NOTES ON J. STOER AND R. BULIRSCH: *INTRODUCTION TO NUMERICAL ANALYSIS*

ALEC WENDLAND

Notes on chapters 2-4 of *Introduction to Numerical Analysis* by J. Stoer and R. Bulirsch [1]. Compiled for MATH 5510: Introduction to Numerical Analysis I at the University of Connecticut.

Contents

1. Interpolation

In this section we consider a family of functions of a single variable x,

$$\Phi(x; a_0, a_1, \dots, a_n),$$

having $n+1$ parameters $a_0, a_1, \dots, a_n$, whose values characterize the individual functions in this family. The interpolation problem for Φ consists of determining these parameters a_i, $i = 0, 1, \dots, n$, so that for $n+1$ given real or complex pairs of numbers (x_i, f_i), $i = 0, 1, \dots, n$, with $x_i \neq x_j$ for $i \neq j$, we have

$$\Phi(x_i; a_0, a_1, \dots, a_n) = f_i, \quad i = 0, 1, \dots, n.$$

Date: March 3, 2023.

Definition 1.0.0.1 (Support Points)**.** *The pairs* (x_i, f_i) *are called* **support points***.*

Definition 1.0.0.2 (Support Abscissas)**.** *The locations* x_i *are called* **support abscissas***.*

Definition 1.0.0.3 (Support Ordinates)**.** *The values* f_i *are called* **support ordinates***.*

Example 1.0.0.4 (Linear Interpolation Problem)**.** *A* **linear interpolation problem** *is an interpolation problem where* Φ *is linear in the parameters* a_i*,* $i = 0, 1, \ldots, n$*, that is,*

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_0\Phi_0(x) + a_1\Phi_1(x) + \cdots + a_n\Phi_n(x).$$

Example 1.0.0.5 (Polynomial Interpolation)**. Polynomial Interpolation** *is a linear interpolation such that*

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_0 + a_1x + +a_2x^2 + \cdots + a_nx^n.$$

Example 1.0.0.6 (Trigonometric Interpolation)**. Trigonometric interpolation** *is a linear interpolation such that*

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_0 + a_1e^{ix} + a_2e^{2ix} + \cdots + a_ne^{nix},$$

where $i^2 := -1$*.*

1.1. **Polynomial Interpolation.**

1.1.1. *The Interpolation Formula of Lagrange.*

Remark. *We denote by* Π^n *the set of all real or complex polynomials* p *of degree* n *or less, that is,*

$$\Pi^n := \{p : p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, a_i, x \in \mathbb{C}, i = 0, 1, \ldots, n\}.$$

Theorem 1.1.1.1 (Existence and Uniqueness of Polynomial Interpolant)**.** *For* $n+1$ *arbitrary support points*

$$(x_i, f_i), \quad i = 0, 1, \ldots, n, \quad x_i \neq x_j \text{ for } i \neq j,$$

there exists a unique polynomial $p \in \Pi^n$ *such that*

$$p(x_i) = f_i, \quad i = 0, 1, \ldots, n.$$

Proof. We first establish existence by construction. Define

$$\omega(x) := \prod_{j=0}^{n}(x - x_j)$$

and note that $\omega(x_i) = 0$ for $i = 0, \ldots, n$. By the product rule, it follows

$$\begin{aligned}
\omega'(x) &= \frac{d}{dx}\left[\prod_{j=0}^{n}(x - x_j)\right] \\
&= (x - x_0)(x - x_1)\ldots(x - x_{n-1}) + (x - x_0)(x - x_1)\ldots(x - x_{n-2})(x - x_n) + \cdots + \\
&\quad (x - x_0)(x - x_2)\ldots(x - x_n) + (x - x_1)(x - x_2)\ldots(x - x_n) \\
&= \sum_{i=0}^{n}\prod_{\substack{j=0 \\ j\neq i}}^{n}(x - x_j).
\end{aligned}$$

Note $\omega'(x_i) \neq 0$ for each $i = 0, \ldots, n$. Define the Lagrange polynomial basis as follows:

$$L_i(x) := \begin{cases} \dfrac{\omega(x)}{(x - x_i)\omega'(x_i)}, & x \neq x_i, \\ 1, & x = x_i. \end{cases} \tag{1.1.1.1}$$

We see that $\{L_i\}_{i=0}^n$ satisfy the conditions

$$L_i(x_k) = \delta_{ik} = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases}$$

In the case that $x \neq x_i$, expanding L_i gives

$$\begin{aligned}
L_i(x) &= \frac{\omega(x)}{(x - x_i)\omega'(x_i)} \\
&= \frac{(x - x_0)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)},
\end{aligned}$$

so that $L_i \in \Pi^n$ for $i = 0, \ldots, n$. Lastly, define the polynomial p as follows:

$$p(x) := \sum_{i=0}^{n} f_i L_i(x) = \sum_{i=0}^{n} f_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{1.1.1.2}$$

Observe, since p is a linear combination of polynomials of degree at most n, we have evidently $p \in \Pi^n$. Moreover,

$$p(x_k) = \sum_{i=0}^{n} f_i L_i(x_k) = f_k, \quad k = 0, 1, \ldots, n,$$

so that p is a polynomial of degree at most n that interpolates f. This establishes existence.

We now show that such an interpolating polynomial p is unique. Suppose that there exist $p_1, p_2 \in \Pi^n$ such that

$$p_1(x_i) = f_i = p_2(x_i), \quad i = 0, \ldots, n.$$

Define

$$p^* := p_1 - p_2$$

and note $p^* \in \Pi^n$. Since $p^*(x_i) = p_1(x_i) - p_2(x_i) = 0$ for each $i = 0, \ldots, n$, we see that p^* has the $(n+1)$ distinct zeros at the support abscissas $\{x_i\}_{i=0}^n$. Since p^* is a polynomial of degree at most n, p^* must vanish identically,

$$p^* \equiv 0.$$

It follows $p_1 \equiv p_2$. □

Remark. *We call the formula given by*

$$p(x) := \sum_{i=0}^{n} f_i L_i(x) = \sum_{i=0}^{n} f_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \tag{1.1.1.3}$$

the **Lagrange interpolation formula**.

Remark. *Note that the coefficients of p depend linearly on the support ordinates f_i.*

Example 1.1.1.2. *Let*

x_i	0	1	3
f_1	1	3	2

We construct the unique polynomial $p \in \Pi^2$ such that $p(x_i) = f_i$, $i = 0, 1, 2$. Observe

$$\begin{aligned}
p(x) &= \sum_{i=0}^{2} f_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \\
&= f_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\
&= (1)\frac{(x-1)(x-3)}{(-1)(-3)} + (3)\frac{x(x-3)}{(1)(-2)} + (2)\frac{x(x-1)}{(3)(2)} \\
&= \frac{1}{3}(x^2 - 4x + 3) - \frac{3}{2}(x^2 - 3x) + \frac{1}{3}(x^2 - x) \\
&= -\frac{5}{6}x^2 + \frac{17}{6}x + 1.
\end{aligned}$$

1.1.2. *Neville's Algorithm.* Lagrange's interpolation formula (1.1.1.3) solves the full interpolation problem all at once. Instead, we can solve the problem for smaller sets of support points first and update the solutions to obtain the solution to the full interpolation problem.

Theorem 1.1.2.1 (Neville's Algorithm)**.** *Let* $\{(x_i, f_i)\}_{i=0}^n$ *be a set of support points and denote by*

$$p_{i_0 i_1 \dots i_k} \in \Pi^k$$

the unique polynomial of degree at most k *such that*

$$p_{i_0 i_1 \dots i_k}(x_{i_j}) = f_{i_j}, \quad j = 0, 1, \dots, k, \quad k = 0, 1, \dots, n.$$

Then each $p_{i_0 i_1 \dots i_k}$ *is given by the following recursion:*

(1) $p_i(x) := f_i$,

(2) $p_{i_0 i_1 \dots i_k}(x) = \dfrac{(x - x_{i_0}) p_{i_1 i_2 \dots i_k}(x) - (x - x_{i_k}) p_{i_0 i_1 \dots i_{k-1}}(x)}{x_{i_k} - x_{i_0}}$.

Proof. First note that, given one support point (x_i, f_i), $i = 0, \dots, n$, the unique polynomial $p_i \in \Pi^0$ that interpolates f_i is simply the constant polynomial $p_i(x) := f_i$.

Consider now the case of at least two support points (x_i, f_i). Define

$$r(x) := \frac{(x - x_{i_0}) p_{i_1 i_2 \dots i_k}(x) - (x - x_{i_k}) p_{i_0 i_1 \dots i_{k-1}}(x)}{x_{i_k} - x_{i_0}}.$$

Noting that r is a linear combination of two polynomials of degree k or less, it follows that $r \in \Pi^k$.

We now show that r has the characteristic properties of $p_{i_0 i_1 \dots i_k}$, in particular, that $r(x_{i_j}) = f_{i_j}$, $j = 0, 1, \dots, k$. Observe that

$$\begin{aligned} r(x_{i_0}) &= \frac{-(x_{i_0} - x_{i_k}) p_{i_0 i_1 \dots i_{k-1}}(x_{i_0})}{x_{i_k} - x_{i_0}} \\ &= p_{i_0 i_1 \dots i_{k-1}}(x_{i_0}) = f_{i_0}, \end{aligned}$$

$$\begin{aligned} r(x_{i_k}) &= \frac{(x_{i_k} - x_{i_0}) p_{i_1 i_2 \dots i_k}(x_{i_k})}{x_{i_k} - x_{i_0}} \\ &= p_{i_1 i_2 \dots i_k}(x_{i_k}) = f_{i_k}, \end{aligned}$$

by the assumption. Moreover, for $j = 1, 2, \dots, k-1$, we have

$$\begin{aligned} r(x_{i_j}) &= \frac{(x_{i_j} - x_{i_0}) p_{i_1 i_2 \dots i_k}(x_{i_j}) - (x_{i_j} - x_{i_k}) p_{i_0 i_1 \dots i_{k-1}}(x_{i_j})}{x_{i_k} - x_{i_0}} \\ &= \frac{(x_{i_j} - x_{i_0}) f_{i_j} - (x_{i_j} - x_{i_k}) f_{i_j}}{x_{i_k} - x_{i_0}} \\ &= f_{i_j}. \end{aligned}$$

That is, r interpolates f at each of i_j, $j = 0, 1, \dots, k$. By the uniqueness of polynomial interpolation, it follows that $r \equiv p_{i_0 i_1 \dots i_k}$. □

Remark. *Neville's algorithm is well-suited for determining the value of the polynomial interpolant p for a single value of x. It is not preferable, however, when multiple evaluations of p are needed.*

We make the observation here that Neville's algorithm produces a symmetric tableau of the values of the (partially) interpolating polynomials $p_{i_0 i_1 \dots i_k}$ for a fixed x : (shown here is the case $k = 3$)

	$k=0$	1	2	3
x_0	$f_0 =: p_0(x)$			
		$p_{01}(x)$		
x_1	$f_1 =: p_1(x)$		$p_{012}(x)$	
		$p_{12}(x)$		$p_{0123}(x)$
x_2	$f_2 =: p_2(x)$		$p_{123}(x)$	
		$p_{23}(x)$		
x_3	$f_3 =: p_3(x)$			

Example 1.1.2.2. *Given*

x_i	0	1	3
f_1	1	3	2

we have

	$k=0$	2	3
$x_0 = 0$	$f_0 =: p_0(2) = 1$		
		$p_{01}(2) = 5$	
$x_1 = 1$	$f_1 =: p_1(2) = 3$		$p_{012}(2) = \frac{10}{3}$
		$p_{12}(2) = \frac{5}{2}$	
$x_2 = 3$	$f_2 =: p_2(2) = 2$		

Note also that this evaluation of $p_{012}(2)$ coincides with the evaluation of $p(2)$ in the example from the previous section.

1.1.3. *Newton's Interpolation Formula: Divided Differences.* Neville's algorithm (1.1.2.1) is aimed at determining specific values of the polynomial interpolant rather than the symbolic polynomial itself. If the polynomial itself is preferred, or if we want to evaluate several arguments ξ_j of the polynomial interpolant simultaneously, then Newton's interpolation formula is preferred.

Given the $n+1$ support points $\{(x_i, f_i)\}_{i=0}^n$, recall from (1.1.1.1) that there exists a unique polynomial $p \in \Pi^n$ that interpolates the points. Write

$$\begin{aligned} p(x) &:= a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \\ &= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \cdots + c_n(x - x_0)(x - x_1)\dots(x - x_{n-1}) \\ &= c_0 + (x - x_0)(c_1 + c_2(x - x_1) + \cdots + (x - x_{n-2})(c_{n-1} + c_n(x - x_{n-1}))\dots). \end{aligned}$$

This setup brings us to the so-called **Horner scheme** for efficiently evaluating polynomials.

Definition 1.1.3.1. *[Horner Scheme] Let $p \in \Pi^n$. Then the Horner scheme for evaluating p at an arbitrary point ξ is given by*

$$p(\xi) = a_0 + (\xi - x_0)(a_1 + +a_2(\xi - x_1) + \cdots + (\xi - x_{n-2})(a_{n-1} + a_n(\xi - x_{n-1}))\dots).$$

It remains to determine the coefficients in (1.1.3.1). One method is as follows:

$$\begin{aligned} f_0 &= p(x_0) = a_0, \\ f_1 &= p(x_1) = a_0 + a_1(x - x_0), \\ f_2 &= p(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1), \\ &\vdots \end{aligned}$$

This can be done with n divisions and $n(n-1)$ multiplications. However, there is a better method which requires only $\frac{n(n+1)}{2}$ divisions and produces useful intermediate results.

Note that, since $p_{i_0 i_1 \dots i_{k-1}}(x)$ and $p_{i_0 i_1 \dots i_k}(x)$ both interpolate the k support points $\{(x_{i_j}, f_{i_j})\}_{j=0}^{k}$, they differ by a polynomial of degree k with the k zeros $x_{i_0}, x_{i_1}, \dots, x_{i_{k-1}}$. Thus there exists a unique coefficient

$$f_{i_0 i_1 \dots i_k}$$

such that

$$p_{i_0 i_1 \dots i_k}(x) = p_{i_0 i_1 \dots i_{k-1}}(x) + f_{i_0 i_1 \dots i_k} \prod_{j=0}^{k-1} (x - x_{i_j}).$$

From this and the fact that $p_{i_0} := f_{i_0}$, it follows

$$p_{i_0 i_1 \dots i_k}(x) = f_{i_0} - f_{i_0 i_1}(x - x_{i_0}) + \cdots + f_{i_0 i_1 \dots i_k}(x - x_{i_0}) \dots (x - x_{i_{k-1}}).$$

We call this form the **Newton representation** of the interpolating polynomial $p_{i_0 i_1 \dots i_k}(x)$. The coefficients are called the $k-$th divided differences.

Theorem 1.1.3.2 (Newton's Interpolation Formula)**.** *Let $\{(x_i, f_i)\}_{i=0}^{n}$ be a set of support points. Define the following recursion*

(1) $f[x_i] := f_i$,

(2) $f[x_i, x_j] = \dfrac{f[x_j] - f[x_i]}{x_j - x_i}$,

(3) $f[x_{i_0}, x_{i_1}, \dots, x_{i_k}] := \dfrac{f[x_{i_1}, x_{i_2}, \dots, x_{i_k}] - f[x_{i_0}, x_{i_1}, \dots, x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}$.

Then the unique polynomial $p \in \Pi^n$ such that

$$p(x_i) = f_i, \quad i = 0, 1, \dots, n$$

is given by

$$p(x) := f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, x_1 \dots, x_n](x - x_0)(x - x_1) \dots (x - x_{n-1}).$$

Proof. We use induction. If $n = 0$, then $p(x) = f[x_0] = f_0$, as desired. In the case that $n = 1$, we have

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0).$$

Observing that

$$p(x_0) = f[x_0] = f_0$$

and

$$p(x_1) = f[x_0] + f[x_0, x_1](x_1 - x_0) = f[x_0] + \frac{f[x_1] - f[x_0]}{x_1 - x_0}(x_1 - x_0) = f_1,$$

we see that (1.1.3.2) holds for $n = 0, 1$.

