0 Preliminaries

The main ingredients of snacks are sugar and fat. The main ingredients of math are logic and set theory. We start from the first three items of Section 0.2.

0.1 First-order logic

Definition 0.1 (Statements of first-order logic). A universal statement is a logic statement of the form

\[ U = (\forall x \in S, A(x)). \]  

(0.1)

An existential statement has the form

\[ E = (\exists x \in S, A(x)), \]  

(0.2)

where \( \forall \) (“for each”) and \( \exists \) (“there exists”) are the quantifiers, \( S \) is a set, “s.t.” means “such that,” and \( A(x) \) is the formula.

A statement of implication/conditional has the form

\[ A \Rightarrow B. \]  

(0.3)

Example 0.1. universal and existential statements:

- \( \forall x \in [2, +\infty), x > 1 \)
- \( \forall x \in \mathbb{R}^+, x > 1 \)
- \( \exists p, q \in \mathbb{Z}, \ s.t. \ p/q = \sqrt{2} \)
- \( \exists p, q \in \mathbb{Z}, \ s.t. \ \sqrt{p} = \sqrt{q} + 1 \)

A logic statement is either true or false. There is no such thing that a logic statement is sometimes true and sometimes false. To prove a universal statement, conceptually we have to verify the statement for all elements in the set. To deny a universal statement, we only need to show a counterexample. To prove an existential statement, we only need to show an instance. To deny an existential statement, conceptually we have to show that the statement holds for none of the elements. \( A(x) \) might also be a logic statement. Hence universal and existential statements might be nested. This observation leads to the next definition.

Definition 0.2. A universal-existential statement is a logic statement of the form

\[ U_E = (\forall x \in S, \exists y \in T \ s.t. \ A(x, y)). \]  

(0.4)

Example 0.2. True or false:

- \( \forall x \in [2, +\infty), \exists y \in \mathbb{Z}^+ \ s.t. \ x^3 < 10^5 \)
- \( \exists y \in \mathbb{R} \ s.t. \ \forall x \in [2, +\infty), x > y \)
- \( \exists y \in \mathbb{R} \ s.t. \ \forall x \in [2, +\infty), x < y \)

Example 0.3 (Translating a English statement into a logic 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 \( \forall a \in 2\mathbb{N}^+, \exists p, q \in \mathbb{P}, \ s.t. \ a = p + q \).

Theorem 0.3. The existential-universal statement implies the corresponding universal-existential statement, but not vice versa.

Example 0.4 (Translating a logic statement to an English statement). Let \( S \) be the set of all human beings.

- \( U_E = (\forall p \in S, \exists q \in S \ s.t. \ q \ is \ p \ s \ mom.) \)
- \( E_U = (\exists q \in S \ s.t. \ \forall p \in S, q \ is \ p \ s \ mom.) \)

\( U_E \) is probably true, but \( E_U \) is certainly false. If \( E_U \) were true, then \( U_E \) would be true. why?

Axiom 0.4 (First-order negation of logical statements). The negations of the statements in Definition 0.1 are

- \( \neg U = (\exists x \in S, \neg A(x)). \)  

(0.6)

- \( \neg E = (\forall x \in S, \neg A(x)). \)  

(0.7)

Example 0.5. The negation of a more complicated logic statement abides by the following rules:

- switch the type of each quantifier until you reach the last formula without quantifiers;
- negate the last formula.

One might need to group quantifiers of like type.

Example 0.6 (The negation of Goldbach’s conjecture).

\[ \exists a \in 2\mathbb{N}^+ \ s.t. \ \forall p, q \in \mathbb{P}, a \neq p + q. \]

This conjecture has been shown to hold up through \( 4 \times 10^{18} \), but no proofs and disproofs have been found.

Example 0.7. Negation of the statement in Definition 0.49.

Axiom 0.5 (Contraposition). A conditional statement is logically equivalent to its contrapositive.

\[ (A \Rightarrow B) \leftrightarrow (\neg B \Rightarrow \neg A) \]

(0.8)

Example 0.8. “If Jack is a man, then Jack is a human being.” is equivalent to “If Jack is not a human being, then Jack is not a man.”

Draw a Euler diagram to illustrate the two sets.

0.2 Sets

Definition 0.6. A set \( S \) is a collection of distinct objects \( x \)’s, often denoted with the following notation

\[ S = \{ x \mid \text{the conditions that } x \text{ satisfies.} \}. \]  

(0.9)

Notation 1. \( \mathbb{R}, \mathbb{Z}, \mathbb{N}, \mathbb{Q}, \mathbb{C} \) denote the sets of real numbers, integers, natural numbers, rational numbers and complex numbers, respectively. \( \mathbb{R}^+, \mathbb{Z}^+, \mathbb{N}^+, \mathbb{Q}^+ \) the sets of positive such numbers.

Definition 0.7. \( S \) is a subset of \( U \), written as \( S \subseteq U \), if and only if (iff) \( x \in S \Rightarrow x \in U \). \( S \) is a proper subset of \( U \), written as \( S \subset U \), if \( S \subseteq U \) and \( \exists x \in U \ s.t. \ x \notin S \).

Definition 0.8. The Cartesian product \( X \times Y \) between two sets \( X \) and \( Y \) is the set of all possible ordered pairs with first element from \( X \) and second element from \( Y \):

\[ X \times Y = \{ (x, y) \mid x \in X, y \in Y \}. \]  

(0.10)
Axiom 0.9 (Fundamental principle of counting). A task consists of a sequence of \( k \) 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

\[
p_{\text{total}} = \prod_{i=1}^{k} n_i = n_1 n_2 \cdots n_k. \tag{0.11}
\]

Example 0.9. Let \( A, E, D \) be the set of appetizers, main entrees, desserts in a restaurants. \( A \times E \times D \) is the set of possible dinner combos. If \( \#A = 10, \#E = 5, \#D = 6, \#(A \times E \times D) = 300 \).

Definition 0.10 (Maximum and minimum). Consider \( S \subseteq \mathbb{R}, S \neq \emptyset \). If \( \exists s_m \in S \) s.t. \( \forall x \in S, x \leq s_m \), then \( s_m \) is the maximum of \( S \) and denoted by \( \text{max} S \). If \( \exists s_m \in S \) s.t. \( \forall x \in S, x \geq s_m \), then \( s_m \) is the minimum of \( S \) and denoted by \( \text{min} S \).

Definition 0.11 (Upper and lower bounds). Consider \( S \subseteq \mathbb{R}, S \neq \emptyset \). \( a \) is an upper bound of \( S \subseteq \mathbb{R} \) if \( \forall x \in S, x \leq a \); then the set \( S \) is said to be bounded above. \( a \) is a lower bound of \( S \) if \( \forall x \in S, x \geq a \); then the set \( S \) is said to be bounded below. \( S \) is bounded if it is bounded above and bounded below.

One difference between a maximum and an upper bound is that the former belongs to the set while the latter might not. Another difference is that, for a bounded interval, the upper bound always exists while the maximum might not exist.

Definition 0.12 (Supremum and infimum). Consider \( S \subseteq \mathbb{R}, S \neq \emptyset \). If \( S \) is bounded above and \( S \) has a least upper bound then we call it the supremum of \( S \) and denote it by \( \text{sup} S \). If \( S \) is bounded below and \( S \) has a greatest lower bound, then we call it the infimum of \( S \) and denote it by \( \text{inf} S \).

Example 0.10. If \( S \) has a maximum, then \( \text{max} S = \sup S \).

Example 0.11. \( \sup[a, b] = \text{sup}(a, b) = \text{sup}(a, b) = \text{sup}(a, b) \).

Axiom 0.13 (Completeness of \( \mathbb{R} \)). Every nonempty subset of \( \mathbb{R} \) that is bounded above has a least upper bound.

In other words, \( \sup S \) for any nonempty \( S \subseteq \mathbb{R} \) exists and is a real number.

Corollary 0.14. Every nonempty subset of \( \mathbb{R} \) that is bounded below has a greatest lower bound.

Definition 0.15. A binary relation between two sets \( X \) and \( Y \) is an ordered triple \( (X, Y, G) \) where \( G \subseteq X \times Y \).

A binary relation on \( X \) is the relation between \( X \) and \( X \).

The statement \( (x, y) \in R \) is read “\( x \) is \( R \)-related to \( y \),” and denoted by \( xRy \) or \( R(x,y) \).

Definition 0.16. A binary relation “\( \leq \)” on some set \( S \) is a total order or linear order on \( S \) iff, \( \forall a, b, c \in S \),

\[ a \leq b \text{ and } b \leq c \Rightarrow a \leq c \] (transitivity);

\[ a \leq b \text{ or } b \leq a \] (totality).

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

Example 0.12. The real numbers with less or equal.

Example 0.13. The English letters of the alphabet with dictionary order.

Example 0.14. The Cartesian product of a set of totally ordered sets with the lexicographical order.

Example 0.15. Sort your book in lexicographical order and save a lot of time. \( \log_{26} N \ll N \).

Definition 0.17. A binary relation “\( \preceq \)” on some set \( S \) is a partial order on \( S \) iff, \( \forall a, b, c \in S \), antisymmetry, transitivity, and reflexivity \( (a \leq a) \) hold. A set equipped with a partial order is called a poset.

Example 0.16. The set of subsets of a set \( S \) ordered by inclusion “\( \subseteq \).”

Example 0.17. The natural numbers equipped with the relation of divisibility.

Example 0.18. The the set of stuff you will put on your body every morning with the time ordered: undershorts, pants, belt, shirt, tie, jacket, socks, shoes, watch.

Example 0.19. Inheritance (“is-a” relation) is a partial order. \( A \rightarrow B \) reads “\( B \) is a special type of \( A \).”

Example 0.20. Composition (“has-a” relation) is also a partial order. \( A \rightarrow B \) reads “\( B \) has an instance/object of \( A \).”

Example 0.21. Implication “\( \Rightarrow \)” is a partial order on the set of logical statements.

Example 0.22. The set of definitions, axioms, propositions, theorems, lemma, etc is a poset with inheritance, composition, and implication. It is helpful to relate them with these partial orderings.

“If syntax sugar does not count, there is nothing left.”

0.3 Functions: limits and continuity

Definition 0.18. A function/mapping \( f \) from \( X \) to \( Y \), written as \( f : X \rightarrow Y \) or \( X \rightarrow Y \), is a subset of the Cartesian product \( X \times Y \) satisfying that \( \forall x \in X \), there is exactly one \( y \in Y \) s.t. \( (x, y) \in X \times Y \). \( X \) and \( Y \) are the domain and range of \( f \), respectively.

The important thing in the above definition is the uniqueness of the pair \( (x, y) \).

Definition 0.19. A function \( f : X \rightarrow Y \) is said to be injective or one-to-one iff

\[ \forall x_1 \in X, \forall x_2 \in X, x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2). \tag{0.12} \]

It is surjective or onto iff

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

It is bijective iff it is both injective and surjective.
draw some diagrams.

**Definition 0.20.** A set $S$ is countable if there exists an injective function $f: S \to \mathbb{N}$ that maps $S$ to $\mathbb{N}$.

**Example 0.23.** Is the integers countable? Is the rationals countable? Is the real numbers countable?

**Definition 0.21.** A scalar function is a function whose range is a subset of $\mathbb{R}$.

**Definition 0.22** (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 as $\lim_{x \to c} f(x) = L$, if

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

The notation reads “as $x$ gets closer to $c$, $f(x)$ gets closer to $L$.” How close is close? As close as you wish. This idea is packaged in the $\varepsilon - \delta$ technique.

**Example 0.24.** show that $\lim_{x \to 2} \frac{1}{x} = \frac{1}{2}$.

**Proof.** If $\varepsilon \geq \frac{1}{2}$, choose $\delta = 1$. Then $x \in (1, 3)$ implies $\left|\frac{1}{x} - \frac{1}{2}\right| < \frac{1}{4}$ since $\frac{1}{2} - \frac{1}{x}$ is a monotonically decreasing function with its supremum at $x = 1$.

If $\varepsilon \in (0, \frac{1}{2})$, choose $\delta = \varepsilon$. Then $x \in (2 - \varepsilon, 2 + \varepsilon) \subset \left(\frac{3}{2}, \frac{5}{2}\right)$. Hence $\left|\frac{1}{x} - \frac{1}{2}\right| = \left|\frac{2 - x}{2x}\right| < \left|2 - x\right| < \varepsilon$. The proof is completed by Definition 0.22.

The philosophy here is that if you can make the difference of two functions as small as you wish, then they have the same limit.

**Definition 0.23.** $f: \mathbb{R} \to \mathbb{R}$ is continuous at $c$ iff

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

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

**Definition 0.24.** Let $I = (a, b)$. A function $f: I \to \mathbb{R}$ is uniformly continuous on $I$ iff

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

**Example 0.25.** On $(a, \infty)$, $f(x) = \frac{1}{x}$ 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 $\forall \varepsilon > 0, \exists \delta = \varepsilon a^2$, s.t.

$$\forall x, y \in I, \ |x - y| < \delta \Rightarrow |f(x) - f(y)| < \frac{|x - y|}{a^2} < \frac{\varepsilon}{a^2} = \varepsilon.
$$

If $a = 0$, negating the condition of uniform continuity, i.e., eq. (0.16), yields $\exists \delta > 0 \text{ s.t. } \forall x, y \in I, \ |x - y| < \delta \Rightarrow |f(x) - f(y)| \geq \varepsilon$.

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

If $\delta \geq \frac{1}{\varepsilon}$, choose $x = \frac{1}{\delta}$, $y = \frac{1}{\delta}$. This choice satisfies $|x - y| < \delta$ since $x - y = \frac{1}{\delta} - \frac{1}{\delta} \leq \delta$. However, $|f(x) - f(y)| = \frac{|x - y|}{xy} = 2\varepsilon > \varepsilon$.

If $\delta \leq \frac{1}{\varepsilon}$, then $2\delta < 1$. Choose $x \in (0, \varepsilon \delta^2)$ and $y \in (2\varepsilon \delta, \delta)$. This choice satisfies $|x - y| < \delta$ and $|x - y| > \varepsilon \delta^2$. However, $|f(x) - f(y)| = \frac{|x - y|}{xy} > \frac{\varepsilon^2 \delta}{\frac{1}{\delta}} > 2\varepsilon > \varepsilon$.

**Exercise 0.26.** On $(a, \infty)$, $f(x) = \frac{1}{x}$ is uniformly continuous if $a > 0$ and is not so if $a = 0$.

**Theorem 0.25.** Uniform continuity implies continuity but the converse is not true.

**Proof.** Exercise.

**Theorem 0.26.** $f: \mathbb{R} \to \mathbb{R}$ is uniformly continuous on $(a, b)$ iff it can be extended to a continuous function $f$ on $[a, b]$.

**Definition 0.27.** The derivative of a function $f: \mathbb{R} \to \mathbb{R}$ at $a$ is the limit

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

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

**Example 0.27.** The derivative of power function $f(x) = x^n$ is $f' = nx^{n-1}$.

**Proof.** use Newton’s generalized binomial theorem.

$$(a + h)^n = \sum_{n=0}^{\infty} \binom{\alpha}{n} a^{\alpha - n} h^n.$$ 

Think about the three cases $\alpha = \frac{1}{2}, \alpha = 1, \alpha = 2$. If $x$ is time and $f(x)$ measures how knowledgeable you are. You probably want to have $f(x) = x^2$ rather than $f(x) = x^\frac{1}{2}$. The reason $f(x) = x^2$ is much better is that its rate of increase also increase.

**Definition 0.28.** 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 $C^k(a, b)$.

In comparison, $C^k[a, b]$ is the space of functions $f$ that $f^{(k)}(x)$ is bounded and uniformly continuous on $(a, b)$.

**Theorem 0.29.** A scalar function $f$ is bounded on $[a, b]$ if $f \in C[a, b]$.

**Theorem 0.30** (Intermediate value). A scalar function $f \in C[a, b]$ satisfies

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

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

This is a more explicit version of that in the book, which states that a continuous function assumes all values between $f(a)$ and $f(b)$ on a closed interval $[a, b]$.

**Theorem 0.31.** 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 $f'(x_0) = \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0} > 0$. By the definition of a limit, $\exists \delta > 0 \text{ s.t. } a < x_0 - \delta < x_0 + \delta < b$ and $|x - x_0| < \delta$ imply $\frac{f(x) - f(x_0)}{x - x_0} > 0$, which is a contradiction to $f(x_0)$ being a maximum when we choose $x \in (x_0, x_0 + \delta)$.


Theorem 0.32 (Rolle’s). If a function \( f : \mathbb{R} \to \mathbb{R} \) satisfies

(i) \( f \in C[a, b] \) and \( f' \) exists on \((a, b)\),
(ii) \( f(a) = f(b) \),

then \( \exists x \in (a, b) \) s.t. \( f'(x) = 0 \).

Proof. By Theorem 0.30, all values between \( \sup f \) and \( \inf f \) will be assumed. If \( f(a) = f(b) = \sup f = \inf f \), then \( f \) is a constant on \([a, b]\) and thus the conclusion holds. Otherwise, Theorem 0.31 completes the proof. \( \square \)

Note that T0.30 is about a closed interval and T0.31 an open interval. Thus we must treat the special case of \( f \) being a constant.

Theorem 0.33 (Mean value). If \( f \in C[a, b] \) and \( f' \) exists on \((a, b)\), then \( \exists \xi \in (a, b) \) s.t. \( f(b) - f(a) = f'()b - a\).

Proof. Construct a linear function \( L : [a, b] \to \mathbb{R} \) such that \( L(a) = f(a), L(b) = f(b) \), then \( \forall r \in (a, b), L'(x) = \frac{f(b) - f(a)}{b-a}. \quad \text{Consider} \quad g(x) = f(x) - L(x) \quad \text{on} \quad [a, b], \quad g(a) = 0, \quad g(b) = 0. \) By Theorem 0.32, \( \exists \xi \in [a, b] \) such that \( g'() = 0 \), which completes the proof. \( \square \)

0.4 Taylor series

Definition 0.34. A sequence is a map on \( \mathbb{N}^+ \) or \( \mathbb{N} \).

Example 0.28. Whether the sequence starts from 0 or 1 is a matter of convention and convenience according to the context.

Definition 0.35 (Limit of a sequence). A sequence \( \{a_n\} \) has the limit \( L \), written as \( \lim_{n \to \infty} a_n = L \), or \( a_n \to L \) as \( n \to \infty \), if

\[
\forall \epsilon > 0, \exists N, \text{ s.t. } \forall n > N, \ |a_n - L| < \epsilon. \quad (0.19)
\]

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

Example 0.29 (A sequence can be used to approximate the limit with increasing accuracy: the story of \( \pi \)).

We human beings have more than a trillion digits. A very good estimate is given by Zu, ChongZhi 1500 years ago:

\[
\pi \approx \frac{355}{113} = 3.14159292,
\]

which is good enough for estimating the length of the perimeter of your backyard foundation.

Theorem 0.36 (Bolzano-Weierstrass). Every bounded sequence has a convergent subsequence.

Definition 0.37. A series associated with an infinite sequence \( \{a_n\} \) is defined as \( \sum_{i=0}^{\infty} a_n \), the sum of all terms of the sequence.

Definition 0.38. The sequence of partial sums \( S_n \) associated to a series \( \sum_{i=0}^{\infty} a_n \) is defined for each \( n \) as the sum of the sequence \( \{a_i\} \) from \( a_0 \) to \( a_n \)

\[
S_n = \sum_{i=0}^{n} a_i. \quad (0.20)
\]

Proposition 0.39. A series converges to \( L \) iff the associated sequence of partial sums converges to \( L \).

Definition 0.40. A power series centered at \( c \) is a series of the form

\[
p(x) = \sum_{n=0}^{\infty} a_n(x-c)^n, \quad (0.21)
\]

where \( a_n \)’s are the coefficients. The interval of convergence is the set of values of \( x \) for which the series converges:

\[
I_c(\rho) = \{ x \mid p(x) \text{ converges} \}. \quad (0.22)
\]

Definition 0.41. 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

\[
T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k \quad (0.23)
\]

called the \( n \)-th Taylor polynomial for \( f(x) \) at \( c \). In particular, the linear approximation for \( f(x) \) at \( c \) is

\[
T_1(x) = f(c) + f'(c)(x-c). \quad (0.24)
\]

Example 0.30. If \( f \in C^{\infty} \), then \( \forall n \in \mathbb{N}, \)

\[
T_n(x) = \left\{ \begin{array}{ll}
\sum_{k=0}^{n} \frac{f^{(k)}(c)}{(k-m)!}(x-c)^{k-m}, & m \in \mathbb{N}, m \leq n \\
0, & m \in \mathbb{N}, m > n
\end{array} \right. \quad (0.25)
\]

Proof. By induction. In the inductive step, regroup the summation into a constant term and another shifted summation. \( \square \)

Definition 0.42. The Taylor series (or Taylor expansion) for \( f(x) \) at \( c \) is

\[
\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!}(x-c)^k. \quad (0.25)
\]

Definition 0.43. The remainder of the \( n \)-th Taylor polynomial in approximating \( f(x) \) is

\[
E_n(x) = f(x) - T_n(x). \quad (0.26)
\]
Theorem 0.44. Let $T_n$ be the $n$th Taylor polynomial for $f(x)$ at $c$.

$$\lim_{n \to \infty} E_n(x) = 0 \iff \lim_{n \to \infty} T_n(x) = f(x). \quad (0.27)$$

Lemma 0.45. $\forall m = 0, 1, 2, \ldots, n$, $E_n^{(m)}(c) = 0$.

Proof. This follows from Definition 0.41 and Example 0.30. □

Theorem 0.46 (Taylor’s theorem with Lagrangian form). Consider a function $f : \mathbb{R} \to \mathbb{R}$. If $f \in C^n [c-d, c+d]$ and $f^{(n+1)}(x)$ exists on $(c-d, c+d)$, then $\forall x \in [c-d, c+d]$, there exists some $\xi$ between $c$ and $x$ such that

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}. \quad (0.28)$$

Proof. Fix $x \neq c$, let $M$ be the unique solution of

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

Consider function

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

Clearly $g(x) = 0$. By Lemma 0.45, $g^{(k)}(c) = 0$ for $k = 0, 1, \ldots, n$. Then Rolle’s theorem implies that

$$\exists x_1 \in (c, x) \text{ s.t. } g'(x_1) = 0. \quad (0.31)$$

If $x < c$, change $(c, x)$ above to $(x, c)$. Apply Rolle’s theorem to $g'(t)$ on $(c, x_1)$ and we have

$$\exists x_2 \in (c, x_1) \text{ s.t. } g''(x_2) = 0. \quad (0.32)$$

Repeatedly using Rolle’s theorem,

$$\exists x_{n+1} \in (c, x_n) \text{ s.t. } g^{(n+1)}(x_{n+1}) = 0. \quad (0.33)$$

Since $T_n^{(n+1)}(t) = 0$, $g^{(n+1)}(t) = M - f^{(n+1)}(t) = 0$, hence $M = f^{(n+1)}(x_{n+1})$. The proof is completed by identifying $\xi$ with $x_{n+1}$. □

Example 0.31. How many terms are needed to compute $e^x$ correctly to four decimal places?

By D0.42, the Taylor series of $e^x$ at $c = 0$ is

$$e^x = \sum_{n=0}^{+\infty} \frac{x^n}{(n+1)!}. \quad (0.34)$$

By T0.46,

$$\exists \xi \in [0, 2] \text{ s.t. } E_n(2) = e^\xi 2^n/(n+1)! < e^2 2^{n+1}/(n+1)! \quad (0.35)$$

Plug in the error criteria and we have $n = 12$, hence 13 terms.

0.5 Riemann integral

Definition 0.47. A partition of an interval $I = [a, b]$ is a finite ordered subset $T_n \subseteq I$ of the form

$$T_n(a, b) = \{a = x_0 < x_1 < \cdots < x_n = b\}. \quad (0.36)$$

The interval $I_i = [x_{i-1}, x_i]$ is the $i$th subinterval of the partition. The norm of the partition is the length of the longest subinterval,

$$h_n = h(T_n) = \max(x_i - x_{i-1}), \quad i = 1, 2, \ldots, n. \quad (0.37)$$

Definition 0.48. The Riemann sum of $f : \mathbb{R} \to \mathbb{R}$ over a partition $T_n$ is

$$S_n(f) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}), \quad (0.38)$$

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

Definition 0.49. $f : \mathbb{R} \to \mathbb{R}$ is integrable (or more precisely Riemann integrable) on $[a, b]$ if:

$$\forall L \in \mathbb{R}, \exists \epsilon > 0, \exists \delta > 0 \text{ s.t. } \forall T_n(a, b) \text{ with } h(T_n) < \delta, |S_n(f) - L| < \epsilon. \quad (0.39)$$

Bind each number to the picture that illustrates Riemann integral.

Example 0.32. The following function $f : [a, b] \to \mathbb{R}$ is not Riemann integrable.

$$f(x) = \begin{cases} 1 & x \text{ is rational;} \\ 0 & x \text{ is irrational.} \end{cases} \quad (0.40)$$

Proof. Negate the equation in Definition 0.49.

$$\forall L \in \mathbb{R}, \exists \epsilon > 0, \exists \delta > 0 \text{ s.t. } \forall T_n(a, b) \text{ with } h(T_n) < \delta, |S_n(f) - L| < \epsilon. \quad (0.41)$$

For any given $L$, one can instantiate $\epsilon = \frac{L-a}{n}$ such that, no matter how small $\delta$ is, one can choose $x_i^* \in S_n(f)$ such that the distance between $S_n(f)$ and $L$ is greater than $\epsilon$. □

Definition 0.50. If $f : \mathbb{R} \to \mathbb{R}$ is integrable on $[a, b]$, then its limit is called the definite integral of $f$ on $[a, b]$:

$$\int_a^b f(x)dx = \lim_{n \to \infty} S_n(f). \quad (0.42)$$

Theorem 0.51. A scalar function $f$ is integrable on $[a, b]$ if $f \in C[a, b]$.

Definition 0.52. 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 $\forall x, y, x \leq y \Rightarrow f(x) \geq f(y)$.

Here ordered sets can be posets, but we will limit our attention to chains.

Theorem 0.53. A scalar function is integrable on $[a, b]$ if it is monotonic on $[a, b]$.
Example 0.33. True or false: a bijective function is either order-preserving or order-reversing?
False. missing continuity. In other words, a continuous bijective function is either order-preserving or order-reversing.

Theorem 0.54 (Integral mean value). Let \( w : [a, b] \to \mathbb{R}^+ \) be integrable on \([a, b]\). For \( f \in C[a, b], \exists \xi \in [a, b] \) s.t.
\[
\int_a^b w(x)f(x)dx = f(\xi) \int_a^b w(x)dx.
\]
(0.34)

Proof. Denote \( m = \inf_{x \in [a, b]} f(x) \), \( M = \sup_{x \in [a, b]} f(x) \), and \( I = \int_a^b w(x)dx \). Then \( mw(x) \leq f(x) \leq Mw(x) \) and
\[
mI \leq \int_a^b w(x)f(x)dx \leq MI.
\]
w > 0 implies \( I \neq 0 \), hence
\[
m \leq \frac{1}{I} \int_a^b w(x)f(x)dx \leq M.
\]
Applying Theorem 0.30 completes the proof. \( \square \)

0.6 Vector spaces

Definition 0.55. A field \( \mathcal{F} \) is a commutative division ring. More commonly, a field \( \mathcal{F} \) is a set together with two binary operations, usually called “addition” and “multiplication” and denoted by “+” and “∗”, such that \( \forall a, b, c \in \mathcal{F} \), the following axioms hold,
- commutativity: \( a + b = b + a, \ a \ast b = b \ast a \);
- associativity: \( (a + b) + c = a + (b + c), \ (a \ast b) \ast c = a \ast (b \ast c) \);
- identity: \( a + 0 = a, \ a \ast 1 = a \);
- invertibility: \( a + (−a) = 0, \ a^{-1}a = 1 \);
- distributivity: \( a(b + c) = ab + ac \).

