f) - L| < \epsilon. \end{aligned} \tag{C.31}$$

Example C.58. 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 (C.31) 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 (C.30) yields $S_n(f) = b - a$. For $\epsilon = \frac{b-a}{4}$, the formula $|S_n(f) - L| \geq \epsilon$ clearly holds.

If $|L| \geq \frac{b-a}{2}$, we choose all x_i^*'s to be irrational so that $f(x_i^*) \equiv 0$; then (C.30) yields $S_n(f) = 0$. For $\epsilon = \frac{b-a}{4}$, the formula $|S_n(f) - L| \geq \epsilon$ clearly holds.

Definition C.59. If $f : \mathbb{R} \to \mathbb{R}$ is integrable on $[a, b]$, then the limit of the Riemann sum of f is called the *definite integral* of f on $[a, b]$:

$$\int_a^b f(x)\mathrm{d}x = \lim_{h_n \to 0} S_n(f). \tag{C.32}$$

Theorem C.60. A scalar function f is integrable on $[a, b]$ if $f \in \mathcal{C}[a, b]$.

Definition C.61. 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 $\forall x, y, \ x \leq y \Rightarrow f(x) \leq f(y)$; $f : \mathbb{R} \to \mathbb{R}$ is *monotonically decreasing* if $\forall x, y, \ x \leq y \Rightarrow f(x) \geq f(y)$.

Theorem C.62. A scalar function is integrable on $[a, b]$ if it is monotonic on $[a, b]$.

Exercise C.63. True or false: a bijective function is either order-preserving or order-reversing?

Theorem C.64 (Integral mean value)**.** Let $w : [a, b] \to \mathbb{R}^+$ be integrable on $[a, b]$. For $f \in \mathcal{C}[a, b]$, $\exists \xi \in [a, b]$ s.t.

$$\int_a^b w(x)f(x)\mathrm{d}x = f(\xi) \int_a^b w(x)\mathrm{d}x. \tag{C.33}$$

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 $mw(x) \leq f(x)w(x) \leq Mw(x)$ and

$$mI \leq \int_a^b w(x)f(x)\mathrm{d}x \leq MI.$$

$w > 0$ implies $I \neq 0$, hence

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

Applying Theorem C.32 completes the proof. □

Theorem C.65 (First fundamental theorem of calculus)**.** Let $a < b$ be real numbers. For a continuous function $f : [a, b] \to \mathbb{R}$ that is Riemann integrable, define a function $F : [a, b] \to \mathbb{R}$ by

$$F(x) := \int_a^x f(y)\mathrm{d}y. \tag{C.34}$$

Then F is differentiable and

$$\forall x_0 \in [a, b], \quad F'(x_0) = f(x_0). \tag{C.35}$$

Theorem C.66 (Second fundamental theorem of calculus)**.** Let $a < b$ be real numbers and let $f : [a, b] \to \mathbb{R}$ be a Riemann integrable function. If $F : [a, b] \to \mathbb{R}$ is the antiderivative of f, i.e. $F'(x) = f(x)$, then

$$\int_a^b f = F(b) - F(a). \tag{C.36}$$

C.7 Metric spaces

Definition C.67. 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: $d(x, y) \geq 0$;

(2) identity of indiscernibles: $x = y \Leftrightarrow d(x, y) = 0$;

(3) symmetry: $d(x, y) = d(y, x)$;

(4) triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A *metric space* is an ordered pair $(\mathcal{X}, d)$ where $\mathcal{X}$ is a set and d is a metric on $\mathcal{X}$.

Example C.68. Set $\mathcal{X}$ to be $C[a, b]$, the set of continuous functions $[a, b] \to \mathbb{R}$. Then the following is a metric on $\mathcal{X}$,

$$d(x, y) = \max_{t \in [a,b]} |x(t) - y(t)|. \tag{C.37}$$

Definition C.69. 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 C.70. Let $\mathcal{X}$ be the set of all bounded and unbounded sequences of complex numbers. Show that the following is a metric on $\mathcal{X}$,

$$d(x, y) = \sum_{j=1}^{\infty} \frac{1}{2^j} \frac{|\xi_j - \eta_j|}{1 + |\xi_j - \eta_j|}, \tag{C.38}$$

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

Definition C.71. For a real number $p \ge 1$, the ℓ^p *space* is the metric space $(\mathcal{X}, d)$ with

$$\mathcal{X} = \left\{ (\xi_j)_{j=1}^{\infty} : \xi_j \in \mathbb{C}; \sum_{j=1}^{\infty} |\xi_j|^p < \infty \right\}; \tag{C.39}$$

$$d(x, y) = \left(\sum_{j=1}^{\infty} |\xi_j - \eta_j|^p \right)^{1/p}, \tag{C.40}$$

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 C.72. Two positive real numbers p, q are called *conjugate exponents* iff they satisfy

$$p > 1, \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{C.41}$$

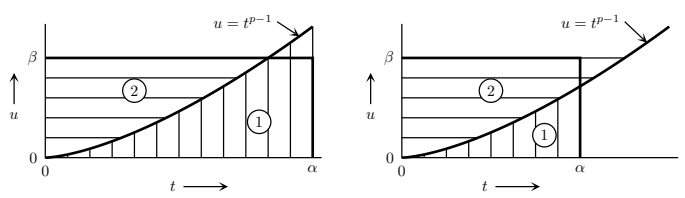
Lemma C.73. Any two positive real numbers α, β satisfy

$$\alpha\beta \le \frac{\alpha^p}{p} + \frac{\beta^q}{q}, \tag{C.42}$$

where p and q are conjugate exponents.