Assume now that (1.1.3.2) holds for all $n = 0, 1, \ldots, k-1$. We show that the result holds for $n = k$.

Recall from (1.1.2.1) that the unique interpolating polynomial $p := p_{0,1,\ldots,k} \in \Pi^k$ is given by the recursion

$$p(x) = \frac{(x - x_0)p_{1,2,\ldots,k}(x) - (x - x_k)p_{0,1,\ldots,k-1}(x)}{x_k - x_0}. \tag{1.1.3.1}$$

Define a polynomial

$$r := p - p_{0,1,\ldots,k-1}.$$

Since both p and $p_{0,1,\ldots,k-1}$ interpolate the k support points $\{(x_i, f_i)\}_{i=0}^{k-1}$, it follows that r is a polynomial with k distinct roots, and, since r is a linear combination of two polynomials of degree k or less, it follows that r is a polynomial of degree precisely k. Therefore, there exists a unique coefficient a_k such that

$$r(x) = a_k \prod_{i=0}^{k-1}(x - x_i).$$

Now note that

$$\begin{aligned} p(x) &= p_{0,1,\ldots,k-1}(x) + r(x) \\ &= \frac{(x - x_0)p_{1,2,\ldots,k-1}(x) - (x - x_{k-1})p_{0,1,\ldots,k-2}(x)}{x_{k-1} - x_0} + a_k \prod_{i=0}^{k-1}(x - x_i). \end{aligned} \tag{1.1.3.2}$$

Noting that $p_{0,1,\ldots,k-1}, p_{1,2,\ldots,k} \in \Pi^{k-1}$, it follows from the induction hypothesis and (1.1.3.1) that

$$\begin{aligned} p(x) &= \frac{(x - x_0)p_{1,2,\ldots,k}(x) - (x - x_k)p_{0,1,\ldots,k-1}(x)}{x_k - x_0} \\ &= \left(\frac{x - x_0}{x_k - x_0}\right) p_{1,2,\ldots,k}(x) - \left(\frac{x - x_{k-1}}{x_k - x_0}\right) p_{0,1,\ldots,k-1}(x) + \left(\frac{x_k - x_{k-1}}{x_k - x_0}\right) p_{0,1,\ldots,k-1}(x) \\ &= \left(\frac{x - x_0}{x_k - x_0}\right) (f[x_1] + f[x_1, x_2](x - x_1) + \cdots + f[x_1, x_2, \ldots, x_k](x - x_1)\ldots(x - x_{k-1})) \\ &\quad - \left(\frac{x - x_{k-1}}{x_k - x_0}\right) (f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, x_1, \ldots, x_{k-1}](x - x_0)\ldots(x - x_{k-2})) \\ &\quad + \left(\frac{x_k - x_{k-1}}{x_k - x_0}\right) p_{0,1,\ldots,k-1}(x) \\ &= \left(\frac{1}{x_k - x_0}\right) (f[x_1, \ldots, x_k](x - x_0)\ldots(x - x_{k-1}) - f[x_0, \ldots, x_{k-1}](x - x_0)\ldots(x - x_{k-1})) \\ &\quad + \left(\frac{x - x_0}{x_k - x_0}\right) (f[x_1] + f[x_1, x_2](x - x_1) + \cdots + f[x_1, \ldots, x_{k-1}](x - x_1)\ldots(x - x_{k-2})) \\ &\quad - \left(\frac{x - x_{k-1}}{x_k - x_0}\right) (f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_{k-2}](x - x_0)\ldots(x - x_{k-3}) \\ &\quad + \left(\frac{x_k - x_{k-1}}{x_k - x_0}\right) p_{0,1,\ldots,k-1}(x) \end{aligned}$$

$$
\begin{aligned}
&= \frac{f[x_1,\dots,x_k]-f[x_0,\dots,x_{k-1}]}{x_k-x_0}\left(\prod_{i=0}^{k-1}(x-x_i)\right)\\
&\quad+\left(\frac{x-x_0}{x_k-x_0}\right)p_{1,2,\dots,k-1}(x)-\left(\frac{x-x_{k-1}}{x_k-x_0}\right)p_{0,1,\dots,k-2}(x)+\left(\frac{x_k-x_{k-1}}{x_k-x_0}\right)p_{0,1,\dots,k-1}(x)\\
&= \frac{f[x_1,\dots,x_k]-f[x_0,\dots,x_{k-1}]}{x_k-x_0}\left(\prod_{i=0}^{k-1}(x-x_i)\right)\\
&\quad+\left(\frac{1}{x_k-x_0}\right)(x-x_0)p_{1,2,\dots,k-1}(x)-(x-x_{k-1})p_{0,1,\dots,k-2}(x)+(x_k-x_{k-1})p_{0,1,\dots,k-1}(x)\\
&= \frac{f[x_1,\dots,x_k]-f[x_0,\dots,x_{k-1}]}{x_k-x_0}\left(\prod_{i=0}^{k-1}(x-x_i)\right)\\
&\quad+\left(\frac{1}{x_k-x_0}\right)(x_{k-1}-x_0)p_{0,1,\dots,k-1}(x)+(x_k-x_{k-1})p_{0,1,\dots,k-1}(x)\\
&= \frac{f[x_1,\dots,x_k]-f[x_0,\dots,x_{k-1}]}{x_k-x_0}\left(\prod_{i=0}^{k-1}(x-x_i)\right)+p_{0,1,\dots,k-1}(x).
\end{aligned}
$$

Finally, it follows from (1.1.3.2) that

$$a_k=\frac{f[x_1,\dots,x_k]-f[x_0,\dots,x_{k-1}]}{x_k-x_0},$$

which completes the proof. □

Recall that the polynomial $p_{i_0 i_1\dots i_k}(x)$ is uniquely determined by the support points in interpolates, so that the polynomial is invariant to any permutation of the indices $i_0, i_1, \dots, i_k$.

Theorem 1.1.3.3. *The divided differences $f[x_{i_0}, x_{i_1}, \dots, x_{i_k}]$ are invariant to permutations of the indices $i_0, i_1, \dots, i_k$. More precisely, if*

$$(j_0, j_1, \dots, j_k) = (i_{s_0}, i_{s_1}, \dots, i_{s_k})$$

is a permutation of the indices $i_0, i_1, \dots, i_k$, then

$$f[x_{j_0}, \dots, x_{j_k}] = f[x_{i_0}, \dots, x_{i_k}].$$

We defer the proof for a later result.

Calculating the divided differences in analogy to Neville's algorithm gives the following tableau, called the **divided–difference scheme**:

	$k=0$	$k=1$	$k=2$	$\dots$
x_0	$f[x_0]$			
		$f[x_0,x_1]$		
x_1	$f[x_1]$		$f[x_0,x_1,x_2]$	
		$f[x_1,x_2]$	$\vdots$	$\ddots$
x_2	$f[x_2]$	$\vdots$		
$\vdots$	$\vdots$			

Note the entries in the second column are given by

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}, \quad f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1}, \quad , \ldots,$$

those in the third column

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}, \quad f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}, \quad \ldots,$$

and obviously

$$\begin{aligned} p(x) &= p_{0,1,\ldots,n}(x) \\ &= f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \ldots (x - x_{n-1}) \end{aligned}$$

is the desired solution to the interpolation problem. The coefficients of the above expression for $p(x)$ are obtained in the top descending diagonal of the divided–difference scheme.

Example 1.1.3.4. *With the same numbers from the previous sections,*

x_i	0	1	3
f_1	1	3	2

we have

	$k=0$	$k=1$	$k=2$
$x_0 = 0$	$f[x_0] = 1$		
		$f[x_0, x_1] = 2$	
$x_1 = 1$	$f[x_1] = 3$		$f[x_0, x_1, x_2] = -\frac{5}{6}$
		$f[x_1, x_2] = -\frac{1}{2}$	
$x = 3$	$f[x_2] = 2$		

Thus

$$\begin{aligned} p(x) &= 1 + 2x - \frac{5}{6}x(x-1) \\ &= -\frac{5}{6}x^2 + \frac{17}{6}x + 1, \end{aligned}$$

which coincides with the results from the previous examples.

Note that frequently the support ordinates f_i are values $f_i := f(x_i)$ of a given function $f(x)$, which we want to approximate by interpolation. The divided differences can then be viewed as multivariate functions of the support abscissas x_i,

$$f[x_{i_0}, x_{i_1}, \ldots, x_{i_k}].$$

We get the following result, which we prove here.

Theorem 1.1.3.5. *The divided differences*

$$f[x_{i_0}, \ldots, x_{i_k}]$$

are symmetric functions of their arguments, that is, they are invariant to permutations of the support abscissas $x_{i_0}, \ldots, x_{i_k}$.

Proof. We use induction.

If $n = 0$, then there is only one support point (x_0, f_0), and clearly

$$f[x_0] = f_0.$$

In the case $n = 1$, we observe

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0} = \frac{f_0 - f_1}{x_0 - x_1} = \frac{f[x_0] - f[x_1]}{x_0 - x_1} = f[x_1, x_0].$$

Suppose that (1.1.3.5) holds for some k, that is,

$$f[x_{i_0}, \ldots, x_{i_k}] = f[x_{j_0}, \ldots, x_{j_k}]$$

for any permutation $(j_0, \ldots, j_k)$ of $(i_0, \ldots, i_k)$. Considering

$$f[x_{i_0}, \ldots, x_{i_k}, x_{i_{k+1}}],$$

we see that

$$\begin{aligned} f[x_{i_0}, \ldots, x_{i_k}, x_{i_{k+1}}] &= \frac{f[x_{i_1}, \ldots, x_{i_{k+1}}] - f[x_{i_0}, \ldots, x_{i_k}]}{x_{i_{k+1}} - x_{i_0}} \\ &= \frac{f[x_{i_0}, \ldots, x_{i_k}] - f[x_{i_1}, \ldots, x_{i_{k+1}}]}{x_{i_0} - x_{i_{k+1}}} \\ &= f[x_{i_0}, \ldots, x_{i_{k+1}}, x_{i_k}]. \end{aligned}$$

Since $f[x_{i_0}, \ldots, x_{i_k}]$ and $f[x_{i_1}, \ldots, x_{i_{k+1}}]$ are invariant to permutations of the indices $i_0, \ldots, i_{k+1}$ by the hypothesis, (1.1.3.5) follows. □

The last result of this section applies when the function $f(x)$ is itself a polynomial.

Theorem 1.1.3.6. *If $f(x)$ is a polynomial of degree N, then*

$$f[x_0, \ldots, x_k] = 0$$

for $k > N$.

Proof. Let $k > N$ and let $\{(x_i, f_i)\}_{i=0}^k$ be any $k + 1$ support points. Then p is a polynomial of degree k or less that interpolates f. Noting that f is itself a polynomial of degree N that interpolates the k support points, it follows by uniqueness that $p \equiv f$.

By uniqueness, it follows that the coefficient of x^k in $p(x)$ must vanish for all $k > N$. This coefficient is given by $f[x_0, \ldots, x_k]$ (1.1.3.2), so that evidently

$$f[x_0, \ldots, x_k] = 0$$

for all $k > N$. □

1.1.4. *The Error in Polynomial Interpolation.* Once again we are given a function $f(x)$ and support points $\{(x_i, f_i)\}_{i=0}^n$, and we interpolate f with the interpolating polynomial $p \in \Pi^n$ such that

$$p(x_i) = f_i, \quad i = 0, 1, \ldots, n.$$

We are interested in how well $p(x)$ reproduces $f(x)$ for arguments x different from the support abscissas x_i, $i = 0, 1, \ldots, n$. Note that the error

$$e(x) := f(x) - p(x)$$

can clearly become arbitrarily large for functions f unless some restrictions are imposed on f.

We first recall Rolle's Theorem, which is essential to the proof of the polynomial interpolation error formula.

Theorem 1.1.4.1 (Rolle's Theorem)**.** *Let f be continuous on $[a,b]$ and differentiable on (a,b), and suppose that $f(a) = f(b)$. Then there exists $\xi \in (a,b)$ such that*

$$f'(\xi) = 0.$$

Proof. If $f(x) \equiv k$ for some $k \in \mathbb{R}$ and all $x \in (a,b)$, then f is constant, and clearly $f'(\xi) \equiv 0$ for all $\xi \in (a,b)$.

We now consider the case that $f(x) \neq f(a)$ for some $x \in (a,b)$. Passing to the consideration of $-f(x)$ as necessary, we assume that $f(x) > f(a)$. Since f is continuous on $[a,b]$, it follows from the extreme value theorem that f attains a maximum value $f(\xi)$ at some $\xi \in [a,b]$. Since $f(x) > f(a)$ for some $x \in (a,b)$, we have evidently $a < \xi < b$, so that $\xi \in (a,b)$. Thus f has a local maximum at $\xi \in (a,b)$. Since f is differentiable, it follows by Fermat's Theorem that $f'(\xi) = 0$. This completes the proof. □

Theorem 1.1.4.2 (Error in Polynomial Interpolation)**.** *Let $p \in \Pi^n$ be the unique polynomial interpolant of f. If the function f has an $(n+1)-st$ derivative, then for every argument $\overline{x}$ there exists a number ξ in the smallest interval $I[x_0, \ldots, x_n; \overline{x}]$ which contains $\overline{x}$ and all support abscissas x_i, $i = 0, 1, \ldots, n$, satisfying*

$$e(\overline{x}) = f(\overline{x}) - p(\overline{x}) = \frac{\omega(\overline{x}) f^{(n+1)}(\xi)}{(n+1)!},$$

where

$$\omega(x) := \prod_{i=0}^{n} (x - x_i).$$

Proof. Let $p \in \Pi^n$ be the unique polynomial that interpolates f at x_i, $i = 0, 1, \ldots, n$.

First, if $\overline{x} = x_i$ for some $i = 0, 1, \ldots, n$, then

$$e(\overline{x}) = e(x_i) = f(x_i) - p(x_i) = 0.$$

Assume now that $\overline{x} \neq x_i$ for all $i = 0, 1, \ldots, n$. Then $\omega(\overline{x}) \neq 0$, so that we may define the constant

$$K := \frac{f(\overline{x}) - p(\overline{x})}{\omega(\overline{x})}.$$

Further, define a function F by

$$F(x) := f(x) - p(x) - K\omega(x).$$

Note that F vanishes for $x = \overline{x}$ and $x = x_i$, $i = 0, 1, \ldots, n$, so consequently F has at least the $n+2$ distinct zeros

$$x_0, x_1, \ldots, x_n, \overline{x}$$

in the interval $I[x_0, x_1, \ldots, x_n; \overline{x}]$. By Rolle's Theorem (1.1.4.1), applied repeatedly, it follows that F' has at least $n+1$ zeros in $I[x_0, x_1, \ldots, x_n; \overline{x}]$, F'' has at least n zeros, and, finally, $F^{(n+1)}$ has at least one zero $\xi \in I[x_0, x_1, \ldots, x_n; \overline{x}]$.

Observe, since $\omega(x)$ is a polynomial of degree precisely $n+1$, we have

$$\omega^{(n+1)}(x) = (n+1)!.$$

Moreover, since p is a polynomial of degree at most n, consequently

$$p^{(n+1)}(x) \equiv 0.$$

Thus

$$F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - p^{(n+1)}(\xi) - K\omega^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K(n+1)! = 0.$$

Rearranging gives

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Finally, since $f(\overline{x}) - p(\overline{x}) - K\omega(\overline{x}) = 0$, we have

$$f(\overline{x}) - p(\overline{x}) = K\omega(\overline{x}) = \frac{\omega(\overline{x})f^{(n+1)}(\xi)}{(n+1)!},$$

which completes the proof. □

The following theorem gives a different error term, derived from Newton's interpolation formula (1.1.3.2).