Definition 0.56. A vector space or linear space over a field \( \mathcal{F} \) is a set \( \mathcal{V} \) together with two binary operations “+” and “∗” respectively called vector addition and scalar multiplication that satisfy the following axioms:

(VSA-1) commutativity
\[
\forall u, v \in \mathcal{V}, \ u + v = v + u;
\]
(VSA-2) associativity
\[
\forall u, v, w \in \mathcal{V}, \ (u + v) + w = u + (v + w);
\]
(VSA-3) compatibility
\[
\forall u \in \mathcal{V}, \forall a, b \in \mathcal{F}, \ (ab)u = a(bu);
\]
(VSA-4) additive identity
\[
\forall u \in \mathcal{V}, \exists 0 \in \mathcal{V}, \ s.t. \ u + 0 = u;
\]
(VSA-5) additive inverse
\[
\forall u \in \mathcal{V}, \exists v \in \mathcal{V}, \ s.t. \ u + v = 0;
\]
(VSA-6) multiplicative identity
\[
\forall u \in \mathcal{V}, \exists 1 \in \mathcal{F}, \ s.t. \ 1u = u;
\]
(VSA-7) distributive laws
\[
\forall u, v \in \mathcal{V}, \forall a, b \in \mathcal{F}, \ 
\begin{cases}
(a + b)u = au + bu, \\
(a(u + v)) = au + av.
\end{cases}
\]
The elements of \( \mathcal{V} \) are called vectors and the elements of \( \mathcal{F} \) are called scalars.

Definition 0.57. A vector space with \( \mathcal{F} = \mathbb{R} \) or \( \mathcal{F} = \mathbb{C} \) is called a real vector space or a complex vector space, respectively.

Example 0.34. We will limit ourselves to these two vector spaces for the whole semester.

Example 0.35. The simplest vector space is \( \{0\} \). Another simple example of a vector space over a field \( \mathbb{F} \) is \( \mathbb{F} \) itself, equipped with its standard addition and multiplication.

Definition 0.58. 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: \( x = (x_1, x_2, \ldots, x_n) \).

Example 0.36. According to the definition, two lists are equal iff they have the same length and the same elements in the same order. Hence the two lists \( (3, 5) \) and \( (5, 3) \) are not equal while the two sets \( \{3, 5\} \) and \( \{5, 3\} \) are equal.

The distinguishing features of a list from a set: finite length, ordering, repetition of elements.

Example 0.37. A vector space composed of all the \( n \)-tuples of a field \( \mathcal{F} \) is known as a coordinate space, denoted by \( \mathcal{F}^n (n \in \mathbb{N}^+) \).

Example 0.38. The properties of forces or velocities in the realworld can be captured by a coordinate space \( \mathbb{R}^2 \) or \( \mathbb{R}^3 \).

Example 0.39. The set of continuous real-valued functions on the interval \([−1, 1]\) forms a real vector space.

Example 0.40. For a set \( \mathcal{S} \), define
\[
\mathbb{F}^\mathcal{S} = \{ f : \mathcal{S} \to \mathbb{F} \}.
\]
Then \( \mathbb{F}^\mathcal{S} \) is a vector space. The above notation \( \mathbb{F}^n \) can be thought of as a special case of \( \mathbb{F}^\mathcal{S} \) because \( n \) can be regarded as \( \{1, 2, \ldots, n\} \) and an element in \( \mathbb{F} \) as a constant function.

Definition 0.59. A linear combination of a list of vectors \( \{v_i\} \) is a vector of the form \( \sum a_i v_i \) where \( a_i \in \mathcal{F} \).

Example 0.41. \((17, −4, 2)\) is a linear combination of \((2, 1, −3), (1, −2, 4)\) because
\[
(17, −4, 2) = 6(2, 1, −3) + 5(1, −2, 4).
\]

Example 0.42. \((17, −4, 5)\) 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 \( a_1 = 6, a_2 = 5 \).
Definition 0.60. The span of a list of vectors \( (v_1) \) is the set of all linear combinations of \( (v_1) \),
\[
\text{span}(v_1, v_2, \ldots, v_m) = \left\{ \sum_{i=1}^{m} a_i v_i \mid a_i \in F \right\} .
\] (0.35)

In particular, the span of the empty set is \( \{0\} \). We say that \( (v_1, v_2, \ldots, v_m) \) spans \( V \) if \( V = \text{span}(v_1, v_2, \ldots, v_m) \).

Example 0.43. (17, −4, 2) ∈ span((2, 1, −3), (1, −2, 4))
(17, −4, 5) ∉ span((2, 1, −3), (1, −2, 4))

Definition 0.61. A vector space \( V \) is called finite dimensional if some list of vectors span \( V \); otherwise it is infinite dimensional.

Example 0.44. Let \( P_m(F) \) denotes the set of all polynomials with coefficients in \( F \) and degree at most \( m \),
\[
P_m(F) = \left\{ p : F \rightarrow F \mid p(z) = \sum_{i=0}^{m} a_i z^i, a_i \in F, a_m \neq 0 \right\} .
\]

Then \( P_m(F) \) is a finite-dimensional vector space for each non-negative integer \( m \). \( P_\infty(F) \) is infinite-dimensional. They are both subspaces of \( F^\mathbb{R} \).

Definition 0.62. A list of vectors \( (v_1, v_2, \ldots, v_m) \) in \( V \) is called linearly independent if
\[
a_1 v_1 + \ldots + a_m v_m = 0 \Rightarrow a_1 = \ldots = a_m = 0. \] (0.36)

Otherwise the list of vectors is called linearly dependent.

Example 0.45. The empty list is declared to be linearly independent. A list of one vector \( (v) \) is linearly independent if \( v \neq 0 \). A list of two vectors is linearly independent if neither vector is a scalar multiple of the other.

Example 0.46. The list \( 1, z, \ldots, z^n \) is linearly independent in \( P_m(F) \) for each \( m \in \mathbb{N} \).

Example 0.47. \( (2, 3, 1), (1, -1, 2), \) and \( (7, 3, 8) \) is linearly dependent in \( \mathbb{R}^3 \) because
\[
2(2, 3, 1) + 3(1, -1, 2) + (-1)(7, 3, 8) = (0, 0, 0).
\]

Example 0.48. Every list of vectors containing the 0 vector is linearly dependent.

Lemma 0.63 (Linear dependence lemma). Suppose \( V = (v_1, v_2, \ldots, v_m) \) is a linearly dependent list in \( V \). Then there exists \( j \in \{1, 2, \ldots, m\} \) such that
- \( v_j \in \text{span}(v_1, v_2, \ldots, v_{j-1}) \);
- if the \( j \)th term is removed from \( V \), the span of the remaining list equals \( \text{span}(v_1, v_2, \ldots, v_m) \).

Proposition 0.64. In a finite-dimensional vector space, the length of every linearly independent list of vectors is less than or equal to the length of every spanning list of vectors.

Definition 0.65. A basis of a vector space \( V \) is a list of vectors in \( V \) that is linearly independent and spans \( V \).

Example 0.49. In particular, the list of vectors
\[
(1, 0, \ldots, 0)^T, (0, 1, 0, \ldots, 0)^T, \ldots, (0, \ldots, 0, 1)^T
\] (0.37)
is called the standard basis of \( F^n \).

Example 0.50. A basis of \( P_m(F) \) in Example 0.44 is \( (z^0, z^1, \ldots, z^n) \).

Proposition 0.66. A list of vectors \( (v_1, \ldots, v_n) \) is a basis of \( V \) if every vector \( u \in V \) can be written uniquely as
\[
u = \sum_{i=1}^{n} a_i v_i,
\] (0.38)
where \( a_i \in F \).

Proposition 0.67. Every spanning list in a vector space \( V \) can be reduced to a basis of \( V \).

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

Definition 0.68. The dimension of a finite-dimensional vector space \( V \), denoted \( \dim V \), is the length of any basis of the vector space.

Example 0.51. The above definition is well defined because any two bases of a finite-dimensional vector space have the same length.

Proposition 0.69. If \( V \) is finite-dimensional, then every spanning list of vectors in \( V \) with length \( \dim V \) is a basis of \( V \).

Proposition 0.70. If \( V \) is finite-dimensional, then every linearly independent list of vectors in \( V \) with length \( \dim V \) is a basis of \( V \).

0.7 Inner product space

Definition 0.71. Let \( F \) be the underlying field of a vector space \( V \). The inner product \( \langle u, v \rangle \) on \( V \) is a function \( V \times V \rightarrow F \) that satisfies
- (IP-1) real positivity: \( \forall v \in V, \langle v, v \rangle \geq 0 \);
- (IP-2) definiteness: \( \langle v, v \rangle = 0 \) iff \( v = 0 \);
- (IP-3) additivity in the first slot:
  \( \langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle \);
- (IP-4) homogeneity in the first slot:
  \( \forall a \in F, \forall v, w \in V, \langle av, w \rangle = a \langle v, w \rangle \);
- (IP-5) conjugate symmetry: \( \forall v, w \in V, \langle v, w \rangle = \overline{\langle w, v \rangle} \).

Remark 0.52. The motivation of defining inner products is to capture notions of length, angle, and so on. With length, we can define the limit of a sequence of vectors so that the idea of “approximation” has a solid footing.

Remark 0.53. Complex conjugate of a complex number \( x = a + ib \) is \( x = a - ib \). For any two complex numbers \( x, y, x + y = \bar{x} + \bar{y}, xy = \bar{x} \bar{y}, \) and \( \bar{x} = |x|^2 \). Then we have additivity in the second slot and conjugate homogeneity in the second slot.
Definition 0.72. The **Euclidean inner product** on $\mathcal{F}^n$ is

$$ (v, w) = \sum_{i=1}^{n} v_i \bar{w}_i. \quad (0.39) $$

Example 0.54. An inner product can be defined on the vector space of continuous real-valued functions on the interval $[-1, 1]$ by

$$ (f, g) = \int_{-1}^{+1} f(x)g(x)dx. $$

Definition 0.73. An **inner product space** is a vector space $V$ equipped with an inner product on $V$.

Definition 0.74. Two vectors $u, v$ are called **orthogonal** if $\langle u, v \rangle = 0$.

Two vectors $u, v$ are orthogonal iff their inner product is the additive identity of the underlying field.

Definition 0.75. Let $\mathcal{F}$ be the underlying field of a vector space $V$. A norm on $V$ is a function $V \to \mathcal{F}$:

$$ \|v\| = \sqrt{\langle v, v \rangle}. \quad (0.40) $$

Definition 0.76. The **Euclidean $\ell_p$ norm** of a vector $v \in \mathcal{F}^n$ is

$$ \|v\|_p = \left( \sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \quad (0.41) $$

and the **Euclidean $\ell_\infty$ norm** is

$$ \|v\|_\infty = \max_i |v_i|. \quad (0.42) $$

**Theorem 0.77** (Equivalence of norms). Any two norms $\|\cdot\|_N$ and $\|\cdot\|_M$ on a finite dimensional vector space $V = \mathbb{C}^n$ satisfy

$$ \exists c_1, c_2 \in \mathbb{R}^+, \text{ s.t. } \forall x \in V, \ c_1 \|x\|_M \leq \|x\|_N \leq c_2 \|x\|_M. \quad (0.43) $$

**Theorem 0.78.** A function $\| \cdot \| : \mathcal{V} \to \mathcal{F}$ is a norm iff it satisfies

- (NRM-1) **real positivity**: $\forall v \in \mathcal{V}$, $\|v\| \geq 0$;
- (NRM-2) **point separation**: $\|v\| = 0 \Rightarrow v = 0$.
- (NRM-3) **absolute homogeneity**: $\forall a \in \mathcal{F}, \forall v \in \mathcal{V}$, $\|av\| = |a|\|v\|$;
- (NRM-4) **triangle inequality**: $\forall u, v \in \mathcal{V}$, $\|u + v\| \leq \|u\| + \|v\|$.

**Remark 0.55.** Theorem 0.78 is another common way to define a norm. The norm defined in (0.40) indeed satisfies conditions (NRM-1,2,3,4). For (NRM-3),

$$ \|av\|^2 = \langle av, av \rangle = a \langle v, av \rangle = a(a\langle v, v \rangle) = |a|^2\|v\|^2. $$

To prove (NRM-4), we have

$$ \|u + v\|^2 = (u + v, u + v) $$

$$ = (u, u) + (v, v) + (u, v) + (v, u) $$

$$ \leq \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| $$

$$ = (\|u\| + \|v\|)^2, $$

where the third step uses Cauchy-Schwarz inequality.

**Theorem 0.79** (Pythagorean). If $u, v$ are orthogonal, then $\|u + v\|^2 = \|u\|^2 + \|v\|^2$.

**Theorem 0.80** (Cauchy-Schwarz inequality).

$$ |\langle u, v \rangle| \leq \|u\|\|v\|, \quad (0.44) $$

where the inequality holds iff one of $u, v$ is a scalar multiple of the other.

**Proof.** For any complex number $\lambda$, (IP-1) implies

$$ \langle x_1 + \lambda x_2, x_1 + \lambda x_2 \rangle \geq 0 $$

$$ \Rightarrow \langle x_1, x_1 \rangle + \lambda \langle x_1, x_2 \rangle + \lambda \langle x_2, x_1 \rangle + \lambda \lambda \langle x_2, x_2 \rangle \geq 0. $$

Set $\lambda = -\langle x_1, x_2 \rangle / \langle x_2, x_2 \rangle$, substitute into the above and we obtain (0.44). \qed

**Example 0.56.** If $x_i, y_i \in \mathbb{R}$, then

$$ \left| \sum_{i=1}^{n} x_i y_i \right|^2 \leq \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2. $$

**Example 0.57.** If $f, g : [-1, 1] \to \mathbb{R}$ are continuous, then

$$ \left| \int_{-1}^{+1} f(x)g(x)dx \right|^2 \leq \left( \int_{-1}^{+1} f^2(x)dx \right) \left( \int_{-1}^{+1} g^2(x)dx \right) $$

**Definition 0.81.** In $\mathbb{C}^n$, a sequence of vectors $\{x_1, x_2, \ldots \}$ is said to **converge** to a vector $x$ iff $\lim_{m \to \infty} \|x_m - x\| = 0$. 
1 Computer Arithmetic

1.1 Floating point number system

If scientific computing were a person, then math would be the soul and computer science the body.

An important principle: hardware determines software.

This is especially true for parallel computing.

World’s fastest computer as of June 2016:

Tianhe-2 with a total of more than 10 million cores and its peak speed at 93 PetaFlops/s.

kilo, mega, giga, tera, peta, exa

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

The word “bit” is a contraction of binary digit.

Definition 1.2. A byte is a unit of information in computing that commonly consists of 8 bits; it is the smallest addressable unit of memory in many computers.

The unit symbols of a bit and a byte are “b” and “B,” respectively. $1kB = 1024B$, $1MB = 1024^2B$, $1GB = 1024^3B$, and $1TB = 1024^4B$.

Definition 1.3. A word is a group of bits with fixed size that are handled as a unit by the instruction set architecture (ISA) and/or hardware of the processor. The word size/width/length is the number of bits in a word and is an important characteristic of processor or computer architecture.

Example 1.1. 32-bit and 64-bit computers are mostly common these days. A 32-bit register can store $2^{32}$ values, hence a processor with 32-bit memory address can directly access 4GB byte-addressable memory.

Holders for floating point numerical values are typically either a word or a multiple of a word. Exactly how many bits are there in a word depends on different architecture.

Definition 1.4. A floating point number system $\mathbb{F} \subset \mathbb{Q}$ is a subset of the rational numbers characterized by

- the base or (radix) $\beta$;
- the precision $p$;
- the exponent range $[L, U]$.

A floating point number (FPN) $y \in \mathbb{F}$ has the form

$$x = \pm m \times \beta^e,$$  \hspace{1cm} (1.1)

where $e \in [L, U]$. The significand/mantissa has the form

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

where the integer $d_i$ satisfies $i \in [0, p-1]$, $d_0 \in [0, \beta - 1]$. $d_0$ and $d_{p-1}$ are called the most significant digit and the least significant digit, respectively. The portion $d_1d_2 \cdots d_{p-1}$ is called the fraction.

Example 1.2. Convert a decimal integer to a binary number:

- divide by 2 and record the remainder,
- repeat until you reach 0,
- concatenate the remainder backwards.

convert 7 and 156 to binary number:

$7 = (111)_{2}$, $156 = (10011110)_2$.

Example 1.3. Convert a decimal fraction to binary number:

- 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.

What is the normalized binary form of $x = 2/3$?

$$x = 0.666 \ldots = (0.10101010 \cdots)_2 = (1.0101010 \cdots)_2 \times 2^{-1}.$$  \hspace{1cm} (1.3)

Example 1.4. Why would the procedures work in the above two examples? Because (1) there are only two possibilities for each bit of a binary number, (2) multiplying by two or dividing by two is a shift of the binary digits.

Also by shifting digits, the FPN form (1.1) is not uniqueness.

Definition 1.5. An FPN is normalized if its mantissa satisfies $1 \leq m < \beta$.

Example 1.5. FPN systems are usually normalized, why?

- The representation of each number is then unique;
- No digits are wasted on leading zeros, thereby maximizing precision;
- In a binary system, the leading bit is always 1 and thus need not be stored, thereby gaining one extra bit of precision for a given field width.

We will assume all FPN systems are normalized unless explicitly stated otherwise.

Definition 1.6. The single precision and double precision FPNs of IEEE standard 754 are normalized FPN systems with the respective characterizations,

$$\beta = 2, \quad p = 23 + 1, \quad e \in [-126, 127],$$ \hspace{1cm} (1.3)

$$\beta = 2, \quad p = 52 + 1, \quad e \in [-1022, 1023].$$ \hspace{1cm} (1.4)

Remark 1.6. 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 mantissa (sketch 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 $d_0$.
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 magnitude are smaller than one. Hence we subtract $1, 2, \ldots, 256$ by $128$ to shift the exponents to $-127, -126, \ldots, 0, 127, 128$. Out of these numbers, $\pm m \times 2^{127}$ is reserved for $\pm 0$ and $\pm m \times 2^{128}$ is reserved for $\pm \infty$, i.e. NaN.

**Definition 1.7.** The machine precision of $\mathbb{F}$ is the distance between 1.0 and the next larger FPN in $\mathbb{F}$,

$$\epsilon_M = 2^{1-p}.$$  \hfill (1.5)

**Corollary 1.8.** If an FPN system $\mathbb{F}$ is normalized, then $$\text{UFL}(\mathbb{F}) = \min |\mathbb{F} \setminus \{0\}| = 2^L,$$ \hfill (1.6) $$\text{OFL}(\mathbb{F}) = \max |\mathbb{F}| = 3 \beta^L (\beta - 2^{1-p}).$$ \hfill (1.7)

**Corollary 1.9.** For a normalized binary FPN system $\mathbb{F}$,

$$\# \mathbb{F} = 2^p (U - L + 1) + 1.$$ \hfill (1.8)

*Proof.* The cardinality can be proved by Axiom 0.9. The factor $2^p$ comes from the sign bit and the mantissa. By Remark 1.6, $U - L + 1$ is the number of elements that represent real numbers. However, the number 0 is missing, so we have to add 1 at the end. \hfill \square

**Definition 1.10.** The range of a normalized FPN system is a subset of $\mathbb{R}$,

$$\mathcal{R}(\mathbb{F}) = \{ x : x \in \mathbb{R}, \text{UFL}(\mathbb{F}) \leq |x| \leq \text{OFL}(\mathbb{F}) \}. $$ \hfill (1.9)

**Example 1.7.** Consider a normalized FPN system with the characterization $\beta = 2, p = 3, L = -1, U = +1$. \hfill \square

**Definition 1.11.** Two normalized FPNs $x, y$ are adjacent to each other in $\mathbb{F}$ iff

$$\forall x \in \mathbb{F} \setminus \{x, y\}, \ |x - y| < |x - z| + |z - y|. $$ \hfill (1.10)

**Lemma 1.12.** Let $x, y$ be two adjacent normalized FPNs satisfying $|x| < |y|$ and $xy > 0$. Then

$$\beta^{-1} \epsilon_M |x| < |x - y| \leq \epsilon_M |x|. $$ \hfill (1.11)

*Proof.* Consider $x > 0$, then $\Delta x = y - x > 0$. By Definitions 1.4 and 1.5, $x = m \times \beta^e$ with $1.0 \leq m < \beta$. $x$ and $y$ only differ from each other at the least significant digit, hence $\Delta x = \epsilon_M \beta^e$. Since $\frac{\Delta x}{x} < \frac{\epsilon_M}{x} \leq \epsilon_M$ and thus $\frac{\Delta x}{x} \in \{ \beta^{-1} \epsilon_M, \epsilon_M \}$. The other cases are similar. \hfill \square

**Definition 1.13.** The subnormal or denormalized numbers are FPNs of the form (1.1) with $e = L$ and $m \in (0, 1)$. A normalized FPN system can be extended by including the subnormal numbers.

**Remark 1.8.** Subnormal numbers have inherently lower precision than normalized FPNs because they have fewer significant digits in their fractional parts.

**Example 1.9.** Adding subnormal FPNs to the previous normalized FPN system with $\beta = 2, p = 3, L = -1, U = +1$ yields

\begin{align*}
\begin{array}{ccccccc}
-3 & -2 & -1 & 0 & 1 & 2 & 3 \\
-2.5 & -2 & -1.5 & 0 & 1 & 1.5 & 2
\end{array}
\end{align*}

\subsection*{1.2 Rounding error analysis}

As a crucial difference between $\mathbb{R}$ and $\mathbb{F}$, $\mathbb{R}$ is continuous and infinite while $\mathbb{F}$ is discrete and finite. To perform computing using $\mathbb{F}$, we have to represent real numbers by the machine numbers in $\mathbb{F}$. Rounding is the act of associating a real number with a suitable machine number.

**Definition 1.14** (rounding). Let $\mathbb{Q}_u \supseteq \mathbb{F}$ denote all rational numbers of the form (1.1) with $e \in \mathbb{Z}$. rounding is a map $\text{fl}(x) : \mathbb{R} \to \mathbb{Q}_u$. The default rounding mode is round to nearest, i.e. $\text{fl}(x)$ is chosen to minimize $|\text{fl}(x) - x|$. In the case of a tie, $\text{fl}(x)$ is chosen by round to even, i.e. $\text{fl}(x)$ is the one with an even last digit $d_{p-1}$.

**Definition 1.15.** A rounded number $\text{fl}(x)$ overflows if $|\text{fl}(x)| > |\text{fl}|$, or underflows if $0 < |\text{fl}(x)| < |\text{fl}|$. A gradual underflow is an underflow of an extended FPN system.

**Definition 1.16.** The unit roundoff of $\mathbb{F}$ is the number $\epsilon_u = \frac{1}{2} \beta^{1-p} = \frac{1}{2} \epsilon_M$. \hfill (1.12)

Some books do not distinguish machine precision from the unit roundoff; we do in this class.

**Theorem 1.17.** If $x \in \mathcal{R}(\mathbb{F})$, then $$\text{fl}(x) = x(1 + \delta), \quad |\delta| < \epsilon_u $$ \hfill (1.13)

*Proof.* By Definition 0.16, $\mathcal{R}(\mathbb{F})$ is a chain. Hence $\forall x \in \mathbb{R}$, $\exists x_L, x_R \in \mathcal{R}(\mathbb{F})$ s.t.

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

If $x = x_L$ or $x_R$, then $\text{fl}(x) - x = 0$, hence we can assume $x_L < x < x_R$.

By Lemma 1.12, Definitions 1.11 and 1.14, $|\text{fl}(x) - x| \leq \frac{1}{2} |x_R - x_L| \leq \epsilon_u \min(|x_L|, |x_R|) < \epsilon_u |x|$. Hence $-\epsilon_u |x| < \text{fl}(x) - x < \epsilon_u |x|$, which yields (1.13). \hfill \square

**Theorem 1.18.** If $x \in \mathcal{R}(\mathbb{F})$, then

$$\text{fl}(x) = \frac{x}{1 + \delta}, \quad |\delta| \leq \epsilon_u $$ \hfill (1.14)

*Proof.* Similar as the previous one, but the equality holds in the case of $x = \frac{1}{2} (x_L + x_R)$ and $x_L$ has $m = 1$. \hfill \square

**Example 1.10.** Find $x_L, x_R$ of $x = 2/3$ in normalized single-precision IEEE 754 standard, which of them is $\text{fl}(x)$?

*Proof.* By the example of Definition 1.4,

$$\frac{2}{3} = (0.1010 \cdots)_2 = (1.0101010 \cdots)_2 \times 2^{-1}. $$

\begin{align*}
x_L &= (1.0101010 \cdots)_2 \times 2^{-1}, \quad x_R = (1.010 \cdots)_2 \times 2^{-1}.
\end{align*}

$x - x_L = \frac{2}{3} \times 2^{-24}$ and $x_R - x_L = 2^{-24}$ so $x_R - x = (x_R - x_L) - (x - x_L) = \frac{1}{2} \times 2^{-24}$. Thus $\text{fl}(x) = x_R$. \hfill \square
Read Definitions 1.21, 1.22 and 1.23 before continuing to Assumption 1.19.

Assumption 1.19 (Model of Machine Arithmetic). For \( x, y \in \mathbb{F} \) and arithmetic operations \( \circ \in +, -, \times, / \),
\[
\text{fl}(x \circ y) = (x \circ y)(1 + \delta), \quad |\delta| \leq \epsilon_u.
\]
(1.15)

Remark 1.11. The model of machine arithmetic assumes that the relative roundoff error is always less than the unit roundoff.

It is emphasized that in the model of machine arithmetic, \( x, y \) are exact machine-representable numbers. So this model should not be confused by the catastrophic cancellation of subtracting two real numbers \( x, y \) not exactly representable by machine numbers. Here the relative error is possibly caused by different exponents and the mantissa having more digits than \( p-1 \).

Example 1.12. Given \( x, y \in \mathbb{R} \), we have
\[
\text{fl}(\text{fl}(x) \circ \text{fl}(y)) = (\text{fl}(x) \circ \text{fl}(y))(1 + \delta_3) = (x(1 + \delta_1) \circ y(1 + \delta_2))(1 + \delta_3)
\]
where \( \delta_3 \leq \epsilon_u \). Although \( 1 + \delta_3 \approx 1 \), \((x(1 + \delta_1) \circ y(1 + \delta_2))\) might have a large relative error.

- Multiplication is benign:
  \[
  x(1 + \delta_1) y(1 + \delta_2) = xy(1 + \delta_1 + \delta_2 + \delta_1 \delta_2).
  \approx xy(1 + \delta_1 + \delta_2).
  \]
- Division is benign:
  \[
  \frac{x(1 + \delta_1)}{y(1 + \delta_2)} = \frac{x}{y}(1 + \delta_1)(1 - \delta_2 + \delta_2^2 - \cdots) \approx \frac{x}{y}(1 + \delta_1 - \delta_2)
  \]
- Addition/subtraction might be catastrophic:
  \[
  x(1 + \delta_1) + y(1 + \delta_2) = x + y + x\delta_1 + y\delta_2
  = (x + y)
  \left(1 + \frac{x\delta_1 + y\delta_2}{x + y}\right).
  \]

Hence the relative error can be arbitrarily large when \( x + y \rightarrow 0 \).