Proof. By (C.41), 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 C.74. Prove that (C.40) is indeed a metric. In particular, prove that (C.40) satisfies the triangular inequality by showing

(a) Lemma C.73 implies the *Hölder inequality*, i.e., for conjugate exponents p, q and for any $(\xi_j) \in \ell^p$, $(\eta_j) \in \ell^q$,

$$\sum_{j=1}^{\infty} |\xi_j \eta_j| \le \left(\sum_{k=1}^{\infty} |\xi_k|^p \right)^{1/p} \left(\sum_{m=1}^{\infty} |\eta_m|^q \right)^{1/q}. \tag{C.43}$$

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

$$\left(\sum_{j=1}^{\infty} |\xi_j + \eta_j|^p \right)^{1/p} \le \left(\sum_{k=1}^{\infty} |\xi_k|^p \right)^{1/p} + \left(\sum_{m=1}^{\infty} |\eta_m|^p \right)^{1/p}. \tag{C.44}$$

(c) The Minkowski inequality implies that the triangular inequality holds for (C.40).

Definition C.75. 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

$$B_r(x) := \{ y \in \mathcal{X} : d(x, y) < r \}. \tag{C.45}$$

Definition C.76. 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

$$\forall r > 0, \ E \cap B_r(x_0) \neq \emptyset. \tag{C.46}$$

Definition C.77. For metric spaces (X, d_1) and (Y, d_2), a function $f : X \to Y$ is *continuous* iff

$$\forall \epsilon > 0 \ \forall x \in X \ \exists \delta > 0 \text{ s.t. } \forall y \in X \quad d_1(x, y) < \delta \ \Rightarrow \ d_2(f(x), f(y)) < \epsilon \tag{C.47}$$

Definition C.78. For metric spaces (X, d_1) and (Y, d_2), a function $f : X \to Y$ is *uniformly continuous* iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \text{ s.t. } \forall x, y \in X \quad d_1(x, y) < \delta \ \Rightarrow \ d_2(f(x), f(y)) < \epsilon. \tag{C.48}$$

Definition C.79. A function $f : X \to Y$ with $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ is *continuous* iff

$$\forall U_a \in \gamma_Y, \exists V_a \in \gamma_X \text{ s.t. } f(V_a) \subset U_a, \tag{C.49}$$

where γ_X and γ_Y are sets of intersections of the open balls to X and Y, respectively,

$$\gamma_X := \{ B(a, \delta) \cap X : a \in X, \delta \in \mathbb{R}^+ \};$$
$$\gamma_Y := \{ B(f(a), \epsilon) \cap Y : f(a) \in Y, \epsilon \in \mathbb{R}^+ \}.$$

Definition C.80. A *basis of neighborhoods* (or a *basis*) on a set X is a collection $\mathcal{B}$ of subsets of X such that

- covering: $\cup \mathcal{B} = X$, and
- refining:

$$\forall U, V \in \mathcal{B}, \forall x \in U \cap V, \ \exists B \in \mathcal{B} \text{ s.t. } x \in B \subset (U \cap V).$$

Definition C.81. For two sets X, Y with bases of neighborhoods $\mathcal{B}_X$, $\mathcal{B}_Y$, a surjective function $f : X \to Y$ is *continuous* iff

$$\forall U \in \mathcal{B}_Y \ \exists V \in \mathcal{B}_X \text{ s.t. } f(V) \subset U. \tag{C.50}$$

Lemma C.82. If a surjective function $f : X \to Y$ is continuous in the sense of Definitions C.79, then it is continuous in the sense of Definition C.81.

Proof. By Definition C.80, the following collections are bases of $X \subseteq \mathbb{R}^m$ and $Y = f(X) \subseteq \mathbb{R}^n$, respectively,

$$\mathcal{B}_X = \{B(a,\delta) \cap X : a \in X, \delta > 0\};$$
$$\mathcal{B}_Y = \{B(b,\epsilon) \cap Y : b \in Y, \epsilon > 0\}.$$

The rest follows from Definition C.81. □

Definition C.83. A subset U of X is *open* (with respect to a given basis of neighborhoods $\mathcal{B}$ of X) iff

$$\forall x \in U \ \exists B \in \mathcal{B} \text{ s.t. } x \in B \subset U. \tag{C.51}$$

Lemma C.84. The intersection of two open sets is open.

Lemma C.85. The union of any collection of open sets is open.

Definition C.86. *The topology of X generated by a basis $\mathcal{B}$* is the collection $\mathcal{T}$ of all open subsets of X in the sense of Definition C.83.

Theorem C.87. The topology of X generated by a basis satisfies

- $\emptyset, X \in \mathcal{T}$;
- $\alpha \subset \mathcal{T} \ \Rightarrow \ \cup_{U\in\alpha} U \in \mathcal{T}$;
- $U, V \in \mathcal{T} \ \Rightarrow \ U \cap V \in \mathcal{T}$.

Definition C.88. For an arbitrary set X, a collection $\mathcal{T}$ of subsets of X is called a *topology on X* iff it satisfies the following conditions,

(TPO-1) $\emptyset, X \in \mathcal{T}$;

(TPO-2) $\alpha \subset \mathcal{T} \ \Rightarrow \ \cup\alpha \in \mathcal{T}$;

(TPO-3) $U, V \in \mathcal{T} \ \Rightarrow \ U \cap V \in \mathcal{T}$.

The pair $(X, \mathcal{T})$ is called a *topological space*. The elements of $\mathcal{T}$ are called *open sets*.

Definition C.89 (Continuous maps between topological spaces)**.** A function $f : X \to Y$ between two topological spaces is *continuous* iff the preimage of each open set $U \subset Y$ is open in X.

Theorem C.90. If a surjective function is continuous in the sense of Definition C.89, it is continuous in the sense of Definition C.81.