Theorem 1.1.4.3. *If the function f has an $(n+1)-st$ derivative, then for every argument $\overline{x}$ there exists a number ξ in the smallest interval $I[x_0, x_1, \ldots, x_n; \overline{x}]$ containing $\overline{x}$ and all support abscissas such that*

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

Proof. In addition to the $n+1$ support points $\{(x_i, f_i)\}_{i=0}^{n}$, introduce an $(n+2)-$nd support point (x_{n+1}, f_{n+1}) by

$$x_{n+1} := \overline{x}, \quad f_{n+1} := f(\overline{x}).$$

Denote by $p_{0,1,\ldots,n}(x) \in \Pi^n$ the unique polynomial interpolant of f. Then, by Newton's interpolation formula (1.1.3.2), we have

$$f(\overline{x}) = p_{0,1,\ldots,n+1}(\overline{x}) = p_{0,1,\ldots,n}(\overline{x}) + f[x_0, x_1, \ldots, x_n; \overline{x}]\omega(\overline{x}).$$

Rearranging,

$$f(\overline{x}) - p_{0,1,\ldots,n}(\overline{x}) = f[x_0, x_1, \ldots, x_n; \overline{x}]\omega(\overline{x}).$$

Since

$$f(\overline{x}) - p(\overline{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)}\omega(\overline{x})$$

for some $\xi \in I[x_0, x_1, \ldots, x_n; \overline{x}]$, it follows

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

This also implies

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

which completes the proof. □

Example 1.1.4.4. *Let $f(x) := \sin(x)$ and let $p \in \Pi^3$ interpolate f at*

$$x_j := \frac{\pi}{3}j, \quad j = 0, 1, 2, 3.$$

We derive an upper bound for the error on the interval $[0, \pi]$. Noting that

$$f^{(4)}(x) = \sin(x),$$

we have

$$\begin{aligned} f(\overline{x}) - p(\overline{x}) &= \frac{f^{(4)}(\xi)}{4!}\omega(\overline{x}) \\ &= \frac{\sin(\xi)}{24} x \left(x - \frac{\pi}{3}\right)\left(x - \frac{2\pi}{3}\right)(x - \pi). \end{aligned}$$

Observe that

$$|x|, |x - \pi| \leq \pi$$

and

$$\left|x - \frac{\pi}{3}\right|, \left|x - \frac{2\pi}{3}\right| \leq \frac{2\pi}{3}$$

for all $\overline{x} \in [0, \pi]$. *Hence,*

$$\begin{aligned} |f(\overline{x}) - p(\overline{x})| &\leq \frac{|\sin(\xi)|}{24}\pi^2 \left(\frac{2\pi}{3}\right)^2 \\ &= \frac{\sin(\xi)}{24}\left(\frac{4\pi^4}{9}\right) \\ &= \frac{\pi^4}{54}|\sin(\xi)| \leq \frac{\pi^4}{54} \approx 1.81. \end{aligned}$$

A common usage for the error formula (1.1.4.2) is to bound $|f(x) - p(x)|$ by bounding $f^{(n+1)}(x)$. If the support abscissas $x_0, \ldots, x_n$ are close, say,

$$\max_{j \neq i} |x_i - x_j| := h < 1,$$

and if $x \in I[x_0, x_1, \ldots, x_n]$, then we have

$$|f(x) - p(x)| = \frac{|f^{(n+1)}(x)|}{(n+1)!}\omega(x).$$

If $f^{(n+1)}$ is uniformly bounded, set

$$M := \sup_{x \in I[x_0, x_1, \ldots, x_n]} \frac{|f^{(n+1)}(x)|}{(n+1)!}.$$

Then

$$|f(x) - p(x)| \leq M\omega(x) \leq Mh^{n+1}.$$

Now as $n \to \infty$, we have $h \to 0$, and evidently then $|f(x) - p(x)| \to 0$.

Definition 1.1.4.5 (Extrapolation)**.** *The use of the interpolating polynomial* $p \in \Pi^n$ *for approximating* f *outside of the interval* $I[x_0, x_1, \ldots, x_n]$ *containing the support abscissas is called* **extrapolation**.

Note that the theory guarantees $|f(x) - p(x)| \to \infty$ as x moves farther and farther outside the interval $I[x_0, x_1, \ldots, x_n]$, since $|\omega(x)| \to \infty$ as $|x| \to \infty$.

On the other hand, it should not be assumed that finer and finer samplings of the function f will lead to better and better approximations through interpolation even within the interval $I[x_0, x_1, \ldots, x_n]$.

Consider for example a real–valued function f that is infinitely often differentiable in a given interval $[a, b]$. To every interval partition

$$\Delta := \{a = x_0 < x_1 < \cdots < x_n = b\}$$

there is an interpolating polynomial $p_\Delta \in \Pi^n$ with $p_\Delta(x_i) = f_i$ for each $x_i \in \Delta$. A sequence of interval partitions

$$\Delta_m := \{a = x_0^{(m)} < x_1^{(m)} < \cdots < x_{n_m}^{(m)} = b\}$$

gives rise to a corresponding sequence of interpolating polynomials p_{Δ_m}. One might expect the polynomials p_{Δ_m} to converge to f if the fineness

$$\|\Delta_m\| := \max_i |x_{i+1}^{(m)} - x_i^{(m)}|$$

of the partitions converges to zero as $m \to \infty$. In general, however, this is not true.

Definition 1.1.4.6 (Runge's Phenomenon)**. Runge's phenomenon** *is a problem of oscillation that occurs near x_0 and x_n when using polynomials of high degree with equispaced support points.*

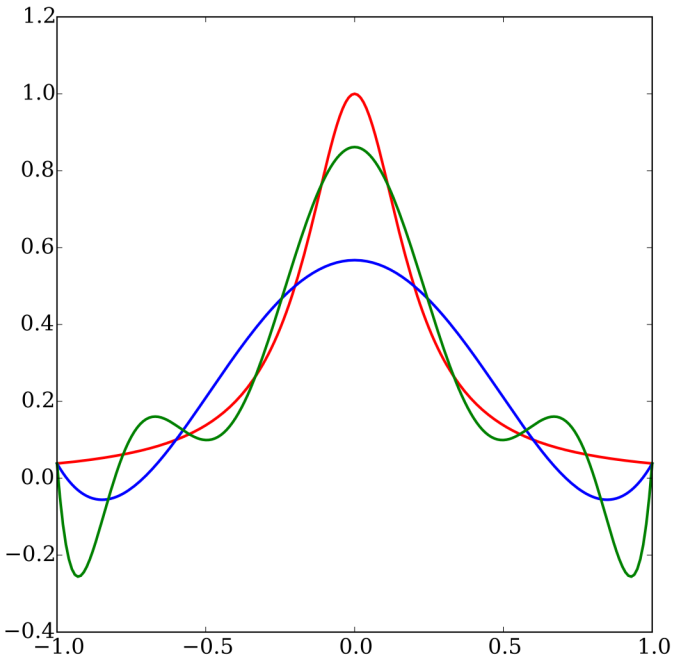

FIGURE 1. Runge's Phenomenon for the Runge Function $f(x) := \dfrac{1}{1+25x^2}$.

Example 1.1.4.7. *Let f be the Runge function*

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

and let $p \in \Pi^n$ interpolate f at the $n+1$ equispaced support points

$$x_j := \frac{2j}{n} - 1, \quad j = 0, 1, \ldots, n.$$

With the equidistant nodes, it may be shown that

$$|\omega(x)| \leq n! h^{n+1},$$

where $h := \frac{2}{n}$ *is defined as the step size.*

Note that the $(n+1)-st$ *derivative of* f *is bounded, so there exists* M_{n+1} *such that*

$$\sup_{-1 \leq x \leq 1} |f^{(n+1)}(x)| \leq M_{n+1}.$$

Thus

$$|f(x) - p(x)| \leq \frac{M_{n+1}}{(n+1)!} n! h^{(n+1)} = \frac{M_{n+1}}{n+1} h^{(n+1)}.$$

But the magnitude of the $(n+1)-st$ *derivative of* f *increases as* n *increases, in particular,* $M_{n+1} \leq (n+1)! 5^{(n+1)}$. *Hence,*

$$|f(x) - p(x)| \leq \frac{(n+1)! 5^{(n+1)}}{n+1} h^{(n+1)} = n! (5h)^{(n+1)} = n! \left(\frac{10}{n}\right)^{n+1},$$

which tends to infinity as n *becomes large.*

1.1.5. *Hermite Interpolation.* We consider the support points

$$\{(x_i, f_i^{(k)})\}_{i=0}^{m}, \quad k = 0, 1, \ldots, n_i - 1,$$

with

$$x_0 < x_1 < \cdots < x_m.$$

The **Hermite interpolation problem** consists of determining a polynomial $p \in \Pi^n$ where

$$n + 1 = \sum_{i=0}^{m} n_i,$$

which satisfies the interpolation conditions

$$p^{(k)}(x_i) = f_i^{(k)}, \quad i = 0, 1, \ldots, m, \quad k = 0, 1, \ldots, n_i - 1.$$

Here we prescribe at each support abscissa not only the value of the function f but also the first $n_i - 1$ derivatives of the polynomial.

We recall the following fundamental result from linear algebra.

Theorem 1.1.5.1 (Invertible Matrix Theorem)**.** *Let* $A : \mathbb{R}^n \to \mathbb{R}^n$ *be linear. Then the following are equivalent:*

(1) $Ax = b$ *is solvable for all* $b \in \mathbb{R}^n$ *(existence),*
(2) $Ax_1 = Ax_2$ *if and only if* $x_1 = x_2$ *(uniqueness),*
(3) $Ax = 0$ *if and only if* $x = 0$ *(*A *is nonsingular),*
(4) A *is invertible.*

The following result establishes existence and uniqueness of the Hermite interpolation problem.

Theorem 1.1.5.2 (Existence and Uniqueness of the Hermite Interpolant)**.** *For arbitrary numbers $x_0 < x_1 < \cdots < x_m$ and $f_i^{(k)}$, $i = 0, 1, \ldots, m$, $k = 0, 1, \ldots, n_i - 1$, there exists a unique polynomial*

$$p \in \Pi^n, \quad n + 1 = \sum_{i=0}^{m} n_i,$$

that satisfies

$$p^{(k)}(x_i) = f_i^{(k)}, \quad i = 0, 1, \ldots, m, \quad k = 0, 1, \ldots, n_i - 1.$$

Proof. Let $p \in \Pi^n$. We may write

$$p(x) := \sum_{l=0}^{n} c_l x^l,$$

for $c_l \in \mathbb{R}$, $l = 0, 1, \ldots, n$. Then

$$\begin{aligned} p^{(k)}(x_i) &= \sum_{l=0}^{n} c_l \left(\frac{d^k}{dx^k} x^l \right) |_{x=x_i} \\ &= \sum_{l=0}^{n} c_l a_{q,l}, \end{aligned}$$

where $q := k + \sum_{j=0}^{i-1} n_i$, and $\sum_{i=0}^{-1} n_i := 0$. This yields an $(n + 1) \times (n + 1)$ linear system. By the invertible matrix theorem (1.1.5.1), it suffices to show uniqueness.

Suppose that $p_1, p_2 \in \Pi^n$ both satisfy the interpolation conditions

$$p_1^{(k)}(x_i) = f^{(k)}(x_i) = p_2^{(k)}(x_i), \quad i = 0, 1, \ldots, m, \quad k = 0, 1, \ldots, n_i - 1.$$

Define the difference polynomial

$$r := p_1 - p_2 \in \Pi^n.$$

Then

$$r^{(k)}(x_i) = 0, \quad i = 0, 1, \ldots, m, \quad k = 0, 1, \ldots, n_i - 1.$$

Thus r has at least $n + 1$ roots, counting multiplicities. Since r is a polynomial of degree at most n, r must vanish identically,

$$r(x) \equiv 0.$$

This proves the theorem. □

The Hermite interpolating polynomials can be given explicitly in a form analogous to the Lagrange interpolation formula (1.1.1.1). The polynomial $p \in \Pi^n$ given by

$$p(x) := \sum_{i=0}^{m} \sum_{k=0}^{n_i - 1} f_i^{(k)} L_{ik}(x)$$

is the desired polynomial (1.1.5.2). The polynomials $L_{ik} \in \Pi^n$ are called the **generalized Lagrange polynomials** and are constructed as follows. Define the auxiliary polynomials

$$l_{ik}(x) := \frac{(x - x_i)^k}{k!} \prod_{\substack{j=0 \\ j \neq i}}^{m} \left(\frac{x - x_j}{x_i - x_j} \right)^{n_j}, \quad 0 \leq i \leq m, \quad 1 \leq k \leq n_i.$$

Put

$$L_{i,n_i-1}(x) := l_{i,n_i-1}(x), \quad i = 0, 1, \ldots, m,$$

and recursively for $k = n_i - 2,\ n_i - 3, \ldots, 0$,

$$L_{ik}(x) := l_{ik}(x) - \sum_{v=k+1}^{n_i-1} l_{ik}^{(v)}(x_i) L_{iv}(x).$$

By induction,

$$L_{ik}(x_j)^{(\sigma)} = \begin{cases} 1, & \text{if } i = j \text{ and } k = \sigma, \\ 0, & \text{otherwise.} \end{cases}$$

Thus the polynomial $p \in \Pi^n$ given by

$$p(x) := \sum_{i=0}^{m} \sum_{j=0}^{n_i-1} f_i^{(k)} L_{ik}(x)$$

is the unique Hermite interpolating polynomial (1.1.5.2).

An alternative way to describe the Hermite interpolation problem is important for Newton– and Neville– type algorithms to construct the Hermite interpolating polynomial. We generalize divided differences to account for repeated abscissas.

Let $x_0 < x_1 < \cdots < x_m$ be a sequence of abscissas. We replace each x_i by n_i copies of itself:

$$\underbrace{x_0 = \cdots = x_0}_{n_0} < \underbrace{x_1 = \cdots = x_1}_{n_1} < \cdots < \underbrace{x_m = \cdots = x_m}_{n_m}.$$

The $n + 1 = \sum_{i=0}^{m} n_i$ elements in this sequence are then defined

$$t_0 = x_0 \le t_1 \le \cdots \le t_n = x_m,$$

where the t_i's, $i = 0, 1, \ldots, n$ are called the **virtual abscissas**.

Note that the virtual abscissas $t_0, t_1, \ldots, t_n$ determine the true abscissas x_i and the integers n_i, $i = 0, 1, \ldots, m$. Recall that the unique polynomial interpolant $p_{01\ldots n}$ is defined by the $n + 1 = \sum_{i=0}^{n} n_i$ interpolation conditions, which are as many as there are index pairs (i, k) with $i = 0, 1, \ldots, m$, $k = 0, 1, \ldots, n_i - 1$, and are as many as there are virtual abscissas $t_0, t_1, \ldots, t_n$.

Observe that the interpolation conditions belonging to the linear ordering of index pairs

$$(0,0), (0,1), \ldots, (0, n_0 - 1), (1,0), \ldots, (1, n_1 - 1), \ldots, (m,0), \ldots, (m, n_m - 1),$$

have the form

$$p_{01\ldots n}^{s_j-1}(t_j) = f^{(s_j-1)}(t_j), \quad j = 0, 1, \ldots, n,$$

if we define s_j, $j = 0, 1, \ldots, n$ to be the number of times each t_j occurs in the subsequence

$$t_0 \le t_1 \le \cdots \le t_j.$$

Also note that

$$x_0 = t_0 = \cdots = t_{n_0-1} < x_1 = t_{n_0} = \cdots = t_{n_0+n_1-1} < \ldots,$$

and

$$s_0 = 1, \ldots, s_{n_0-1} = n_0; s_{n_0} = 1, \ldots, s_{n_0+n_1-1} = n_1, \ldots,$$

establishes the equivalence of the above form and (1.1.5.2).

We move to establish (1.1.5.2) algebraically. Note that any polynomial $p(t) \in \Pi^n$ can be written in the form

$$p(t) = \sum_{j=0}^{n} b_j \frac{t^j}{j!} = \Pi(t)b, \quad b := [b_0, b_1, \ldots, b_n]^\top,$$

where $\Pi(t)$ is defined to be the row vector

$$\Pi(t) := [1, t, \ldots, \frac{t^n}{n!}].$$

Thus by Theorem (1.1.5.2), the system

$$\Pi^{(s_j-1)}(t_j)b = f^{(s_j-1)}(t_j), \quad j = 0, 1, \ldots, n,$$

has a unique solution b. We get the following corollary, which is equivalent to (1.1.5.2).