Theorem 1.20. If \( \forall i = 0, 1, \ldots, n, x_i \in \mathbb{F}, x_i > 0 \), then
\[
\text{fl}\left(\sum_{i=0}^{n} x_i\right) = (1 + \delta)\sum_{i=0}^{n} x_i,
\]
(1.16)
where \( |\delta| < (1 + \epsilon_u)^n - 1 \approx n\epsilon_u \).

Proof. See page 49 of Kincaid and Cheney 3rd Ed.

1.3 Stability and Conditioning

When using computer to calculate approximation of a function, we can think of the algorithm as a black box. Two issues of paramount significance are stability and conditioning.

Definition 1.21. Let \( \hat{x} \) be an approximation to \( x \in \mathbb{R} \). The accuracy of \( \hat{x} \) can be measured by its absolute error
\[
E_{\text{abs}}(\hat{x}) = |\hat{x} - x|.
\]
(1.17)

and its relative error
\[
E_{\text{rel}}(\hat{x}) = \frac{|\hat{x} - x|}{|x|}.
\]
(1.18)

Remark 1.13. It is usually the relative error that is of interest because it is scale dependent and the answers to scientific computing vary enormously in magnitude.

Definition 1.22. 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 \( \hat{f}(\hat{x}) = \hat{f}(x) \), assuming such an \( \hat{x} \) exists.

Remark 1.14. Why is it called “backward” error? Given \( \hat{y} \), we want to know for which input have we solved the problem. Or the question: “for which input have we solved the problem?” In the picture below, \( x^*, f^*, y^* \) represent \( \hat{x}, \hat{f}, \hat{y} \), respectively.

Definition 1.23. An algorithm \( \hat{y} = \hat{f}(x) \) for computing \( y = f(x) \) is accurate if its forward error is small for all \( x \), i.e. \( \forall x \in \text{dom}(f), E_{\text{rel}}(\hat{f}(x)) \leq \epsilon c \), where \( c \) is a small constant.

Theorem 1.24 (Loss of precision). If \( x, y \in \mathbb{F}, x > y > 0 \), and
\[
\beta^{-t} \leq 1 - \frac{y}{x} \leq \beta^{-s},
\]
(1.19)
then at most \( t \) and at least \( s \) significant digits are lost in the subtraction \( x - y \).

Proof. Rewrite \( x = m_x \cdot \beta^m \) and \( y = m_y \cdot \beta^m \) with \( 1 \leq m_x, m_y < \beta \). Since \( x > y \), the computer shifts \( y \) so that it has the same exponent as \( x \) before performing \( x - y \). Then
\[
y = (m_y \cdot \beta^{m-n}) \times \beta^n
\Rightarrow x - y = (m_x - m_y \cdot \beta^{m-n}) \times \beta^n
\Rightarrow m_{x-y} = m_x - m_y \left(1 - \frac{m_y \cdot \beta^m}{m_x \cdot \beta^m}\right) = m_x \left(1 - \frac{y}{x}\right)
\Rightarrow \beta^{-t} \leq m_{x-y} < \beta^{-s} \Rightarrow \beta^{-t} \leq m_{x-y} \leq \beta^{-s}.
\]
To normalize \( m_{x-y} \) into the interval \([1, \beta]\), the significant digits of \( m_{x-y} \) has to be shifted to the left and spurious 0’s are attached to the right, which means that at least \( s \) digits and at most \( t \) digits are lost.
**Rule 1.25.** Catastrophic cancellation should be avoided whenever possible.

**Example 1.15.** Calculate \( y = f(x) = x - \sin x \) for \( x \to 0 \).

Since \( x \approx \sin x \) when \( x \) is small, the calculation involves a loss of significance. The solution is to use the Taylor series

\[
x - \sin x = x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \right)
\]

\[
= \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots
\]

**Definition 1.26.** An algorithm \( \hat{f}(x) \) for computing \( y = f(x) \) is backward stable if its backward error is small for all \( x \), i.e.

\[
\forall x \in \text{dom}(f), \exists \hat{x} \text{ s.t. } \hat{f}(x) = f(\hat{x}) \Rightarrow E_{\text{rel}}(\hat{x}) \leq c \epsilon_u, \quad (1.20)
\]

where \( c \) is a small constant.

**Definition 1.27.** An algorithm \( \hat{f}(x_1, x_2) \) for computing \( y = f(x_1, x_2) \) is backward stable if

\[
\forall (x_1, x_2) \in \text{dom}(f), \exists (\hat{x}_1, \hat{x}_2) \text{ s.t. } \hat{f}(x_1, x_2) = f(\hat{x}_1, \hat{x}_2) \Rightarrow \begin{cases} E_{\text{rel}}(\hat{x}_1) \leq c_1 \epsilon_u, \\ E_{\text{rel}}(\hat{x}_2) \leq c_2 \epsilon_u, \end{cases} \quad (1.21)
\]

where \( c_1, c_2 \) are two small constants.

**Example 1.16.** For \( f(x_1, x_2) = x_1 - x_2, x_1, x_2 \in \mathbb{R}(\mathbb{F}) \), the algorithm \( \hat{f}(x_1, x_2) = \text{fl}(f(x_1) - \text{fl}(x_2)) \) is backward stable.

**Proof.** \( \hat{f}(x_1, x_2) = (\text{fl}(x_1) - \text{fl}(x_2))(1 + \delta_3) \) by Assumption 1.19. Then by Theorem 1.17, \( f(x_1, x_2) = (x_1(1 + \delta_1) - x_2(1 + \delta_2))(1 + \delta_3) = (x_1 + \delta_1 - x_2 - \delta_2)(1 + \delta_3) = \hat{x}_1 - \hat{x}_2 \). Hence \( E_{\text{rel}}(x_1) = (\delta_1 + \delta_3 + \delta_1 \delta_3), \) and \( E_{\text{rel}}(x_2) = (\delta_2 + \delta_3 + \delta_2 \delta_3) \).

Note that the above does not hold if the condition \( x_1, x_2 \in \mathbb{R}(\mathbb{F}) \) fails to hold. The next example illustrates this. \( \square \)

**Example 1.17.** For \( f(x) = 1 + x, x \in \mathbb{R}^+ \), the algorithm \( \hat{f}(x) = \text{fl}(1 + \text{fl}(x)) \) is not backward stable.

**Proof.** It suffices to give a counterexample. \( \forall x \in (0, \epsilon_u), \hat{f}(x) = 1.0 = f(\hat{x}) \) with \( \hat{x} = 0 \). However, \( E_{\text{rel}}(\hat{x}) = 1. \) \( \square \)

**Definition 1.28.** An algorithm is stable or numerically stable if

\[
\forall x \in \text{dom}(f), \exists \hat{x} \text{ s.t. } \frac{|\hat{f}(x) - f(x)|}{|f(x)|} \leq c_f \epsilon_u, \quad (1.22)
\]

where \( c_f, c \) are two small constants.

**Remark 1.18.** By Example 1.17, the condition of backward stability might be too strong to rule out reasonable algorithms, the errors of which are dominated by rounding errors. Definition 1.28 weakens the condition on stability and hence applies to a wider range of algorithms. However, it is Definition 1.26 that is the one usually adopted for stability analysis, particularly for condition numbers.

The next picture is a shift of the dashed arrow of the upper triangular in the previous one, where \( f^*, y^* \) represent \( f \) and \( y \), respectively.

**Example 1.19.** Stability means Nearly the right answer to nearly the right question. The computed value \( y \) hardly differs from the value \( y + \Delta y \) that would have been produced by an input \( x + \Delta x \) that hardly differs from the actual input \( x \).

In comparison, backward stability implies Exactly the right answer to nearly the right question.

**Remark 1.20.** If an algorithm is backward stable, then it is numerically stable.

**Proof.** By Definition 1.26, \( f(\hat{x}) = \hat{f}(x) \), hence \( c_f = 0 \). The other condition also follows trivially. \( \square \)

**Example 1.21.** For \( f(x) = 1 + x, x \in \mathbb{R}^+ \), the algorithm \( \hat{f}(x) = \text{fl}(1.0 + \text{fl}(x)) \) is stable.

**Proof.** If \( x < \epsilon_u \), then \( \text{fl}(x) = 0, \hat{f}(x) = 1.0 \). Choose \( \hat{x} = x \), then \( f(\hat{x}) - x = \hat{f}(x) \) and \( \left| \frac{f(x) - f(\hat{x})}{f(x)} \right| = \left| \frac{x}{1+x} \right| < 2 \epsilon_u \).

Otherwise \( x \geq \epsilon_u \). Definitions 1.16 and 1.10 yield \( x \in \mathbb{R}(\mathbb{F}) \). By Theorem 1.17, \( f(x) = (1 + x(1 + \delta_1))(1 + \delta_2), \) 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

\[
E_{\text{rel}}(\hat{x}) = \delta_1 + \delta_2 + \delta_1 \delta_2 < 3 \epsilon_u
\]

\[
\Rightarrow \left| \frac{\hat{f}(x) - f(\hat{x})}{f(\hat{x})} \right| = \left| \frac{\delta_2}{1 + x(1 + \delta_1 + \delta_2 + \delta_1 \delta_2)} \right| \leq \epsilon_u.
\]

\( \square \)

**Remark 1.22.** Stability qualitatively describes the sensitivity of the answer to small changes of the input for an algorithm. The notion of condition numbers quantify it.

**Definition 1.29.** 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,

\[
C_f(x) = \left| \frac{f'(x)}{f(x)} \right|. \quad (1.23)
\]

**Remark 1.23.** Definition 1.29 is derived from a Taylor expansion,

\[
y' - y = f(x + \Delta x) - f(x) = f'(x) \Delta x + O((\Delta x)^2)
\]

\[
\Rightarrow \frac{\Delta y}{y} = \left( \frac{x f'(x)}{f(x)} \right) \frac{\Delta x}{x} + O((\Delta x)^2).
\]
Definition 1.30. 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.

Rule 1.31. The forward error, backward error, and the condition number for \( y = f(x) \) is related by

\[
E_{\text{rel}}(\hat{y}) \lesssim C_f E_{\text{rel}}(\hat{x}).
\]  

Remark 1.24. This follows directly from the definition of condition numbers. The approximation mark “\( \approx \)” refers to the fact that the quadratic term \( (\Delta x)^2 \) has been ignored. One way to interpret the above rule is to say that the computed solution to an ill-conditioned problem may have a large forward error.

Example 1.25. Consider the condition number for \( f(x) = \arcsin(x) \).

By Definition 1.29,

\[
C_f(x) = \left| \frac{x f'(x)}{f(x)} \right| = \frac{x}{\sqrt{1 - x^2} \arcsin x}.
\]

Hence \( C_f(x) \to +\infty \) as \( x \to \pm 1 \). Draw the graph of the function and show that the slope of the graph is vertical at \( x = \pm 1 \).

Example 1.26. Consider solving the equation \( f(x) = 0 \). Let \( r \) be a simple root with \( f(r) = 0 \) and \( f'(r) \neq 0 \). If we perturb the function \( f \) to \( F = f + \epsilon g \), \( g \in C^2 \), where is the new root?

Suppose \( r + h \) is the new root, i.e. \( F(r + h) = 0 \), or,

\[
f(r + h) + \epsilon g(r + h) = 0.
\]

Taylor’s series of \( F(r + h) \) yields

\[
f(r) + h f'(r) + \epsilon [g(r) + h g'(r)] = O(h^2).
\]

Hence we have

\[
h \approx -\epsilon \frac{g(r)}{f'(r) + \epsilon g'(r)} \approx -\epsilon \frac{g(r)}{f'(r)}.
\]

Example 1.27 (Wilkinson). Continuing Example 1.26, we define

\[
f(x) = \prod_{k=1}^{20} (x - k),
\]

\[
g(x) = x^{20}.
\]

How is the root \( r = 20 \) affected by perturbing \( f \) to \( f + \epsilon g \)?

The answer is

\[
h \approx -\epsilon \frac{g(20)}{f'(20)} = -\epsilon \frac{20^{20}}{19!} \approx -\epsilon 10^9.
\]

Hence a small change of the coefficient in the monomial \( x^{20} \) would cause a large change of the root.

Definition 1.32. The componentwise condition number of a vector function \( f : \mathbb{R}^m \to \mathbb{R}^n \) is

\[
\text{cond}_f(x) = \left\{ \frac{\|\Gamma(x)\|}{\|f(x)\|} \right\},
\]

where the matrix \( \Gamma(x) = [\gamma_{\nu\mu}(x)] \) and each component is

\[
\gamma_{\nu\mu}(x) = \left| \frac{x_{\nu} \frac{\partial f_{\mu}}{\partial x_{\nu}}}{f_{\nu}(x)} \right|.
\]

Definition 1.33. The condition number of a vector function \( f : \mathbb{R}^m \to \mathbb{R}^n \) is

\[
\text{cond}_f(x) = \frac{||x||_{\infty} ||\frac{\partial f}{\partial x}||_{\infty}}{||f(x)||_{\infty}}.
\]

Example 1.28. Definition 1.32 might be more accurate than Definition 1.33 in characterizing the condition of an algorithm.

\[
f(x) = \left[ \frac{1}{1 + \frac{1}{x_1}}, \frac{1}{1 + \frac{1}{x_2}} \right]
\]

See [Gautschi 2012 NA, p.16].

Example 1.29. In solving the linear system \( Au = b \), the algorithm can be viewed as taking the input \( b \) and returning the output \( A^{-1}b \), i.e. \( f(b) = A^{-1}b \). Clearly \( \frac{\partial f}{\partial x} = A \).

Then Definition 1.33 becomes

\[
\text{cond}_f(x) = \frac{||b||_{\infty} ||A^{-1}||_{\infty}}{||u||_{\infty} ||A||_{\infty} ||A^{-1}||_{\infty}}.
\]

In practice the input \( b \) can take any value, hence the condition number of \( A \) (in the sense of linear algebra) is

\[
\max(\text{cond}_f(x)) = \frac{||Au||_{\infty} ||A^{-1}||_{\infty}}{||u||_{\infty} ||A||_{\infty} ||A^{-1}||_{\infty}}.
\]

Definition 1.34. The Hilbert matrix \( H_n \in \mathbb{R}^{n \times n} \) is

\[
h_{i,j} = \frac{1}{i + j - 1}.
\]

Remark 1.30. The condition number of Hilbert matrices is

\[
\text{cond}_2 H_n \sim \frac{(\sqrt{2} + 1)^{4n+4}}{2^n n^{4n+4}},
\]

which is \( 1.6 \times 10^{13} \) for \( n = 10 \).

Definition 1.35. The Vandermonde matrix \( V_n \in \mathbb{R}^{n \times n} \) is

\[
v_{i,j} = t_j^{-1},
\]

where \( t_1, t_2, \ldots, t_n \) are parameters.

Remark 1.31. If the parameters \( t_1, t_2, \ldots, t_n \) are equally spaced in \([-1, 1]\), the condition number of Vandermonde matrices is

\[
\text{cond}_2 V_n \sim \frac{1}{\pi} e^{-\pi/4} e^{n/4(\pi + 2\ln 2)},
\]

which is \( 1.05 \times 10^9 \) for \( n = 20 \).

Remark 1.32. The above condition numbers measure the sensitivity of the math problem to the change of the input. The following measures that of an algorithm. If the condition number of the math problem is very large, any numerical computation of the problem is hopeless. In comparison, for an unstable algorithm, there might exist another algorithm that computes the same underlying problem in a stable manner.
Definition 1.36. Consider approximating a function \( f : \mathbb{R}^m \to \mathbb{R}^n \) with an algorithm \( f_A : \mathbb{F}^m \to \mathbb{F}^n \). Assume
\[
\forall \mathbf{x} \in \mathbb{R}^m, \exists \mathbf{x}_A \in \mathbb{R}^m \text{ s.t. } f_A(x) = f(x_A),
\]
the condition number of the algorithm \( f_A \) is defined as
\[
(\text{cond } A)(x) = \frac{1}{\epsilon_u} \inf_{(\mathbf{x} - \mathbf{x}_A)} \frac{||\mathbf{x} - \mathbf{x}_A||}{||\mathbf{x}||}.
\]
Remark 1.33. Obviously, (1.31) implies
\[
\forall \mathbf{x}_A \text{ satisfying (1.30), (cond } A)(x) \leq \frac{1}{\epsilon_u} ||\mathbf{x}_A - \mathbf{x}||.
\]

By Definitions 1.26 and 1.36, to say that (cond \( A \))(\( x \)) is small for an algorithm \( A \) is equivalent to say that \( A \) is backward stable.

Example 1.34. Consider an algorithm \( A \) for calculating \( y = \ln x \). For any positive number \( x \), this program produces a \( y_A \) satisfying \( y_A = (1 + \delta) \ln x \) where \( \delta \leq 5\epsilon_u \).

What is the condition number of the algorithm?

We clearly have
\[
y_A = \ln x_A \quad \text{where } x_A = x^{1+\delta},
\]
and consequently
\[
E_{\text{rel}}(x_A) = \left| \frac{x^{1+\delta} - x}{x} \right| = |x^{\delta} - 1| = |\epsilon^{\delta} \ln x - 1| \
\]
\[\approx |\delta \ln x| \leq 5|\ln x|\epsilon_u.
\]
Hence \( A \) is well conditioned except when \( x \to 0^+ \).

Remark 1.35. We have quantified the sensitivities of the math problem and the corresponding algorithm. When we use an algorithm to solve a math problem, both the condition number of the problem and that of the algorithm influence the overall error.

Theorem 1.37. Consider using computer arithmetic to solve a math problem
\[
f : \mathbb{R}^m \to \mathbb{R}^n, \quad y = f(x).
\]
Denote the computer input and output as
\[
x^* \approx x, \quad y^* = f_A(x^*),
\]
where \( f_A \) is the algorithm that approximates \( f \).

The relative error of approximating \( y \) with \( y_A \) can be bounded as
\[
E_{\text{rel}}(y_A) \leq E_{\text{rel}}(x^*) (\text{cond } f)(x) + \epsilon_u (\text{cond } f)(x^*) (\text{cond } A)(x^*),
\]
where the relative error is defined in (1.18).

Proof. By the triangle inequality, we have
\[
\frac{||y_A - y||}{||y||} \leq \frac{||f_A(x^*) - f(x)||}{||f(x)||} + \frac{||f_A(x^*) - f(x^*)||}{||f(x)||} .
\]
By (1.24), the first term is
\[
\frac{||f(x^*) - f(x)||}{||f(x)||} \leq (\text{cond } f)(x) ||x^* - x|| ||x|| = E_{\text{rel}}(x^*) (\text{cond } f)(x).
\]
By (1.24) and Definition 1.36, the second term is
\[
\left| \frac{||f_A(x^*) - f(x^*)||}{||f(x)||} \right| \leq (\text{cond } f)(x^*) ||x^* - x|| ||x|| = E_{\text{rel}}(x^*) (\text{cond } f)(x).
\]

Furthermore, by Assumption 1.19 and the assumptions on \( \sin x \) and \( \cos x \), we have
\[
f_A(x) = \frac{(1 - \cos x)(1 + \delta_1)(1 + \delta_2)(1 + \delta_3)}{(\sin x)(1 + \delta_4)}
\]
where \( |\delta_i| \leq \epsilon_u \) for \( i = 1, 2, 3, 4 \). Neglecting the quadratic terms of \( O(\delta_i^2) \), the above equation is equivalent to
\[
f_A(x) = \frac{1}{\sin x} \left( 1 + \delta_2 + \delta_4 - \delta_3 - \delta_1 \frac{\cos x}{1 - \cos x} \right)
\]
hence we have \( \varphi(x) = 3 + \frac{\cos x}{1 - \cos x} \) and
\[
(\text{cond } A)(x) \leq \frac{\sin x}{\epsilon_u} \left( 3 + \frac{\cos x}{1 - \cos x} \right).
\]
Hence, (cond \( A \))(\( x \)) \( \to +\infty \) as \( x \to 0 \). On the other hand, (cond \( A \))(\( x \)) \( \to 6/\pi \) as \( x \to \pi/2 \).

Analysis of the other algorithm is left as an exercise.
2 Solving Nonlinear Equations

Remark 2.1. From high school we know how to solve the equation $ax^2 + bx + c = 0$. In a more general context, the analytic solution of the equation $f(x) = 0$ may not be available. Hence we need to study how to obtain the solution numerically. A practical example is the Kepler’s equation

$$x - a \sin x - b = 0,$$

where $a$ and $b$ vary largely.

Definition 2.1. An algorithm is a step-by-step procedure that takes some set of values as its input and produces some set of values as its output.

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

If a precondition is violated, the behavior or the output of the algorithm is undefined. We will use check or require to enforce preconditions.

Definition 2.3. A postcondition is a condition that holds for the output after the execution of an algorithm.

When testing a program, one essentially tests that the postconditions hold after feeding test cases whose input satisfies the preconditions.

Definition 2.4. An invariant is a condition that holds during the execution of an algorithm.

2.1 Three basic methods

Algorithm 2.5. The bisection method finds a root of a continuous function $f : \mathbb{R} \to \mathbb{R}$ by repeatedly bisecting the interval and selecting the subinterval in which 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: $f \in C[a, b]$,

$s\text{gn}(f(a)) \neq s\text{gn}(f(b))$

Output: $c, h, k$

Postconditions: $|f(c)| < \epsilon$ or $|b| < \delta$ or $k = M$

1. $u \leftarrow f(a)$
2. $v \leftarrow f(b)$
3. $h \leftarrow b - a$
4. for $k = 1 : M$
5. $h \leftarrow h/2$
6. $c \leftarrow a + h$
7. $w \leftarrow f(c)$
8. if $|h| < \delta$ or $|w| < \epsilon$
9. break
10. end
11. if $\text{sgn}(w) \neq \text{sgn}(u)$
12. $b \leftarrow c$
13. $v \leftarrow w$
14. else
15. $a \leftarrow c$
16. $u \leftarrow w$
17. end
18. end

Remark 2.2. An algorithm, just as a theorem, is a contract. If we personalize a contract, the contract says “you (the user) gives me the input that satisfies preconditions, I will give you the output that satisfies the postconditions.”

Example 2.3. What are the invariants in Algorithm 2.5? Which quantities do $a, b, c, u, v$ represent? These representation invariants facilitate the proof of Theorem 2.9.

Remark 2.4. Why do we need $M, \delta, \epsilon$?

Each of them is related to a practical stopping criteria.

Definition 2.6 (Q-order of convergence). A convergent sequence $\{x_n\}$ is said to converge to $L$ with $Q$-order $p$ ($p \geq 1$) if

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

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

Remark 2.5. The order of convergence measures the speed of convergence. Take root finding of $f(x) = \sin(x) = 0$ close to 3. If the asymptotic factor $c = 0.1$, then each iteration of a linearly convergent method will yield one more correct digit. If $c = 1/2$, then each iteration of a linearly convergent method will yield one more correct bit. For a quadratically convergent method, each iteration will roughly double the number of correct digits/bits.

Definition 2.7. A sequence of iterates $\{x_n\}$ is said to converge linearly to $L$ if

$$\exists c \in (0, 1), \exists d > 0, \text{ s.t. } \forall n \in \mathbb{N}, |x_n - L| \leq c^n d.$$  \hspace{1cm} (2.1)

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

$$\exists c > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n > N, |x_{n+1} - L| \leq c|x_n - L|^p.$$  \hspace{1cm} (2.2)

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

Remark 2.6. Definition 2.7 follows from Definition 2.6. (Use Definition 0.35 to show this!)

Note that (2.1) also guarantees convergence while (2.2) does not. As an example, the sequence $\{1, -1, 1, -1, \ldots\}$ satisfies (2.2) for $L = 0$, $p = 1$, and $c = 1$, but does not converge.

Theorem 2.8 (monotonic sequence theorem). Every bounded monotonic sequence is convergent.

Theorem 2.9. For a continuous function $f : [a_0, b_0] \to \mathbb{R}$ satisfying $\text{sgn}(f(a_0)) \neq \text{sgn}(f(b_0))$, the sequence of iterates in the bisection method converges linearly with asymptotic factor $\frac{1}{2}$.

$$\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n = \lim_{n \to \infty} c_n = \alpha,$$  \hspace{1cm} (2.3)

$$f(\alpha) = 0,$$  \hspace{1cm} (2.4)

$$|c_n - \alpha| \leq 2^{-n/2}(b_0 - a_0),$$  \hspace{1cm} (2.5)

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

15
Proof. It follows from the bisection method that
\[
\begin{align*}
  a_0 &\leq a_1 \leq a_2 \leq \cdots \leq b_0, \\
  b_0 &\geq b_1 \geq b_2 \geq \cdots \geq a_0, \\
  b_{n+1} - a_{n+1} &\geq \frac{1}{2} (b_n - a_n).
\end{align*}
\]
By Theorem 2.8, \(\{a_n\}\) and \(\{b_n\}\) both converge. Also, \(\lim_{n \to \infty} (b_n - a_n) = \lim_{n \to \infty} \frac{1}{2^n} (b_0 - a_0) = 0\), hence \(\lim b_n = \lim a_n = r\). By the given condition and the algorithm, the invariant \(f(a_n)f(b_n) \leq 0\) always holds. Since \(f\) is continuous, \(\lim f(a_n)f(b_n) = f(\lim a_n)f(\lim b_n)\), then \(f^2(r) \leq 0\) implies \(f(r) = 0\). (2.5) is another important invariant that can be proven by induction. Comparing (2.5) to (2.1) yields convergence of the bisection method. Also, the convergence is linear with asymptotic factor as \(c = \frac{1}{2}\). \(\square\)

Remark 2.7. Pros of the bisection method include guaranteed convergence, clear final error, and ease in parallel computing. Its cons are the slow convergence and potential difficulty to find two points \(a, b\) that satisfy \(f(a)f(b) < 0\). This observations are based on the fact that we might not know the exact form of \(f(x)\) in many practical situations.

Algorithm 2.10. Newton’s method finds the root of \(f: \mathbb{R} \to \mathbb{R}\) near an initial guess \(x_0\) by the iteration formula
\[
x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \in \mathbb{N}. \tag{2.6}
\]

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 \(k = M\)

1. \(x \leftarrow x_0\)
2. for \(k = 0 : M\)
3. \(u \leftarrow f(x)\)
4. if \(|u| < \epsilon\) then \(\mid\) break \(\mid\)
5. \(x \leftarrow x - u/f'(x)\)
6. end
7. end

Remark 2.8. The independent parameters of a function may include a function. This special type of function is called a functional. In matlab, a functional is realized by using the function pointer “@.”

Remark 2.9. Newton’s method converges so fast that stopping criterion \(\delta\) in the bisection method is not needed.

Theorem 2.11. 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 \(\alpha\), then the sequence of iterates \(\{x_n\}\) in the Newton’s method converges quadratically to the root \(\alpha\), i.e.
\[
\lim_{n \to \infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^2} = -\frac{f''(\alpha)}{2f'(\alpha)}. \tag{2.7}
\]

Proof. By Taylor’s theorem T0.46 and 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 \(\xi\) is between \(\alpha\) and \(x_n\). \(f(\alpha) = 0\) yields
\[
-\alpha = -x_n + f(x_n) + \frac{(\alpha - x_n)^2}{2} f''(\xi).
\]
By (2.6), we have
\[(*) : x_{n+1} - \alpha = x_n - \frac{f(x_n)}{f'(x_n)} + \frac{(\alpha - x_n)^2}{2} \frac{f''(\xi)}{f'(x_n)}.
\]
Define
\[M = \frac{\max_{x \in \mathcal{B}} |f''(x)|}{2 \min_{x \in \mathcal{B}} |f'(x)|}\]
and pick \(x_0\) sufficiently close to \(\alpha\) such that
(i) \(|x_0 - \alpha| = \delta_0 < \delta\);
(ii) \(M\delta_0 < 1\);
(iii) \(\forall x \in \mathcal{B}_0 = [\alpha - \delta_0, \alpha + \delta_0], \ f'(x) \neq 0\).