C.8 Uniform convergence

Definition C.91 (Limiting value of a function)**.** Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be metric spaces. Let E be a subset of $\mathcal{X}$ and $x_0 \in \mathcal{X}$ be an adherent point of E. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to *converge* to $L \in \mathcal{Y}$ as x converges to $x_0 \in E$, written

$$\lim_{x\to x_0; x\in E} f(x) = L, \tag{C.52}$$

iff

$$\forall \epsilon > 0, \exists \delta > 0 \text{ s.t. } \forall x \in E,$$
$$|x - x_0|_\mathcal{X} < \delta \ \Rightarrow \ |f(x) - L|_\mathcal{Y} < \epsilon. \tag{C.53}$$

Notation 17. In Definition C.91 we used the synonym notation

$$|u - v|_\mathcal{X} := d_\mathcal{X}(u, v). \tag{C.54}$$

Definition C.92 (Pointwise convergence)**.** Let $(f_n)_{n=1}^\infty$ be a sequence of functions from one metric space $(\mathcal{X}, d_\mathcal{X})$ to another $(\mathcal{Y}, d_\mathcal{Y})$, and let $f : \mathcal{X} \to \mathcal{Y}$ be another function. We say that $(f_n)_{n=1}^\infty$ *converges pointwise* to f on $\mathcal{X}$ iff

$$\forall x \in \mathcal{X}, \qquad \lim_{n\to\infty} f_n(x) = f(x), \tag{C.55}$$

or, equivalently,

$$\forall \epsilon > 0, \forall x \in \mathcal{X}, \ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall n > N, \ |f_n(x) - f(x)|_\mathcal{Y} < \epsilon. \tag{C.56}$$

Example C.93. Consider $f_n : [0,1] \to \mathbb{R}$ defined by $f_n(x) := x^n$ and $f : [0,1] \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} 1 & \text{if } x = 1; \\ 0 & \text{if } x \in [0,1). \end{cases}$$

The functions f_n are continuous and converge pointwise to f, which is discontinuous. Hence pointwise convergence does not preserve continuity.

Example C.94. For the functions in Example C.93, 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 C.95. Consider the interval $[a, b] = [0, 1]$, and the function sequence $f_n : [a, b] \to \mathbb{R}$ given by

$$f_n(x) := \begin{cases} 2n & \text{if } x \in \left[\frac{1}{2n}, \frac{1}{n}\right]; \\ 0 & \text{otherwise.} \end{cases}$$

Then (f_n) converges pointwise to $f(x) = 0$. However, $\int_a^b f_n = 1$ for every n while $\int_a^b f = 0$. Hence

$$\lim_{n\to\infty} \int_a^b f_n \neq \int_a^b \lim_{n\to\infty} f_n.$$

Hence pointwise convergence does not preserve integral.

Example C.96. Pointwise convergence does not preserve boundedness. For example, the function sequence

$$f_n(x) = \begin{cases} \exp(x) & \text{if } \exp(x) \le n; \\ n & \text{if } \exp(x) > n \end{cases} \tag{C.57}$$

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

$$f_n(x) = \begin{cases} \frac{1}{x} & \text{if } x \ge \frac{1}{n}; \\ 0 & \text{if } x \in (0, \frac{1}{n}) \end{cases} \tag{C.58}$$

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

$$f_n(x) = n \sin \frac{x}{n} \tag{C.59}$$

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

Definition C.97 (Uniform convergence)**.** Let $(f_n)_{n=1}^{\infty}$ be a sequence of functions from one metric space $(\mathcal{X}, d_{\mathcal{X}})$ to another $(\mathcal{Y}, d_{\mathcal{Y}})$, and let $f : \mathcal{X} \to \mathcal{Y}$ be another function. We say that $(f_n)_{n=1}^{\infty}$ *converges uniformly* to f on $\mathcal{X}$ iff

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N}^+ \text{ s.t. } \forall x \in \mathcal{X}, \forall n > N,\ |f_n(x) - f(x)|_{\mathcal{Y}} < \epsilon. \tag{C.60}$$

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 C.98. Uniform convergence implies pointwise convergence.

Proof. This follows directly from (C.56), (C.60), and Theorem A.10. □

Example C.99 (Uniform convergence of Taylor series)**.** Consider $f : \mathbb{R} \to \mathbb{R}$ and the sequence of its Taylor polynomial $(T_n)_{n=1}^{\infty}$ in Definition C.47. For any interval $I_r := (a - r, a + r)$, $(T_n)_{n=1}^{\infty}$ converges locally uniformly to $f|_{I_r}$ if r is less or equal to the radius of convergence of f at a. In particular, $(T_n)_{n=1}^{\infty}$ converges locally uniformly to f if the radius of convergence of f is $+\infty$.

Theorem C.100. Consider $b_{ij} \ge 0$ in $\mathbb{R}^* := \mathbb{R} \cup \{+\infty, -\infty\}$. If b_{ij} is monotone increasing in i for each j and is monotone increasing in j for each i, then we have

$$\lim_{i=0}^{\infty} \lim_{j=0}^{\infty} b_{ij} = \lim_{j=0}^{\infty} \lim_{i=0}^{\infty} b_{ij} \tag{C.61}$$

with all the indicated limits existing in $\mathbb{R}^*$.

Theorem C.101. Suppose that $\{f_n : [a, b] \to \mathbb{R}\}$ is a sequence of continuous functions satisfying

- $\{f_n(x_0)\}$ converges for some $x_0 \in [a, b]$,
- each f_n is differentiable on (a, b),
- $\{f_n'\}$ converges uniformly on (a, b).