Corollary 1.1.5.3. *For any nondecreasing finite sequence*

$$t_0 \le t_1 \le \cdots \le t_n$$

of $n+1$ real numbers, the $(n+1) \times (n+1)$ matrix

$$V_n(t_0, t_1, \ldots, t_n) := \begin{bmatrix} \Pi^{(s_0-1)}(t_0) \\ \Pi^{(s_1-1)}(t_1) \\ \vdots \\ \Pi^{(s_n-1)}(t_n) \end{bmatrix}$$

is nonsingular.

Example 1.1.5.4. *For $t_0 = t_1 < t_2$, we have*

$$V_2(t_0, t_1, t_2) = \begin{bmatrix} 1 & t_0 & \frac{t_0^2}{2} \\ 0 & 1 & t_1 \\ 1 & t_2 & \frac{t_2^2}{2} \end{bmatrix}.$$

We now formulate a Neville–type algorithm for Hermite interpolation. We associate with each segment

$$t_i \le t_{i+1} \le \cdots \le t_{i+k}, \quad 0 \le i \le i+k \le n$$

of virtual abscissas the solution $p_{i,i+1,\ldots,i+k} \in \Pi^k$ of the partial Hermite interpolation problem belonging to this subsequence, that is, the solution of

$$p_{i,i+1,\ldots,i+k}^{(s_j-1)}(t_j) = f^{(s_j-1)}(t_j), \quad j = i, i+1, \ldots, i+k.$$

Recall that the integers s_j, $i \le j \le i+k$ are defined with respect to the subsequence, that is, s_j is the number of times the value of t_j occurs within the sequence $t_i, t_{i+1}, \ldots, t_j$.

Example 1.1.5.5. *Suppose that $n_0 = 2$, $n_1 = 3$, and*

$$\begin{cases} x_0 = 0, & f_0^{(0)} = -1, \quad f_0^{(1)} = -2, \\ x_1 = 1, & f_1^{(0)} = 0, \quad f_1^{(1)} = 10, \quad f_1^{(2)} = 40. \end{cases}$$

We get the virtual abscissas t_j, $j = 0, 1, 2, 3, 4$, with

$$t_0 = t_1 := x_0 = 0, \quad t_2 = t_3 = t_4 := x_1 = 1.$$

For the subsequence $t_1 \le t_2 \le t_3$, that is, $i = 1$ and $k = 2$, we have

$$t_1 = x_0 < t_2 = t_3 = x_1, \quad s_1 = s_2 = 1, s_3 = 2.$$

Now the interpolating polynomial $p_{123} \in \Pi^2$ satisfies

$$p_{123}^{(s_1-1)}(t_1) = p_{123}(0) = f^{(0)}(0) = -1,$$
$$p_{123}^{(s_2-1)}(t_2) = p_{123}(1) = f^{(0)}(1) = 0,$$
$$p_{123}^{(s_3-1)}(t_3) = p'_{123}(1) = f^{(1)}(1) = 10.$$

The following analogs to Neville's algorithm (1.1.2.1) hold. We find, if $t_i = t_{i+1} = \cdots = t_{i+k} = x_l$, then

$$p_{i,i+1,\ldots,i+k}(x) = \sum_{r=0}^{k} \frac{f_l^{(r)}}{r!}(x - x_l)^r,$$

and, if $t_i < t_{i+k}$,

$$p_{i,i+1,\ldots,i+k}(x) = \frac{(x - t_i)p_{i+1,i+2,\ldots,i+k}(x) - (x - t_{i+k})p_{i,i+1,\ldots,i+k-1}(x)}{t_{i+k} - t_i}.$$

In analogy to (1.1.3.2) we now define the **generalized divided differences**

$$f[t_i, t_{i+1}, \ldots, t_{i+k}]$$

as the coefficient of x^k in the polynomial $p_{i,i+1,\ldots,i+k} \in \Pi^k$. We find, if $t_i = t_{i+1} = \cdots = t_{i+k} = x_l$,

$$f[t_i, t_{i+1}, t_{i+k}] = \frac{1}{k!} f_l^{(k)},$$

and if $t_i < t_{i+k}$,

$$f[t_i, t_{i+1}, \ldots, t_{i+k}] = \frac{f[t_{i+1}, t_{i+2}, \ldots, t_{i+k}] - f[t_i, t_{i+1}, \ldots, t_{i+k-1}]}{t_{i+k} - t_i}.$$

As a restatement,

$$f[t_i, t_{i+1}, \ldots, t_{i+k}] = \begin{cases} \dfrac{1}{k!} f^{(k)}(x_l), & \text{if } t_i = t_{i+1} = \cdots = t_{i+k} = x_l, \\ \dfrac{f[t_{i+1}, t_{i+2}, \ldots, t_{i+k}] - f[t_i, t_{i+1}, \ldots, t_{i+k-1}]}{t_{i+k} - t_i}, & \text{otherwise.} \end{cases}$$

Using the generalized divided differences

$$a_k := f[t_0, t_1, \ldots, t_k], \quad k = 0, 1, \ldots, n,$$

the solution $p \in \Pi^n$ of the Hermite interpolation problem can be represented in Newton form by

$$p(x) = a_0 + a_1(x - t_0) + a_2(x - t_0)(x - t_1) + \cdots + a_n(x - t_0)(x - t_1)\ldots(x - t_{n-1}).$$

Example 1.1.5.6. *We illustrate the use of generalized divided differences by finding the Hermite interpolant of the following data.*

x	$f(x)$	$f'(x)$	$f''(x)$
0	-1	-2	
1	0	10	40

We get the following scheme:

$$
\begin{array}{l|llll}
t_0 := 0 & f[t_0] = -1 & & & \\
 & & f[t_0,t_1] = -2 & & \\
t_1 := 0 & f[t_1] = -1 & & f[t_0,t_1,t_2] = 3 & \\
 & & f[t_1,t_2] = 1 & & f[t_0,t_1,t_2,t_3] = 6 \\
t_2 := 1 & f[t_2] = 0 & & f[t_1,t_2,t_3] = 9 & \\
 & & f[t_2,t_3] = 10 & & f[t_1,t_2,t_3,t_4] = 11 \\
t_3 := 1 & f[t_3] = 0 & & f[t_2,t_3,t_4] = 20 & \\
 & & f[t_3,t_4] = 10 & & \\
t_4 := 1 & f[t_4] = 0 & & &
\end{array}
\qquad f[t_0,t_1,t_2,t_3,t_4] = 5
$$

Thus the Hermite interpolating polynomial is then

$$p(x) = -1 - 2x + 3x^2 + 6x^2(x-1) + 5x^2(x-1)^2.$$

We give a result for the interpolation error incurred by Hermite interpolation.

Theorem 1.1.5.7 (Error in Hermite Interpolation)**.** *Let the real function f by $n+1$ times differentiable on the interval $[a,b]$, and consider $m+1$ support abscissas $x_i \in [a,b]$,*

$$x_0 < x_1 < \cdots < x_m.$$

If the polynomial $p(x)$ is of degree at most n,

$$n+1 = \sum_{i=0}^{m} n_i,$$

and satisfies the interpolation conditions

$$p^{(k)}(x_i) = f^{(k)}(x_i), \quad i = 0,1,\ldots,m, \quad k = 0,1,\ldots,n_i - 1,$$

then for every $\overline{x} \in [a,b]$ there exists $\xi \in I[x_0,x_1,\ldots,x_m;\overline{x}]$ such that

$$e(\overline{x}) = f(\overline{x}) - p(\overline{x}) = \frac{\omega(\overline{x}) f^{(n+1)}(\xi)}{(n+1)!},$$

where

$$\omega(x) := \prod_{i=0}^{m} (x - x_i)^{n_i}.$$

Proof. The proof of (1.1.5.7) is entirely analogous to that of the error for standard polynomial interpolation, see (1.1.4.2). □

1.2. **Trigonometric Interpolation.**

1.2.1. *Basic Facts.* Trigonometric interpolation uses linear combinations of the trigonometric functions $\sin(kx)$ and $\cos(kx)$ for integer k. We restrict our attention to linear interpolation. For $N = 2M$ (even) support points $\{(x_k, f_k)\}_{k=0}^{N-1}$,

$$\Psi(x) := \frac{A_0}{2} + \sum_{k=1}^{M}(A_k \cos(kx) + B_k \sin(kx)) + \frac{A_M}{2}\cos(Mx), \tag{1.2.1.1}$$

and for $N = 2M + 1$ (odd) support points $\{(x_k, f_k)\}_{k=0}^{N-1}$,

$$\Psi(x) := \frac{A_0}{2} + \sum_{k=1}^{M}(A_k \cos(kx) + B_k \sin(kx)). \tag{1.2.1.2}$$

Trigonometric interpolation is frequently used for data which are periodic with known period.

We recall the following result from complex analysis.

Theorem 1.2.1.1 (De Moivre's Formula)**.** *For every integer k,*

$$e^{ikx} = \cos(kx) + i\sin(kx).$$

We consider uniform partitions of the interval $[0, 2\pi]$,

$$x_l := \frac{2\pi l}{N}, \quad l = 0, 1, \ldots, N-1.$$

For such partitions, the trigonometric interpolation problem becomes a problem of finding a so–called **phase polynomial** of order N

$$p(x) := \beta_0 + \beta_1 e^{ix} + \beta_2 e^{2ix} + \cdots + \beta_{N-1} e^{(N-1)x}, \tag{1.2.1.3}$$

with complex coefficients β_j such that

$$p(x_l) = f_l, \quad l = 0, 1, \ldots, N-1.$$

By definition of x_l, we note that

$$e^{-ilx_k} = e^{-il\left(\frac{2\pi k}{N}\right)} = e^{2\pi i\left(-\frac{lk}{N}\right)} = e^{2\pi ik} \cdot e^{2\pi i\left(-\frac{lk}{N}\right)} = e^{2\pi i(k-\frac{lk}{N})} = e^{2\pi i\frac{(N-l)k}{N}} = e^{i(N-l)x_k}.$$

Therefore

$$\cos(kx_l) = \frac{e^{ikx_l} + e^{i(N-k)x_l}}{2}, \quad \sin(kx_l) = \frac{e^{ikx_l} - e^{i(N-k)x_l}}{2i}. \tag{1.2.1.4}$$

Making these substitutions into (1.2.1.1) and (1.2.1.2) for $\Psi(x)$ produces the phase polynomial $p(x)$ with coefficients β_j, $j = 0, 1, \ldots, N-1$. Observe for $N = (2M+1)$ (odd), we have

$$\begin{aligned}
\Psi(x) &= \frac{A_0}{2} + \sum_{k=1}^{M}\left[\frac{A_k}{2}\left(e^{ikx} + e^{i(N-k)x}\right) + \frac{B_k}{2i}\left(e^{ikx} - e^{i(N-k)x}\right)\right] \\
&= \frac{A_0}{2} + \sum_{k=1}^{M}\left[\frac{A_k}{2}\left(e^{ikx} + e^{i(N-k)x}\right) + \frac{B_k}{2}\left(-ie^{ikx} + ie^{i(N-k)x}\right)\right] \\
&= \frac{A_0}{2} + \sum_{k=1}^{M}\left[\frac{1}{2}\left(A_k - iB_k\right)e^{ikx} + \frac{1}{2}(A_k + iB_k)e^{i(N-k)x}\right]
\end{aligned}$$

$$= \frac{A_0}{2} + \sum_{k=1}^{M} \frac{1}{2}(A_k - iB_k)e^{ikx} + \sum_{k=M+1}^{2M} \frac{1}{2}(A_k + iB_k)e^{ikx}.$$

Similarly for $N = 2M$ (even), we get

$$\begin{aligned}\Psi(x) &= \frac{A_0}{2} + \sum_{k=1}^{M-1} \left[\frac{A_k}{2} \left(e^{ikx} + e^{i(N-k)x}\right) + \frac{B_k}{2i} \left(e^{ikx} - e^{i(N-k)x}\right)\right] + \frac{A_M}{2} \left(\frac{1}{2} \left(e^{iMx} + e^{i(N-M)x}\right)\right) \\ &= \frac{A_0}{2} + \sum_{k=1}^{M-1} \left[\frac{1}{2}(A_k - iB_k)e^{ikx} + \frac{1}{2}(A_k + iB_k)e^{i(N-k)x}\right] + \frac{A_M}{2}e^{iMx} \\ &= \frac{A_0}{2} + \sum_{k=1}^{M-1} \frac{1}{2}(A_k - iB_k)e^{ikx} + \frac{A_M}{2}e^{iMx} + \sum_{k=M+1}^{2M-1} \frac{1}{2}(A_k + iB_k)e^{ikx}.\end{aligned}$$

We arrive at the following result.

Lemma 1.2.1.2. *If N is odd, then $N = 2M + 1$ and*

$$\beta_0 = \frac{A_0}{2}, \quad \beta_j = \frac{1}{2}(A_j - iB_j), \quad \beta_{N-j} = \frac{1}{2}(A_j + iB_j), \quad j = 1, 2, \ldots, M,$$
$$A_0 = 2\beta_0, \quad A_k = \beta_k + \beta_{N-k}, \quad B_k = i(\beta_k - \beta_{N-k}), \quad k = 1, 2, \ldots, M.$$

If N is even, then $N = 2M$ and

$$\beta_0 = \frac{A_0}{2}, \quad \beta_j = \frac{1}{2}(A_j - iB_j), \quad \beta_M = \frac{A_M}{2}, \quad \beta_{N-j} = \frac{1}{2}(A_j + iB_j), \quad j = 1, 2, \ldots, M-1,$$
$$A_0 = 2\beta_0, \quad A_k = \beta_k + \beta_{N-k}, \quad A_M = 2\beta_M, \quad B_k = i(\beta_k - \beta_{N-k}), \quad k = 1, 2, \ldots, M-1.$$

The trigonometric expression $\Psi(x)$ and its corresponding phase polynomial p agree for all support abscissas $x_k = 2\pi k/N$ of an equispaced partition of the interval $[0, 2\pi]$,

$$f_k = \Psi(x_k) = p(x_k), \quad k = 0, 1, \ldots, N-1.$$

However $\Psi(x) = p(x)$ need not hold at intermediate points $x \neq x_k$, $k = 0, 1, \ldots, N-1$. The two interpolation problems are equivalent in the sense that a solution to one problem will produce a solution to the other by means of the coefficient relations in (1.2.1.2).

We note here that the phase polynomials p in (1.2.1.3) are structurally simpler than the trigonometric expressions $\Psi(x)$. Introduce the notation

$$\omega := e^{ix}, \quad \omega_k := e^{ix_k} = e^{2ki\pi/N},$$
$$p(\omega) := \beta_0 + \beta_1\omega + \beta_2\omega^2 + \cdots + \beta_{N-1}\omega^{N-1}.$$

Since $\omega_j \neq \omega_k$ for $j \neq k$, $0 \leq j, k \leq N-1$, we see that we are faced with a standard polynomial interpolation problem, that is, we seek to find the complex polynomial p of degree $N-1$ or less such that

$$p(\omega_k) = f_k, \quad k = 0, 1, \ldots, N-1.$$

We get the following result.

Theorem 1.2.1.3 (Existence and Uniqueness of Phase Polynomial)**.** *For any support points* $\{(x_k, f_k)\}_{k=0}^{N-1}$*, with* f_k *complex and* $x_k := 2\pi k/N$*, there exists a unique phase polynomial*

$$p(x) = \beta_0 + \beta_1 e^{ix} + \beta_2 e^{2ix} + \cdots + \beta_{N-1} e^{(N-1)x}$$

with

$$p(x_k) = f_k$$

for $k = 0, 1, \ldots, N-1$.