(iii) is realizable by the continuity of \(f'\) and the assumption \(f'(\alpha) \neq 0\). The definition of \(M\) and (*) imply
\[|x_{n+1} - \alpha| \leq M|x_n - \alpha|^2.
\]
Comparing the above to (2.2) 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 \(\alpha\).

By (i) and (ii), we have \(M|x_0 - \alpha| < 1\). Then it is easy to obtain the following via induction,
\[|x_n - \alpha| \leq \frac{1}{M} (M|x_0 - \alpha|)^{2^n},
\]
which shows both (a) and (b) and completes the proof. \(\square\)

Remark 2.10. The main advantage of Newton’s method is its fast convergence.

As a minor disadvantage, Newton’s method require that we know the derivative \(f'(x)\), which may be difficult or time-consuming to compute. The major disadvantage of Newton’s method is that we have no idea whether \(x_0\) is close enough to the real root. Hence the convergence is not guaranteed. The following picture gives an example.
Theorem 2.12. A function \( f : [a, b] \rightarrow [c, d] \) is bijective \( \iff \) it is continuous and strictly monotonic.

Theorem 2.13. If a \( C^2 \) function \( f : \mathbb{R} \rightarrow \mathbb{R} \) satisfies \( f(\alpha) = 0 \), \( f' > 0 \) and \( f'' > 0 \), then \( \alpha \) 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 \( \alpha \).

Proof. By Theorem 2.12, \( f \) is a bijection since \( f \) is continuous and strictly monotonic. With 0 in its range, \( f \) must have one and only one root.

In the proof of Theorem 2.11, we have

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

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, \( x_{n+1} - \alpha = x_n - \alpha - \frac{f(x_n)}{f'(x_n)} \). Therefore \( x_{n+1} \) is closer to \( \alpha \) than \( x_n \); this shows convergence.

Remark 2.11. Sometimes the evaluation of \( f'(x) \) is very time consuming. In this case we can reuse function values at \( x_n \) and \( x_{n-1} \) to approximate the derivative. This is the main idea of the secant method, which can be quite cost effective.

Algorithm 2.14. The secant method finds a root of \( f : \mathbb{R} \rightarrow \mathbb{R} \) near initial guesses \( x_0, x_1 \) by the iteration

\[
x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad n \in \mathbb{N}^+.
\]

**Definition 2.15.** The sequence \( \{F_n\} \) of Fibonacci numbers is defined as

\[
F_0 = 0, \quad F_1 = 1, \quad F_{n+1} = F_n + F_{n-1}.
\]

**Theorem 2.16 (Binet’s formula).** Denote the golden ratio by \( r_0 = \frac{1+\sqrt{5}}{2} \approx 1.618 \) and let \( r_1 = 1 - r_0 = \frac{1-\sqrt{5}}{2} \). Then

\[
F_n = \frac{r_0^n - r_1^n}{\sqrt{5}}.
\]

**Theorem 2.17.** Consider a \( C^2 \) function \( f : \mathbb{R} \rightarrow \mathbb{R} \) on \( B = [a - \delta, a + \delta] \) satisfying \( f(\alpha) = 0 \) and \( f'(\alpha) \neq 0 \). If both \( x_0 \) and \( x_1 \) are chosen sufficiently close to \( \alpha \), then the iterates \( \{x_n\} \) in the secant method converges to the root \( \alpha \) with order \( p = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618 \).

Proof. Define \( f[a, b] = \frac{f(a) - f(b)}{a - b} \) as the divided difference. Then formula (2.8) is equivalent to

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

By the mean value theorem T0.33, there exist \( \xi_n \) between \( x_{n-1} \) and \( x_n \) and \( \zeta_n \) between \( x_{n-1} \) and \( x_n \), and \( \alpha \) such that

\[
x_{n+1} - \alpha = (x_n - \alpha)(x_{n-1} - \alpha) \frac{f''(\xi_n)}{f''(\zeta_n)}.
\]

Define \( E_i = |x_i - \alpha| \),

\[
M = \frac{\max_{x \in B} |f''(x)|}{2 \min_{x \in B} |f'(x)|}.
\]
and we have

\[ ME_{n+1} \leq ME_n ME_{n-1}. \]

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, \( ME_0 < \eta; ME_1 < \eta, ME_2 < ME_1 ME_0 < \eta^2, ME_3 < ME_2 ME_1 < \eta^3, ME_{n+1} < ME_n ME_{n-1} < \eta^{n+\eta_{n-1}} = \eta^{q_n+1}, \) i.e.

\[ E_n < B_n = \frac{1}{M} \eta^{q_n}. \]

\( \{ q_n \} \) is a Fibonacci sequence starting from \( q_0 = 1, q_1 = 1. \) Eqn (2.10) then yields \( q_n \to \frac{\log \frac{1}{M}}{\log \eta} \) as \( n \to \infty. \) Hence \( \lim_{n\to\infty} E_n = 0. \)

Consider the worst case of \( E_n \) being close to its upper bound. Then

\[ B_{n+1} = \frac{1}{M} \eta^{q_{n+1}} = \frac{1}{M} \eta^{q_{n+1}} = M^{\eta^{q_{n+1}}/\eta^{q_{n+2}}} \leq M^{\eta^{q_{n+1}}/\eta^{q_n}} \]

since \( q_{n+1} - q_n = r_{n+1}^2 > -1. \) This proves the order of convergence.

**Corollary 2.18.** Consider solving \( f(x) = 0 \) near a root \( \alpha. \) Let \( m \) and \( sm \) be the time to evaluate \( f(x) \) and \( f'(x) \) respectively. The minimum time to obtain the desired absolute accuracy \( \epsilon \) with Newton’s method and the secant method are respectively

\[ T_N = (1 + s)m \log_2 K, \]

\[ T_S = m \log_\eta K, \]

where \( r_0 = \frac{1 + \sqrt{5}}{2}, \) \( \epsilon = \frac{|f''(\alpha)|}{27(\alpha^3)}, \) and

\[ K = \frac{\log \epsilon}{\log |f(\alpha)|}. \]

**Proof.** We showed \( |x_n - \alpha| \leq \frac{1}{M} (M|x_0 - \alpha|)^{2^n} \) in proving Theorem 2.11. Denote \( E_n = |x_n - \alpha|, \) we have

\[ ME_n \leq (ME_0)^{2^n}. \]

The desired accuracy \( \epsilon \) can be satisfied in \( n \) iterations if \((ME_0)^{2^n} \leq M \epsilon. \) When \( \epsilon \) is sufficiently small, \( M \to c. \) Hence we have

\[ n = \log_2 K. \]

For each iteration, Newton’s method incurs one function evaluation and one derivative evaluation, which cost time \( m \) and \( sm, \) respectively. Therefore, (2.11) holds.

For the secant method, the proof of Theorem 2.17 has

\[ ME_n \leq (ME_0)^{2^{n-1}}/\sqrt{5}. \]

The desired accuracy \( \epsilon \) can be satisfied in \( n \) iterations if \((ME_0)^{2^{n-1}}/\sqrt{5} \leq M \epsilon, i.e. \) \( r_0^2 \leq \frac{\sqrt{5}}{2}K. \) Hence

\[ n \leq \log_{r_0} K + \log_{r_0} \frac{\sqrt{5}}{2} < \log_{r_0} K + 1. \]

Since the first two values \( x_0 \) and \( x_1 \) are given in the secant method, the number of iterations is \( \log_{r_0} K \) (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.

**Remark 2.12.** The secant method is faster than the Newton method if the ratio \( T_S/T_N < 1, \) which implies \( s > \frac{\log 2}{\log 5} - 1 \approx 0.44. \) Although Newton’s method has a faster convergence, the secant method may be less expansive per iteration. Hence both convergence and cost should be considered when choosing the most appropriate method.

### 2.2 Fixed-point iterations

**Definition 2.19.** A fixed point of a function \( g \) is an independent parameter of \( g \) satisfying \( g(\alpha) = \alpha. \)

**Definition 2.20.** A fixed-point iteration is a method for finding a fixed point of \( g \) with a formula of the form

\[ x_{n+1} = g(x_n), \quad n \in \mathbb{N}. \]

**Example 2.13.** Newton’s method is a fixed-point iteration.

**Example 2.14.** To calculate the square root of some positive real number \( a, \) we can formulate the problem as find the root of \( f(x) = x^2 - a = 0. \) For \( a = 1, \) the initial guess of \( x_0 = 2, \) and the three choices of \( g_1(x) = x = x^2 + x - a, \) \( g_2(x) = x = \frac{a}{x}, \) and \( g_3(x) = x = \frac{1}{2}(x + \frac{a}{x}), \) verify that \( g_1 \) diverges, \( g_2 \) oscillates, \( g_3 \) converges. The theorems in this section will explain why.

**Lemma 2.21.** If \( g : [a, b] \to [a, b] \) is continuous, then \( g \) has at least one fixed point in \([a, b].\)

**Proof.** If \( a = g(a) \) or \( b = g(b), \) we are done. Otherwise \( g(x) \in (a, b). \) Consider \( f(x) = g(x) - x. \) \( f(a) > 0 \) and \( f(b) < 0. \) The proof is completed by the intermediate value theorem T0.30.

**Definition 2.22.** A function \( g : [a, b] \to [a, b] \) is a contraction or contractive mapping on \([a, b] \) if

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

**Example 2.15.** Any linear function \( f(x) = \lambda x + c \) with \( 0 \leq \lambda < 1 \) is a contraction.

**Theorem 2.23.** If \( g(x) \) is a continuous contraction on \([a, b], \) then it has a unique fixed point \( \alpha \in [a, b]. \) Furthermore, the fixed-point iteration (2.14) converges to \( \alpha \) for any choice \( x_0 \in [a, b] \) and

\[ |x_n - \alpha| \leq \frac{\lambda^n}{1 - \lambda}|x_1 - x_0|. \]
Proof. By Lemma 2.21, $g$ has at least one fixed point in $[a, b]$. Suppose there are two distinct fixed points $\alpha$ and $\beta$, then $|\alpha - \beta| = |f(\alpha) - f(\beta)| \leq \lambda|\alpha - \beta|$, which implies $|\alpha - \beta| \leq 0$, i.e., the two fixed points are identical.

By Definition 2.22, $x_{n+1} = g(x_n)$ implies that all $x_n$’s stay in $[a, b]$. To prove convergence,

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

By induction and the triangle inequality,

$$|x_n - \alpha| \leq \lambda^n|x_0 - \alpha|$$

$$\leq \lambda^n(|x_1 - x_0| + |x_1 - \alpha|)$$

$$= \lambda^n(|x_1 - x_0| + \lambda|x_0 - \alpha|).$$

From the first and last right-hand sides (RHSs), we have $|x_0 - \alpha| \leq \frac{1}{1-\lambda}|x_1 - x_0|$, which yields (2.16). \hfill \Box

Remark 2.16. Sometimes it is difficult to check whether a function is a contraction. Alternatively, we can check that the function derivative is uniformly less than one.

Theorem 2.24. Consider $g : [a, b] \rightarrow [a, b]$. If $g \in C^1[a, b]$ and $\lambda = \max_{x \in [a, b]} |g'(x)| < 1$, then $g$ has a unique fixed point $\alpha$ in $[a, b]$. Furthermore, the fixed-point iteration (2.14) converges to $\alpha$ for any choice $x_0 \in [a, b]$, the error bound (2.16) holds, and

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

(2.17)

Proof. The mean value theorem T0.33 implies that, for all $x, y \in [a, b]$, $g'(\alpha) \leq \lambda|x - y|$. Theorem 2.23 yields all the results except (2.17), 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 $\xi$ is between $x_n$ and $\alpha$. \hfill \Box

Corollary 2.25. Let $\alpha$ be a fixed point of $g : \mathbb{R} \rightarrow \mathbb{R}$ with $|g'(\alpha)| < 1$ and $g \in C^1(B)$ on $B = [\alpha - \delta, \alpha + \delta]$ with some $\delta > 0$. If $x_0$ is chosen sufficiently close to $\alpha$, then the results of Theorem 2.23 hold.

Proof. Choose $\lambda$ so that $|g'(\alpha)| < \lambda < 1$. Choose $\delta_0 \leq \delta$ so that $\max_{x \in B_0} |g'(x)| \leq \lambda < 1$ on $B_0 = [\alpha - \delta_0, \alpha + \delta_0]$. Then $g(B_0) \subset B_0$ and applying Theorem 2.24 completes the proof. \hfill \Box

Remark 2.17. It is now a good point to go back to Example 2.14 and explain the different behaviors of the three methods.

Corollary 2.26. Consider $g : [a, b] \rightarrow [a, b]$ with a fixed point $g(\alpha) = \alpha \in [a, b]$. If $\exists p \in \mathbb{N}^*$, s.t.

$$\left\{ \begin{array}{l} g \in C^p[a, b], \\
\forall k = 1, 2, \ldots, p - 1, g^{(k)}(\alpha) = 0, \\
g^{(p)}(\alpha) \neq 0, \end{array} \right.$$ 

(2.18)

then the fixed-point iteration (2.14) converges to $\alpha$ with $p$th-order accuracy for any choice $x_0 \in [a, b]$.

Proof. By Corollary 2.25, the fixed-point iteration converges uniquely to $\alpha$ because $g'(\alpha) = 0$. The Taylor expansion of $g$ at $\alpha$ is

$$E_{abs}(x_{n+1}) = x_{n+1} - \alpha = g(x_n) - g(\alpha)$$

$$= \sum_{i=1}^{p-1} \frac{(x_n - \alpha)^i}{i!} g^{(i)}(\alpha) + \frac{(x_n - \alpha)^p}{p!} g^{(p)}(\alpha).$$

Since $g^{(p)}$ is continuous on $[a, b]$, by Theorem 0.29, it is bounded on $[a, b]$. Hence there exists a constant $M$ such that $E_{abs}(x_{n+1}) < ME^p_{abs}(x_n)$. \hfill \Box

Example 2.18. 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 $F(x) = \frac{x^2 + 3R}{3x^2 + R}$.

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

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

<table>
<thead>
<tr>
<th>$n$</th>
<th>$F^{(n)}(x)$</th>
<th>$F^{(n)}(\sqrt{R})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-\frac{8R(x^2 + R)}{(3x^2 + R)^2}</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-\frac{48R(9x^4 + 18x^2 + R^2)}{(3x^2 + R)^3}</td>
<td>-\frac{48R(9x^4 + 18x^2 + R^2)}{(4R)^3} = \frac{3}{2R} \neq 0</td>
</tr>
</tbody>
</table>

The rest follows from Corollary 2.26.
3 Interpolation

**Definition 3.1.** Interpolation constructs new data points within the range of a discrete set of known data points, usually by generating an interpolating function whose graph goes through all known data points.

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

3.1 Polynomial interpolation

**Theorem 3.2.** 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

$$p_n(x_i) = f_i, \quad \forall i = 0, 1, \ldots, n.$$  \hspace{1cm} (3.1)

**Proof.** Set up a polynomial $\sum_{i=0}^{n} a_i x^i$ with $n + 1$ undetermined coefficients $a_i$. The condition (3.1) leads to the system of $n + 1$ equations:

$$a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n = f_1,$$

where $i = 0, 1, \ldots, n$. The determinant of the system is the Vandermonde determinant:

$$V(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix}. \hspace{1cm} (3.2)$$

To evaluate $V$, consider the function

$$V(x) = V(x_0, x_1, \ldots, x_{n-1}, x)$$

$$= \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix}. \hspace{1cm} (3.3)$$

Clearly, $V(x) \in \mathbb{P}_n$ 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. Hence it can be expressed as

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

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

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

and consequently the recursion

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

An induction based on $V(x_0, x_1) = x_1 - x_0$ then yields

$$V(x_0, x_1, \ldots, x_{n-1}, x) = \prod_{i>j} (x_i - x_j).$$

The proof is completed by the distinctness of the points and Cramer’s rule. \hspace{1cm} $\square$

**Remark 3.2.** Suppose the values $f_i$ comes from a function $f$ as $f_i = f(x_i)$. Apart from the fact that $p_n(x)$ agrees with $f(x)$ on the given points $x_i$, how much would $p_n(x)$ differ from $f(x)$ for $x \neq x_i$?

**Theorem 3.3** (Generalized Rolle). Let $n \geq 2$. Suppose that $f \in \mathcal{C}[a, b]$ 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 to the $n - 1$ intervals $(x_i, x_{i+1})$ yields $n - 1$ points $\zeta_i$ where $f^{(n)}(\xi) = 0$. Consider $f, f', f'', \ldots, f^{(n-2)}$ as new functions. Repeatedly applying the above arguments completes the proof. \hspace{1cm} $\square$

**Theorem 3.4** (Cauchy remainder for polynomial interpolation). Let $f \in \mathcal{C}^n[a, b]$ and suppose that $f^{(n+1)}(x)$ exists at each point of $(a, b)$. If $a \leq x_0 < x_1 < \cdots < x_n \leq b$, then

$$R_n(f; x) := f(x) - p_n(f; x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \hspace{1cm} (3.4)$$

where $p_n(f; x)$ is the unique polynomial in $\mathbb{P}_n$ that coincides with $f$ at $x_0, x_1, \ldots, x_n$, $\xi \in (a, b)$ and $\xi$ depends on $x, x_0, x_1, \ldots, x_n$, and $f$.

**Proof.** Since $f(x_k) = p_n(f; x_k)$, the remainder $R_n(f; x)$ vanished 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 3.3, $W^{(n+1)}(\xi) = 0$ 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 (3.4) holds. \hspace{1cm} $\square$

**Corollary 3.5.** Suppose $f \in \mathcal{C}^{n+1}[a, b]$. Then

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

where $M_{n+1} = \max_{x \in [a, b]} |f^{(n+1)}(x)|$. \hspace{1cm} (3.5)
Example 3.3. 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 $(x) = \text{arcsin}(x)$. Then

\[
f''(x) = x(1 - x^2)^{-\frac{3}{2}}, \quad f'''(x) = (1 + 2x^2)(1 - x^2)^{-\frac{5}{2}}.
\]

Since the third derivative is positive over $[0.5330, 0.5340]$, the maximum value of $f'''$ occurs at 0.5340. By Corollary 3.5 we have $|R_1| \leq 4.42 \times 10^{-7}$. The true error is about $1.10 \times 10^{-7}$.

3.2 The Lagrange formula

Remark 3.4. The proof of Theorem 3.2 gives one way to construct an interpolating polynomial. However, it requires solving a linear system. Can we construct interpolation polynomials explicitly? The answer is yes and the Lagrange formula is such a well-known one.

Definition 3.6. To interpolate given values $f_0, f_1, \ldots, f_n$ at distinct points $x_0, x_1, \ldots, x_n$, the Lagrangian formula is

\[
p_n(x) = \sum_{k=0}^{n} f_k \ell_k(x), \quad (3.6)
\]

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

\[
\ell_k(x) = \prod_{i \neq k, i=0}^{n} \frac{x - x_i}{x_k - x_i}. \quad (3.7)
\]

Example 3.5. 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 $p_2(x) = 3x^2 - 16x + 21$.

Proposition 3.7. Define

\[
v(x) = \prod_{i=0}^{n} (x - x_i). \quad (3.8)
\]

Then the fundamental polynomial for pointwise interpolation can be expressed as

\[
\ell_k(x) = \frac{v(x)}{(x - x_k)v'(x_k)}. \quad (3.9)
\]

Proof. $v'(x)$ is the summation of $n + 1$ terms. When $x$ is replaced with $x_k$, all terms vanish except one.

Proposition 3.8. The fundamental polynomials $\ell_k(x)$ satisfy the Cauchy relations as follows.

\[
\sum_{k=0}^{n} \ell_k(x) \equiv 1 \quad \forall j = 1, \ldots, n, \quad (3.10)
\]

\[
\sum_{k=0}^{n} (x_k - x_j) \ell_k(x) \equiv 0 \quad (3.11)
\]

Proof. Since $q(x) \in \mathbb{P}_n$, Theorems 3.2 and 3.4 imply $p_n(q; x) \equiv q(x)$. For $j = 0, 1, \ldots, n$, define $q(x) = (x - u)^j$ and apply the Lagrange formula.

3.3 The Newton formula

Remark 3.6. The Lagrange formula has one drawback. If we desire to pass from degree $n$ to degree $n + 1$, we must determine an entirely new set of fundamental polynomials. In the Newton formula, constructing the degree-$(n+1)$ interpolating polynomial only entails adding one more term to the degree-$n$ interpolating polynomial.

For this purpose, we express the unique interpolating polynomial $p_n \in \mathbb{P}_n$ as

\[
p_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots + a_n \prod_{i=0}^{n-1} (x-x_i),
\]

which motivates the following definition.

Definition 3.9. To interpolate given values $f_0, f_1, \ldots, f_n$ at distinct points $x_0, x_1, \ldots, x_n$, the Newton formula is

\[
p_n(x) = a_0 + \sum_{k=1}^{n} \left( a_k \prod_{i=0}^{k-1} (x - x_i) \right), \quad (3.12)
\]

where the $k$th 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 3.10. Suppose $(i_0, i_1, i_2, \ldots, i_k)$ is a permutation of $(0, 1, 2, \ldots, k)$. Then

\[
f[x_0, x_1, \ldots, x_k] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_k}]. \quad (3.13)
\]

Proof. This follows directly from the uniqueness of the interpolating polynomial in Theorem 3.2.

Corollary 3.11. The $k$th divided difference can be expressed as

\[
f[x_0, x_1, \ldots, x_k] = \frac{\sum_{i=0}^{k} \frac{f_i}{\prod_{j \neq i; j=0}^{k} (x_i - x_j)}}{\prod_{i=0}^{k} v'(x_i)}. \quad (3.14)
\]

where $v(x)$ is defined in (3.8).

Proof. Use Definition 3.6 and Proposition 3.7.

Theorem 3.12. Divided differences satisfy the recursion

\[
f[x_0, x_1, \ldots, x_k] = f[x_1, x_2, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}], \quad x = x_0 \quad (3.15)
\]

Proof. By Definition 3.9, $f[x_1, x_2, \ldots, x_k]$ 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 $i = 1, 2, \ldots, k - 1$, $P(x_k) = P_2(x_k)$. Therefore, $P(x)$ as above is the interpolating polynomial for given values at the $k + 1$ points. The rest follows from the definitions of $P(x)$ and the kth divided difference.
Remark 3.7. Theorem 3.12 can be used to generate the table of divided differences.

Definition 3.13. The $k$th divided difference $(k > 0)$ on the table of divided differences

\[
\begin{array}{c|cccc}
  x_0 & f[x_0] & f[x_0_1] & f[x_0_2] & f[x_0_n] \\
  x_1 & f[x_1] & f[x_0_1] & f[x_1_2] & f[x_0_2] \\
  x_2 & f[x_2] & f[x_1_2] & f[x_2_3] & f[x_0_3] \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  x_n & f[x_n] & f[x_{n-1}] & f[x_{n-2}] & f[x_{n-1}] \\
\end{array}
\]

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

Example 3.8. Redo Example 3.5 with the Newton formula. Observe that it is easy to go from linear interpolation to quadratic interpolation.

Example 3.9 (Page 96 of NAG2012). Generate the table of divided differences for $f(0) = 3$, $f(1) = 4$, $f(2) = 7$, $f(4) = 10$.

Theorem 3.14. For distinct points $x_0, x_1, \ldots, x_n$ and an arbitrary $x$, we have

\[
f(x) = f[x_0] + f[x_0, x_1] (x - x_0) + \cdots + f[x_0, \ldots, x_n] \prod_{i=0}^{n-1} (x - x_i) \quad (3.16)
\]

Proof. Take another point $z \neq x$. Applying the Newton formula to $x_0, x_1, \ldots, x_n, z$ yields

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

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

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

Replacing the dummy variable $z$ with $x$ yields (3.16).

The above argument assumes $x \neq x_i$. Now consider the case of $x = x_i$ for some fixed $j$. Rewrite (3.16) as $f(x) = p_n(f; x) + R(x)$. We need to show

\[
p_n(f; x_j) + R(x_j) - f(x_j) = 0,
\]

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 3.15. Suppose $f \in \mathcal{C}[a, b]$ 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

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

where $\xi \in (a, b)$.

Proof. This follows from Theorems 3.14 and 3.4.

Corollary 3.16. If $x_0 < x_1 < \cdots < x_n$, $f \in \mathcal{C}[x_0, x_n]$ and suppose that $f^{(n)}(x)$ exists at each point of $(x, x_n)$, then

\[
\lim_{x \to x_0} f[x_0, x_1, \ldots, x_n, x] = \frac{1}{n!} f^{(n)}(x_0).
\]

Proof. Set $x = x_{n+1}$ in Corollary 3.15 and replace $n + 1$ by $n$. Then we have $\xi \to x_0$ as $x_i \to x_0$.

Remark 3.10. We can think of Theorem 3.14 as a generalization of Taylor’s expansion.

### 3.4 Hermite interpolation

Definition 3.17. 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 $M = \max m_i$, the Hermite interpolation problem seeks to find a polynomial $p$ of the lowest degree such that

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

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

Definition 3.18. The $n$th divided difference at $n+1$ “confluent” (i.e. identical) points is defined as

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

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

Remark 3.11. With Definition 3.18, we can build a table of divided difference for Hermite interpolation, in the same way as we did in Newton interpolation. The only difference here is to apply (3.20) whenever possible; see the examples on pages 98-99 on NAG2012.

Theorem 3.19. For the Hermite interpolation problem in Definition 3.17, denote $N = k + \sum m_i$. Denote by $p_N(f; x)$ the unique element of $\mathbb{P}_N$ for which (3.19) holds. Suppose $f^{(N+1)}(x)$ exists in $(a, b)$. Then

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

Proof. The proof is similar to that of Theorem 3.4. Pay attention to the difference caused by the multiple roots of the polynomial $\prod_{i=0}^{N} (x - x_i)^{m_i+1}$.
3.5 Piecewise-polynomial splines

**Remark 3.12.** The points \( x_0, x_1, \ldots, x_n \) in Theorem 3.2 are usually given \textit{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

\[
\lim_{n \to +\infty} p_n(f; x) = f(x). \tag{3.22}
\]

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

The above plot is created by interpolating

\[
f(x) = \frac{1}{1 + x^2} \tag{3.23}
\]
on \( x_i = -5 + 10 \frac{i}{n}, i = 0, 1, \ldots, n \) with \( n = 2, 4, 6, 8. \)

**Remark 3.13.** There are two ways to achieve (3.22): spline interpolation and Chebyshev approximation. The gist of spline interpolation is to use low-degree polynomials but increase the number of subintervals.

**Definition 3.20.** Given nonnegative integers \( n, k \), and a strictly increasing sequence \( \{x_i\} \) that partitions \([a, b]\),

\[
a = x_1 < x_2 < \cdots < x_N = b, \tag{3.24}
\]
the spline functions of degree \( n \) and smoothness class \( k \) relative to the partition \( \{x_i\} \) is

\[
S_n^k = \{s : s \in C^k[a, b]; \forall i \in [1, N - 1], s|_{[x_i, x_{i+1}]} \in P_n\}. \tag{3.25}
\]
The \( x_i \)'s are called \textit{knots} of the spline.