Then we have

- $\{f_n\}$ converges uniformly on $[a, b]$ to a function f,
- both $f'(x)$ and $\lim_n f_n'(x)$ exist for any $x \in (a, b)$,
- $f'(x) = \lim_n f_n'(x)$ for any $x \in (a, b)$.

C.9 Vector calculus

Lemma C.102. For $E \subset \mathbb{R}$, $f : E \to \mathbb{R}$, $x_0 \in E$, and $L \in \mathbb{R}$, the following two statements are equivalent,

(a) f is differentiable at x_0 and $f'(x_0) = L$;

(b) $\lim_{x\to x_0, x\in E\setminus\{x_0\}} \frac{|f(x)-f(x_0)-L(x-x_0)|}{|x-x_0|} = 0$.

Exercise C.103. Prove Lemma C.102.

Definition C.104 (Total derivative)**.** For $E \subset \mathbb{R}^n$, $f : E \to \mathbb{R}^m$, $x_0 \in E$, f is *differentiable at x_0 with derivative* $L : \mathbb{R}^n \to \mathbb{R}^m$ if

$$\lim_{x\to x_0, x\in E\setminus\{x_0\}} \frac{\|f(x) - f(x_0) - L(x - x_0)\|_2}{\|x - x_0\|_2} = 0. \tag{C.62}$$

We denote the derivative of f with $f'(x_0) = L$ and also call it the *total derivative of f*.

Example C.105. For $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $L : \mathbb{R}^2 \to \mathbb{R}^2$,

$$f(x, y) := (x^2, y^2), \quad L(x, y) := (2x, 4y), \tag{C.63}$$

we claim that f is differentiable at $(1, 2)$ with $f'(x_0) = L$. To show this, we compute

$$\begin{aligned}
&\lim_{\substack{(x,y)\to(1,2)\\(x,y)\neq(1,2)}} \frac{\|f(x, y) - f(1, 2) - L((x, y) - (1, 2))\|_2}{\|(x, y) - (1, 2)\|_2} \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \frac{\|f(1 + a, 2 + b) - f(1, 2) - L(a, b)\|_2}{\|(a, b)\|_2} \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \frac{\|\left((1 + a)^2, (2 + b)^2\right) - (1, 4) - (2a, 4b)\|_2}{\|(a, b)\|_2} \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \frac{\|(a^2, b^2)\|_2}{\|(a, b)\|_2} \\
\le&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \left(\frac{\|(a^2, 0)\|_2}{\|(a, b)\|_2} + \frac{\|(0, b^2)\|_2}{\|(a, b)\|_2}\right) \\
=&\lim_{\substack{(a,b)\to(0,0)\\(a,b)\neq(0,0)}} \sqrt{a^2 + b^2} \\
=&0.
\end{aligned}$$

Lemma C.106. Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, and $x_0 \in E$ an interior point of E. Suppose f is differentiable at x_0 with derivative L_1 and also differentiable at x_0 with derivative L_2. Then $L_1 = L_2$.

Exercise C.107. Prove Lemma C.106.

Definition C.108 (Directional derivative)**.** Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, $x_0 \in E$ an interior point of E, and $\mathbf{v} \in \mathbb{R}^n$ a vector. If the limit

$$\lim_{t\to 0; t>0, x_0+t\mathbf{v}\in E} \frac{f(x_0 + t\mathbf{v}) - f(x_0)}{t}$$

exists, we say that f is *differentiable in the direction* $\mathbf{v}$ *at* x_0, and we denote this limit as

$$D_{\mathbf{v}} f(x_0) := \lim_{t\to 0; t>0, x_0+t\mathbf{v}\in E} \frac{f(x_0 + t\mathbf{v}) - f(x_0)}{t}. \tag{C.64}$$

Example C.109. For $\mathbf{v} = (3, 4)$ and $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined in (C.63), we have $D_{\mathbf{v}} f(1, 2) = (6, 16)$.

Example C.110. For $f : \mathbb{R} \to \mathbb{R}$, $D_{+1} f(x)$ is the right derivative of f at x (if it exists), and similarly $D_{-1} f(x)$ is the left derivative of f at x (if it exists).

Lemma C.111. Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, $x_0 \in E$ an interior point of E, and $\mathbf{v} \in \mathbb{R}^n$ a vector. If f is differentiable at x_0, then f is also differentiable in the direction $\mathbf{v}$ at x_0, and

$$D_{\mathbf{v}} f(x_0) = f'(x_0)\mathbf{v}. \tag{C.65}$$

Definition C.112 (Partial derivative)**.** Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, $x_0 \in E$ an interior point of E, and $1 \le j \le n$. The *partial derivative of f with respect to the x_j variable at x_0* is defined by

$$\begin{aligned} \tfrac{\partial f}{\partial x_j}(x_0) &:= \lim_{t \to 0; t>0, x_0 + te_j \in E} \tfrac{f(x_0 + te_j) - f(x_0)}{t} \\ &= \tfrac{\mathrm{d}}{\mathrm{d}t} f(x_0 + te_j)|_{t=0} \end{aligned} \tag{C.66}$$

provided that the limit exists. Here e_j is the jth standard basis vector of $\mathbb{R}^n$.