Proof. The proof follows immediately from the existence and uniqueness of polynomial interpolation, see (1.1.1.1). □

The coefficients β_j, $j = 0, 1, \ldots, N-1$ of the interpolating phase polynomial p can be expressed in closed form. Note, for $0 \le j, k \le N-1$, we have

$$\omega_j^k = e^{ijx_k} = e^{2jki\pi/N} = e^{ikx_j} = \omega_k^j,$$

and, by De Moivre's formula (1.2.1.1),

$$\begin{aligned}\omega_k^{-j} &= e^{-ijx_k} = e^{-2jki\pi/N} = e^{ik(-2j\pi/N)} \\ &= \cos(-2kj\pi/N) + i\sin(-2kj\pi/N) = \cos(2kj\pi/N) - i\sin(2kj\pi/N) \\ &= \overline{\omega_k^j}.\end{aligned}$$

More importantly, however, we define the inner product

$$\langle \omega^j, \omega^l \rangle := \sum_{k=0}^{N-1} \omega_k^j \overline{\omega_k^l} = \sum_{k=0}^{N-1} e^{ijx_k} \cdot e^{-ilx_k}.$$

Lemma 1.2.1.4. *If* $0 \le j, l \le N-1$*, then we have*

$$\langle \omega^j, \omega^l \rangle = \begin{cases} N, & j = l, \\ 0, & j \ne l. \end{cases}$$

Proof. If $j = l$, then

$$\langle \omega^j, \omega^l \rangle = \sum_{k=0}^{N-1} \omega_k^j \omega_k^{-j} = \sum_{k=0}^{N-1} e^{ijx_k} \cdot e^{-ijx_k} = \sum_{k=0}^{N-1} 1 = N.$$

Now suppose that $j \ne l$. Then

$$\begin{aligned}\langle \omega^j, \omega^l \rangle &= \sum_{k=0}^{N-1} \omega_k^j \omega_k^{-l} = \sum_{k=0}^{N-1} e^{i(j-l)x_k} = \sum_{k=0}^{N-1} e^{ikx_{j-l}} = \sum_{k=0}^{N-1} \left(e^{ix_{j-l}}\right)^k \\ &= \sum_{k=0}^{N-1} \omega_{j-l}^k.\end{aligned}$$

Since $j \ne l$ with $0 \le j, l \le N-1$, we have

$$\omega_{j-l} = e^{2\pi i(j-l)/N} \ne 1.$$

Note that $(\omega - 1)\sum_{k=0}^{N-1} \omega^k = \omega^N - 1$. Thus

$$\omega^N - 1 = \left(e^{2\pi i(j-l)/N}\right)^N - 1 = e^{2\pi i(j-l)} - 1 = 1 - 1 = 0,$$

so that

$$\langle \omega^j, \omega^l \rangle = 0.$$

This completes the proof. □

Lemma (1.2.1.4) says that the vectors ω_k form an orthogonal basis for $\mathbb{C}^N$. From this orthogonality follows the following result.

Theorem 1.2.1.5 (Closed form of Phase Polynomial Coefficients)**.** *The phase polynomial* $p(x) = \sum_{j=0}^{N-1} \beta_j e^{ijx}$ *satisfies*

$$p(x_k) = f_k, \quad k = 0, 1, \ldots, N-1,$$

for f_k *complex and* $x_k = 2\pi k/N$, $k = 0, 1, \ldots, N-1$, *if and only if*

$$\beta_j = \frac{1}{N} \sum_{k=0}^{N-1} f_k \omega_k^{-j} = \frac{1}{N} \sum_{k=0}^{N-1} f_k e^{-2\pi i j k/N},$$

for $j = 0, 1, \ldots, N-1$.

Proof. Because of $f_k = p(x_k)$, $k = 0, 1, \ldots, N-1$, we have

$$\begin{aligned}
\frac{1}{N} \sum_{k=0}^{N-1} f_k e^{-ijx_k} &= \frac{1}{N} \sum_{k=0}^{N-1} p(x_k) e^{-ijx_k} \\
&= \frac{1}{N} \sum_{k=0}^{N-1} e^{-ijx_k} \sum_{l=0}^{N-1} \beta_l e^{ilx_k} \\
&= \frac{1}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \beta_l e^{ilx_k} \cdot e^{-ijx_k} \\
&= \frac{1}{N} \sum_{l=0}^{N-1} \beta_l \sum_{k=0}^{N-1} e^{ilx_k} \cdot e^{-ijx_k} \\
&= \frac{1}{N} \sum_{l=0}^{N-1} \beta_l \sum_{k=0}^{N-1} \omega_k^l \overline{\omega_k^j} \\
&= \frac{1}{N} \sum_{l=0}^{N-1} \beta_l \langle \omega^l, \omega^j \rangle \\
&= \frac{1}{N} (\beta_j N) = \beta_j.
\end{aligned}$$

□

We return to the original trigonometric expressions $\Psi(x)$ and state the main result for this section.

Theorem 1.2.1.6 (Solution of Trigonometric Interpolation Problem)**.** *The trigonometric expressions*

$$\Psi(x) = \frac{A_0}{2} + \sum_{k=1}^{M}(A_k \cos(kx) + B_k \sin(kx)),$$

$$\Psi(x) = \frac{A_0}{2} + \sum_{k=1}^{M-1}(A_k \cos(kx) + B_k \sin(kx)) + \frac{A_M}{2}\cos(Mx),$$

where $N = 2M + 1$ (odd) and $N = 2M$ (even), respectively, satisfy

$$\Psi(x_k) = f_k, \quad k = 0, 1, \ldots, N-1,$$

for $x_k = 2\pi k/N$ if and only if the coefficients of $\Psi(x)$ are given by

$$A_k = \frac{2}{N}\sum_{l=0}^{N-1} f_l \cos(kx_l) = \frac{2}{N}\sum_{l=0}^{N-1} f_l \cos\left(\frac{2\pi kl}{N}\right),$$

$$B_k = \frac{2}{N}\sum_{l=0}^{N-1} f_l \sin(kx_l) = \frac{2}{N}\sum_{l=0}^{N-1} f_l \sin\left(\frac{2\pi kl}{N}\right).$$

Proof. In all cases,

$$A_0 = 2\beta_0 = \frac{2}{N}\sum_{l=0}^{N-1} f_l e^0 = \frac{2}{N}\sum_{l=0}^{N-1} f(x_l).$$

If $N = 2M$, then

$$\begin{aligned} A_m = 2\beta_M &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)e^{-2\pi iMl/2M} \\ &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)e^{-\pi il} \\ &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos(\pi l) \\ &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos\left(\frac{2\pi lM}{2M}\right) \\ &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos(Mx_l). \end{aligned}$$

Finally, consider

$$A_k = \beta_k + \beta_{N-k} = \frac{1}{N}\sum_{l=0}^{N-1} f(x_l)(e^{-ikx_l} + e^{-i(N-k)x_l}).$$

By Euler's formula (1.2.1.4),

$$A_k = \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos(kx_l),$$

and

$$\begin{aligned} B_k &= i(\beta_k - \beta_{N-k}) \\ &= \frac{i}{N}\sum_{l=0}^{N-1} f(x_l)(e^{-ikx_l} - e^{-i(N-k)x_l}) \\ &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\sin(kx_l). \end{aligned}$$

This completes the proof. □

Example 1.2.1.7. *We construct the trigonometric polynomial of degree $M = 2$ given the following data points:*

$$\left\{(0,1), \left(\frac{\pi}{2}, 3\right), (\pi, -5), \left(\frac{3\pi}{2}, 2\right)\right\}.$$

Note

$$\begin{aligned} A_0 &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos(0) \\ &= \frac{1}{2}\sum_{l=0}^{3} f(x_l) \\ &= \frac{1}{2}[1 + 3 - 5 + 2] = \frac{1}{2} \end{aligned}$$

,

$$\begin{aligned} A_1 &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos(x_l) \\ &= \frac{1}{2}\sum_{l=0}^{3} f(x_l)\cos(x_l) \\ &= \frac{1}{2}\left[(1\cdot\cos(0)) + \left(3\cdot\cos\left(\frac{\pi}{2}\right)\right) + (-5\cdot\cos(\pi)) + \left(2\cdot\cos\left(\frac{3\pi}{2}\right)\right)\right] \\ &= \frac{1}{2}[1 + 5] = 3, \end{aligned}$$

$$\begin{aligned} A_2 &= \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\cos(2x_l) \\ &= \frac{1}{2}\sum_{l=0}^{3} f(x_l)\cos(2x_l) \end{aligned}$$

$$= \frac{1}{2}[(1 \cdot \cos(0)) + (3 \cdot \cos(\pi)) + (-5 \cdot \cos(2\pi)) + (2 \cdot \cos(3\pi))]$$
$$= \frac{1}{2}[1 - 3 - 5 - 2] = -\frac{9}{2},$$

$$B_1 = \frac{2}{N}\sum_{l=0}^{N-1} f(x_l)\sin(x_l)$$
$$= \frac{1}{2}\sum_{l=0}^{N-1} f(x_l)\sin(x_l)$$
$$= \frac{1}{2}\left[(1 \cdot \sin(0)) + \left(3 \cdot \sin\left(\frac{\pi}{2}\right)\right) + (-5 \cdot \sin(\pi)) + \left(2 \cdot \sin\left(\frac{3\pi}{2}\right)\right)\right]$$
$$= \frac{1}{2}[3 - 2] = \frac{1}{2}.$$

Hence, the interpolating trigonometric polynomial is

$$\Psi(x) = \frac{A_0}{2} + \sum_{k=1}^{1}(A_k\cos(kx) + B_k\sin(kx)) + \frac{A_2}{2}\cos(2x)$$
$$= \frac{A_0}{2} + A_1\cos(x) + B_1\sin(x) + \frac{A_2}{2}\cos(2x)$$
$$= \frac{1}{4} + 3\cos(x) + \frac{1}{2}\sin(x) - \frac{9}{4}\cos(2x).$$

Moreover, the coefficients of the corresponding phase polynomial are

$$\beta_0 = \frac{A_0}{2} = \frac{1/2}{2} = \frac{1}{4},$$
$$\beta_1 = \frac{1}{2}(A_1 - iB_1) = \frac{1}{2}\left(3 - \frac{1}{2}i\right) = \frac{3}{2} - \frac{1}{4}i,$$
$$\beta_3 = \frac{1}{2}(A_1 + iB_1) = \frac{1}{2}\left(3 + \frac{1}{2}i\right) = \frac{3}{2} + \frac{1}{4}i,$$
$$\beta_2 = \frac{A_2}{2} = \frac{-9/2}{2} = -\frac{9}{4}.$$

Thus, the corresponding phase polynomial is

$$p(x) = \beta_0 + \beta_1 e^{ix} + \beta_2 e^{2ix} + \beta_3 e^{3ix}$$
$$= \frac{1}{4} + \left(\frac{3}{2} - \frac{1}{4}i\right)e^{ix} - \frac{9}{4}e^{2ix} + \left(\frac{3}{2} + \frac{1}{4}i\right)e^{3ix}.$$

1.3. **Interpolation by Spline Functions.** Spline functions yield smooth interpolating curves which are less likely to exhibit the large oscillations characteristic of polynomials of high degree.

Definition 1.3.0.1 (Knot)**.** *Let*

$$\Delta : a = x_0 < x_1 < \cdots < x_n = b$$

be a partition of an interval $[a, b]$*. The points* x_i*,* $i = 0, 1, \ldots, n$*, are called* **knots***.*

Given a partition $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ of $[a, b]$, splines are piecewise polynomial functions $S : [a, b] \to \mathbb{R}$, with certain smoothness properties that are composed of polynomials, namely, the restrictions $S|I_i$ of S to $I_i := (x_{i-1}, x_i)$, $i = 1, 2, \ldots, n$, are polynomials.

In the following sections we describe the case of cubic splines, which are composed of cubic polynomials, $S|I_i \in \Pi^3$.

1.3.1. *Theoretical Foundations.* Throughout this section we let

$$\Delta := \{a = x_0 < x_1 < \cdots < x_n\}$$

be a partition of the interval $[a, b]$. We first give the definition of a cubic spline function.

Definition 1.3.1.1 (Cubic Spline Function)**.** *A* **cubic spline function** S_Δ *on* Δ *is a real-valued function* $S_\Delta : [a, b] \to \mathbb{R}$ *with the properties:*

(1) $S_\Delta \in C^2[a, b]$*, that is,* S_Δ *is twice continuously differentiable on* $[a, b]$*.*
(2) S_Δ *coincides on each subinterval* $[x_{i-1}, x_i]$*,* $i = 1, 2, \ldots, n$ *with a polynomial of degree at most three.*

We see that a cubic spline function consists of cubic polynomials pieced together so that their values and those of their first two derivatives coincide at the interior knots x_i, $i = 1, 2, \ldots, n - 1$.

Definition 1.3.1.2 (Interpolating Spline Function)**.** *Let* $\{f_i\}_{i=0}^n$ *be a finite sequence of* $n+1$ *real numbers. An* **interpolating spline function** *is a spline function*

$$S_\Delta(f, \cdot)$$

such that $S_\Delta(f_i, x_i) = f_i$ *for each* $i = 0, 1, \ldots, n$*.*

Note that an interpolating spline function $S_\Delta(f, \cdot)$ is not uniquely determined by the sequence f of support ordinates. There are two degrees of freedom left, so we impose additional requirements on S_Δ, known as the **side/spline conditions**.

Definition 1.3.1.3 (Common Side Conditions)**.** *Three common side conditions for an interpolating cubic spline function are as follows:*

(1) $S''_\Delta(f, a) = S''_\Delta(f, b) = 0$ *(natural);*
(2) $S^{(k)}_\Delta(f, a) = S^{(k)}_\Delta(f, b)$*, for* $k = 0, 1, 2$ *(periodic);*
(3) $S'_\Delta(f, a) = f'_0$*,* $S'_\Delta(f, b) = f'_n$*, for given numbers* f'_0*,* f'_n *(clamped).*

We show later that each of these conditions ensures uniqueness of the interpolating spline function $S_\Delta(f, \cdot)$.

We now present definitions that will allow us to establish the above result as well as a characteristic minimum property of cubic spline functions.

Definition 1.3.1.4 (Absolutely Continuous)**.** *A real–valued function $f : [a, b] \to \mathbb{R}$ is said to be* **absolutely continuous** *on the interval $[a, b]$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$\sum_i |f(b_i) - f(a_i)| < \epsilon$$

for every finite set of intervals $[a_i, b_i]$ with

$$a \leq a_1 < b_1 < a_2 < b_2 < \cdots < a_n < b_n = b$$

and $\sum_i |b_i - a_i| < \delta$.

We get the following properties for absolutely continuous functions.

Lemma 1.3.1.5. *If $f : [a, b] \to \mathbb{R}$ is absolutely continuous, then*

(1) f is continuous;
(2) f' exists almost everywhere;
(3) $f(x) = f(a) + \int_a^x f'(t)\ dt$ for all $x \in [a, b]$ (fundamental theorem of calculus);
(4) $\int_a^b f(x)g'(x)\ dx = f(x)g(x)|_a^b - \int_a^b f'(x)g(x)\ dx$ (integration by parts).

Definition 1.3.1.6 (L^p-space)**.** *The set $L^2[a, b]$ denotes the set of all real–valued square integrable functions on $[a, b]$, that is,*

$$\int_a^b |f(t)|^2\ dt$$

exists and is finite.

Definition 1.3.1.7 (κ^m)**.** *For a positive integer m, we define*

$$\kappa^m[a, b]$$

to be the set of all real–valued functions $f : [a, b] \to \mathbb{R}$ for which $f^{(m-1)}$ is absolutely continuous on $[a, b]$ and $f^{(m)} \in L^2[a, b]$.

Definition 1.3.1.8 (κ_p^m)**.** *We denote by*

$$\kappa_p^m[a, b]$$

the subset of all functions $f \in \kappa^m[a, b]$ such that $f^{(k)}(a) = f^{(k)}(b)$ for each $k = 0, 1, \ldots, m-1$, that is, $f^{(k)}$, $k = 0, 1, \ldots, m - 1$ is periodic.

Note that $S_\Delta \in \kappa^3[a, b]$, and moreover $S_\Delta(f, \cdot) \in \kappa_p^3[a, b]$ if the periodic side conditions are satisfied.