**Notation 2.** 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 3.14.** As an extreme, \( S_n^0 = P_n \), i.e. all the pieces of \( s \in S_n^2 \) belong to a single polynomial. On the other end, \( S_n^0 \) is the class of piecewise linear interpolating functions; see Fig. 2.10 on page 103 of NAG2012. The most popular splines are the cubic splines in \( S_3^2 \).

**Lemma 3.21.** Denote \( m_i = s'(f; x_i) \) for \( s \in S_3^2 \). Then, for each \( i = 2, 3, \ldots, N - 1 \), we have

\[
\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = 3\lambda_i f[x_i, x_{i+1}] + 3\lambda_i f[x_{i-1}, x_i], \tag{3.26}
\]

where

\[
\mu_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}, \quad \lambda_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}. \tag{3.27}
\]

**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 \( p_i(x) = f_i, \ p_i(x_{i+1}) = f_{i+1}, \ p'_i(x_i) = m_i, \ p'_i(x_{i+1}) = m_{i+1} \) is

\[
\begin{array}{c|ccc}
    & f_i & f_{i+1} & m_i & m_{i+1} \\
    x_i & f_i & f_{i+1} & m_i & m_{i+1} \\
    x_{i+1} & f_{i+1} & K_i & \frac{K_i - m_i}{x_{i+1} - x_i} & \frac{m_{i+1} - K_i}{x_{i+1} - x_i} \\
    \end{array}
\]

Then the Newton formula yield

\[
p_i(x) = f_i + (x - x_i)m_i + (x - x_i)^2 \frac{K_i - m_i}{x_{i+1} - x_i} + (x - x_i)^2 (x - x_{i+1}) \frac{m_{i+1} - 2K_i}{(x_{i+1} - x_i)^2},
\]
or equivalently

\[
\begin{align*}
p_i(x) &= c_{i,0} + c_{i,1}(x - x_i) + c_{i,2}(x - x_i)^2 + c_{i,3}(x - x_i)^3, \\
& \quad c_{i,0} = f_i, \\
& \quad c_{i,1} = m_i, \\
& \quad c_{i,2} = \frac{3K_i - 2m_i - m_{i+1}}{(x_{i+1} - x_i)^2}, \\
& \quad c_{i,3} = \frac{m_{i+1} - 2K_i}{(x_{i+1} - x_i)^2}.
\end{align*}
\]

Then \( p_i(x) \in C^2 \) implies that \( p''_i(x_i) = p''_i(x_{i+1}) \), i.e.

\[
3c_{i-1,3}(x_i - x_{i-1}) = c_{i,2} - c_{i-1,2}.
\]
The substitution of the coefficients \( c_{i,j} \) into the above equation yields (3.26).

**Lemma 3.22.** Denote \( M_i = s''(f; x_i) \) for \( s \in S_3^2 \). Then, for each \( i = 2, 3, \ldots, N - 1 \), we have

\[
\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = 6f[x_{i-1}, x_i, x_{i+1}] \tag{3.29}
\]
where \( \mu_i \) and \( \lambda_i \) are the same as those in (3.27). Also,

\[
2M_i + M_{i+1} = 6f[x_i, x_{i+1}], \tag{3.30}
\]

\[
M_{N-1} + 2M_N = 6f[x_{N-1}, x_N, x_N]. \tag{3.31}
\]

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

\[
s(x) = f_i + s'(x_i)(x - x_i) + \frac{M_i}{2}(x - x_i)^2 + \frac{s''(x_i)}{6}(x - x_i)^3. \tag{3.32}
\]

Differentiate (3.32) twice, set \( x = x_{i+1} \), and we have

\[
s'''(x_i) = \frac{M_{i+1} - M_i}{x_{i+1} - x_i}. \tag{3.33}
\]

Substitute (3.33) into (3.32), set \( x = x_{i+1} \), and we have

\[
s'(x_i) = f[x_i, x_{i+1}] - \frac{1}{6}(M_{i+1} + 2M_i)(x_{i+1} - x_i). \tag{3.34}
\]

Differentiate (3.32) once, set \( x = x_{i-1} \), and we have

\[
s'(x_{i-1}) = s'(x_{i-1}) = M_i(x_i - x_{i-1}) + \frac{s''(x_{i-1})}{2}(x_i - x_{i-1})^2.
\]
The substitution of (3.33) and (3.34) into the above equation yields (3.29).

As for (3.30), the cubic polynomial on \([x_1, x_2]\) can be written as

\[
s_1(x) = f[x_1] + f[x_1, x_1](x - x_1) + \frac{M_1}{2} (x - x_1)^2 + \frac{s''''(x_1)}{6} (x - x_1)^3.
\]

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

\[
s_1(x) = f[x_1] + f[x_1, x_1](x - x_1) + \frac{M_1}{2} (x - x_1)^2 + \frac{M_2 - M_1}{6(x_2 - x_1)} (x - x_1)^3.
\]

Set \(x = x_2\), divide both sides by \(x_2 - x_1\), and we have

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

which yields (3.30). (3.31) can be proven similarly.

**Remark 3.15.** In either (3.26) or (3.29), there are \(N\) unknowns but only \(N - 2\) equations. Two more conditions are needed to determine the spline. How these two conditions are given leads to different types of splines.

**Definition 3.23** (Types of splines).

- A complete cubic spline \(s \in S^3_2\) satisfies boundary conditions \(s'(a) = f'(a)\) and \(s'(b) = f'(b)\).
- A cubic spline with specified second derivatives at its end points: \(s''(a) = f''(a)\) and \(s''(b) = f''(b)\).
- A natural cubic spline \(s \in S^3_2\) satisfies boundary conditions \(s''(a) = 0\) and \(s''(b) = 0\).
- A not-a-knot cubic spline \(s \in S^3_2\) satisfies that \(s'''(x)\) exists at \(x = x_2\) and \(x = x_{N-1}\).
- A periodic cubic spline \(s \in S^3_2\) is obtained from replacing \(s_f(b) = f(b)\) with \(s_f(b) = s_f(a)\), \(s''(f; b) = s''(f; a)\), and \(s''''(f; b) = s''''(f; a)\).

**Remark 3.16.** The natural cubic spline is seldom used in reality because the second derivative of the function to be interpolated is usually nonzero at the end points.

**Theorem 3.24.** For a given function \(f : [a, b] \to \mathbb{R}\), there exists a unique complete/natural/periodic cubic spline \(s(f; x)\) that interpolates \(f\).

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

By the proof of Lemma 3.21, \(s\) is uniquely determined if all the \(m_i\)’s are uniquely determined on all intervals. For a complete cubic spline we already have \(m_1 = f'(a)\) and \(m_N = f'(b)\). Assemble (3.26) into a linear system

\[
\begin{pmatrix}
\frac{2}{\lambda_2} & \frac{\mu_2}{2} & \mu_3 \\
\vdots & \ddots & \ddots \\
\lambda_{i-1} & 2 & \mu_i \\
\lambda_{N-2} & \frac{\mu_{N-2}}{2} & \lambda_{N-1} \frac{m_{N-2}}{2} & m_{N-1}
\end{pmatrix}
\begin{pmatrix}
m_2 \\
m_3 \\
\vdots \\
m_N
\end{pmatrix} = b,
\]

where the vector \(b\) is determined from the known information. (3.27) implies that the matrix in the above equation is diagonally dominant. Therefore its determinant is nonzero and the \(m_i\)’s can be uniquely determined.

**Remark 3.17.** To determine a complete cubic spline, we reduce (3.26) to the \((N-2)\)-by-\((N-2)\) linear system (3.35) by applying end conditions, solve the linear system, calculate divided differences, and finally use (3.28).

To determine a cubic spline with specified 2nd derivatives, we reduce (3.29) to a \((N-2)\)-by-\((N-2)\) linear system by the end conditions, solve the linear system, compute divided differences, and finally use (3.32), (3.33), and (3.34).

**Example 3.18.** 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 \(s(5)\). Verify Theorem 3.28. \(s(5) \approx 1.60977, \ln(5) \approx 1.60944\).

\[
\int_a^b [s''(x)]^2 dx \leq \int_a^b [g''(x)]^2 dx, \tag{3.36}
\]

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 \(\eta \in C^2[a, b]; \eta(a) = \eta(b) = 0\), and \(\forall i = 1, 2, \ldots, N, \eta(x_i) = 0\). Then

\[
\int_a^b [g''(x)]^2 dx = \int_a^b [s''(x) + \eta''(x)]^2 dx = \int_a^b [s''(x)]^2 dx + \int_a^b [\eta''(x)]^2 dx + 2 \int_a^b s''(x)\eta''(x) dx.
\]

It follows from

\[
\int_a^b s''(x)\eta''(x) dx = \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} s''(x)\eta''(x) dx
\]

\[
= \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} \eta''(x) s''(x) dx
\]

\[
= \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} s''(x)\eta''(x) dx
\]

\[
= - \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} \eta''(x) s''(x) dx = \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} \eta(x) s''(x) dx
\]

\[
= 0
\]

that

\[
\int_a^b [g''(x)]^2 dx = \int_a^b [s''(x)]^2 dx + \int_a^b [\eta''(x)]^2 dx,
\]

which completes the proof.
Theorem 3.26. For any function \( g \in C^2[a, b] \) that satisfies \( g''(a) = 0, g''(b) = 0 \), and \( g(x_i) = f(x_i) \) for each \( i = 1, 2, \ldots, N \), the natural cubic spline \( s = s(f; x) \) satisfies
\[
\int_a^b |s''(x)|^2 \, dx \leq \int_a^b |g''(x)|^2 \, dx, \tag{3.37}
\]
where the equality holds only when \( g(x) = s(f; x) \).

Proof. The proof is similar to that of Theorem 3.25. Although \( q'(a) = q'(b) = 0 \) does not hold, we do have \( s''(a) = s''(b) = 0 \).

Lemma 3.27. Suppose a \( C^2 \) 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
\[
\forall x \in [a, b], \quad |s''(x)| \leq 3 \max_{s \in [a, b]} |f''(x)|. \tag{3.38}
\]

Proof. Since \( s''(x) \) is linear on \([x_j, x_{j+1}]\), \( |s''(x)| \) attains its maximum at \( x_j \) for some \( j \). If \( j = 2, \ldots, N-1 \), it follows from Lemma 3.22 and Corollary 3.15 that
\[
2M_j = 6|f[x_j-1, x_j, x_{j+1}]| - \mu_jM_{j-1} - \lambda_jM_{j+1} \Rightarrow 2|M_j| \leq 6|f[x_j-1, x_j, x_{j+1}]| + (\mu_j + \lambda_j)|M_j| \Rightarrow \exists \xi \in [x_{j-1}, x_{j+1}] \text{ s.t. } |M_j| \leq 3|f''(\xi)|.
\]
\[
|s''(x)| \leq 3 \max_{x \in [a, b]} |f''(x)|. \tag{3.39}
\]
If \( |s''(x)| \) attains its maximum at \( x_1 \) or \( x_N \), (3.39) clearly holds for a cubic spline with specified second derivatives at these end points. Due to symmetry, it suffices to prove (3.39) 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 (3.30), we have
\[
2|M_1| \leq 6|f[x_1, x_1, x_2]| + |M_2| \leq 6|f[x_1, x_1, x_2]| + |M_1|
\]
which, together with Corollary 3.15, implies
\[
\exists \xi \in [x_1, x_2] \text{ s.t. } |M_1| \leq 3|f''(\xi)|.
\]
This completes the proof.

Remark 3.19. Lemma 3.27 cannot be generalized to individual intervals. In other words,
\[
\forall x \in [x_i, x_{i+1}], \quad |s''(x)| \leq 3 \max_{x \in [a, b]} |f''(x)|
\]
does not hold because the maximum amplitude of \( f'' \) may occur inside another interval.

Theorem 3.28. Suppose a \( C^4 \) 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
\[
\forall j = 0, 1, 2, \quad |f^{(j)}(x) - s^{(j)}(x)| \leq c_j h^{4-j} \max_{x \in [a, b]} |f^{(j)}(x)|, \tag{3.40}
\]
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. Consider an auxiliary function \( \hat{s} \in C^2[a, b] \) that satisfies
\[
\forall i = 1, 2, \ldots, N, \quad \hat{s}_{|x_i, x_{i+1}} \in \mathbb{P}_3, \quad \hat{s}''(x) = f''(x_i).
\]
We can obtain such an \( \hat{s} \) by interpolating \( f''(x) \) with some \( \hat{s} \in S_2^4 \) and integrating \( \hat{s} \) twice. Then Theorem 3.4 implies
\[
\exists \xi_i \in [x_i, x_{i+1}], \quad \forall x \in [x_i, x_{i+1}], \quad |f''(x) - \hat{s}''(x)| \leq \frac{1}{2} |f''(\xi_i)|(x-x_i)(x-x_{i+1}),
\]

hence we have
\[
|f''(x) - s''(x)| \leq \frac{1}{8} \max_{x \in [a, b]} |f^{(4)}(x)| (x_{i+1} - x_i)^2
\]
and
\[
|f''(x) - s''(x)| \leq \frac{1}{8} \max_{x \in [a, b]} |f^{(4)}(x)|. \tag{3.41}
\]

Now consider interpolating \( f(x) - \hat{s}(x) \) with a cubic spline. Since \( \hat{s}(x) \in S_2^4 \), the interpolant must be \( s(x) - \hat{s}(x) \). Then Lemma 3.27 yields
\[
\forall x \in [a, b], \quad |s''(x) - \hat{s}''(x)| \leq 3 \max_{x \in [a, b]} |f''(x) - \hat{s}''(x)|,
\]

which, together with (3.41), leads to (3.40) for \( j = 2 \):
\[
|f''(x) - s''(x)| \leq |f''(x) - \hat{s}''(x)| + |\hat{s}''(x) - s''(x)| \leq 4|f''(x) - \hat{s}''(x)| \leq \frac{1}{2} h^2 \max_{x \in [a, b]} |f^{(4)}(x)|. \tag{3.42}
\]

For \( j = 0 \), we have \( f(x) - s(x) = 0 \) for \( x = x_i, x_{i+1} \). Then Rolle’s theorem T0.32 implies \( f'((\xi_i)) - s'((\xi_i)) = 0 \) for some \( \xi_i \in [x_i, x_{i+1}] \). It follows from the fundamental theorem of calculus that
\[
\forall x \in [x_i, x_{i+1}], \quad |f'(x) - s'(x)| = \int_{\xi_i}^x (f''(t) - s''(t)) \, dt,
\]

which, together with the integral mean value theorem T0.54 and (3.42), yields
\[
|f'(x) - s'(x)| \leq \frac{1}{2} h^3 \max_{x \in [a, b]} |f^{(4)}(x)|.
\]

This proves (3.40) for \( j = 1 \). Finally, consider interpolating \( f(x) - s(x) \) with some linear spline \( \bar{s} \in S_0^2 \). The interpolation conditions dictate \( \forall x \in [a, b], \bar{s}(x) \equiv 0 \). Hence
\[
|f(x) - s(x)|_{x \in [x_i, x_{i+1}]} = |f(x) - s(x) - \bar{s}(x)|_{x \in [x_i, x_{i+1}]}
\]
\[
\leq \frac{1}{8} (x_{i+1} - x_i)^2 \max_{x \in [x_i, x_{i+1}]} |f''(x) - s''(x)|
\]
\[
\leq \frac{1}{16} h^4 \max_{x \in [a, b]} |f^{(4)}(x)|,
\]

where the second step follows from Theorem 3.4 and the third step from (3.42).
3.6 The Chebyshev polynomials

Remark 3.20. In estimating the remainder in Corollary 3.5, we cannot do much about \( \max |f^{(n+1)}(x)| \) as \( f \) is usually given \textit{a priori}. However, if we have the freedom to choose positions of the interpolation points, we can further reduce the remainder. More precisely, how can we choose \( x_0, x_1, \ldots, x_n \) to minimize the quantity

\[
\max_{x \in [a,b]} |(x-x_0)(x-x_1) \cdots (x-x_n)|?
\]

Such \( x_i \)'s are the zeros of the Chebyshev polynomial.

Definition 3.29. The \textit{Chebyshev polynomial} of degree \( n \) of the first kind is

\[ T_n(x) = \cos(n \arccos x). \]  

(3.43)

Remark 3.21. First, \( |T_n(x)| \leq 1 \). Second, (3.43) indeed defines a polynomial of degree \( n \). Set \( x = \cos \theta \). Then

\[
e^{i \theta} = \cos \theta + i \sin \theta \Rightarrow (e^{i \theta})^n = e^{in \theta} = \cos(n \theta) + i \sin(n \theta) = (x + i \sqrt{1-x^2})^n
\]

\[
= \cos n \theta = x^n + \left( \frac{n}{2} \right) x^{n-2}(x^2 - 1) + \cdots,
\]

where the first line is the Euler’s formula and the last line follows from the binomial theorem. The first five Chebyshev polynomials are

\[
\begin{align*}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 - 1
\end{align*}
\]

Aside from the above explicit formula, Chebyshev polynomials can also be generated by a recurrence relation.

Theorem 3.30. \( \forall n \in \mathbb{N}^+ \), \( T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \). 

(3.44)

Proof. By trigonometric identities, we have

\[
\begin{align*}
cos(n+1) \theta &= \cos n \theta \cos \theta - \sin n \theta \sin \theta, \\
cos(n-1) \theta &= \cos n \theta \cos \theta + \sin n \theta \sin \theta.
\end{align*}
\]

Adding up the two equations and setting \( \cos \theta = x \) complete the proof. \( \square \)

Corollary 3.31. The coefficient of \( x^n \) in \( T_n \) is \( 2^{n-1} \).

Proof. Use (3.44), \( T_0 = 1 \), and \( T_1 = x \) in an induction. \( \square \)

Theorem 3.32. \( T_n(x) \) has simple zeros at the \( n \) points

\[ x_k = \cos \left( \frac{2k-1}{2n} \pi \right), \quad (3.45) \]

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

\[ x_k' = \cos \left( \frac{k}{n} \pi \right), \quad k = 0, 1, \ldots, n, \quad (3.46) \]

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

Proof. (3.43) and (3.45) yield

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

Differentiate (3.43) 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, \( \forall k = 1, 2, \ldots, n-1 \),

\[
T'_n(x_k) = n \left( 1 - \cos^2 \left( \frac{k \pi}{n} \right) \right)^{-\frac{1}{2}} \sin(k \pi) = 0.
\]

For \( k = 0, n \), \( T'_n(x_k') \) attains its extreme values since \( T_n(x_n') = 1 \), \( T_n(x_0') = -1 \), and by (3.43) \( |T_n(x)| \leq 1 \). Clearly these are the only extrema of \( T_n(x) \) on \([-1,1]\). \( \square \)

Remark 3.22. According to Theorem 3.32, the zeros of Chebyshev polynomials are the projections onto the real line of equally spaced points on the unit circle; see Fig. 2.7 in NAG2012.

Theorem 3.33 (Chebyshev). Denote by \( \tilde{P}_n \) the class of all polynomials of degree \( n \) with leading coefficient 1. Then

\[
\forall p \in \tilde{P}_n, \quad \max_{x \in [-1,1]} \left| \frac{T_n(x)}{2^{n-1}} \right| \leq \max_{x \in [-1,1]} |p(x)|. \quad (3.47)
\]

Proof. By Theorem 3.32, \( T_n(x) \) assumes its extrema \( n+1 \) times at the points \( x_k' \) defined in (3.46). Suppose (3.47) does not hold. Then Corollary 3.31 implies that

\[
\exists p \in \tilde{P}_n \text{ s.t. } \max_{x \in [-1,1]} |p(x)| < \frac{1}{2^{n-1}}. \quad (3.48)
\]

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

\[
Q(x_k') = \left( \frac{1}{2} \right)^{n-k} - p(x_k'), \quad k = 0, 1, \ldots, n.
\]

By (3.48), \( Q(x) \) 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 (3.48). \( \square \)

Corollary 3.34.

\[
\max_{x \in [-1,1]} |x^n + a_1 x^{n-1} + \cdots + a_n| \geq \frac{1}{2^{n-1}}. \quad (3.49)
\]

Corollary 3.35. Suppose polynomial interpolation is performed for \( f \) on the \( n+1 \) zeros of \( T_{n+1}(x) \) as in Theorem 3.32. The Cauchy remainder in Theorem 3.4 satisfies

\[
|R_n(f; x)| \leq \frac{1}{2^n (n+1)!} \max_{x \in [-1,1]} \left| f^{(n+1)}(x) \right|. \quad (3.50)
\]

Proof. Theorem 3.4 and Corollary 3.31 yield

\[
|R_n(f; x)| = \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} \sum_{k=0}^{n} (x-x_k) \right| = \left| \frac{f^{(n+1)}(\xi)}{2^n (n+1)!} |T_{n+1}| \right|.
\]

Definition 3.29 completes the proof as \( |T_{n+1}| \leq 1 \). \( \square \)
### 3.7 B-Splines

**Remark 3.23.** Splines in piecewise-polynomial (pp) form has one drawback: the derivation has to be repeated for different degrees of splines. In this subsection, we approach splines from the viewpoint of a linear space so that different degrees of splines can be derived by simple recurrence relations. The letter “B” in B-splines stands for basis of the vector space of splines.

**Notation 3.** In the notation $S_n^{-1}(t_1, t_2, \cdots, t_N)$, $t_i$’s in the parentheses represent knots of a spline. When there is no danger of ambiguity, we also use the shorthand notation $S_{n,N}^{-1} := S_n^{-1}(t_1, t_2, \cdots, t_N)$ or simply $S_n^{-1}$.

**Theorem 3.36.** The set of splines $S_n^{-1}(t_1, t_2, \cdots, t_N)$ is a linear space with dimension $n + N - 1$.

**Proof.** It is easy to verify from Definition 0.56 that $S_n^{-1}(t_1, t_2, \cdots, t_N)$ is indeed a linear space. Note that the additive identity is the zero function not the zero number. 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, \ldots, $(n - 1)$th derivatives of adjacent polynomials match. Hence the dimension is $(N - 1)(n + 1) - n(N - 2) = n + N - 1$. 

**Definition 3.37.** The truncated power function with exponent $n$ is defined as

$$x_n^+ = \begin{cases} x^n & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

**Example 3.24.** According to Definition 3.37, we have

$$\forall t \in [a, b], \int_a^t (t - x)^n dx = \int_a^t (t - x)^{n+1} dx = \frac{(t-a)^{n+1}}{n + 1}. \quad (3.52)$$

**Remark 3.25.** For $n \in \mathbb{N}^+$, $x_n^+ \in C^{n-1}(-\infty, +\infty)$. Hence $S_n^{-1}$ is $C^{n-1}(-\infty, 0, +\infty)$.

**Lemma 3.38.** The following is a basis of $S_n^{-1}(t_1, \ldots, t_N)$,

$$1, x, x^2, \ldots, x^n, (x - t_2)^n_+, (x - t_3)^n_+, \ldots, (x - t_{N-1})^n_+, \quad (3.53)$$

**Proof.** $\forall i = 2, 3, \ldots, N - 1$, $(x - t_i)^n_+ \in S_n^{-1}(t_1, t_2, \ldots, t_N)$. Also, $\forall i = 0, 1, \ldots, n$, $x^i \in S_n^{-1}(t_1, t_2, \ldots, t_N)$. Suppose

$$\sum_{i=0}^{n} a_i x^i + \sum_{j=2}^{N-1} a_{n+j}(x - t_j)^n_+ = 0(x). \quad (3.54)$$

To satisfy (3.54) for all $x < t_2$, $a_i$ must be 0 for each $i = 0, 1, \ldots, n$. To satisfy (3.54) for all $x \in (t_2, t_3)$, $a_{n+2}$ must be 0. Similarly, all $a_{n+j}$’s must be zero. Hence, the functions in (3.53) are linearly independent by Definition 0.62. The proof is completed by Theorem 3.36, Proposition 0.70, and the fact that there are $n + N - 1$ functions in (3.53).

**Corollary 3.39.** Any $s \in S_{n,N}^{-1}$ can be expressed as

$$s(x) = \sum_{i=0}^{n} a_i(x - t_i)^i + \sum_{j=2}^{N-1} a_{n+j}(x - t_j)^n_+, \quad x \in [t_1, t_N]. \quad (3.55)$$

**Proof.** By Lemma 3.38, it suffices to point out that span$\{1, x, \ldots, x^n\} = \text{span}\{1, (x - t_1), \ldots, (x - t_1)^n\}$. 

**Example 3.26.** (3.55) with $n = 1$ is the linear spline interpolation. Imagine an iron rod that is initially straight. Place one of its end at $(t_1, f_1)$ and let it go through $(t_2, f_2)$. In general $(t_3, f_3)$ will be off the rod, but we can bend the rod at $(t_2, f_2)$ to make the rod go through $(t_3, f_3)$. This “bending” process corresponds to adding the first truncated power function in (3.55).

**Remark 3.27.** (An informative proof of Corollary 3.39). Denote $p_i(x) = s(x)|_{x=[t_1, t_i]}$. Clearly $p_{N-1}(x) = s(x)$.

$$p_1(x) = a_0 + a_1(x - t_1) + \cdots + a_n(x - t_1)^n.$$

For $x \in [t_2, t_3]$, $p_2(x)$ only differs from the extension of $p_1(x)$ for the $n$th derivative, hence it can be expressed as

$$p_2(x) = p_1(x) + a_{n+1}(x - t_2)^n_+.$$

Repeating the argument for $i = 3, 4, \ldots, N - 1$ completes the proof.

**Remark 3.28.** The basis in (3.53) has several drawbacks. First, an coefficient $a_{n+i}$ in (3.55) depends on all of $a_{n+2}, a_{n+3}, \ldots, a_{n+i-1}$. Thus errors in determining $a_{n+2}, a_{n+3}, \ldots, a_{n+i-1}$ propagate into the calculation of $a_{n+i}$, which makes the whole process ill-conditioned. Second, inserting a new knot would change all coefficients of those knots on the right of the new knot. A much better basis of $S_{n,N}^1$ is the hat functions. See Fig. 2.11 at p.105 of NAG2012.

**Definition 3.40.** The hat function at $t_i$ is

$$\hat{B}_i(x) = \begin{cases} \frac{x-t_i}{t_i-t_{i-1}} & x \in (t_{i-1}, t_i], \\ \frac{t_{i+1} - x}{t_{i+1} - t_i} & x \in (t_i, t_{i+1}], \\ 0 & \text{otherwise}. \end{cases} \quad (3.56)$$

**Theorem 3.41.** The hat functions form a basis of $S_{n}^1$.

**Proof.** By Definition 3.40, we have

$$\hat{B}_i(t_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \quad (3.57)$$

Suppose $\sum_{i=1}^{N} c_i \hat{B}_i(x) = 0(x)$. Then we have $c_i = 0$ for each $i = 1, 2, \ldots, N$ by setting $x = t_i$ and applying (3.57). Hence by Definition 0.62 the hat functions are linearly independent. It suffices to show that $\text{span}\{\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_N\} = S_{n}^1$, which is true because

$$\forall s(x) \in S_{n}^1, \exists s_B(x) = \sum_{i=1}^{N} s(t_i) \hat{B}_i(x) \quad s(x) = s_B(x).$$

On each interval $[t_i, t_{i+1}]$, (3.57) 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. Thus Definition 0.65 completes the proof. 

---

Qinghai Zhang Lectures on Numerical Analysis Fall 2016
Remark 3.29. The hat functions are analogous to the elementary Lagrange interpolation polynomials. How do we pass from the first degree to higher degrees? This question leads to the more general B-splines.