Exercise C.113. Show that the existence of partial derivatives at x_0 does not imply that the function is differentiable at x_0 by considering the differentiability of the following function $f : \mathbb{R}^2 \to \mathbb{R}$ at (0,0).

$$f(x, y) = \begin{cases} \frac{x^3}{x^2+y^2} & \text{if } (x, y) \neq (0, 0); \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

Theorem C.114. Let E be a subset of $\mathbb{R}^n$, $f : E \to \mathbb{R}^m$ a function, F a subset of E, and $x_0 \in E$ an interior point of F. If all the partial derivatives $\frac{\partial f}{\partial x_j}$ exist on F and are continuous at x_0, then f is differentiable at x_0, and the linear transformation $f'(x_0) : \mathbb{R}^n \to \mathbb{R}^m$ is defined by

$$f'(x_0)(\mathbf{v}) = \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(x_0). \tag{C.67}$$

Definition C.115. The *derivative matrix* or *differential matrix* or *Jacobian matrix* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a $m \times n$ matrix,

$$Df := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}. \tag{C.68}$$

Appendix D

Fourier Analysis of Linear PDEs

Definition D.1. The *L^2-norm* of a Lebesgue-measurable function $u : \mathbb{R} \to \mathbb{C}$ is a nonnegative or infinite real number

$$\|u\| = \left[\int_{-\infty}^{\infty} |u(x)|^2 \mathrm{d}x\right]^{\frac{1}{2}}. \tag{D.1}$$

Notation 18. Denote by L^2 the set of all functions whose L^2-norms are finite, i.e.,

$$L^2 = \{u : \|u\| < \infty\}. \tag{D.2}$$

Similarly, L^1 and L^∞ respectively denote the sets of functions with finite L^1- and L^∞- norms,

$$\|u\|_1 = \int_{-\infty}^{\infty} |u(x)|\mathrm{d}x, \quad \|u\|_\infty = \sup_{-\infty<x<\infty} |u(x)|. \tag{D.3}$$

Since the L^2 norm is the norm used in most applications, we have reserved the symbol $\|\cdot\|$ without a subscript for it.

D.1 Fourier transform

Definition D.2. The *Fourier transform* of a function $u \in L^2$ is the function $\hat{u} : \mathbb{R} \to \mathbb{C}$ given by

$$\hat{u}(\xi) = (\mathcal{F}u)(\xi) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\xi x} u(x)\mathrm{d}x. \tag{D.4}$$

Theorem D.3. If $u \in L^2$, then its Fourier transform (D.4) also belongs to L^2, and u can be recovered from $\hat{u}$ by the *inverse Fourier transform*

$$u(x) = (\mathcal{F}^{-1}\hat{u})(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi x} \hat{u}(\xi)\mathrm{d}\xi. \tag{D.5}$$

Theorem D.4. Let $u, v \in L^2$ have Fourier transforms $\hat{u} = \mathcal{F}u, \hat{v} = \mathcal{F}v$. Then

(a) Linearity. If $c \in \mathbb{R}$, then

$$\mathcal{F}\{u + v\}(\xi) = \hat{u}(\xi) + \hat{v}(\xi), \tag{D.6}$$
$$\mathcal{F}\{cu\}(\xi) = c\hat{u}(\xi). \tag{D.7}$$

(b) Translation. If $x_0 \in \mathbb{R}$, then

$$\mathcal{F}\{u(x + x_0)\}(\xi) = e^{i\xi x_0}\hat{u}(\xi). \tag{D.8}$$

(c) Modulation. If $\xi_0 \in \mathbb{R}$, then

$$\mathcal{F}\{e^{i\xi_0 x}u(x)\}(\xi) = \hat{u}(\xi - \xi_0). \tag{D.9}$$

(d) Dilation. If $c \in \mathbb{R}$ with $c \neq 0$, then

$$\mathcal{F}\{u(cx)\}(\xi) = \frac{1}{|c|}\hat{u}\Big(\frac{\xi}{c}\Big). \tag{D.10}$$

(e) Conjugation.

$$\mathcal{F}\{\bar{u}\}(\xi) = \overline{\hat{u}(-\xi)}. \tag{D.11}$$

(f) Differentiation. If $u_x \in L^2$, then

$$\mathcal{F}\{u_x\}(\xi) = i\xi\hat{u}(\xi). \tag{D.12}$$

(g) Inversion.

$$\mathcal{F}^{-1}\{u\}(\xi) = \hat{u}(-\xi). \tag{D.13}$$

Proof. Most of the conclusions follow directly from the definition (D.4). We only show (D.12) by

$$\begin{aligned}\widehat{u_x}(\xi) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\xi x} u_x(x)\mathrm{d}x \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-i\xi)e^{-i\xi x} u(x)\mathrm{d}x = i\xi\hat{u}(\xi),\end{aligned}$$

where we have applied integration by parts with assumptions that $u(x)$ is smooth and decays at ∞. □

Definition D.5. A function $u(x)$ is *even, odd, real, or imaginary* if $u(x) = u(-x)$, $u(x) = -u(-x)$, $u(x) = \overline{u(x)}$, or $u(x) = -\overline{u(x)}$, respectively; $u(x)$ is *Hermitian or skew-Hermitian* if $u(x) = \overline{u(-x)}$ or $u(x) = -\overline{u(-x)}$, respectively.

Theorem D.6. Let $u \in L^2$ have Fourier transform $\hat{u} = \mathcal{F}u$. Then

(a) $u(x)$ is even (odd) $\Leftrightarrow$ $\hat{u}(\xi)$ is even (odd).

(b) $u(x)$ is real (imaginary) $\Leftrightarrow$ $\hat{u}(\xi)$ is hermitian (skew-hermitian) and therefore

(c) $u(x)$ is real and even $\Leftrightarrow$ $\hat{u}(\xi)$ is real and even.

(d) $u(x)$ is real and odd $\Leftrightarrow$ $\hat{u}(\xi)$ is imaginary and odd.

(e) $u(x)$ is imaginary and even $\Leftrightarrow$ $\hat{u}(\xi)$ is imaginary and even.

(f) $u(x)$ is imaginary and odd $\Leftrightarrow$ $\hat{u}(\xi)$ is real and odd.