The structure of $\kappa^m[a, b]$ allows us to endow the function space with a seminorm.

Definition 1.3.1.9 (κ^2 Seminorm)**.** *Let $f \in \kappa^2[a, b]$. We define the $\kappa^2[a, b]$ seminorm $|\cdot|_{\kappa^2}$ by*

$$|f|_{\kappa^2}^2 := \int_a^b |f''(x)|^2\ dx.$$

Note that $|f|_{\kappa^2} \geq 0$ for all $f \in \kappa^2[a, b]$. However, $|\cdot|_{\kappa^2}$ is not a full norm but only a seminorm, for $|f|_{\kappa^2} = 0$ may hold for functions f that are not identically zero, for example, for all linear functions $f(x) := mx + b$.

Theorem 1.3.1.10. *If $f \in \kappa^2[a,b]$, $\Delta := \{a = x_0 < x_1 < \cdots < x_n = b\}$ is a partition of the interval $[a,b]$, and if S_Δ is a spline function with knots $x_i \in \Delta$, then*

$$|f - S_\Delta|_{\kappa^2}^2 = |f|_{\kappa^2}^2 - |S_\Delta|_{\kappa^2}^2 - 2\left[(f'(x) - S'_\Delta(x))S''_\Delta(x)|_a^b - \sum_{i=1}^{n}(f(x) - S_\Delta(x))S'''_\Delta(x)|_{x_{i-1}^+}^{x_i^-}\right].$$

Note that $S'''_\Delta(x)$ is piecewise constant, with possible discontinuities at the interior knots $x_1, x_2, \ldots, x_{n-1}$. We indicate by x_i^- and x_{i-1}^+ in the above theorem the left and right limits of S'''_Δ at x_i and x_{i-1}, respectively.

Proof. By the definition of $|\cdot|$, we have

$$\begin{aligned}
|f - S_\Delta|^2 &= \int_a^b |f''(x) - S''_\Delta(x)|^2 \, dx \\
&= \int_a^b |f''(x)|^2 - 2f''(x)S''_\Delta(x) + |S''_\Delta(x)|^2 \, dx \\
&= |f|^2 + |S_\Delta|^2 - 2\int_a^b f''(x)S''_\Delta(x) \, dx \\
&= |f|^2 - |S_\Delta|^2 - 2\int_a^b (f''(x) - S''_\Delta(x))S''_\Delta(x) \, dx.
\end{aligned}$$

Recalling that $S_\Delta^{(k)}$, $k = 0, 1, 2, 3$ is defined piecewise, integrating by parts gives

$$\begin{aligned}
&\int_a^b (f''(x) - S''_\Delta(x))S''_\Delta(x) \, dx = \\
&\quad = \sum_{i=1}^{n}(f'(x) - S'_\Delta(x))S''_\Delta(x)|_{x_{i-1}}^{x_i} - \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}(f'(x) - S'_\Delta(x))S'''_\Delta(x) \, dx \\
&\quad = \sum_{i=1}^{n}(f'(x) - S'_\Delta(x))S''_\Delta(x)|_{x_{i-1}}^{x_i} - \sum_{i=1}^{n}\Big[(f(x) - S_\Delta(x))S'''_\Delta(x)|_{x_{i-1}^+}^{x_i^-} - \\
&\qquad \int_{x_{i-1}}^{x_i}(f(x) - S_\Delta(x))S_\Delta^{(4)}(x) \, dx\Big] \\
&\quad = (f'(x) - S'_\Delta(x))S''_\Delta(x)|_a^b - \sum_{i=1}^{n}(f(x) - S_\Delta(x))S'''_\Delta(x)|_{x_{i-1}^+}^{x_i^-},
\end{aligned}$$

Since $S_\Delta^{(4)} \equiv 0$ on $[a,b]$ and

$$\sum_{i=1}^{n}(f'(x) - S'_\Delta(x))S''_\Delta(x)|_{x_{i-1}}^{x_i}$$

telescopes by the continuity of S''_Δ. This completes the proof. □

We arrive at the minimum–norm property of spline functions.

Theorem 1.3.1.11 (Minimum–Norm Property and Uniqueness of Cubic Spline)**.** *Given a partition*

$$\Delta := \{a = x_0 < x_1 < \cdots < x_n = b\}$$

of the interval $[a, b]$*, values* $\{f_i\}_{i=0}^n$*, and a function* $\phi \in \kappa^2[a, b]$ *with* $\phi(x_i) = f_i$ *for each* $i = 0, 1, \ldots, n$*, then*

$$|\phi|_{\kappa^2} \geq |S_\Delta(f, \cdot)|_{\kappa^2},$$

and, more precisely,

$$|\phi - S_\Delta(f, \cdot)|_{\kappa^2}^2 = |\phi|_{\kappa^2}^2 - |S_\Delta(f, \cdot)|_{\kappa^2}^2 \geq 0$$

holds for every interpolating spline function $S_\Delta(f, \cdot)$*, provided that one of the following conditions is satisfied:*

(1) $S''_\Delta(f, a) = S''_\Delta(f, b) = 0$ *(natural),*
(2) $S^{(k)}_\Delta(f, a) = S^{(k)}_\Delta(f, b)$ *for* $k = 0, 1, 2$ *(periodic),*
(3) $S'_\Delta(f, a) = \phi'(a)$*,* $S'_\Delta(f, b) = \phi'(b)$ *(clamped).*

In each of these cases, the interpolating spline function $S_\Delta(f, \cdot)$ *is uniquely determined.*

Proof. We handle existence of the interpolating spline function by construction in the following section.

In each of the above three cases of side conditions, the expressions

$$(\phi'(x) - S_\Delta(x))S''_\Delta(x)\big|_a^b = 0$$

and

$$\sum_{i=1}^{n} (\phi(x) - S_\Delta(x))S'''_\Delta(x)\big|_{x_{i-1}^+}^{x_i^-} = 0$$

vanish in the identity (1.3.1.10). Thus

$$|\phi - S_\Delta(f, \cdot)|_{\kappa^2}^2 = |\phi|_\kappa^2 - |S_\Delta(f, \cdot)|_{\kappa^2}^2 \geq 0,$$

so that evidently

$$|\phi|_{\kappa^2}^2 \geq |S_\Delta(f, \cdot)|_{\kappa^2}^2.$$

This proves the minimum norm property of the interpolating spline function $S_\Delta(f, \cdot)$.

To show uniqueness, assume that $\overline{S}_\Delta(f, \cdot)$ is another interpolating spline function having the same properties as $S_\Delta(f, \cdot)$. Then $\overline{S}_\Delta(f, \cdot)$ satisfies the same properties as ϕ in the statement of (1.3.1.11), so letting $\overline{S}_\Delta(f, \cdot)$ play the role of ϕ, the minimum norm property of $S_\Delta(f, \cdot)$ implies that

$$|\overline{S}_\Delta(f, \cdot) - S_\Delta(f, \cdot)|_{\kappa^2}^2 = |\overline{S}_\Delta(f, \cdot)|_{\kappa^2}^2 - |S_\Delta(f, \cdot)|_{\kappa^2}^2 \geq 0.$$

Since $\overline{S}_\Delta(f, \cdot)$ and $S_\Delta(f, \cdot)$ may switch roles, we have similarly

$$|S_\Delta(f, \cdot) - \overline{S}_\Delta(f, \cdot)|_{\kappa^2}^2 = |S_\Delta(f, \cdot)|_{\kappa^2}^2 - |\overline{S}_\Delta(f, \cdot)|_{\kappa^2}^2 \geq 0.$$

Evidently $|\overline{S}_\Delta(f, \cdot)|_{\kappa^2}^2 = |S_\Delta(f, \cdot)|_{\kappa^2}^2$. Thus

$$|\overline{S}_\Delta(f, \cdot) - S_\Delta(f, \cdot)|_{\kappa^2}^2 = \int_a^b \left|\overline{S}''_\Delta(f, x) - S''_\Delta(f, x)\right|^2 \, dx = 0.$$

By the continuity of $S''_\Delta(f,\cdot)$ and $\overline{S}''_\Delta(f,\cdot)$, we have

$$\overline{S}''_\Delta(f,x) \equiv S''_\Delta(f,x)$$

for all $x \in [a,b]$. Integrating, we obtain

$$\overline{S}_\Delta(f,x) = S_\Delta(f,x) + cx + d$$

for some $c, d \in \mathbb{R}$. But observe that

$$\overline{S}_\Delta(f,a) - S_\Delta(f,a) = 0 = cx + d|_{x=a},$$
$$\overline{S}_\Delta(f,b) - S_\Delta(f,b) = 0 = cx + d|_{x=b},$$

and this implies that $c = d = 0$. Hence,

$$\overline{S}_\Delta(f,x) \equiv S_\Delta(f,x)$$

for all $x \in [a,b]$, which completes the proof. □

1.3.2. *Determining Interpolating Cubic Spline Functions.* In this section we construct cubic spline functions which assume prescribed values at their knots and satisfy one of the side conditions (1.3.1.11). In doing this we will have proved the existence of such spline functions. We have already established uniqueness in (1.3.1.11).

Throughout this section,

$$\Delta := \{x_i : i = 0, 1, \ldots, n\}$$

will be a fixed partition of the interval $[a,b]$ by knots $a = x_0 < x_1 < \cdots < x_n = b$. We put

$$Y := \{y_i\}_{i=0}^n,$$

a sequence of $n+1$ prescribed real numbers. Finally, we denote by I_j the subinterval

$$I_j := [x_{j-1}, x_j], \quad j = 1, 2, \ldots, n,$$

and

$$h_j := x_j - x_{j-1}, \quad j = 1, 2, \ldots, n$$

will denote the length of each I_j.

Definition 1.3.2.1 (Moments)**.** *We call the values of the second derivatives at knots $x_j \in \Delta$,*

$$M_j := S''_\Delta(Y, x_j), \quad j = 0, 1, \ldots, n$$

of the interpolating spline function $S_\Delta(Y,\cdot)$ the **moments** *M_j of $S_\Delta(Y,\cdot)$.*

We will show that interpolating spline functions are characterized by their moments.

Recall that the second derivative $S''_\Delta(Y,\cdot)$ coincides with a linear function in each subinterval $I_{j+1} = [x_j, x_{j+1}]$, $j = 0, 1, \ldots, n-1$, and that we can express these linear functions in terms of the moments M_j by

$$S''_\Delta(Y,x) = M_j \frac{x_{j+1} - x}{h_{j+1}} + M_{j+1} \frac{x - x_j}{h_{j+1}},$$

for $x \in [x_j, x_{j+1}]$. By integration, we obtain

$$S'_\Delta(Y,x) = -M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}} + A_j, \tag{1.3.2.1}$$

$$S_\Delta(Y,x) = M_j \frac{(x_{j+1} - x)^3}{6h_{j+1}} + M_{j+1} \frac{(x - x_j)^3}{6h_{j+1}} + A_j(x - x_j) + B_j,$$

for all $x \in [x_j, x_{j+1}]$, $j = 0, 1, \ldots, n-1$, and where A_j, B_j are constants of integration. Recalling that $S_\Delta(Y, x_j) = y_j$ and $S_\Delta(Y, x_{j+1}) = y_{j+1}$ by supposition, we have

$$y_j = S_\Delta(Y, x_j) = M_j \frac{(x_{j+1} - x_j)^3}{6h_{j+1}} + B_j = M_j \frac{h_{j+1}^2}{6} + B_j,$$

$$y_{j+1} = S_\Delta(Y, x_{j+1}) = M_{j+1} \frac{(x_{j+1} - x_j)^3}{6h_{j+1}} + A_j(x_{j+1} - x_j) + B_j = M_{j+1} \frac{h_{j+1}^2}{6} + A_j h_{j+1} + B_j,$$

so that we arrive at the following equations for A_j and B_j :

$$B_j = y_j - M_j \frac{h_{j+1}^2}{6}, \tag{1.3.2.2}$$

$$\begin{aligned} A_j &= \frac{1}{h_{j+1}} \left(y_{j+1} - M_{j+1} \frac{h_{j+1}^2}{6} - B_j \right) \\ &= \frac{y_{j+1}}{h_{j+1}} - M_{j+1} \frac{h_{j+1}}{6} - \frac{y_j}{h_{j+1}} + M_j \frac{h_{j+1}}{6} \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j). \end{aligned} \tag{1.3.2.3}$$

This gives the following representation of the interpolating spline function $S_\Delta(Y, \cdot)$ in terms of its moments M_j :

$$S_\Delta(Y, x) = \alpha_j + \beta_j(x - x_j) + \gamma_j(x - x_j)^2 + \delta_j(x - x_j)^3, \tag{1.3.2.4}$$

for all $x \in [x_j, x_{j+1}]$, and where

$$\begin{aligned} \alpha_j &:= S_\Delta(Y, x_j) = y_j, \\ \beta_j &:= S'_\Delta(Y, x_j) = -M_j \frac{(x_{j+1} - x_j)^2}{2h_{j+1}} + A_j = -\frac{M_j h_{j+1}}{2} + A_j \\ &= \frac{-M_j h_{j+1}}{2} + \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j) \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{2M_j + M_{j+1}}{6} h_{j+1}, \\ \gamma_j &:= \frac{1}{2} S''_\Delta(Y, x_j) = \frac{1}{2} M_j, \\ \delta_j &:= \frac{1}{6} S'''_\Delta(Y, x_j^+) = \frac{-M_j}{6h_{j+1}} + \frac{M_{j+1}}{6h_{j+1}} = \frac{M_{j+1} - M_j}{6h_{j+1}}. \end{aligned}$$

This characterizes $S_\Delta(Y, \cdot)$ in terms of its moments M_j. It remains to calculate these moments M_j.

Recall from (1.3.2.2) that

$$\begin{aligned} S'_\Delta(Y, x) &= -M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}} + A_j \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j) - M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}}. \end{aligned}$$

The continuity of $S'_\Delta(Y, \cdot)$ at the interior knots $x = x_j$, $j = 1, 2, \ldots, n-1$, namely, the relations $S'_\Delta(Y, x_j^-) = S'_\Delta(Y, x_j^+)$, yields $n - 1$ equations for the moments M_j. Inserting the

values from (1.3.2.2) into (1.3.2.1) gives for $x \in [x_j, x_{j+1}]$ the following:

$$\begin{aligned} S'_\Delta(Y, x_j^-) &= \frac{y_j - y_{j-1}}{h_j} - \frac{h_j}{6}(M_j - M_{j-1}) + M_j \frac{(x_j - x_{j-1})^2}{2h_j} \\ &= \frac{y_j - y_{j-1}}{h_j} - \frac{h_j M_j}{6} + \frac{h_j M_{j-1}}{6} + \frac{M_j h_j}{2} \\ &= \frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3} M_j + \frac{h_j}{6} M_{j-1}, \end{aligned}$$

and

$$\begin{aligned} S'_\Delta(Y, x_j^+) &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j) - M_j \frac{(x_{j+1} - x_j)^2}{2h_{j+1}} \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6} M_{j+1} + \frac{h_{j+1}}{6} M_j - \frac{h_{j+1}}{2} M_j \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1}. \end{aligned}$$

Since $S'_\Delta(Y, x_j^-) = S'_\Delta(Y, x_j^+)$,

$$\frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3} M_j + \frac{h_j}{6} M_{j-1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1},$$

which implies

$$\frac{h_j}{6} M_{j-1} + \frac{h_{j+1} - h_j}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \tag{1.3.2.5}$$

for the interior moments, $j = 1, 2, \ldots, n-1$. These are $n-1$ equations for the $n+1$ unknown moments. We gain two further equations from each of the side conditions (1.3.1.11).

Case (1) [Natural spline]: $S''_\Delta(Y, a) = M_0 = 0 = M_n = S''_\Delta(Y, b)$.