Definition 3.42. B-splines are defined recursively by
\[ B_{i}^{n+1}(x) = \frac{x-t_{i-1}}{t_{i+n}-t_{i-1}} B_{i}^{n}(x) + \frac{t_{i+n+1}-x}{t_{i+n+1}-t_{i}} B_{i+1}^{n}(x). \] (3.58)

The base of the recursion is the B-spline of degree zero,
\[ B_{i}^{0}(x) = \begin{cases} 1 & \text{if } x \in (t_{i-1}, t_{i}], \\ 0 & \text{otherwise}. \end{cases} \] (3.59)

Example 3.30. The hat functions in Definition 3.40 are clearly the B-splines of degree one:
\[ B_{1}^{1} = B_{1}. \] (3.60)

In (3.58), B-splines of higher degrees are defined by generalizing the idea of hat functions.

Example 3.31. The quadratic B-splines \( B_{i}^{2}(x) \) are given by
\[
\begin{cases}
(x-t_{i-1})^2 \quad & x \in (t_{i-1}, t_{i}); \\
(x-t_{i-1})(t_{i+n}-t_{i}) & x \in (t_{i}, t_{i+1}); \\
(x-t_{i-1})(t_{i+n}-t_{i}) + (t_{i+n+1}-x)(x-t_{i+1}) & x \in (t_{i+1}, t_{i+2}); \\
(t_{i+n+1}-t_{i+1}) & x \in (t_{i+2}, t_{i+n-1}); \\
0, & \text{otherwise}.
\end{cases}
\] (3.61)

Definition 3.43. The support of a function \( f : X \to \mathbb{R} \) is
\[ \text{supp}(f) = \text{closure}(\{x \in X \mid f(x) \neq 0\}). \] (3.62)

Lemma 3.44. 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 (3.59) and (3.58). \( \square \)

Remark 3.32. Lemma 3.44 is helpful in memorizing the recursive definition (3.58). The knots in the fractional coefficient of \( B_{i}^{n}(x) \) are exactly the end points of its interval of support; the same is true for \( B_{i}^{n+1}(x) \).

Remark 3.33. As one advantage of B-splines over truncated power functions, they have local support. When \( n \) is increased to \( n+1 \), the interval of support covers one more knot to its right while the left endpoint \( t_{i-1} \) remains fixed. Another common convention is to choose \( t_{i} \) as the fixed left end.

Definition 3.45. 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
\[ L(\alpha x + \beta y) = \alpha L(x) + \beta L(y), \] (3.63)
then \( L \) is called a linear functional over \( X \).

Example 3.34. \( X = C[a, b] \) then the elements of \( X \) are functions continuous over \([a, b]\). \( L(f) = \int_{a}^{b} f(x) dx \) and \( L(f) = \int_{a}^{b} x^2 f(x) dx \) are both linear functionals over \( X \).

Notation 4. We have used the notation \( f[x_{0}, \ldots, x_{k}] \) for the \( k \)th divided difference of \( f \), inline with considering \( f[x_{0}, \ldots, x_{k}] \) as a generalization of the Taylor expansion. Hereafter, for analysis of 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 \( C[x_{0}, x_{k}] \).

Theorem 3.46 (Leibniz formula). For \( k \in \mathbb{N} \), the \( k \)th divided difference of a product of two functions satisfies
\[ [x_{0}, \ldots, x_{k}]fg = \sum_{i=0}^{k} [x_{0}, \ldots, x_{i}]f \cdot [x_{i+1}, \ldots, x_{k}]g. \] (3.64)

Proof. The induction basis \( k = 0 \) holds because (3.64) reduces to \( x_{0}fg = f(x_{0})g(x_{0}) \). Now suppose (3.64) holds. For the induction step, we have from Theorem 3.12 that
\[ [x_{0}, \ldots, x_{k+1}]fg = \frac{[x_{1}, \ldots, x_{k+1}]fg - [x_{0}, \ldots, x_{k}]fg}{x_{k+1} - x_{0}}. \]
By the induction hypothesis, we have
\[ [x_{1}, \ldots, x_{k+1}]fg = \sum_{i=0}^{k} [x_{1}, \ldots, x_{i+1}]f \cdot [x_{i+1}, \ldots, x_{k+1}]g. \]

Remark 3.35. For memorization, it is helpful to compare Theorem 3.46 to the Leibniz rule for differentiation
\[ (fg)^{(n)} = \sum_{i=0}^{n} \binom{n}{i} f^{(n-i)}g^{(i)}. \] (3.65)
Remark 3.36. There exists a relation between B-splines and truncated power functions, e.g.,

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

The algebra is illustrated by the figures below,

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

Theorem 3.47. For any \( n \in \mathbb{N} \), we have

\[ B^n_i(x) = (t_{i+n} - t_{i-1}) \cdot [t_{i-1}, \ldots, t_{i+n}](t - x)^n_+ \quad (3.66) \]

Proof. For \( n = 0 \), (3.66) reduces to

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

which is the same as (3.59). Hence the induction basis holds. Now assume the induction hypothesis (3.66) hold.

By Definition 3.37, \((t - x)^{n+1}_+ = (t - x)(t - x)^n_+ \). Then Theorem 3.46 yields

\[ [t_{i-1}, \ldots, t_{i+n}](t - x)^{n+1}_+ = (t_{i-1} - x) \cdot [t_{i-1}, \ldots, t_{i+n}](t - x)^n_+ + [t_{i-1}, \ldots, t_{i+n}](t - x)^n_+ \quad (3.67) \]

where the last step follows from (3.67). Similarly,

\[ \gamma(x) = \frac{t_{i+n+1} - x}{t_{i+n+1} - t_i} B^n_{i+1}(x) = (t_{i+n+1} - x) \cdot [t_{i}, \ldots, t_{i+n}] \cdot (t - x)^n_+ \\
= (t_{i+n+1} - t_i) \cdot [t_{i}, \ldots, t_{i+n}](t - x)^n_+ + (t_i - x) \cdot [t_{i}, \ldots, t_{i+n}](t - x)^n_+ \\
= [t_{i+1}, \ldots, t_{i+n+1}](t - x)^n_+ + [t_{i}, \ldots, t_{i+n+1}](t - x)^n_+ \\
- [t_{i+1}, \ldots, t_{i+n+1}](t - x)^n_+ - [t_{i}, \ldots, t_{i+n+1}](t - x)^n_+ = [t_{i}, \ldots, t_{i+n+1}](t - x)^n_+ - [t_{i}, \ldots, t_{i+n+1}](t - x)^n_+ \]

which completes the inductive proof.

Remark 3.37. Some authors define B-splines by (3.66) and then deduce (3.58) from it.

Corollary 3.48. The average of a B-spline over its support only depends on its degree,

\[ \frac{1}{t_{i+n} - t_{i-1}} \int_{t_{i-1}}^{t_{i+n}} B^n_i(x) \, dx = \frac{1}{n + 1} \quad (3.68) \]

Proof. The left-hand side (LHS) of (3.68) is

\[ \frac{1}{t_{i+n} - t_{i-1}} \int_{t_{i-1}}^{t_{i+n}} B^n_i(x) \, dx = \int_{t_{i-1}}^{t_{i+n}} \frac{1}{t_{i+n} - t_{i-1}} \int_{t_{i-1}}^{t_{i+n}} B^n_i(x) \, dx. \]

where the first step follows from Theorem 3.47, the second step from the commutativity of integration and taking divided difference, the third step from (3.52), and the last step from Corollary 3.15.

Theorem 3.49. For \( n \geq 2 \), we have, \( \forall x \in \mathbb{R} \),

\[ \frac{d}{dx} B^n_i(x) = \frac{n B^n_{i-1}(x)}{t_{i+n} - t_i - t_{i-1}} - \frac{n B^n_{i+1}(x)}{t_{i+n} - t_{i+1}} \quad (3.69) \]

For \( n = 1 \), (3.69) holds for all \( x \) except at the three knots \( t_{i-1}, t_i, t_{i+1} \), where the derivative of \( B^n_i \) is not defined.

Proof. We first show that (3.69) holds for all \( x \) except at the knots \( t_j \). By (3.60), (3.56), and (3.59), we have

\[ \forall x \in \mathbb{R} \setminus \{t_{i-1}, t_i, t_{i+1}\}, \frac{d}{dx} B^n_i(x) = \frac{1}{t_{i+1} - t_i} B^n_i(x) - \frac{1}{t_{i+1} - t_i} B^n_{i+1}(x). \]
Hence the induction hypothesis holds. Now suppose (3.69) holds \( \forall x \in \mathbb{R} \setminus \{t_{i-1}, \ldots, t_{i+n} \} \). Differentiate (3.58), apply the induction hypothesis (3.69), and we have
\[
\frac{d}{dx} B_{n+1}^i(x) = \frac{B_i^0(x)}{t_{i+n} - t_{i-1}} - \frac{B_i^n(x)}{t_{i+n+1} - t_i} + nC(x),
\] (3.70)
where \( C(x) \) is
\[
\frac{x - t_{i-1}}{t_{i+n} - t_{i-1}} \left[ \frac{B_{n-1}^i(x)}{t_{i+n-1} - t_{i-1}} - \frac{B_{n}^i(x)}{t_{i+n} - t_i} \right] + \frac{t_{i+n+1} - x}{t_{i+n+1} - t_i} \left[ \frac{B_{n}^i(x)}{t_{i+n+1} - t_i} - \frac{B_{n+1}^i(x)}{t_{i+n+1} - t_{i+1}} \right] = \frac{1}{t_{i+n} - t_{i-1}} \left[ \frac{(x - t_{i-1})B_{n-1}^i(x)}{t_{i+n-1} - t_{i-1}} + (t_{i+n} - x)B_{n+1}^i(x) \right] + \frac{1}{t_{i+n+1} - t_i} \left[ \frac{(x - t_i)B_{n}^i(x)}{t_{i+n+1} - t_i} - (t_{i+n+1} - x)B_{n+1}^i(x) \right]
\]
\[
= \frac{B_i^n(x)}{t_{i+n} - t_{i-1}} - \frac{B_i^{n+1}(x)}{t_{i+n+1} - t_i}.
\]
Then (3.70) can be written as
\[
\frac{d}{dx} B_{n+1}^i(x) = \frac{(n+1)B_i^0(x)}{t_{i+n} - t_{i-1}} - \frac{(n+1)B_i^n(x)}{t_{i+n+1} - t_i},
\]
which completes the inductive proof of (3.69) except at the knots. Since \( B_i^1 = B_i \) is continuous, an easy induction with (3.58) shows that \( B_i^n \) is continuous for all \( n \geq 1 \). Hence the right-hand side of (3.69) is continuous for all \( n \geq 2 \). Therefore, if \( n \geq 2 \), \( \frac{d}{dx} B_i^n(x) \) exists for all \( x \in \mathbb{R} \). This completes the proof.

**Corollary 3.50.** \( B_i^n \in S_n^{-1} \).

**Proof.** For \( n = 1 \), the induction basis \( B_i^1(x) \in S_i^0 \) holds because of (3.60). The rest of the proof follows from (3.58) and Theorem 3.49 via an easy induction.

**Theorem 3.51** (Marsden’s identity). For any \( n \in \mathbb{N} \),
\[
(t - x)^n = \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n-1}) B_i^n(x),
\] (3.71)
where the product \((t - t_i) \cdots (t - t_{i+n-1})\) is defined as \(1\) for \( n = 0 \).

**Proof.** For \( n = 0 \), (3.71) follows from Definition 3.42. Now suppose (3.71) holds. A linear interpolation of the linear function \( f(t) = t - x \) is the function itself,
\[
t - x = \frac{t - t_{i+n}}{t_{i-1} - t_{i+n}} (t_{i-1} - x) + \frac{t - t_{i-1}}{t_{i+n} - t_{i-1}} (t_{i+n} - x).
\] (3.72)
Hence for the inductive step we have
\[
(t - x)^{n+1} = (t - x) \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n-1}) B_i^n(x)
\]
\[
= \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) \frac{t_{i+n} - x}{t_{i-1} - t_{i+n}} B_i^n(x)
\]
\[
+ \sum_{i=-\infty}^{+\infty} (t - t_{i-1}) \cdots (t - t_{i+n-1}) \frac{t_{i+n} - x}{t_{i-1} - t_{i+n-1}} B_i^n(x)
\]
\[
= \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) \frac{t_{i+n} - x}{t_{i-1} - t_{i+n}} B_i^n(x)
\]
\[
+ \sum_{i=-\infty}^{+\infty} (t - t_{i-1}) \cdots (t - t_{i+n-1}) \frac{t_{i+n} - x}{t_{i-1} - t_{i+n-1}} B_i^n(x)
\]
\[
= \sum_{i=-\infty}^{+\infty} (t - t_i) \cdots (t - t_{i+n}) B_i^{n+1}(x),
\]
where the first step follows from the induction hypothesis, the second step from (3.72), the third step from replacing \( i \) with \( i + 1 \) in the second summation, and the last step from (3.58).

**Corollary 3.52.** For any \( j \in \mathbb{Z} \) and \( n \in \mathbb{N} \),
\[
(t_j - x)^n = \sum_{i=-\infty}^{+\infty} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x).
\] (3.73)

**Proof.** We need to show that the RHS is \((t_j - x)^n\) if \( x \leq t_j \) and 0 otherwise. Set \( t = t_j \) in (3.71) and we have
\[
\sum_{i=-\infty}^{+\infty} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x) = (t_j - x)^n.
\]
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 3.44 implies that \( B_i^n(x) = 0 \) for all \( x \leq t_j \). Hence
\[
x \leq t_j \Rightarrow \sum_{i=-\infty}^{j-1} (t_j - t_i) \cdots (t_j - t_{i+n-1}) B_i^n(x) = (t_j - x)^n.
\]
Otherwise \( x > t_j \), then Lemma 3.44 implies \( B_i^n(x) = 0 \) for each \( i \leq j - n \). This completes the proof.

**Definition 3.53.** The elementary symmetric polynomial of degree \( k \) in \( n \) variables is the sum of all products of \( k \) distinct variables chosen from the \( n \) variables,
\[
\sigma_k(x_1, \ldots, x_n) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} x_{i_1}x_{i_2} \cdots x_{i_k}.
\] (3.74)
In particular, \( \sigma_0(x_1, \ldots, x_n) = 1 \) and \( \forall k > n, \quad \sigma_k(x_1, \ldots, x_n) = 0 \).

If the distinctiveness condition is dropped, we have the complete symmetric polynomial of degree \( k \) in \( n \) variables,
\[
\tau_k(x_1, \ldots, x_n) = \sum_{1 \leq i_1 \leq \cdots \leq i_k \leq n} x_{i_1}x_{i_2} \cdots x_{i_k}.
\] (3.75)
Example 3.38. $\sigma_2(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 + x_2 x_3$. In comparison, $\tau_2(x_1, x_2, x_3) = \sigma(x_1, x_2, x_3) + x_1^2 + x_2^2 + x_3^2$.

Lemma 3.54. For $k \leq n$, the elementary symmetric polynomials satisfy a recursion,
\[
\sigma_{k+1}(x_1, \ldots, x_n, x_{n+1}) = \sigma_{k+1}(x_1, \ldots, x_n) + x_{n+1} \sigma_k(x_1, \ldots, x_n).
\] (3.76)

Proof. The terms in $\sigma_{k+1}(x_1, \ldots, x_n, x_{n+1})$ can be assorted into two groups: (a) those that contain the factor $x_{n+1}$ and (b) those that do not. By the symmetry in (3.75), group (a) must be $\sigma_{k+1}(x_1, \ldots, x_n)$ and group (b) must be $x_{n+1} \sigma_k(x_1, \ldots, x_n)$.

Example 3.39. $\sigma_2(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 + x_2 x_3$.

Definition 3.55. The generating function for the elementary symmetric polynomials is
\[
g_{\sigma, n}(z) = \prod_{i=1}^{n} (1 + x_i z) = (1 + x_1 z) \cdots (1 + x_n z). \] (3.77)

while that for the complete symmetric polynomials is
\[
g_{\tau, n}(z) = \prod_{i=1}^{n} \frac{1}{1 - x_i z} = \frac{1}{1 - x_1 z} \cdots \frac{1}{1 - x_n z}. \] (3.78)

Lemma 3.56. The elementary and complete symmetric polynomials are related to their generating functions as
\[
g_{\sigma, n}(z) = \sum_{k=0}^{n} \sigma_k(x_1, \ldots, x_n) z^k. \] (3.79)
\[
g_{\tau, n}(z) = \sum_{k=0}^{+\infty} \tau_k(x_1, \ldots, x_n) z^k. \] (3.80)

Proof. With Lemma 3.54, we prove (3.79) by induction. For (3.80), the geometric series $1/(1-x) = \sum_{k=0}^{+\infty} x^k$ and (3.78) yield
\[
g_{\tau, n}(z) = \prod_{i=1}^{n} \sum_{k=0}^{+\infty} x_i^k z^k
= (1 + x_1 z + x_1^2 z^2 + \cdots)(1 + x_2 z + x_2^2 z^2 + \cdots)
\cdots (1 + x_n z + x_n^2 z^2 + \cdots).
\]
The coefficient of the monomial $z^k$, is the sum of all possible products of $k$ variables from $x_1, x_2, \ldots, x_n$. Definition 3.53 then completes the proof.

Example 3.40.
\[
(1 + x_1 z)(1 + x_2 z)(1 + x_3 z)
= 1 + (x_1 + x_2 + x_3) z
+ (x_1 x_2 + x_1 x_3 + x_2 x_3) z^2 + x_1 x_2 x_3 z^3.
\]

Lemma 3.57. For $k \leq n$, the complete symmetric polynomials satisfy a recursion,
\[
\tau_{k+1}(x_1, \ldots, x_n, x_{n+1})
= \tau_{k+1}(x_1, \ldots, x_n) + x_{n+1} \tau_k(x_1, \ldots, x_n, x_{n+1}). \] (3.81)

Proof. (3.78) implies that
\[
g_{\tau, n+1} - g_{\tau, n} = x_n z g_{\tau, n+1}.
\] (3.82)
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.

Remark 3.41. We can deduce Lemma 3.54 from Lemma 3.57 by removing each term that contains a factor of repeated $x_i$.

Theorem 3.58. The complete symmetric polynomial of degree $n - m$ in $n + 1$ variables is the $n$th divided difference of the monomial $z^m$,
\[
\forall m \in \mathbb{N}^+, i \in \mathbb{N}, \quad \tau_{m-n}(x_1, \ldots, x_{i+n}) = [x_1, \ldots, x_{i+n}] z^m.
\] (3.83)

Theorem 3.59. 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
\[
\left( \frac{n}{k} \right) x^k = \sum_{i=-\infty}^{+\infty} \sigma_k(t_i, \ldots, t_{i+n-1}) B_i^n(x). \] (3.84)

Proof. Lemma 3.56 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
\[
(t - t_i) \cdots (t - t_{i+n-1}) = \sum_{k=0}^{n} \sigma_k(t_i, \ldots, t_{i+n-1}) (-1)^k t^n.
\]

Substituting the above into (3.71) yields
\[
(t - x)^n = \sum_{i=-\infty}^{+\infty} \sigma_k(t_i, \ldots, t_{i+n-1}) (-1)^k t^n B_i^n(x)
= \sum_{k=0}^{n} \left\{ t^{n-k} (-1)^k \sum_{i=-\infty}^{+\infty} \sigma_k(t_i, \ldots, t_{i+n-1}) B_i^n(x) \right\}.
\]

On the other hand, the binomial theorem states that
\[
(t - x)^n = \sum_{k=0}^{n} \left( \frac{n}{k} \right) (-1)^k x^k
= \sum_{k=0}^{n} (-1)^k \left( \frac{n}{k} \right) x^k.
\]

Comparing the last two equations completes the proof.

Corollary 3.60 (Partition of Unity).
\[
\forall n \in \mathbb{N}, \quad \sum_{i=-\infty}^{+\infty} B_i^n = 1. \] (3.85)

Proof. Setting $k = 0$ in Theorem 3.59 yields (3.85).
Theorem 3.61. The following list of B-splines is a basis of \( S_n^{-1}(t_1, t_2, \ldots, t_N) \),
\[
B_{2-n}(x), B_{3-n}(x), \ldots, B_N(x).
\] (3.86)

Proof. It is easy to verify that
\[
\forall t_i \in \mathbb{R}, \ (x - t_i)^n_+ = (-1)^n \{(x - t_i)^n - (t_i - x)^n_+\}.
\] (3.87)

Then it follows from Theorem 3.51 and Corollary 3.52 that each truncated power function \((x - t_i)_+^n\) can be expressed as a linear combination of B-splines. By Lemma 3.38, each element in \( S_n^{-1}(t_1, t_2, \ldots, t_N) \) can be expressed as a linear combination of
\[1, x, x^2, \ldots, x^n, (x - t_2)_+^n, (x - t_3)_+^n, \ldots, (x - t_{N-1})_+^n.\]

Theorem 3.59 states that each monomial \( x^3 \) can also be expressed as a linear combination of B-splines. Since the domain is restricted to \([t_i, t_{j}]\), we know from Lemma 3.44 that only those B-splines in the list of (3.86) appear in the linear combination. Therefore, these B-splines form a spanning list of \( S_n^{-1}(t_1, t_2, \ldots, t_N) \). The proof is completed by Proposition 0.69, Theorem 3.36, and the fact that the length of the list (3.86) is also \( n + N - 1 \).

Remark 3.42. Now that B-splines form a basis of \( S_n^{-1} \), they can be employed to solve interpolation problems. We illustrate the process by a special type of B-splines: those with uniformly spaced knots. Without loss of generality, we focus on the cardinal splines in Definition 3.67, where the uniform spacing is one. However, before we analyze these splines, we need to know more about forward difference and backward difference.

Definition 3.62. For \( n \in \mathbb{N}^+ \), the \( n \)th forward difference associated with a sequence of values \( \{f_0, f_1, \ldots\} \) is
\[
\Delta f_i = f_{i+1} - f_i, \\
\Delta^{n+1} f_i = \Delta^n f_i = \Delta f_{i+1} - \Delta f_i, \quad (3.88)
\]
and the \( n \)th backward difference is
\[
\nabla f_i = f_i - f_{i-1}, \\
\nabla^{n+1} f_i = \nabla^n f_i = \nabla f_i - \nabla f_{i-1}. \quad (3.89)
\]

Theorem 3.63. The forward difference and backward difference are related as
\[
\forall n \in \mathbb{N}^+, \quad \Delta^n f_i = \nabla^n f_{i+n}. \quad (3.90)
\]

Proof. An easy induction.

Remark 3.43. In light of Theorem 3.63, hereafter we only study forward differences since similar conclusions on backward differences can be deduced by Theorem 3.63.

Theorem 3.64. The forward difference can be expressed explicitly as
\[
\Delta^n f_i = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k}. \quad (3.91)
\]

Proof. For \( n = 1 \), (3.91) reduces to \( \Delta f_i = f_{i+1} - f_i \). The rest of the proof is an induction utilizing the identity
\[
\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}. \quad (3.92)
\]

Suppose (3.91) holds. For the inductive step, we have
\[
\Delta^{n+1} f_i = \Delta \Delta^n f_i = \Delta \left( \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k} \right) \\
= \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k+1} - \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_{i+k} \\
= \sum_{k=1}^{n} (-1)^{n-1-k} \binom{n}{k-1} f_{i+k} + f_{i+n+1} \\
+ \sum_{k=1}^{n} (-1)^{n-1-k} \binom{n}{k} f_{i+k} + (-1)^{n+1} f_i \\
= \sum_{k=0}^{n-1} (-1)^{n-1-k} \binom{n+1}{k} f_{i+k},
\]
which completes the induction proof.

Theorem 3.65. Suppose a sequence of values \( f_i \) is obtained by evaluating \( f_i = f(x_i) \) at points \( x_i \) with uniform spacing \( h \). Then
\[
\Delta^n f_0 = n! h^n [x_0, x_1, \ldots, x_n] f. \quad (3.92)
\]

Proof. See the solution of homework 6 for an inductive proof. A more informative proof is provided here. For the Newton polynomial \( v(x) = \prod_{i=0}^{n} (x - x_i) \) defined in (3.8), we have \( v'(x_k) = \prod_{i=0, i \neq k}^{n} (x_k - x_i) \). It follows from \( x_k - x_i = (k - i)h \) that
\[
v'(x_k) = \prod_{i=0, i \neq k}^{n} (k - i)h = h^n k!(n - k)!(1)^{n-k}. \quad (3.93)
\]

Then we have
\[
[x_0, x_1, \ldots, x_n] f = \sum_{k=0}^{n} \frac{f_k}{h^n} = \sum_{k=0}^{n} (-1)^{n-k} \frac{f_k}{h^n (n - k)!} \\
= \frac{1}{h^n} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_k = \Delta^n f_0 \frac{1}{h^n n!},
\]
where the first step follows from Corollary 3.11, the second from (3.93), and the last from Theorem 3.64.

Theorem 3.66 (Newton’s forward difference formula). Suppose \( p_n(f; x) \in P_n \) interpolates \( f(x) \) at \( x_0, x_1, \ldots, x_n \) with \( x_i = a + ih \). Then
\[
p_n(f; x) = f(a) + \frac{\Delta f_0}{h} (x - a) + \frac{\Delta^2 f_0}{2h^2} (x - a)(x - a - h) \\
+ \cdots + \frac{\Delta^n f_0}{n! h^n} \prod_{i=0}^{n-1} (x - a - ih). \quad (3.94)
\]

Proof. In Theorem 3.14, we set \( f(x) = p_n(f; x) \) on the left-hand side of (3.16), apply Theorem 3.66, and note that fact that any \( (n + 1) \)th divided difference applied to a degree \( n \) polynomial is zero.
Definition 3.67. The cardinal B-spline of degree $n$, denoted by $B^n_{i,Z}$, is the B-spline in Definition 3.42 on the knot set $Z$.

Corollary 3.68. Cardinal B-splines of the same degree are translates of one another, i.e.

$$\forall x \in \mathbb{R}, \quad B^n_{i,Z}(x) = B^n_{i+1,Z}(x+1). \quad (3.95)$$

Proof. The recurrence relation (3.58) reduces to

$$B^n_{i,Z}(x) = \frac{x-i+1}{n+1} B^n_{i,Z}(x) + \frac{i+n-1-x}{n+1} B^n_{i+1,Z}(x).$$

(3.96)

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

Corollary 3.69. A cardinal B-spline is symmetric about the center of its interval of support, i.e.

$$\forall x \in \mathbb{R}, \quad B^n_{i,Z}(x) = B^n_{j,Z}(2i+n-1-x). \quad (3.97)$$

Proof. The proof is similar with that of Corollary 3.68.

Example 3.44. For $t_i = i$, the quadratic B-spline in Example 3.31 simplifies to

$$B^2_{i,Z}(x) = \begin{cases} 
\frac{(x-i+1)^2}{6}, & x \in (i-1,i]; \\
\frac{3}{2} - \frac{1}{2}(x-(i+\frac{1}{2}))^2, & x \in (i,i+1]; \\
\frac{(i+2-x)^2}{2}, & x \in (i+1,i+2]; \\
n_0, & \text{otherwise.}
\end{cases}$$

(3.98)

It is straightforward to verify Corollaries 3.68 and 3.69. It also follows from (3.98) that

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

(3.99)

Example 3.45. For $t_i = i$, the cubic cardinal B-spline is

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

(3.100)

It follows that

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

(3.101)

Theorem 3.70. The cardinal B-spline of degree $n$ can be explicitly expressed as

$$B^n_{i,Z}(x) = \frac{1}{n!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n+1}{k} (k+i-x)^n. \quad (3.102)$$

Proof. Theorems 3.47, 3.65, and 3.64 yield

$$B^n_{i,Z}(x) = (n+1)[i-1,\ldots,i+n][-x]_n = \frac{n+1}{n!} \Delta^{n+1}(i-1-x)_n = \frac{1}{n!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n+1}{k} (i-1+k-x)_n.$$ 

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

Remark 3.46. Theorem 3.70 generalizes Examples 3.44 and 3.45. It also helps plotting of cardinal B-splines.