Definition D.7. The *convolution* of two functions u, v is the function $u * v$ defined by

$$\begin{aligned}&(u*v)(x) = (v*u)(x)\\ &= \int_{-\infty}^{\infty} u(x-y)v(y)\mathrm{d}y = \int_{-\infty}^{\infty} u(y)v(x-y)\mathrm{d}y,\end{aligned}$$

assuming these integrals exist.

Theorem D.8. If $u \in L^2$ and $v \in L^1$ (or vice versa), then $u * v \in L^2$, and $\widehat{u*v}$ satisfies

$$\widehat{u*v}(\xi) = \sqrt{2\pi}\hat{u}(\xi)\hat{v}(\xi). \tag{D.14}$$

Proof. By Definition D.7 and D.2, we calculate

$$\begin{aligned}\widehat{u*v}(\xi) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} u(y)v(x-y)\mathrm{d}y\Big)e^{-i\xi x}\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(y)v(x-y)e^{-i\xi x}\mathrm{d}x\mathrm{d}y\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(y)v(z)e^{-i\xi(y+z)}\mathrm{d}z\mathrm{d}y\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u(y)e^{-i\xi y}\mathrm{d}y\int_{-\infty}^{\infty} u(z)e^{-i\xi z}\mathrm{d}z\\ &= \sqrt{2\pi}\hat{u}(\xi)\hat{v}(\xi),\end{aligned}$$

where the third step follows from the variable substitution $z = x - y$. □

Theorem D.9. The L^2-norms of u and $\hat{u}$ are related by *Parseval's equality*, a.k.a. the *Plancherel theorem*,

$$\forall u \in L^1(\mathbb{R}^N)\cap L^2(\mathbb{R}^N), \qquad \|\hat{u}\| = \|u\|. \tag{D.15}$$

Proof. We calculate

$$\begin{aligned}\int |u(x)|^2\mathrm{d}x &= \int \overline{u(x)}u(x)\mathrm{d}x = \Big(\overline{u(x)} * u(-x)\Big)(0)\\ &= \frac{1}{\sqrt{2\pi}}\int \widehat{\overline{u(x)} * u(-x)}(\xi)\mathrm{d}\xi\\ &= \int \widehat{\overline{u(x)}}(\xi)\widehat{u(-x)}(\xi)\mathrm{d}\xi\\ &= \int \overline{\widehat{u}(-\xi)}\,\widehat{u}(-\xi)\mathrm{d}\xi = \int |\widehat{u}(\xi)|^2\mathrm{d}\xi,\end{aligned}$$

where in the first step $\overline{u(x)}$ denotes the conjugate of $u(x)$, the second step follows from Definition D.7, the third from (D.5), the fourth from Theorem D.8, the fifth from (D.10) and (D.11), and the last from the symmetry of the integral limits. This is almost a proof except that we need to verify that $\overline{u} * u$ is in L^2 and that $\widehat{\overline{u} * u} = |\hat{u}(\xi)|^2$ is in L^1. These are out of the scope of this course and we refer the reader to a standard text on Fourier analysis. □

Example D.10 (B-splines)**.** For the function

$$u(x) = \begin{cases}\sqrt{\frac{\pi}{2}}, & \text{for } -1 \le x \le 1;\\ 0, & \text{otherwise,}\end{cases} \tag{D.16}$$

(D.1) yields $\|u\| = \sqrt{\pi}$, and (D.4) gives

$$\hat{u}(\xi) = \frac{1}{2}\int_{-1}^{1} e^{-i\xi x}\mathrm{d}x = \frac{e^{-i\xi x}}{-2i\xi}\Big|_{-1}^{1} = \frac{\sin\xi}{\xi}, \tag{D.17}$$

where the function $\xi \mapsto \frac{\sin\xi}{\xi}$ is known as the *sinc function*. From (D.1) and the indispensable identity

$$\int_{-\infty}^{\infty}\frac{\sin^2 s}{s^2}\mathrm{d}s = \pi, \tag{D.18}$$

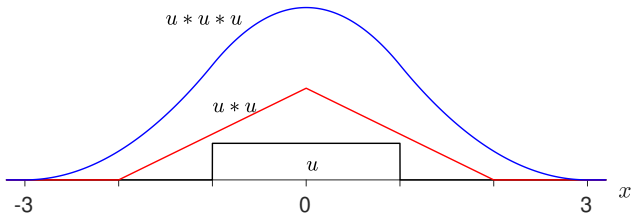
which can be derived by complex contour integration, we calculate $\|\hat{u}\| = \sqrt{\pi}$, which confirms (D.15).

By Definition D.7, it is readily verified that

$$(u*u)(x) = \begin{cases}\frac{\pi}{2}(2-|x|), & \text{for } -2 \le x \le 2,\\ 0, & \text{otherwise,}\end{cases} \tag{D.19}$$

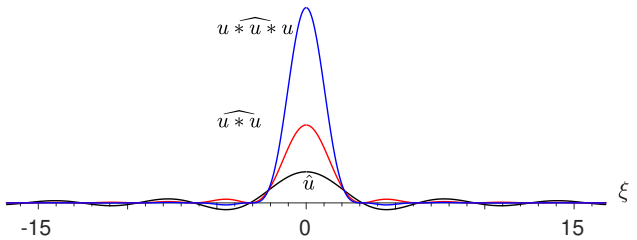
and

$$(u*u*u)(x) = \begin{cases}\frac{3(\sqrt{2\pi})^3}{8} - \frac{(\sqrt{2\pi})^3}{8}x^2, & \text{for } -1 \le x \le 1,\\ \frac{(\sqrt{2\pi})^3}{16}(9-6|x|+x^2), & \text{for } 1 \le |x| \le 3,\\ 0, & \text{otherwise.}\end{cases}$$