Case (2) [Periodic spline]: Since $S''_\Delta(Y, a) = S''_\Delta(Y, b)$, evidently $M_0 = M_n$, so that

$$\begin{aligned} S'_\Delta(Y, b) &= \frac{h_n}{6} M_{n-1} + \frac{h_n + h_1}{3} M_n + \frac{h_1}{6} M_1 \\ &= \frac{y_1 - y_n}{h_1} - \frac{y_n - y_{n-1}}{h_n}. \end{aligned}$$

The condition $S'_\Delta(Y, a)$ is similar, recalling that the periodic case requires $y_0 = y_n$.

Case (3) [Clamped spline]: Since $S'_\Delta(Y, a) = f'(a)$,

$$\begin{aligned} \frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 &= \frac{y_1 - y_0}{h_1} - S_\Delta(Y, a) \\ &= \frac{y_1 - y_0}{h_1} - y'_0, \end{aligned}$$

and likewise

$$\begin{aligned} \frac{h_n}{6} M_{n-1} + \frac{h_n}{3} M_n &= S'_\Delta(Y, b) - \frac{y_n - y_{n-1}}{h_n} \\ &= y'_n - \frac{y_n - y_{n-1}}{h_n}. \end{aligned}$$

We can write these equations as well as those in (1.3.2.5) in the following format:

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j, \quad j = 1, 2, \ldots, n-1,$$

where we define

$$\begin{aligned} \lambda_j &:= \frac{h_{j+1}}{h_j + h_{j+1}}, \\ \mu_j &:= 1 - \lambda_j = \frac{h_j}{h_j + h_{j+1}} \\ d_j &:= \frac{6}{h_j + h_{j+1}} \left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right) \end{aligned} \tag{1.3.2.6}$$

For the side conditions, we proceed as follows for each of the three cases:

Case (1) [Natural spline]:

$$\lambda_0 := 0, \quad d_0 := 0, \quad \mu_n := 0, \quad d_n := 0. \tag{1.3.2.7}$$

Case (2) [Periodic spline]:

$$\begin{aligned} \lambda_0 := \frac{h_1}{h_n + h_1}, \quad \mu_0 := \frac{h_n}{h_n + h_1}, \quad \lambda_n := \lambda_0, \quad \mu_n := \mu_0, \\ d_0 := \frac{6}{h_n + h_1} \left(\frac{y_1 - y_n}{h_1} - \frac{y_n - y_{n-1}}{h_n} \right) =: d_n. \end{aligned} \tag{1.3.2.8}$$

Case (3) [Clamped spline]:

$$\begin{aligned} \lambda_0 := 1, \quad d_0 := \frac{6}{h_1} \left(\frac{y_1 - y_1}{h_1} - y_0' \right), \quad \mu_0 := 0, \\ \mu_n := 1, \quad d_n := \frac{6}{h_n} \left(y_n' - \frac{y_n - y_{n-1}}{h_n} \right), \quad \lambda_n := 0. \end{aligned} \tag{1.3.2.9}$$

In cases (1) and (3) we get the following $(n+1) \times (n+1)$ system of linear equations for the moments M_j :

$$\begin{bmatrix} 2 & \lambda_0 & 0 & \cdot & \cdot & 0 \\ \mu_1 & 2 & \lambda_1 & & & \cdot \\ 0 & \mu_2 & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & 0 \\ \cdot & & & \cdot & 2 & \lambda_{n-1} \\ 0 & \cdot & \cdot & 0 & \mu_n & 2 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ \cdot \\ \cdot \\ \cdot \\ M_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \cdot \\ \cdot \\ \cdot \\ d_n \end{bmatrix}. \tag{1.3.2.10}$$

To avoid singularity in the periodic case (2), we have the following $n \times n$ linear system for case (2):

$$\begin{bmatrix} 2 & \lambda_1 & 0 & \ldots & 0 & \mu_1 \\ \mu_2 & 2 & \lambda_2 & 0 & \ldots & 0 \\ 0 & \mu_3 & \cdot & \cdot & & \vdots \\ \vdots & & \cdot & \cdot & \cdot & 0 \\ 0 & & & \cdot & 2 & \lambda_{n-1} \\ \lambda_n & 0 & \ldots & 0 & \mu_n & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \cdot \\ \cdot \\ \cdot \\ M_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_n \end{bmatrix}. \tag{1.3.2.11}$$

Solving the systems (1.3.2.10) and (1.3.2.11) gives the moments M_j, $j = (0,)1, 2, \ldots, n$.

Note in particular that in (1.3.2.10) and (1.3.2.11) we have

$$\mu_j \geq 0, \quad \lambda_j \geq 0,$$

and moreover

$$\mu_j + \lambda_j = 1$$

for $j = 0, 1 \ldots, n$. Also, these coefficients μ_j and λ_j depend only on the partition Δ and not on the prescribed values $y_j \in Y$.

The following result guarantees that these systems are always (uniquely) solvable.

Theorem 1.3.2.2 (Existence of Interpolating Spline Function)**.** *The systems* (1.3.2.10) *and* (1.3.2.11) *of linear equations are nonsingular for any partition* Δ *of* $[a, b]$.

Proof. Consider the $(n+1) \times (n+1)$ matrix

$$A := \begin{bmatrix} 2 & \lambda_0 & 0 & \cdot & \cdot & 0 \\ \mu_1 & 2 & \lambda_1 & & & \cdot \\ 0 & \mu_2 & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & 0 \\ \cdot & & & \cdot & 2 & \lambda_{n-1} \\ 0 & \cdot & \cdot & 0 & \mu_n & 2 \end{bmatrix}$$

of the linear system (1.3.2.10). Since $\lambda_j, \mu_j \geq 0$ for each $j = 0, 1, \ldots, n$, the matrix A has the following property:

$$Az = w \implies \max_{j=0,1,\ldots,n} |z_j| \leq \max_{j=0,1,\ldots,n} |w_j| \tag{1.3.2.12}$$

for all vectors $z, w \in \mathbb{R}^{n+1}$, $z := [z_0, z_1, \ldots, z_n]^\top$, $w := [w_0, w_1, \ldots, w_n]^\top$. Let r be such that $|z_r| = \max_{j=0,1,\ldots,n} |z_j|$. From $Az = w$, we have

$$\mu_r z_{r-1} + 2z_r + \lambda_r z_{r+1} = w_r,$$

where $\mu_0 := 0$ and $\lambda_n := 0$ if necessary. By the definition of r and the fact that $\mu_r + \lambda_r = 1$, it follows

$$\begin{aligned} \max_{j=0,1,\ldots,n} |w_j| &\geq |w_r| \\ &\geq 2|z_r| - \mu_r|z_{r-1}| - \lambda_r|z_{r+1}| \\ &\geq 2|z_r| - \mu_r|z_r| - \lambda_r|z_r| \\ &= (2 - \mu_r - \lambda_r)|z_r| \\ &= |z_r| = \max_{j=0,1,\ldots,n} |z_j|. \end{aligned}$$

By contradiction, suppose that the matrix A were singular. Then there exists a nontrivial solution $z \neq 0$ of the homogeneous system $Az = 0$, from which (1.3.2.12) yields the contradiction

$$0 < \max_{j=0,1,\ldots,n} |z_j| \leq 0,$$

so that evidently $z = 0$. In other words, $z = 0$ and $z \neq 0$ simultaneously, and you should really be ashamed of yourself for ever supposing the conclusion was not so.

The nonsingularity of the matrix in (1.3.2.11) is shown similarly. This completes the proof. □

1.3.3. *Convergence Properties of Cubic Spline Functions.* Recall that interpolating polynomials may not converge to a function f whose values they interpolate, even if the partitions Δ are chosen such that the fineness $\|\Delta\|$ of the partition converges to zero (see section 1.1.4). On the other hand, interpolating spline functions do converge towards f as $\|\Delta\|$ approaches zero, provided mild conditions on f and the partitions Δ are satisfied.

We first show that the moments of the interpolating spline function $S_\Delta(Y,\cdot)$ converge to the second derivatives of the given function f. For concreteness, fix a partition

$$\Delta := \{a = x_0 < x_1 < \ldots, x_n = b\}$$

of the interval $[a,b]$, and let

$$M := [M_0, M_1, \ldots, M_n]^\top$$

be the vector of moments M_j of the interpolating spline function $S_\Delta(Y,\cdot)$ with $y_j := f(x_j)$, $j = 0,1,\ldots,n$, as well as the clamped side condition

$$S'_\Delta(Y,a) = f'(a), \quad S'_\Delta(Y,b) = f'(b).$$

Note that the vector M of moments satisfies the equation

$$AM = d,$$

where

$$A := \begin{bmatrix} 2 & \lambda_0 & 0 & \cdot & \cdot & 0 \\ \mu_1 & 2 & \lambda_1 & & & \cdot \\ 0 & \mu_2 & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & 0 \\ \cdot & & & \cdot & 2 & \lambda_{n-1} \\ 0 & \cdot & \cdot & 0 & \mu_n & 2 \end{bmatrix}$$

as in (1.3.2.10) and

$$d_0 := \frac{6}{h_1}\left(\frac{y_1 - y_0}{h_1} - y'_0\right), \quad d_n := \frac{6}{h_n}\left(y'_n - \frac{y_n - y_{n-1}}{h_n}\right),$$
$$d_j := \frac{6}{h_j + h_{j+1}}\left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}\right), \quad j = 1,2,\ldots,n-1,$$

as in (1.3.2.6) of the previous section. Let F and r be the vectors

$$F := \begin{bmatrix} f''(x_0) \\ f''(x_1) \\ \vdots \\ f''(x_n) \end{bmatrix}, \quad r := d - AF = A(M - F).$$

Denoting by $\|z\|_\infty := \max_{j=0,1,\ldots,n} |z_j|$ the infinity norm for vectors $z \in \mathbb{R}^n$ and

$$\|h_\Delta\|_\infty := \max_{j=0,1,\ldots,n-1} |x_{j+1} - x_j|$$

the fineness of the partition Δ, we get the following result.

Lemma 1.3.3.1 (Convergence of Moments)**.** *If $f \in C^4[a,b]$ and $|f^{(4)}(x)| \le L$ for all $x \in [a,b]$, then*

$$\|M - F\|_\infty \le \|r\|_\infty \le \frac{3}{4} L \|h_\Delta\|_\infty^2.$$

Proof. We start with r_0. Note that

$$\begin{aligned} r_0 &= d_0 - (AF)_0 \\ &= d_0 - 2f''(x_0) - \lambda_0 f''(x_1) \\ &= \frac{6}{h_1}\left(\frac{y_1 - y_0}{h_1} - y_0'\right) - 2f''(x_0) - f''(x_1), \end{aligned}$$

since $\lambda_0 + \mu_0 = \lambda_0 = 1$. Using Taylor's theorem to express $y_1 = f(x_1)$ and $f''(x_1)$ in terms of f about x_0 gives

$$\begin{aligned} r_0 &= \frac{6}{h_1}\left[\frac{y_1}{h_1} - \frac{y_0}{h_1} - y_0'\right] - 2f''(x_0) - f''(x_1) \\ &= \frac{6}{h_1}\left[\frac{f(x_0) + h_1 f'(x_0) + \frac{h_1^2}{2}f''(x_0) + \frac{h_1^3}{6}f'''(x_0) + \frac{h_1^4}{24}f^{(4)}(\tau_1)}{h_1} - \frac{f(x_0)}{h_1} - \frac{h_1 f'(x_0)}{h_1}\right] - \\ &\qquad 2f''(x_0) - \left[f''(x_0) + h_1 f'''(x_0) + \frac{h_1^2}{2}f^{(4)}(\tau_2)\right] \\ &= \frac{6}{h_1}\left[\frac{h_1}{2}f''(x_0) + \frac{h_1^2}{6}f'''(x_0) + \frac{h_1^3}{24}f^{(4)}(\tau_1)\right] - 2f''(x_0) - f''(x_0) - h_1 f'''(x_0) - \frac{h_1^2}{2}f^{(4)}(\tau_2) \\ &= \left[3f''(x_0) + h_1 f'''(x_0) + \frac{h_1^2}{4}f^{(4)}(\tau_1)\right] - 3f''(x_0) - h_1 f'''(x_0) - \frac{h_1^2}{2}f^{(4)}(\tau_2) \\ &= \frac{h_1^2}{4}f^{(4)}(\tau_1) - \frac{h_1^2}{2}f^{(4)}(\tau_2), \end{aligned}$$

for some $\tau_1, \tau_2 \in [x_0, x_1]$. Hence,

$$\begin{aligned} |r_0| &\le \left|\frac{h_1^2}{4}f^{(4)}(\tau_1)\right| + \left|\frac{h_1^2}{2}f^{(4)}(\tau_2)\right| \\ &\le \frac{h_1^2}{4}L + \frac{h_1^2}{2}L \\ &\le \frac{3}{4}L\|h_\Delta\|_\infty^2. \end{aligned}$$

Analogously, we find for

$$r_n = d_n - (AF)_n = d_n - f''(x_{n-1}) - 2f''(x_n)$$

that

$$|r_n| \le \frac{3}{4}L\|h_\Delta\|_\infty^2.$$

We now turn to the consideration of r_j for $j = 1, 2, \ldots, n-1$. Observe

$$\begin{aligned} r_j &= d_j - (AF)_j = d_j - \mu_j F_{j-1} - 2F_j - \lambda_j F_{j+1} = d_j - \mu_j f''(x_{j-1}) - 2f''(x_j) - \lambda_j f''(x_{j+1}) \\ &= \frac{6}{h_j + h_{j+1}}\left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}\right) - \frac{h_j}{h_j + h_{j+1}}f''(x_{j-1}) - 2f''(x_j) - \frac{h_{j+1}}{h_j + h_{j+1}}f''(x_{j+1}) \\ &= \frac{6}{h_j + h_{j+1}}\left(\frac{f(x_{j+1})}{h_{j+1}} - \frac{f(x_j)}{h_{j+1}} - \frac{f(x_j)}{h_j} - \frac{f(x_{j-1})}{h_j}\right) - \frac{h_j}{h_j + h_{j+1}}f''(x_{j-1}) - \\ &\qquad 2f''(x_j) - \frac{h_{j+1}}{h_j + h_{j+1}}f''(x_{j+1}). \end{aligned}$$

Applying Taylor's theorem about x_j then gives

$$\begin{aligned}
r_j = \frac{6}{h_j+h_{j+1}} &\left[\left(\frac{f(x_j)+h_{j+1}f'(x_j)+\frac{h_{j+1}^2}{2}f''(x_j)+\frac{h_{j+1}^3}{6}f'''(x_j)+\frac{h_{j+1}^4}{24}f^{(4)}(\tau_1)}{h_{j+1}}\right)-\right.\\
&\left.\frac{f(x_j)}{h_{j+1}}-\frac{f(x_j)}{h_j}+\left(\frac{f(x_j)-h_jf'(x_j)+\frac{h_j^2}{2}f''(x_j)-\frac{h_j^3}{6}f'''(x_j)+\frac{h_j^4}{24}f^{(4)}(\tau_2)}{h_j}\right)\right]-\\
&\frac{h_j}{h_j+h_{j+1}}\left[f''(x_j)-h_jf'''(x_j)+\frac{h_j^2}{2}f^{(4)}(\tau_3)\right]-2f''(x_j)-\\
&\frac{h_{j+1}}{h_j+h_{j+1}}\left[f''(x_j)+h_{j+1}f'''(x_j)+\frac{h_{j+1}^2}{2}f^{(4)}(\tau_4)\right]\\
=\frac{6}{h_j+h_{j+1}}&\left[\frac{h_{j+1}}{2}f''(x_j)+\frac{h_{j+1}^2}{6}f'''(x_j)+\frac{h_{j+1}^3}{24}f^{(4)}(\tau_1)+\frac{h_j}{2}f''(x_j)-\frac{h_j^2}{6}f'''(x_j)+\right.\\
&\left.\frac{h_j^3}{24}f^{(4)}(\tau_2)\right]-\frac{h_j}{h_j+h_{j+1}}\left[f''(x_j)-h_jf'''(x_j)+\frac{h_j^2}{2}f^{(4)}(\tau_3)\right]-2f''(x_j)-\\
&\frac{h_{j+1}}{h_j+h_{j+1}}\left[f''(x_j)+h_{j+1}f'''(x_j)+\frac{h_{j+1}^2}{2}f^{(4)}(\tau_4)\right]\\
=\frac{1}{h_j+h_{j+1}}&\left[\frac{h_{j+1}^3}{4}f^{(4)}(\tau_1)+\frac{h_j^3}{4}f^{(4)}(\tau_2)-\frac{h_j^3}{2}f^{(4)}(\tau_3)-\frac{h_{j+1}^3}{2}f^{(4)}(\tau_4)\right],
\end{aligned}$$

for some $\tau_i \in [x_{j-1}, x_{j+1}]$, $i = 1, 2, 3, 4$. Thus

$$\begin{aligned}
|r_j| &\le \frac{1}{h_j+h_{j+1}}\left[\frac{h_{j+1}^3}{4}L+\frac{h_j^3}{4}L+\frac{h_j^3}{2}L+\frac{h_{j+1}^3}{2}L\right]\\
&= \frac{1}{h_j+h_{j+1}}\left[\frac{3h_j^3}{4}L+\frac{3h_{j+1}^3}{4}L\right] = \frac{3}{4}L\left[\frac{h_j^3+h_{j+1}^3}{h_j+h_{j+1}}\right]\\
&\le \frac{3}{4}L\|h_\Delta\|_\infty^2\left[\frac{h_j+h_{j+1}}{h_j+h_{j+1}}\right]\\
&= \frac{3}{4}L\|h_\Delta\|_\infty^2.
\end{aligned}$$

This shows

$$\|r\|_\infty \le \frac{3}{4}L\|h_\Delta\|_\infty^2.$$

Recall that since $\mu_j, \lambda_j \ge 0$ for all $j = 0, 1, \ldots, n$ and $\mu_j + \lambda_j = 1$ for $j = 1, 2, \ldots, n-1$, we have

$$\max_{j=0,1,\ldots,n}|M_j - F_j| \le \max_{j=0,1,\ldots,n}|r_j|.$$

That is,

$$\|M - F\|_\infty \le \|r\|_\infty.$$

Hence,

$$\|M - F\|_\infty \le \|r\|_\infty \le \frac{3}{4}L\|h_\Delta\|_\infty^2.$$

This completes the proof. □

We arrive at the main convergence result from this section.