Corollary 3.71. The value of a cardinal B-spline at an integer $j$ is

$$B^n_{i,Z}(j) = \frac{1}{n!} \sum_{k=j-i+1}^{n} (-1)^{n-k} \binom{n+1}{k} (k+i-j)^n \quad (3.103)$$

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

Proof. This follows directly from Theorem 3.70 and Definition 3.37.

Corollary 3.72. There is a unique B-spline $S(x) \in \mathbb{S}_3^2$ that interpolates $f(x)$ at $1,2,\ldots,N$ with $S'(1) = f'(1)$ and $S'(N) = f'(N)$. Furthermore, this B-spline is

$$S(x) = \sum_{i=-1}^{N} a_i B^n_{i,Z}(x), \quad (3.104)$$

where

$$a_{-1} = a_1 - 2f'(1), \quad a_N = a_{N-2} + 2f'(N), \quad (3.105)$$

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

$$\mathbf{b}^T = [6f(1) + 2f'(1), 6f(2), \ldots, 6f(N-1), 6f(N) - 2f'(N)],$$

$$M = \begin{bmatrix} 
4 & 2 & \cdots & \cdots \\
1 & 4 & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
\cdots & \cdots & \cdots & 1 \\
\cdots & \cdots & \cdots & 2 \\
\cdots & \cdots & \cdots & 4 
\end{bmatrix}.$$ 

Proof. By Theorems 3.61 and 3.44, we have

$$\forall i = 1,2,\ldots,N, \quad f(i) = a_{i-2} B^2_{i-2,Z}(i) + a_{i-1} B^2_{i-1,Z}(i) + a_i B^2_{i,Z}(i).$$

Then (3.101) and Corollary 3.68 yield

$$\forall i = 1,2,\ldots,N, \quad a_{i-2} + 4a_{i-1} + a_i = 6f(i), \quad (3.106)$$

which proves the middle $N-2$ equations of $M \mathbf{a} = \mathbf{b}$. By Theorem 3.49, we have

$$\frac{d}{dx} B^n_{i,Z}(x) = B^{n+1}_{i,Z}(x) - B^n_{i+1,Z}(x). \quad (3.107)$$

Differentiate (3.104), apply (3.107), set $x = 1$, apply (3.99) and we have the first identity in (3.105), which, together with (3.106), 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 (3.105) can be shown similarly. The diagonal dominance of $M$ implies a nonzero determinant of $M$ and therefore $\mathbf{a}$ is uniquely determined. The uniqueness of $S(x)$ then follows from (3.105).
4 Approximation

**Definition 4.1.** Given a normed vector space \( Y \) of functions and its subspace \( X \subseteq Y \). A function \( \hat{f} \in X \) is called the **best approximation** to \( f \in Y \) from \( X \) with respect to the norm \( \| \cdot \| \) iff

\[
\forall \varphi \in X, \quad \| f - \varphi \| \leq \| f - \varphi \|. \tag{4.1}
\]

**Remark 4.1.** Given \( n \) linearly independent elements \( u_1, u_2, \ldots, u_n \in X \), we are interested in a particular type of best approximation \( \hat{f} = \sum_{i=1}^{n} a_i u_i \). This is called the fundamental problem of linear approximation.

**Definition 4.2** \((L^p \text{ functions})\). 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 4.3.** For a weight function \( \rho(x) \in L[a, b] \), define

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

Then \( L^2_\rho[a, b] \) is a vector space. If \( \int_a^b \rho(x)dx > 0 \) and \( \forall x \in [a, b], \rho(x) \geq 0 \), then \( L^2_\rho[a, b] \) with

\[
\langle u, v \rangle = \int_a^b \rho(t)u(t)v(t)dt \tag{4.3}
\]

is an inner product space over \( \mathbb{R} \) and \( L^2_\rho[a, b] \) with

\[
\| u \|_2 = \left( \int_a^b \rho(t)|u(t)|^2dt \right)^{\frac{1}{2}} \tag{4.4}
\]

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

**Proof.** This follows from Definitions 0.56, 0.71, and 0.73, and 0.75. \( \square \)

**Remark 4.2.** \(|u(t)|^2\) may be different from \( u^2(t) \) because the underlying field may be the complex numbers!

**Definition 4.4.** Setting the norm in (4.1) to that in (4.4) yields the **least square approximation** on \( L^2_\rho[a, b] \), one particular type of best approximation on which we focus.

### 4.1 Orthonormal systems

**Definition 4.5.** A subset \( S \) of an inner product space \( X \) is called **orthonormal** if

\[
\forall u, v \in S, \quad \langle u, v \rangle = \begin{cases} 0 & \text{if } u \neq v, \\ 1 & \text{if } u = v. \end{cases} \tag{4.5}
\]

**Example 4.3.** The unit vectors in \( \mathbb{R}^n \) are orthonormal.

**Example 4.4.** The Chebyshev polynomials of the first kind as in Definition (3.43) are orthogonal with respect to (4.3) where \( a = -1, b = 1, \rho = \sqrt{1-x^2} \). However, they do not satisfy the second case in (4.5).

**Theorem 4.6.** Any finite set of nonzero orthogonal elements \( u_1, u_2, \ldots, u_n \) is linearly independent.

**Proof.** by contradiction using Definitions 0.62 and 4.5. \( \square \)

**Definition 4.7.** 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

\[
v_{n+1} = u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u^*_k \rangle u^*_k, \tag{4.6}
\]

\[
\| u^*_{n+1} \| = \| v_{n+1} \|, \tag{4.7}
\]

with the recursion basis as \( v_1 = u_1, u^*_1 = v_1/\|v_1\| \).

**Theorem 4.8.** For a finite or infinite independent list \( (u_1, u_2, \ldots) \), the Gram-Schmidt process yields constants

\[
\begin{align*}
a_{11} & = 1, \\ a_{21} & = a_{22}, \\ a_{31} & = a_{32} + a_{33}, \\ & \vdots
\end{align*}
\]

such that \( a_{kk} = \frac{1}{\|v_k\|} > 0 \) and the elements \( u^*_1, u^*_2, \ldots \)

\[
\begin{align*}
u^*_1 & = a_{11}u_1, \\ u^*_2 & = a_{21}u_1 + a_{22}u_2, \\ u^*_3 & = a_{31}u_1 + a_{32}u_2 + a_{33}u_3, \\ & \vdots
\end{align*}
\]

are orthonormal.

**Proof.** Apply the Gram-Schmidt process in Definition 4.7. It is clear that \( u^*_n+1 \) is normal. To show \( u^*_n+1 \) is orthogonal to \( u^*_n, u^*_{n-1}, \ldots, u^*_1 \), we have

\[
\langle v_{n+1}, u^*_j \rangle = \left( u_{n+1} - \sum_{k=1}^{n} \langle u_{n+1}, u^*_k \rangle u^*_k \right) \langle u^*_j, u^*_i \rangle \\
= \langle u_{n+1}, u^*_j \rangle - \sum_{k=1}^{n} \langle u_{n+1}, u^*_k \rangle \langle u^*_k, u^*_j \rangle \\
= \langle u_{n+1}, u^*_j \rangle - \langle u_{n+1}, u^*_j \rangle = 0.
\]

Finally, \( a_{kk} = \frac{1}{\|v_k\|} \) follows directly from (4.7). \( \square \)

**Corollary 4.9.** We can find constants

\[
\begin{align*}
b_{11} & = a_{11}, \\ b_{21} & = a_{21}, \\ b_{31} & = a_{31}, \\ & \vdots
\end{align*}
\]

such that \( b_{ii} > 0 \) and

\[
\begin{align*}
u_1 & = b_{11}u^*_1, \\ u_2 & = b_{21}u^*_1 + b_{22}u^*_2, \\ u_3 & = b_{31}u^*_1 + b_{32}u^*_2 + b_{33}u^*_3, \\ & \vdots
\end{align*}
\]
Proof. A lower-triangular matrix with positive diagonal elements is invertible.

Corollary 4.10. In Theorem 4.8, we have \( \langle u_n^*, u_i \rangle = 0 \) for each \( i = 1, 2, \ldots, n - 1 \).

Proof. By Corollary 4.9, 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 4.11. Using the Gram-Schmidt orthonormalizing process with the inner product (4.3), we obtain from the independent list of monomials \((1, x, x^2 \ldots)\) the following classic orthonormal polynomials:

<table>
<thead>
<tr>
<th>Type</th>
<th>( a )</th>
<th>( b )</th>
<th>( \rho(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chebyshev polynomials of the first kind</td>
<td>-1</td>
<td>1</td>
<td>( \frac{1}{\sqrt{1-x^2}} )</td>
</tr>
<tr>
<td>Chebyshev polynomials of the second kind</td>
<td>-1</td>
<td>1</td>
<td>( \sqrt{1-x^2} )</td>
</tr>
<tr>
<td>Legendre polynomials</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Jacobi polynomials</td>
<td>-1</td>
<td>1</td>
<td>( (1-x)^{\alpha}(1+x)^{\beta} )</td>
</tr>
<tr>
<td>Laguerre polynomials</td>
<td>0</td>
<td>+\infty</td>
<td>( x^\alpha e^{-x} )</td>
</tr>
<tr>
<td>Hermite polynomials</td>
<td>-\infty</td>
<td>+\infty</td>
<td>( e^{-x^2} )</td>
</tr>
</tbody>
</table>

where \( \alpha, \beta > -1 \) for Jacobi polynomials and \( \alpha > -1 \) for Laguerre polynomials.

Example 4.5. We compute the first 3 Legendre polynomials using the Gram-Schmidt process.

\[
u_1 = 1, \quad v_1 = 1, \quad \|v_1\|^2 = \int_{-1}^{1} dz = 2, \quad u_1^* = \frac{1}{\sqrt{2}}.
\]

\[
u_2 = x, \quad v_2 = x - \left(x, \frac{1}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} = x, \quad \|v_2\|^2 = \frac{2}{3},
\]

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

\[
u_3 = x^2 - \left(x^2, \frac{3}{\sqrt{2}} \right) \frac{3}{\sqrt{2}} x = x^2, \quad \|v_3\|^2 = \int_{-1}^{1} \left( x^2 - \frac{3}{\sqrt{2}} \right) dx = 2 - \frac{2}{3},
\]

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

4.2 Fourier (or orthogonal) expansions

Definition 4.12. 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

\[
w \sim \sum_{n=1}^{\infty} \langle w, u_n^* \rangle u_n^* \quad \text{(4.10)}
\]

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 \( \sum_{n=1}^{\infty} \langle w, u_n^* \rangle u_n^* - w \).

Remark 4.6. We do not write "=\) in (4.10) because for an infinite sequence the series might not converge.

Example 4.7. With the Euclidean inner product in Definition 0.72, we select orthonormal vectors in \( \mathbb{R}^3 \) as

\[
u_1^* = (1,0,0)^T, \quad u_2^* = (0,1,0)^T, \quad u_3^* = (0,0,1)^T.
\]

For the vector \( w = (a,b,c)^T \), the Fourier coefficients are

\[
\langle w, u_1^* \rangle = a, \quad \langle w, u_2^* \rangle = b, \quad \langle w, u_3^* \rangle = c,
\]

and the projections of \( w \) onto \( u_1^* \) and \( u_2^* \) are

\[
\langle w, u_1^* \rangle u_1^* = (a,0,0)^T, \quad \langle w, u_2^* \rangle u_2^* = (0,b,0)^T.
\]

The Fourier expansion of \( w \) is

\[
w = \sum_{i=1}^{3} \langle w, u_i^* \rangle u_i^* + \langle w, u_3^* \rangle u_3^* = (a, b, 0)^T + (b, a, 0)^T = (a+b, a+b, 0)^T
\]

with the error of Fourier expansion as 0; see Theorem 4.13.

Remark 4.8. For an orthonormal list of vectors \( u_i \), we can construct an orthonormal list by \( u_i^* = u_i/\|u_i\| \) to arrive at the Fourier expansion

\[
w \sim \sum_{i} \langle w, u_i^* \rangle u_i^*/\|u_i\| = \sum_{i} \langle w, u_i \rangle u_i, \quad \text{(4.11)}
\]

Example 4.9. With the following orthonormal list

\[
\frac{1}{\sqrt{2\pi}} \sin x, \frac{1}{\sqrt{\pi}} \cos x, \frac{1}{\sqrt{2\pi}} \sin(nx), \frac{1}{\sqrt{\pi}} \cos(nx), \ldots
\]

in \( L^2_{\rho=1}[-\pi, \pi] \), we obtain from Definition 4.12 the Fourier series of a function \( f(x) \).

\[
f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx), \quad \text{(4.12)}
\]

\[
a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, \quad b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx.
\]

Theorem 4.13. 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

\[
w = \sum_{i=1}^{n} \langle w, u_i^* \rangle u_i^*, \quad \text{(4.13)}
\]

i.e. \( w \) is equal to its Fourier expansion.

Proof. By the condition \( w = \sum_{i=1}^{n} a_i u_i \) and Corollary 4.9, 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 \( u_i^* \) implies

\[
\forall k = 1, 2, \ldots, n, \quad \langle u_k^*, w \rangle = c_k,
\]

which completes the proof.
Theorem 4.14 (Minimum properties of Fourier expansions). Let $u_1^*, u_2^*, \ldots$ be an orthonormal system and let $w$ be arbitrary. Then
\[
\left\| w - \sum_{i=1}^{N} (w, u_i^*) u_i^* \right\| \leq \left\| w - \sum_{i=1}^{N} a_i u_i^* \right\|, \tag{4.14}
\]
for any selection of constants $a_1, a_2, \ldots, a_N$.

Proof. With the shorthand notation \( \sum_i = \sum_{i=1}^{N} \), we deduce from Definition 0.71 and properties of inner products
\[
\left\| w - \sum_i a_i u_i^* \right\|^2 = (w, w) - \sum_i a_i^2 (w, u_i^*)^2 - 2 \sum_i a_i \langle w, u_i^* \rangle \langle u_i^*, w \rangle
\]
\[
= (w, w) - \sum_i a_i^2 (w, u_i^*)^2 - \sum_i \langle w, u_i^* \rangle \langle u_i^*, w \rangle
\]
\[
+ \sum_i \sum_j a_i a_j \langle u_i^*, u_j^* \rangle
\]
\[
= (w, w) - \sum_i a_i (w, u_i^*) - \sum_i a_i^2 \langle w, u_i^* \rangle^2 + \sum_i |a_i|^2
\]
\[
= \|w\|^2 - \sum_i |\langle w, u_i^* \rangle|^2 + \sum_i |a_i - \langle w, u_i^* \rangle|^2, \tag{4.15}
\]
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 4.15. 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 \( \hat{u} = \sum_k \langle w, u_k^* \rangle u_k^* \) where \( u_k^* \)'s are the \( u_k \)'s orthonormalized by the Gram-Schmidt process. The error norm is
\[
\|w - \hat{u}\|^2 := \min_{a_k} \left\| w - \sum_{k=1}^{n} a_k u_k \right\|^2 = \|w\|^2 - \sum_{k=1}^{n} |\langle w, u_k^* \rangle|^2. \tag{4.16}
\]
Proof. This follows directly from (4.15).

Corollary 4.16 (Bessel inequality). If $u_1^*, u_2^*, \ldots, u_N^*$ are orthonormal, then, for an arbitrary \( w \),
\[
\sum_{i=1}^{N} |\langle w, u_i^* \rangle|^2 \leq \|w\|^2. \tag{4.17}
\]
Proof. This follows directly from Corollary 4.15 and the real positivity of a norm.

Corollary 4.17. The Gram-Schmidt process in Definition 4.7 satisfies
\[
\forall n \in \mathbb{N}^+, \|v_{n+1}\|^2 = \|u_{n+1}\|^2 - \sum_{k=1}^{n} |\langle u_{n+1}, u_k^* \rangle|^2. \tag{4.18}
\]

Proof. By (4.6), 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 4.15, identifying $w$ with $u_{n+1}$ completes the proof.

Remark 4.10. The Fourier expansion helps our understanding of the Gram-Schmidt process in two aspects:

- Definition 4.7 is a special type of Fourier expansion; this helps the memorization.
- As a complement to Definition 4.7 and Theorem 4.8, Corollary 4.17 reveals additional relations between vectors $u_k$, $v_k$, and $u_k^*$.}

Example 4.11. The least square approximation problem of $e^x$ on $[-1, 1]$ with a linear and quadratic polynomial with the weight function $\rho = 1$.

First note that the problem is equivalent to
\[
\min_{a_i} \int_{-1}^{+1} \left( e^x - \sum_{i=0}^{n} a_i x^i \right)^2 dx \tag{4.19}
\]
for $n = 1, 2$. Use the Legendre polynomials derived in Example 4.5:
\[
u_1 = \frac{1}{\sqrt{2}}, \quad \nu_2 = \sqrt{\frac{3}{2}} x, \quad \nu_3 = \frac{1}{4} \sqrt{10}(3x^2 - 1).
\]
The Fourier coefficients of $e^x$ are
\[
\begin{align*}
b_0 &= \int_{-1}^{+1} \frac{1}{\sqrt{2}} e^x dx = \frac{1}{\sqrt{2}} \left( e - \frac{1}{e} \right), \\
b_1 &= \int_{-1}^{+1} \sqrt{\frac{3}{2}} x e^x dx = \sqrt{6} e^{-1}, \\
b_2 &= \int_{-1}^{+1} \frac{1}{4} \sqrt{10} (3x^2 - 1) e^x dx = \frac{\sqrt{10}}{2} \left( e - \frac{7}{e} \right).
\end{align*}
\]
The minimizing polynomials are thus
\[
\hat{\varphi}_n = \begin{cases} \frac{1}{2} (e^2 - 1) + \frac{3}{2} x & n = 1; \\ \hat{\varphi}_1 + \frac{3}{2} (e^2 - 7)(3x^2 - 1) & n = 2. \end{cases} \tag{4.20}
\]
This illustrates the permanence advantage of orthonormal systems in solving least square approximation: Fourier coefficient for the best approximation of a lower degree can be reused in constructing that of a higher degree.

4.3 The normal equations

Remark 4.12. With an orthonormal list, we can easily find the best linear approximation of a given function. If the list is not orthonormal but only independent, one way to the best approximation is via the normal equations.

Theorem 4.18. 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 $w$,
\[
\forall j = 1, 2, \ldots, n, \quad \left( w - \sum_{k=1}^{n} \langle w, u_k^* \rangle u_k^* \right) \perp u_j^*, \tag{4.21}
\]
where \( \perp \) denotes orthogonality.
Proof. Take the inner product of the two vectors and apply the conditions on orthonormal systems.

Remark 4.13. \( w \) may or may not be in \( X \).

Corollary 4.19. 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

\[
\forall j = 1, 2, \ldots, n, \quad \left( w - \sum_{k=1}^{n} a_k u_k \right) \perp u_j, \quad (4.22)
\]

Proof. Since \( \hat{\varphi} = \sum_{k=1}^{n} a_k u_k \) is the best linear approximant to \( w \), Theorem 4.14 implies that

\[
\sum_{k=1}^{n} a_k u_k = \sum_{k=1}^{n} (w, u_k^\ast) u_k^\ast.
\]

Corollary 4.9 and Theorem 4.18 complete the proof.

Definition 4.20. Let \( u_1, u_2, \ldots, u_n \) be a sequence of elements in an inner product space. The \( n \times n \) matrix

\[
G = G(u_1, u_2, \ldots, u_n) = \begin{pmatrix} (u_1, u_1) & \langle u_1, u_2 \rangle & \cdots & \langle u_1, u_n \rangle \\ \\ (u_2, u_1) & (u_2, u_2) & \cdots & \langle u_2, u_n \rangle \\ & & \vdots & \vdots \\ & & \langle u_n, u_1 \rangle & (u_n, u_2) \cdots \langle u_n, u_n \rangle \end{pmatrix}
\]

is the Gram matrix of \( u_1, u_2, \ldots, u_n \). Its determinant

\[
g = g(u_1, u_2, \ldots, u_n) = \det((u_1, u_j)) \quad (4.24)
\]

is the Gram determinant of the elements \( u_i \)'s.

Lemma 4.21. Let \( w_i = \sum_{j=1}^{n} a_{ij} u_j \) for \( i = 1, 2, \ldots, n \). Let \( A = (a_{ij}) \) and its conjugate transpose \( A^H = (\overline{a_{ji}}) \).

Then we have

\[
G(w_1, w_2, \ldots, w_n) = AG(u_1, u_2, \ldots, u_n)A^H \quad (4.25)
\]

and

\[
g(w_1, w_2, \ldots, w_n) = \det(A)^2 g(u_1, u_2, \ldots, u_n). \quad (4.26)
\]

Proof. The inner product of \( u_i \) and \( w_j \) yields

\[
G(u_1, w_1) \begin{pmatrix} (u_1, u_1) & \langle u_1, u_2 \rangle & \cdots & \langle u_1, u_n \rangle \\ \\ (u_2, u_1) & (u_2, u_2) & \cdots & \langle u_2, u_n \rangle \\ & & \vdots & \vdots \\ & & \langle u_n, u_1 \rangle & (u_n, u_2) \cdots \langle u_n, u_n \rangle \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} a_{1i} u_i \\ \\ \sum_{i=1}^{n} a_{2i} u_i \\ & \vdots \\ \sum_{i=1}^{n} a_{ni} u_i \end{pmatrix} = G(u_1, u_2, \ldots, u_n)A^H.
\]

Therefore (4.25) holds since

\[
G(w_1, w_2, \ldots, w_n) = \begin{pmatrix} (w_1, w_1) & \langle w_1, w_2 \rangle & \cdots & \langle w_1, w_n \rangle \\ \\ (w_2, w_1) & (w_2, w_2) & \cdots & \langle w_2, w_n \rangle \\ & & \vdots & \vdots \\ & & \langle w_n, w_1 \rangle & (w_n, w_2) \cdots \langle w_n, w_n \rangle \end{pmatrix} = AG(u_1, u_2, \ldots, u_n)A^H.
\]

Then (4.26) follows from taking the determinant of (4.25) and applying the identities \( \det(AB) = \det(A) \det(B) \) and \( \det(A) = \det(A^T) \).

Theorem 4.22. For nonzero elements \( u_1, u_2, \ldots, u_n \),

\[
0 \leq g(u_1, u_2, \ldots, u_n) \leq \prod_{k=1}^{n} \| u_k \|^2, \quad (4.27)
\]

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

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

\[
w_k = \begin{pmatrix} \sum_{i=1}^{n} c_i u_i = 0, & k = j; \\ u_k, & k \neq j. \end{pmatrix}
\]

We have \( g(w_1, w_2, \ldots, w_n) = 0 \) because \( \langle w_j, w_k \rangle = 0 \) for each \( k \). By the Laplace theorem, we expand the determinant of \( C = (c_{ij}) \) according to minors of its \( j \)th row:

\[
det(C) = \begin{vmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \ddots & \vdots & \vdots \\ c_1 & c_2 & \cdots & c_j & c_{n} \\ 0 & 0 & \cdots & 0 & 1 \end{vmatrix} = 0 + \cdots + 0 + c_j + 0 + \cdots + 0 = c_j \neq 0.
\]

Then Lemma 4.21 yields \( g(u_1, u_2, \ldots, u_n) = 0 \).

Now suppose \( u_1, u_2, \ldots, u_n \) are linearly independent. Theorem 4.8 yields constants \( a_{ij} \) such that \( a_{kk} > 0 \) and the following vectors are orthonormal:

\[
u_k^* = \sum_{i=1}^{k} a_{ik} u_i.
\]

Then Definition 4.20 implies \( g(u_1^*, u_2^*, \ldots, u_n^*) = 1 \). Also, we have \( \det(a_{ij}) = \prod_{k=1}^{n} a_{kk} \) because the matrix \( (a_{ij}) \) is triangular. It then follows from Lemma 4.21 that

\[
g(u_1, u_2, \ldots, u_n) = \prod_{i=k}^{n} \frac{1}{a_{kk}} > 0. \quad (4.28)
\]

The above arguments show that \( g(u_1, u_2, \ldots, u_n) = 0 \) if and only if \( u_1, u_2, \ldots, u_n \) are linearly dependent.
Suppose \( u_1, u_2, \ldots, u_n \) are orthogonal. By Definition 4.20, \( G(u_1, u_2, \ldots, u_n) \) is a diagonal matrix with \( \| u_k \|^2 \) on the diagonals. Hence the orthogonality of \( u_k \)'s implies

\[
g(u_1, u_2, \ldots, u_n) = \prod_{k=1}^{n} \| u_k \|^2. \tag{4.29}
\]

For the converse statement, suppose (4.29) holds. From Theorem 4.8 we know \( \frac{1}{a_{kk}} = \| v_k \| \). Then (4.28) and (4.29) yield

\[
\forall k = 1, 2, \ldots, n, \quad \| u_k \|^2 = \| v_k \|^2. \tag{4.30}
\]

Then Corollary 4.17 and (4.30) imply

\[
\forall k = 1, 2, \ldots, n, \quad \sum_{j=1}^{k-1} \| u_k, u_j \|^2 = 0,
\]

which further implies

\[
\forall k = 1, 2, \ldots, n, \forall j = 1, 2, \ldots, k-1, \quad \langle u_k, u_j^* \rangle = 0,
\]

which, together with Corollary 4.9, 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 (4.28), \( \frac{1}{a_{kk}} = \| v_k \| \), and Corollary 4.17.

**Remark 4.14.** An alternative proof of Theorem 4.22.

**Proof.** Consider a nonzero vector \( x = (x_1, x_2, \ldots, x_n)^T \) with each \( x_i \) in the field that underlies span \((u_1, \ldots, u_n)\). Let \( x^H \) denote the conjugate transpose of \( x \). By properties of inner products in Definition 0.71, we have

\[
xG(u_1, u_2, \ldots, u_n)x^H = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \langle u_i, u_j \rangle x_j
\]

\[
= \sum_{i=1}^{n} \left( \sum_{j=1}^{n} x_i u_i \right) \langle x_j \rangle = \left( \sum_{j=1}^{n} x_j u_i \right) \sum_{j=1}^{n} x_j u_j
\]

\[
= \left\| \sum_{i=1}^{n} x_i u_i \right\|^2 \geq 0,
\]

where, by (NRM-2) in Theorem 0.78, the equality holds only when \( \sum_{i=1}^{n} x_i u_i = 0 \), i.e. the \( u_i \)'s are linearly dependent. Hence \( G(u_1, u_2, \ldots, u_n) \) is positive definite if and only if \( (u_1, u_2, \ldots, u_n) \) is linearly independent. From linear algebra, we know that the determinant of a positive definite matrix is positive. The above argument proves the first inequality in (4.27).

For a positive definite matrix \( A \) satisfying \( A^H = A \), consider its block form,

\[
A = \begin{bmatrix} a_{11} & b^H \\ b & A_{n-1} \end{bmatrix}.
\]

Let \( I \) denote an identity matrix. Define

\[
P = \begin{bmatrix} \frac{1}{A_{n-1}} & 0^T \\ -b^T & I \end{bmatrix}.
\]

Since \( \det(P) = 1 \), the determinant of \( A \) is the same as

\[
P^HAP = \begin{bmatrix} a_{11} - b^H A_{n-1}^{-1}b & 0^T \\ 0 & A_{n-1} \end{bmatrix}.
\]

Without changing the determinant of \( A \), we can repeat the above process to diagonalize \( A \). Consequently, we have

\[
det(A) = \prod_{i=1}^{n} (a_{ii} - \alpha_i), \tag{4.31}
\]

where \( a_{ii} \) is the \( i \)th diagonal entry of \( A \), \( \alpha_n = 0 \), and the positive-definiteness of \( A \) implies that \( \alpha_i > 0 \) for \( i = 1, 2, \ldots, n-1 \).