By (D.14) and (D.17), their Fourier transforms are

$$\widehat{u*u}(\xi) = \sqrt{2\pi}\frac{\sin^2\xi}{\xi^2}, \quad \widehat{u*u*u}(\xi) = 2\pi\frac{\sin^3\xi}{\xi^3}. \tag{D.20}$$



In general, a convolution $u_{(p)}$ of p copies of u has the Fourier transform

$$\widehat{u_{(p)}}(\xi) = \mathcal{F}\{u*u*\cdots*u\}(\xi) = (2\pi)^{\frac{p-1}{2}}\Big(\frac{\sin\xi}{\xi}\Big)^p.$$

Example D.11. The function

$$u(x) = \begin{cases}\frac{\pi}{2}, & \text{for } -2 \le x < 0,\\ -\frac{\pi}{2}, & \text{for } 0 < x \le 2,\\ 0, & \text{otherwise,}\end{cases} \tag{D.21}$$

has its Fourier transform as

$$\begin{aligned}\hat{u}(\xi) &= \frac{\sqrt{2\pi}}{4}\int_{-2}^{0} e^{-i\xi x}\mathrm{d}x - \frac{\sqrt{2\pi}}{4}\int_{0}^{2} e^{-i\xi x}\mathrm{d}x \\ &= \frac{\sqrt{2\pi}}{-4i\xi}(1 - e^{2i\xi} - e^{-2i\xi} + 1) \\ &= \frac{\sqrt{2\pi}}{4i\xi}(e^{i\xi} - e^{-i\xi})^2 = \sqrt{2\pi}\frac{i\sin^2\xi}{\xi}.\end{aligned}$$

Hence $\hat{u}(\xi)$ is $i\xi$ times the Fourier transform (D.20) of the triangular hat function (D.19), which confirms (D.12).

Definition D.12. A function u defined on $\mathbb{R}$ is said to have *bounded variation* if there is a constant M such that for any finite m and any points $x_0 < x_1 < \cdots < x_m$,

$$\sum_{j=1}^{m} |u(x_j) - u(x_{j-1})| \le M. \tag{D.22}$$

Theorem D.13. Let u be a function in L^2.

(a) If u has $p-1$ continuous derivatives in L^2 for some $p \ge 0$, and a pth derivative in L^2 that has bounded variation, then

$$\hat{u}(\xi) = O(|\xi|^{-p-1}) \quad \text{as } |\xi| \to \infty. \tag{D.23}$$

(b) If u has infinitely many continuous derivatives in L^2, then we have

$$\hat{u}(\xi) = O(|\xi|^{-M}) \quad \text{as } |\xi| \to \infty \text{ for all } M. \tag{D.24}$$

The converse also holds.

(c) If u can be extended to an analytic function of $z = x + iy$ in the complex strip $|\mathrm{Im} z| < a$ for some $a > 0$, with $\|u(x+iy)\| \le$ const uniformly for each constant $-a < y < a$, then

$$e^{a|\xi|}\hat{u}(\xi) \in L^2, \tag{D.25}$$

and conversely.

(d) If u can be extended to an entire function of $z = x + iy$ with $|u(z)| = O(e^{a|z|})$ as $|z| \to \infty (z \in \mathbb{C})$ for some $a > 0$, then $\hat{u}$ has compact support contained in $[-a, a]$, i.e.,

$$\hat{u}(\xi) = 0 \quad \text{for all } |\xi| > a, \tag{D.26}$$

and conversely.

Example D.14. The square wave u of Example D.10 satisfies condition (a) of Theorem D.13 with $p = 0$, so its Fourier transform should satisfy

$$|\hat{u}(\xi)| = O(|\xi|^{-1}),$$

as is verified by (D.17). On the other hand, suppose we interchange the role of u and $\hat{u}$ and apply the theorem again. The function $u(\xi) = \sin\xi/\xi$ is entire, and since $\sin(\xi) = (e^{i\xi} - e^{-i\xi})/2i$, it satisfies

$$u(\xi) = O(e^{|\xi|}) \quad \text{as } |\xi| \to \infty$$

(with ξ now taking complex values). By part (d) of Theorem D.13, it follows that $u(x)$ must have compact support contained in $[-1, 1]$, as indeed it does.

Repeating the example for $u * u$, condition (a) now applies with $p = 1$, and the Fourier transform (D.20) is indeed of magnitude $O(|\xi|^{-2})$, as required. Interchanging u and $\hat{u}$, we note that $\sin^2\xi/\xi^2$ is an entire function of magnitude $O(e^{2|\xi|})$ as $|\xi| \to \infty$, and $u * u$ has support contained in $[-2, 2]$.

D.2 Fourier analysis

Lemma D.15. For a linear PDE of the form

$$\frac{\partial u}{\partial t} + \sum_{n=1}^{N} a_n \frac{\partial^n u}{\partial x^n} = 0, \tag{D.27}$$

the evolution of a single Fourier mode of wave number ξ satisfies the ODE

$$\frac{\partial \hat{u}}{\partial t}(\xi, t) + \sum_{n=1}^{N} a_n (i\xi)^n \hat{u}(\xi, t) = 0. \tag{D.28}$$

Proof. Differentiating (D.5) with respect to t and x yields

$$\begin{aligned}\frac{\partial u(x,t)}{\partial t} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{u}_t(\xi, t) e^{i\xi x}\mathrm{d}\xi, \\ \frac{\partial^n u(x,t)}{\partial x^n} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{u}(\xi, t)(i\xi)^n e^{i\xi x}\mathrm{d}\xi.\end{aligned}$$

Plug these equations into (D.27) and we have (D.28). □

Lemma D.16. The solution to the linear PDE (D.27) with initial condition $u(x, 0) = \eta(x)$ is

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i(\xi x - \omega t)}\mathrm{d}\xi, \tag{D.29}$$

where $\omega := \sum_{n=1}^{N} a_n \xi^n i^{n-1}$.