Apply the arguments in the previous paragraph to \( G(u_1, u_2, \ldots, u_n) \) and we have

\[
g(u_1, u_2, \ldots, u_n) \leq \prod_{k=1}^{n} \| u_k \|^2,
\]

where the equality holds only when all off-diagonal entries in \( G(u_1, u_2, \ldots, u_n) \) are zero, i.e. \( \langle u_i, u_j \rangle = 0 \) for all \( i \neq j \), i.e. \( (u_1, u_2, \ldots, u_n) \) is orthogonal. Conversely, if \( (u_1, u_2, \ldots, u_n) \) is orthogonal, then \( G(u_1, u_2, \ldots, u_n) \) is clearly a diagonal matrix with \( \| u_k \|^2 \) on the diagonal.

**Theorem 4.23.** Let \( \hat{\varphi} = \sum_{i=1}^{n} a_i u_i \) be the best approximation to \( w \) constructed from the independent list \((u_1, u_2, \ldots, u_n)\). Then the coefficients

\[
a = [a_1, a_2, \ldots, a_n]^T
\]

are uniquely determined from the linear system of normal equations,

\[
G(u_1, u_2, \ldots, u_n)a = c, \tag{4.32}
\]

where \( c = [(w, u_1), (w, u_2), \ldots, (w, u_n)]^T. \)

**Proof.** Corollary 4.19 yields

\[
\langle w, u_j \rangle = \sum_{k=1}^{n} a_k \langle u_k, u_j \rangle,
\]

which is simply the \( j \)th equation of (4.32). The uniqueness of the coefficients follows from Theorem 4.22 and Cramer’s rule.

**Example 4.15.** Solve Example 4.11 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 (4.23), (4.3), and \( p = 1 \):

\[
G(1, x, x^2) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & x & x^2 \\ 1 & x^2 & x^2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}.
\]

We then calculate the vector

\[
c = \begin{bmatrix} e^{x}, e^x, e^{2x} \end{bmatrix} = \begin{bmatrix} e^{-1/e} \\ 2/e \\ e^{-5/e} \end{bmatrix}.
\]

The normal equations then yields

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

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 (4.20).
**Remark 4.16.** Comparing Examples 4.15 and 4.11, we observe several advantages of the Fourier-expansion (FE) approach over the normal-equations (NE) approach:

- the number of calculating inner products in FE is only $O(n)$ while that in NE is $O(n^2)$,
- the coefficients in FE have the property of permanence while those in NE change for a different $n$ (verify this by solving the normal equations with $n = 2$),
- FE is well-conditioned while NE is ill-conditioned for large $n$.

### 4.4 Discrete least squares

**Definition 4.24.** Define a function $\lambda : \mathbb{R} \to \mathbb{R}$

$$\lambda(t) = \begin{cases} 0 & \text{if } t \in (-\infty, a), \\ \int_a^b \rho(\tau)d\tau & \text{if } t \in [a, b], \\ \int_a^b \rho(\tau)d\tau & \text{if } t \in (b, +\infty). \end{cases} \quad (4.33)$$

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

$$d\lambda = \begin{cases} \rho(t)d\tau & \text{if } t \in [a, b], \\ 0 & \text{otherwise}, \end{cases} \quad (4.34)$$

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

**Definition 4.25.** The *discrete measure* or the *Dirac measure* associated with the point set $\{t_1, t_2, \ldots, t_N\}$ is a measure $d\lambda$ that is nonzero only at the points $t_i$ and has the value $\rho_i$ there. The *support of the discrete measure* is the set $\{t_1, t_2, \ldots, t_N\}$.

**Definition 4.26.** The *Heaviside function* is the truncated power function with exponent 0,

$$H(x) = x^0_+ = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \quad (4.35)$$

**Remark 4.17.** The Dirac Delta function, $\delta(x)$, is roughly a generalized function that satisfies

$$\delta(x) = \begin{cases} +\infty & x = 0, \\ 0 & x \neq 0. \end{cases} \quad (4.36)$$

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

Useful properties of $\delta(x)$ include

$$\int_{-\infty}^{+\infty} \delta(x)dx = 1, \quad (4.37)$$

$$\int_{-\infty}^{x} \delta(t)dt = H(x), \quad (4.38)$$

$$\int_{-\infty}^{+\infty} f(t)\delta(t - t_0)dt = f(t_0). \quad (4.39)$$

**Lemma 4.27.** For a function $u : \mathbb{R} \to \mathbb{R}$, define

$$\lambda(t) = \sum_{i=1}^{N} \rho_i H(t - t_i), \quad (4.40)$$

and we have

$$\int_{\mathbb{R}} u(t)d\lambda = \sum_{i=1}^{N} \rho_i u(t_i). \quad (4.41)$$

**Proof.** $(4.40)$, $(4.38)$, and $(4.39)$ yield

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


**Remark 4.18.** By Definition 4.24 and Lemma 4.27, we can solve the discrete least square problem by reusing the procedures in Examples 4.11 and 4.15, but the definition of inner product should be switched to

$$\langle u(t), v(t) \rangle = \sum_{i=1}^{N} \rho(t_i)u(t_i)v(t_i), \quad (4.42)$$

which follows directly from $(4.41)$.

**Example 4.19.** Consider a table of sales record.

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>256</td>
<td>201</td>
<td>159</td>
<td>61</td>
<td>77</td>
<td>40</td>
</tr>
<tr>
<td>x</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>y</td>
<td>17</td>
<td>25</td>
<td>103</td>
<td>156</td>
<td>222</td>
<td>345</td>
</tr>
</tbody>
</table>

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,

$$\sum_{i=1}^{12} \left( y_i - \sum_{j=0}^{2} a_j x^j \right)^2.$$ 

Reusing the procedures in Example 4.15, we have

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

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

Then the normal equations yield

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

The corresponding polynomial is plotted in the figure.
5 Numerical Integration

Definition 5.1. A weighted quadrature formula \( I_n(f) \) for a function \( f \in L[a,b] \) is a formula

\[
I_n(f) := \sum_{k=1}^{n} w_k f(x_k) \tag{5.1}
\]

that approximates the definite integral of \( f \) on \([a,b]\)

\[
I(f) := \int_{a}^{b} f(x) \rho(x) dx, \tag{5.2}
\]

where \( \rho \in L[a,b] \) is a non-negative weight function that satisfies \( \int_{a}^{b} \rho(x) dx > 0 \). 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.

Remark 5.1. If \( a \) and/or \( b \) are infinite, \( I(f) \) and \( I_n(f) \) in (5.1) may still be well defined if the moment of weight \( \mu_j := \int_{a}^{b} x^j \rho(x) dx \) exists and is finite for all \( j \in \mathbb{N} \).

5.1 Accuracy and convergence

Definition 5.2. The remainder, or error, of \( I_n(f) \) is

\[
E_n(f) := I(f) - I_n(f). \tag{5.4}
\]

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

\[
\forall f \in C[a,b], \lim_{n \to \infty} I_n(f) = I(f). \tag{5.5}
\]

Definition 5.3. A weighted quadrature formula (5.1) has (polynomial) degree of exactness \( d_E \) iff

\[
\left\{ \begin{array}{l}
\forall f \in P_{d_E}, \quad E_n(f) = 0, \\
\exists g \in P_{d_E+1}, \text{s.t. } E_n(g) \neq 0.
\end{array} \right. \tag{5.6}
\]

Lemma 5.4. 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

\[
\forall k = 1, \ldots, n, \quad w_k = \int_{a}^{b} \rho(x) \ell_k(x) dx, \tag{5.7}
\]

where \( \ell_k(x) \) is the fundamental polynomial for pointwise interpolation in (3.7) applied to the given nodes,

\[
\ell_k(x) := \prod_{i \neq k, i=1}^{n} \frac{x - x_i}{x_k - x_i}. \tag{5.8}
\]

Proof. Let \( p_{n-1}(f;x) \) be the unique polynomial that interpolates \( f \) at the distinct nodes, as in Theorem 3.2. Then

\[
\sum_{k=1}^{n} \int_{a}^{b} w_k p_{n-1}(x_k) = \int_{a}^{b} p_{n-1}(f;x) \rho(x) dx
\]

\[
= \int_{a}^{b} \sum_{k=1}^{n} \left( \ell_k(x) f(x_k) \right) \rho(x) dx = \sum_{k=1}^{n} \int_{a}^{b} w_k f(x_k),
\]

where the first step follows from \( d_E \geq n-1 \) and the second step from the interpolation conditions (3.1), the Lagrange formula, and the uniqueness of \( p_{n-1}(f;x) \). The proof is completed by setting \( f \) to be the hat function \( \hat{B}_k(x) \) (see Definition 3.40) for each \( x_k \).

Definition 5.5. A subset \( \mathbb{V} \subset C[a,b] \) is dense in \( C[a,b] \) iff \( \forall f \in C[a,b], \forall \epsilon > 0, \exists \mathbb{V} \in \mathbb{V}, \text{s.t. } \max_{x \in [a,b]} |f(x) - f_\mathbb{V}(x)| \leq \epsilon. \)

Theorem 5.6 (Weierstrass). The set of polynomials is dense in \( C[a,b] \). In other words, for any given \( f(x) \in C[a,b] \) and given \( \epsilon > 0 \), one can find a polynomial \( p_n(x) \) (of sufficiently high degree) such that

\[
\forall x \in [a,b], \quad |f(x) - p_n(x)| \leq \epsilon. \tag{5.10}
\]

Proof. Not required.

Theorem 5.7. 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 (5.1). Let \( \mathbb{V} \) be a dense subset of \( C[a,b] \). \( I_n(f) \) is convergent for \( C[a,b] \) if and only if

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

(b) \( B := \sup_{n \in \mathbb{N}^+} \sum_{k=1}^{n} |w_k| < +\infty. \)

Proof. For necessity, it is trivial to deduce (a) from (5.5). In contrast, it is highly nontrivial to deduce (b) from (5.5). 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. page 121 of [Cryer, C. (1982) Numerical Functional Analysis. Oxford Univ. Press].

For the sufficiency, find \( f_\mathbb{V} \in \mathbb{V} \) such that (5.9) holds, define \( K := \max_{x \in [a,b]} |f(x) - f_\mathbb{V}(x)| \). Then we have

\[
|E_n(f)| \leq |I(f) - I_n(f)| + |I_n(f) - I_\mathbb{V}(f)| + |I_\mathbb{V}(f) - I(f)|
\]

\[
= \left| \int_{a}^{b} (f(x) - f_\mathbb{V}(x)) \rho(x) dx \right|
\]

\[
+ |I_\mathbb{V}(f) - I_n(f)| + \sum_{k=1}^{n} \int_{a}^{b} w_k |f(x_k) - f_\mathbb{V}(x_k)|
\]

\[
\leq K \left| \int_{a}^{b} \rho(x) dx + \sum_{k=1}^{n} |w_k| \right| + |I_\mathbb{V}(f) - I_n(f)|,
\]

where the first step follows from the triangular inequality, the second from Definition 5.1, and the third from the integral mean value theorem 0.54. The terms inside the brackets is bounded because of \( \rho \in L[a,b] \) and condition (b). By condition (a), \( |I_\mathbb{V}(f) - I_n(f)| \) can be made arbitrarily small. The proof is completed by the fact that \( K \) can also be arbitrarily small.

Remark 5.2. By the Weierstrass theorem, condition (a) in Theorem 5.7 is satisfied if we derive integration formulas via interpolating the integrand by polynomials at the nodes. Whether or not such a formula is convergent boils down to checking condition (b).
5.2 Newton-Cotes formulas

Definition 5.8. A Newton-Cotes formula is a formula (5.1) based on approximating \( f(x) \) by interpolating it on equally spaced nodes \( x_1, \ldots, x_n \in [a, b] \).

Remark 5.3. Two special cases of Newton-Cotes formulas for \( n = 2, 3 \) are also known as the trapezoidal rule and Simpson’s rule.

Definition 5.9. The trapezoidal rule is a formula (5.1) 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

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

Remark 5.4. By Definition 5.9, (5.11) can be derived as

\[
I(f) \approx \int_a^b \left[ \frac{x-a}{b-a} f(a) + \frac{x-b}{a-b} f(b) \right] \, dx
= \frac{1}{b-a} \left[ f(a) \frac{(x-a)^2}{2} - f(b) \frac{(x-b)^2}{2} \right]_a^b
= \frac{b-a}{2} [f(a) + f(b)] = I_T(f).
\]

Also by Definition 5.9, the trapezoidal rule has \( d_E = 1 \).

Example 5.5. Derive the trapezoidal rule for the weight function \( \rho(x) = x^{-1/2} \) on the interval \([0, 1]\). Note that (5.11) is not applicable since \( \rho(0) = \infty \). (5.7) yields

\[
w_1 = \int_0^1 x^{-1/2} \, dx = \frac{4}{3},
\]

\[
w_2 = \int_0^1 x^{-1/2} x \, dx = \frac{2}{3}.
\]

Hence the formula is

\[
I_T(f) = \frac{2}{3} [2f(0) + f(1)].
\]  

Theorem 5.10. For \( f \in C^2[a, b] \), the remainder of the trapezoidal rule satisfies

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

Proof. By Theorem 3.2, the interpolating polynomial \( p_n(f; x) \) is unique. Then we have

\[
E_T(f) = -\int_a^b \frac{f'''(\xi(x))}{2} (x-a)(b-x) \, dx
= -\frac{f'''(\xi)}{2} \int_a^b (x-a)(b-x) \, dx = -\frac{(b-a)^3}{12} f'''(\xi),
\]

where the first step follows from Theorem 3.4 and the second step from the integral mean value theorem (Theorem 0.54). Here we can apply Theorem 0.54 because \( w(x) = (x-a)(b-x) \) is always positive on \((a, b)\). Also note that \( \xi \) is a function of \( x \) while \( \xi \) is a constant depending only on \( f, a, \) and \( b \).

Definition 5.11. Simpson’s rule is a formula (5.1) based on approximating \( f(x) \) by a quadratic polynomial that goes through \((a, f(a))^T, (b, f(b))^T, \) and \((a+b)/2, f((a+b)/2))^T\).

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

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

Theorem 5.12. For \( f \in C^4[a, b] \), the remainder of Simpson’s rule satisfies

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

Proof. It is difficult to imitate the proof of Theorem 5.10, since \((x-a)(x-b)(x-a+b)\) changes sign over \([a, b]\) and the integral mean value theorem is not applicable. To overcome this difficulty, we can formulate the interpolation via a Hermite problem so that Theorem 0.54 can be applied. See homework 11 for the details.

Example 5.6. Consider the integral

\[
I = \int_{-4}^{4} \frac{dx}{1+x^2} = 2 \tan^{-1}(4) = 2.6516 \cdots
\]

As shown below, the Newton-Cotes formula appears to be non-convergent.

<table>
<thead>
<tr>
<th>( n-1 )</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I_{n-1} )</td>
<td>5.4902</td>
<td>2.2776</td>
<td>3.3288</td>
<td>1.9411</td>
<td>3.5956</td>
</tr>
</tbody>
</table>

Remark 5.7. For equally spaced nodes, the interpolating polynomials have wilder and wilder oscillations as the degree increases. Consequently, condition \( (b) \) of Theorem 5.7 does not hold. Hence Newton-Cotes formulas are not convergent even for well-behaved functions in \( C[a, b] \). In practice, Newton-Cotes formula with \( n > 8 \) is seldom used.

Remark 5.8. The non-convergence problem of Newton-Cotes formulas can be overcome via

- fixing \( d_E \), subdividing \([a, b]\), and summing up, or
- choosing the nodes \( x_k \)’s carefully.

The above two approaches lead to composite rules and Gauss formulas, respectively.

5.3 Composite formulas

Remark 5.9. Divide \([a, b]\) into \( n \) subintervals of equal length, apply the trapezoidal rule to each subinterval, sum up the results and we have the following composite trapezoidal rule.

Definition 5.13. The composite trapezoidal rule for approximating \( I(f) \) in (5.2) with \( \rho(x) \equiv 1 \) is

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

where \( h = \frac{b-a}{n} \) and \( x_k = a + kh \).
For \( f \in C^2[a,b] \), the remainder of the composite trapezoidal rule satisfies
\[
\exists \xi \in (a,b) \text{ s.t. } E^T_n(f) = -\frac{b-a}{12} h^2 f''(\xi). \tag{5.18}
\]

**Proof.** Apply Theorem 5.10 to the subintervals, sum up the errors, and we have
\[
E^T_n(f) = -\frac{b-a}{12} h^2 \left[ \frac{1}{n} \sum_{k=0}^{n-1} f''(\xi_k) \right]. \tag{5.19}
\]

If \( f \in C^2[a,b] \) implies \( f'' \in C[a,b] \). The proof is completed by (5.19) and the intermediate value Theorem 0.30. \( \square \)

**Definition 5.15.** The composite Simpson’s rule for approximating \( I(f) \) in (5.2) with \( \rho(x) \equiv 1 \) is
\[
I^S_n(f) = \frac{b-a}{3} \left[ f(x_0) + 4f(x_1) + 2f(x_2) + \ldots + 4f(x_{n-1}) + f(x_n) \right]. \tag{5.20}
\]

**Theorem 5.16.** For \( f \in C^4[a,b] \) and \( n \in 2\mathbb{N}^+ \), the remainder of the composite Simpson’s rule satisfies
\[
\exists \xi \in (a,b) \text{ s.t. } E^S_n(f) = -\frac{b-a}{180} h^4 f^{(4)}(\xi). \tag{5.21}
\]

**Proof.** homework. \( \square \)

**Remark 5.10.** Compared to the trapezoidal rule, Simpson’s rule is usually much more efficient because it gains two more orders of accuracy with roughly the same computational cost. However, the composite trapezoidal rule does have its advantages: it is much better than composite Simpson’s rule for trigonometric polynomials (4.12). See page 167 of NAG2012.

### 5.4 Gauss formulas

**Lemma 5.17.** Let \( n, m \in \mathbb{N}^+ \) and \( m \leq n \). Given polynomials \( p = \sum_{i=0}^{n+m} p_i x^i \in P_{n+m} \) and \( s = \sum_{i=0}^{n} s_i x^i \in P_n \) satisfying \( p_{n+m} \neq 0 \) and \( s_n \neq 0 \), there exist unique polynomials \( q \in P_m \) and \( r \in P_{n-1} \) such that
\[
p = qs + r. \tag{5.22}
\]

**Proof.** Rewrite (5.22) as
\[
\sum_{i=0}^{n+m} p_i x^i = \left( \sum_{i=0}^{m} q_i x^i \right) \left( \sum_{i=0}^{n} s_i x^i \right) + \sum_{i=0}^{n-1} r_i x^i. \tag{5.23}
\]

Since monomials are linearly independent, (5.23) consists of \( n + m + 1 \) equations, the last \( m + 1 \) of which are
\[
p_{n+m} = q_m s_n,
p_{n+m-1} = q_m s_{n-1} + q_{m-1} s_n,
\ldots
\]
\[
p_n = q_m s_{n-m} + \ldots + q_0 s_n.
\]

This is a nonsingular triangular linear system in terms of the coefficients of \( q \), which can be determined uniquely from coefficients of \( p \) and \( s \). Then \( r \) can be determined uniquely by \( p - qs \) from (5.23). \( \square \)

**Definition 5.18.** The node polynomial associated with the nodes \( x_k \)’s of a weighted quadrature formula is
\[
v_n(x) = \prod_{k=1}^{n} (x - x_k). \tag{5.24}
\]

**Theorem 5.19.** Suppose a quadrature formula (5.1) has \( d_E \geq n-1 \). Then it can be improved to have \( d_E \geq n+j-1 \) where \( j \in (0,n) \) by and only by imposing the additional conditions on its node polynomial and weight function,
\[
\forall p \in \mathbb{P}_{j-1}, \quad \int_a^b v_n(x) p(x) \rho(x) dx = 0. \tag{5.25}
\]

**Proof.** For the necessity, we have
\[
\int_a^b v_n(x) p(x) \rho(x) dx = \sum_{k=1}^{n} w_k v_n(x_k)p(x_k) = 0,
\]
where the first step follows from \( d_E \geq n+j-1 \) and \( v_n(x)p(x) \in \mathbb{P}_{n+j-1} \), and the second step from (5.24).

To prove the sufficiency, we must show that \( E_n(p) = 0 \) for any \( p \in \mathbb{P}_{n+j-1} \). Lemma 5.17 yields
\[
\forall p \in \mathbb{P}_{n+j-1}, \exists q \in \mathbb{P}_j, \quad r \in \mathbb{P}_{n-1}, \text{ s.t. } p = qv_n + r. \tag{5.26}
\]

Consequently, we have
\[
\int_a^b p(x) \rho(x) dx = \int_a^b q(x)v_n(x) \rho(x) dx + \int_a^b r(x) \rho(x) dx
\]
\[
= \int_a^b r(x) \rho(x) dx = \sum_{k=1}^{n} w_k r(x_k)
\]
\[
= \sum_{k=1}^{n} w_k [p(x_k) - q(x_k)v_n(x_k)] = \sum_{k=1}^{n} w_k p(x_k),
\]
where the first step follows from (5.26), the second from condition (b), the third from condition (a), the fourth from (5.26), and the last from (5.24). \( \square \)

**Remark 5.11.** The index \( j \) in (5.25) 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 5.19 is optimal: the formula (5.1) achieves the highest degree of exactness \( 2n-1 \). From an algebraic viewpoint, the \( 2n \) degrees of freedom of nodes and weights in (5.1) determine a polynomial of degree at most \( 2n-1 \). This optimal case is known as the Gaussian quadrature.

**Definition 5.20.** A Gaussian quadrature formula is a formula (5.1) whose nodes are the zeros of the polynomial \( v_n(x) \) in (5.24) that satisfies (5.25) for \( j = n \).

**Corollary 5.21.** A Gauss formula has \( d_E = 2n - 1 \).

**Proof.** This follows from Theorem 5.19 and Remark 5.11 that \( d_E \) can not be improved to \( 2n \). \( \square \)
Corollary 5.22. Weights of a Gauss formula \( I_n(f) \) are
\[
\forall k = 1, \cdots, n, \quad w_k = \int_a^b \frac{v_n(x)}{x - x_k} c_n'(x_k) \rho(x)dx, \tag{5.27}
\]
where \( v_n(x) \) is the node polynomial that defines \( I_n(f) \).

Proof. This follows from Lemma 5.4; also see (3.9).

Remark 5.12. Given \( [a, b] \) and the weight function \( \rho(x) \), we can determine a Gauss formula \( I_n(f) \) as follows,
(a) determine a monic polynomial \( v_n(x) \) that satisfies
\[
\forall p \in \mathbb{P}_{n-1}, \quad (v_n, p) := \int_a^b p(x)\rho(x)\rho(x)dx = 0,
\]
which is equivalent to \( (1, \pi(x)) = 0 \) and \( (x, \pi(x)) = 0 \) because \( \mathbb{P}_1 = \text{span}(1, x) \). These two conditions yield
\[
\int_0^1 (c_1 - c_1 x + x^2) x^{-1/2}dx = \frac{2}{5} + 2c_0 - \frac{2}{3}c_1 = 0, \\
\int_0^1 x(c_1 - c_1 x + x^2) x^{-1/2}dx = \frac{2}{7} + \frac{2}{3}c_0 - \frac{2}{5}c_1 = 0.
\]
Hence \( c_1 = \frac{6}{7}, c_0 = \frac{3}{35} \), and the orthogonal polynomial is
\[
\pi(x) = \frac{3}{35} \frac{6}{7} + x^2
\]
with its zeros at
\[
x = \frac{1}{7} \left( 3 - 2 \frac{6}{\sqrt{5}} \right), \quad x = \frac{1}{7} \left( 3 + 2 \frac{6}{\sqrt{5}} \right).
\]
Remark 5.14 (A comparison of Examples 5.5 and 5.13). The degree of exactness of the trapezoidal rule is 1 while that of the two-point Gauss formula is 3. Hence we expect that the Gauss formula is much more accurate. Indeed, calculate errors of the two formulas (5.12) and (5.28) for \( f(x) = \cos \left( \frac{1}{2} \pi x \right) \) and we have
\[
E_1^T = 0.226453 \ldots, \\
E_2^G = 0.002197 \ldots.
\]

Theorem 5.23. The zeros of real orthogonal polynomials over \( [a, b] \) are real, simple, and are inside \( (a, b) \).

Proof. For fixed \( n \geq 1 \), suppose \( p_n(x) \) does not change sign in \( [a, b] \). Then \( \int_a^b \rho(x)p_n(x)dx = (p_n, p_0) > 0 \). 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 = \( p_n(x), \frac{p_n(x)}{(x-x_1)^2} \). \( \langle (1, \rho(x)), 0 \rangle > 0 \), which is false. Therefore every zero is simple.

Suppose that only \( j \leq n \) zeros of \( p_n \), say \( x_1, x_2, \ldots, x_j \), are in \( [a, b] \) and all others zeros are out of \( [a, b] \). Let \( v_1(x) = \prod_{i=1}^j (x - x_i) \in \mathbb{P}_{n-j} \). Then \( p_n v_j = \prod_{n-j} v_j \) where \( P_{n-j} \) is a polynomial of degree \( n \). \( P_{n-j} \) does not change sign on \( [a, b] \). Hence \( \langle (P_{n-j}, v_j), 0 \rangle > 0 \), which contradicts the orthogonality of \( p_n(x) \) and \( v_j(x) \).

Corollary 5.24. All nodes of a Gauss formula are real, distinct, and contained in \( (a, b) \).

Proof. This follows from Definition 5.20 and Theorem 5.23.

Lemma 5.25. Gauss formulas have positive weights.

Proof. For each \( j = 1, 2, \ldots, n \), the definition of \( \ell_j(x) \) in (5.7) implies \( \ell_j^2 \in \mathbb{P}_{2n-2} \), then \( d_k = 2n - 1 \) implies
\[
0 < \int_a^b \ell_j^2(x)w(x)dx = \sum_{k=1}^n w_k \ell_j^2(x_k) = w_j,
\]
which completes the proof.

Lemma 5.26. Gauss formulas satisfy \( \sum_{k=1}^n w_k = \mu_0 > 0 \).

Proof. This follows from setting \( j = 0 \) in (5.3) and applying the condition on \( \rho \) in Definition 5.1.

Theorem 5.27. Gauss formulas are convergent for \( C[a, b] \).

Proof. Denote by \( \mathbb{P} \) the set of real polynomials. Theorem 5.6 states that \( \mathbb{P} \) is dense in \( C[a, b] \), i.e. condition (a) in Theorem 5.7 holds. Condition (b) also holds because of Lemma 5.26, (5.3), and \( \rho \in L[a, b] \). The rest of the proof follows from Theorem 5.7.

Theorem 5.28. For \( f \in C^n[a, b] \), the remainder of a Gauss formula \( I_n(f) \) satisfies
\[
\exists \xi \in [a, b] \text{ s.t. } \quad E_n^G(f) = \int_a^b \rho(x)c_n^2(x)dx,
\]
where \( c_n \) is the node polynomial that defines \( I_n \).

Proof. Not required.