Proof. Rewrite (D.28) as

$$\frac{\partial \hat{u}(\xi, t)}{\partial t} = -i\omega \hat{u}(\xi, t),$$

and we have from Duhamel's principle

$$\hat{u}(\xi, t) = e^{-i\omega t}\hat{\eta}(\xi).$$

Then the inverse Fourier transform yields

$$\begin{aligned}u(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{u}(\xi, t) e^{i\xi x}\mathrm{d}\xi \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i(\xi x - \omega t)}\mathrm{d}\xi.\end{aligned}$$

□

Definition D.17. The *dispersion relation* of a PDE or a wave problem is the relation between the frequency ω and the wave number ξ, i.e.,

$$\omega = \omega(\xi). \tag{D.30}$$

Example D.18. The beam equation

$$\varphi_{tt} + \gamma^2 \varphi_{xxxx} = 0 \tag{D.31}$$

is characterized by its dispersion relation $\omega = \pm\gamma\xi^2$.

Example D.19. The linear Korteweg-deVries (KdV) equation

$$\varphi_t + c_0\varphi_x + \nu\varphi_{xxx} = 0 \tag{D.32}$$

is characterized by its dispersion relation $\omega = c_0\xi - \nu\xi^3$.

Definition D.20. The system (D.27) is said to be *hyperbolic* if the PDE is hyperbolic; it is *dissipative* if ω is purely imaginary; it is *dispersive* if $\omega(\xi)$ is real and $\omega'(\xi)$ is not a constant.

Definition D.21. The *phase velocity* of a monochromatic wave with wave number ξ is

$$C_p(\xi) := \frac{\omega(\xi)}{\xi}. \tag{D.33}$$

Definition D.22. The *group velocity* of a monochromatic wave with wave number ξ is

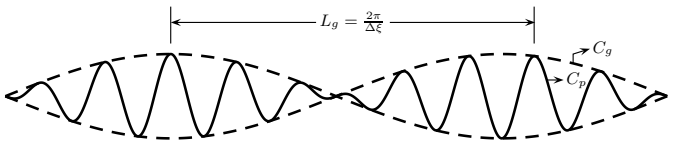
$$C_g(\xi) := \frac{\mathrm{d}\omega(\xi)}{\mathrm{d}\xi}. \tag{D.34}$$

Example D.23. For the linear PDE (D.27), the phase velocity of a single Fourier mode is

$$C_p(\xi) = \sum_{n=1}^{N} a_n \xi^{n-1} i^{n-1}$$

while the group velocity is

$$C_g(\xi) = \sum_{n=1}^{N} n a_n \xi^{n-1} i^{n-1}.$$



Example D.24. For the advection equation, we have $a_1 = a$ and $a_i = 0$ for all $i > 1$. Consequently, we have

$$\omega = a\xi, \qquad C_p = a = C_g,$$

hence Lemma D.16 yields

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i\xi(x-at)} \mathrm{d}\xi = \eta(x-at).$$

Thus all wave modes that constitute the initial data $\eta(x)$ move at the same phase speed, which is also the moving speed of energy.

Example D.25. For the heat equation

$$u_t = \nu u_{xx},$$

we have $a_2 = -\nu < 0$, $a_1 = a_3 = a_4 = \ldots = 0$, and thus

$$\omega = a_2\xi^2 i, \qquad C_p = a_2\xi i, \quad C_g = 2a_2\xi i.$$

Lemma D.16 yields

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i\xi x} e^{a_2\xi^2 t} \mathrm{d}\xi,$$

where the term "$e^{i\xi x}$" denotes the initial mode ξ that does not move while "$e^{-\nu\xi^2 t}$" represents the exponential decay with respect to time.

Example D.26. For the dispersion equation

$$u_t = u_{xxx},$$

we have $a_3 = -1$, $a_1 = a_2 = a_4 = a_5 = \ldots = 0$,

$$\omega = \xi^3, \qquad C_p = \xi^2, \quad C_g = 2\xi^2.$$

Lemma D.16 yields

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta} e^{i\xi(x-\xi^2 t)} \mathrm{d}\xi,$$

thus there is no damping, but different phases move with different speed ξ^2.

Example D.27. For the equation

$$u_t + au_x + bu_{xxx} = 0,$$

we have $a_1 = a$, $a_3 = b$, $a_2 = a_4 = a_5 = \ldots = 0$, and

$$\omega = a\xi - b\xi^3, \qquad C_p = a - b\xi^2, \quad C_g = a - 3b\xi^2.$$

Lemma D.16 yields

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\eta}(\xi) e^{i\xi(x-(a-b\xi^2)t)} \mathrm{d}\xi.$$

Bibliography

W. L. Briggs, V. E. Henson, and S. F. McCormick. *A multigrid tutorial*. SIAM, 2nd edition, 2000. ISBN: 0-89871-462-1.

M. P. Calvo, J. de Frutos, and J. Novo. Linearly implicit Runge-Kutta methods for advection-reaction-diffusion equations. *Appl. Numer. Math.*, 37:535–549, 2001.

C. W. Cryer. *Numerical Functional Analysis*. Monographs on Numerical Analysis. Oxford University Press, 1982. ISBN:9780198534105.

E. Hairer, S. P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, second revised edition, 1993. ISBN: 978-3-540-56670-0.

C. A. Kennedy and M. H. Carpenter. Additive Runge-Kutta schemes for convection-diffusion-reaction equations. *Appl. Numer. Math.*, 44:139–181, 2003.