Theorem 1.3.3.2 (Convergence of Interpolating Spline Functions)**.** *Suppose that* $f \in C^4[a,b]$ *and* $|f^{(4)}(x)| \leq L$ *for all* $x \in [a,b]$*. Let* Δ *be a partition* $\Delta := \{a = x_0 < x_1 < \cdots < x_n = b\}$ *of the interval* $[a,b]$*, and* $K \in \mathbb{R}$ *a constant such that*

$$\frac{\|h_\Delta\|_\infty}{|x_{j+1} - x_j|} \leq K, \quad j = 0, 1, \ldots, n-1.$$

If S_Δ *is the spline function that interpolates the values of the function* f *at the knots* $x_j \in \Delta$*,* $j = 0, 1, \ldots, n$*, and satisfies*

$$S'_\Delta(a) = f'(a), \quad S'_\Delta(b) = f'(b),$$

then there exist constants $c_r \leq 2$*, which do not depend on the partition* Δ*, such that*

$$|f^{(r)}(x) - S^{(r)}_\Delta(x)| \leq c_r L K \|h_\Delta\|_\infty^{4-r}, \quad r = 0, 1, 2, 3.$$

We note that the constant $K \geq 1$ bounds the deviation of the partition Δ from uniformity, that is, K guarantees that there is no clustering of knots.

Proof. We begin with the case $r = 3$. Recall that, for $x \in [x_{j-1}, x_j]$,

$$S''_\Delta(x) = M_{j-1}\frac{x_j - x}{h_j} + M_j\frac{x - x_{j-1}}{h_j},$$

so that

$$S'''_\Delta(x) = -\frac{M_{j-1}}{h_j} + \frac{M_j}{h_j} = \frac{M_j - M_{j-1}}{h_j}.$$

Thus for all $x \in [x_{j-1}, x_j]$,

$$\begin{aligned} S'''_\Delta(x) - f'''(x) &= \frac{M_j - M_{j-1}}{h_j} - f'''(x) \\ &= \frac{M_j - f''(x_j)}{h_j} - \frac{M_{j-1} - f''(x_{j-1})}{h_j} + \frac{1}{h_j}\left[f''(x_j) - f''(x) \; + \right. \\ &\qquad \left. (f''(x) - f''(x_{j-1}))\right] - f'''(x). \end{aligned}$$

Using Taylor's theorem to express the derivatives of f about x, we have

$$\begin{aligned} S'''_\Delta(x) - f'''(x) &= \frac{M_j - f''(x_j)}{h_j} - \frac{M_{j-1} - f''(x_{j-1})}{h_j} + \frac{1}{h_j}\left[f''(x) + (x_j - x)f'''(x) + \right. \\ &\qquad \left. \frac{1}{2}(x_j - x)^2 f^{(4)}(\tau_1) - f''(x)\right] + \frac{1}{h_j}\left[f''(x) - (f''(x) + \right. \\ &\qquad \left. (x_{j-1} - x)f'''(x) + \frac{1}{2}(x_{j-1} - x)^2 f^{(4)}(\tau_2)\right)\Bigg] - f'''(x) \\ &= \frac{M_j - f''(x_j)}{h_j} - \frac{M_{j-1} - f''(x_{j-1})}{h_j} + \frac{1}{h_j}\left[(x_j - x)f'''(x) + \frac{1}{2}(x_j - x)^2 f^{(4)}(\tau_1) - \right. \\ &\qquad \left. (x_{j-1} - x)f'''(x) - \frac{1}{2}(x_{j-1} - x)^2 f^{(4)}(\tau_2) - h_j f'''(x)\right] \\ &= \frac{M_j - f''(x_j)}{h_j} - \frac{M_{j-1} - f''(x_{j-1})}{h_j} + \frac{1}{h_j}\left[(x_j - x_{j-1})f'''(x) - h_j f'''(x) + \right. \end{aligned}$$

$$\frac{1}{2}(x_j - x)^2 f^{(4)}(\tau_1) - \frac{1}{2}(x_{j-1} - x)^2 f^{(4)}(\tau_2)\Big],$$

for some $\tau_1, \tau_2 \in [x_{j-1}, x_j]$. From (1.3.3.1) and the fact that $h_j = x_j - x_{j-1}$, as well as $x - x_j \le h_j$, we have

$$\begin{aligned}
|f'''(x) - S'''_\Delta(x)| &\le \frac{3}{4}L\frac{\|h_\Delta\|_\infty^2}{h_j} + \frac{3}{4}L\frac{\|h_\Delta\|_\infty^2}{h_j} + \frac{1}{h_j}\left[\frac{1}{2}L\|h_\Delta\|_\infty^2\right] \\
&= \frac{3}{2}L\frac{\|h_\Delta\|_\infty^2}{h_j} + \frac{L}{2}\frac{\|h_\Delta\|_\infty^2}{h_j} = 2L\frac{\|h_\Delta\|_\infty^2}{h_j} \\
&\le 2L\|h_\Delta\|_\infty.
\end{aligned}$$

Since $K \ge 1$, we conclude that

$$|f'''(x) - S'''_\Delta(x)| \le 2LK\|h_\Delta\|_\infty.$$

We now show the case $r = 2$. Let $x \in [a, b]$. There exists a closest knot x_j. Without loss of generality, assume that $x < x_j$, so that $x \in [x_{j-1}, x_j]$. We may also assume that $|x_j - x| \le \frac{h_j}{2} \le \frac{1}{2}\|h_\Delta\|_\infty$. From (1.3.3.1) and the result for $r = 3$, we have

$$\begin{aligned}
|f''(x) - S''_\Delta(x)| &= f''(x_j) - S''_\Delta(x_j) + \int_{x_j}^{x} (f'''(t) - S'''_\Delta(t))\, dt \\
&\le \frac{3}{4}L\|h_\Delta\|_\infty^2 + \int_{x_j}^{x} 2LK\|h_\Delta\|_\infty\, dt \\
&= \frac{3}{4}L\|h_\Delta\|_\infty^2 + (2LK\|h_\Delta\|_\infty)|_{t=x_j}^{x} \\
&= \frac{3}{4}L\|h_\Delta\|_\infty^2 + 2LK\|h_\Delta\|_\infty^2(x - x_j) \\
&\le \frac{3}{4}L\|h_\Delta\|_\infty^2 + LK\|h_\Delta\|_\infty^2 \\
&\le \frac{3}{4}LK\|h_\Delta\|_\infty^2 + LK\|h_\Delta\|_\infty^2 \\
&= \frac{7}{4}LK\|h_\Delta\|_\infty^2,
\end{aligned}$$

since $K \ge 1$. Hence,

$$|f''(x) - S''_\Delta(x)| \le \frac{7}{4}LK\|h_\Delta\|_\infty^2.$$

We next consider $r = 1$. In addition to the boundary points $\xi_0 := a$, $\xi_{n+1} := b$, it follows by Rolle's theorem (1.1.4.1) that there exist n further points $\xi_j \in [x_{j-1}, x_j]$, $j = 1, 2, \ldots, n$, such that

$$f'(\xi_j) = S'_\Delta(x_j), \quad j = 0, 1, \ldots, n+1,$$

by the side conditions. Let $x \in [a, b]$. There exists a closest point ξ_j, for which we have that

$$|x - \xi_j| \le \|h_\Delta\|_\infty.$$

Thus

$$f'(x) - S'_\Delta(x) = \int_{\xi_j}^{x} (f''(t) - S''_\Delta(t))\, dt.$$

By the result for $r = 2$,

$$\begin{aligned}|f'(x) - S'_\Delta(x)| &\leq \left| \int_{\xi_j}^{x} (f''(t) - S''_\Delta(t))\ dt \right| \\ &\leq \int_{\xi_j}^{x} \frac{7}{4} LK \|h_\Delta\|_\infty^2\ dt \\ &= \frac{7}{4} LK \|h_\Delta\|_\infty^2 (x - \xi_j) \\ &\leq \frac{7}{4} LK \|h_\Delta\|_\infty^3.\end{aligned}$$

This proves

$$|f'(x) - S'_\Delta(x)| \leq \frac{7}{4} LK \|h_\Delta\|_\infty^3.$$

Finally, we show $r = 0$. Let $x \in [a, b]$ and recall that there exists a closest knot x_j. Without loss of generality, say $x \leq x_j$ such that $x \in [x_{j-1}, x_j]$ and $|x_j - x| \leq \frac{1}{2}\|h_\Delta\|_\infty$. Note that, by the fundamental theorem of calculus,

$$f(x) - S_\Delta(x) = \int_{x_j}^{x} (f'(t) - S'_\Delta(t)\ dt.$$

By the result for $r = 1$, it follows

$$\begin{aligned}|f(x) - S_\Delta(x)| &\leq \int_{x_j}^{x} \frac{7}{4} LK \|h_\Delta\|_\infty^3 \\ &= \frac{7}{4} LK \|h_\Delta\|_\infty t|_{t=x_j}^{x} \\ &\leq \frac{7}{8} LK \|h_\Delta\|_\infty^4.\end{aligned}$$

Hence,

$$|f(x) - S_\Delta(x) \leq \frac{7}{8} LK \|h_\Delta\|_\infty^4,$$

which completes the proof. □

Note that Theorem (1.3.3.2) implies that for sequences

$$\Delta_m := \{a = x_0^{(m)} < x_1^{(m)} < \cdots < x_{n_m}^{(m)} = b\}$$

of partitions with $\|h_{\Delta_m}\|_\infty \to 0$ which satisfy the conditions of (1.3.3.2), the corresponding interpolating spline functions S_{Δ_m} that satisfy the hypotheses of (1.3.3.2) and their first two derivatives converge uniformly to f and its first two derivatives on $[a, b]$. This is much different from the case of standard polynomial interpolation, where the polynomial interpolant may not converge even pointwise to f for an arbitrary partition.

2. Function Approximation

2.1. **Least Squares Approximation.** Least squares function approximation seeks to find for a given $f \in C[a,b]$ and positive integer n the polynomial $p_n \in \Pi^n$ which minimizes the least squares ($L^2[a,b]$) error

$$\|f - p_n\|_2 := \left\{ \int_a^b (f(x) - p_n(x))^2 \, dx \right\}^{1/2}, \tag{2.1.0.1}$$

so that for all $q \in \Pi^n$ we have

$$\left\{ \int_a^b (f(x) - p_n(x))^2 \, dx \right\}^{1/2} \leq \left\{ \int_a^b (f(x) - q(x))^2 \, dx \right\}^{1/2}.$$

2.1.1. *Orthogonal Polynomials and Least Squares Approximation.* Let $f \in C[a,b]$ and let $p_n \in \Pi^n$ be the polynomial of degree at most n that minimizes the $L^2[a,b]$ error

$$\int_a^b (f(x) - p_n(x))^2 \, dx.$$

Since $p_n \in \Pi^n$, we may write

$$p_n(x) = \sum_{k=0}^{n} c_k x^k = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n.$$

Define the **minimizer**

$$E(c_0, c_1, \ldots, c_n) = \int_a^b (f(x) - p_n(x))^2 \, dx. \tag{2.1.1.1}$$

Thus, the problem becomes one of finding coefficients $c_0, c_1, \ldots, c_n$ for p_n such that

$$E(c_0, c_1, \ldots, c_n) \leq \int_a^b (f(x) - q(x))^2 \, dx$$

for all $q \in \Pi^n$. Evidently, a necessary condition for the coefficients $c_0, c_1, \ldots, c_n$ to minimize (2.1.1.1) is that

$$\frac{\partial E(c_0, c_1, \ldots, c_n)}{\partial c_j} = 0, \quad \text{for each } j = 0, 1, \ldots, n.$$

Since

$$\begin{aligned}
E(c_0, c_1, \ldots, c_n) &= \int_a^b (f(x) - p_n(x))^2 \, dx \\
&= \int_a^b [f(x)]^2 \, dx - 2\int_a^b f(x) p_n(x) \, dx + \int_a^b [p_n(x)]^2 \, dx \\
&= \int_a^b [f(x)]^2 \, dx - 2\int_a^b f(x) \left[\sum_{k=0}^{n} c_k x^k\right] dx + \int_a^b \left[\sum_{k=0}^{n} c_k x^k\right]^2 dx \\
&= \int_a^b [f(x)]^2 \, dx - 2\sum_{k=0}^{n} c_k \int_a^b x^k f(x) \, dx + \int_a^b \left[\sum_{k=0}^{n} c_k x^k\right]^2 dx,
\end{aligned}$$

we have for each $j = 0, 1, \ldots, n$ that

$$\frac{\partial E(c_0, c_1, \ldots, c_n)}{\partial c_j} = -2\int_a^b x^j f(x) \, dx + 2\sum_{k=0}^{n} c_k \int_a^b x^{j+k} \, dx. \tag{2.1.1.2}$$

Rearranging (2.1.1.2), we find $p_n(x)$ by setting $\frac{\partial E(c_0, c_1, \ldots, c_n}{\partial c_j} = 0$ for each $j = 0, 1, \ldots, n$ to get the $n+1$ **normal equations**

$$\sum_{k=0}^{n} c_k \int_a^b x^{j+k} \, dx = \int_a^b x^j f(x) \, dx, \quad j = 0, 1, \ldots, n, \tag{2.1.1.3}$$

that must be solved to obtain the $n+1$ coefficients $c_0, c_1, \ldots, c_n$. The normal equations always have a unique solution, provided that $f \in C[a,b]$.

Example 2.1.1.1. *Given $f(x) = x^2$, we calculate the polynomial $p \in \Pi^1$ that satisfies*

$$\int_0^2 (f(x) - p(x))^2 \, dx \leq \int_0^2 (f(x) - q(x))^2 \, dx$$

for all $q \in \Pi^1$. The normal equations for

$$p(x) := a_0 + a_1 x$$

are:

$$a_0 \int_0^2 dx + a_1 \int_0^2 x \, dx = \int_0^2 x^2 \, dx,$$
$$a_0 \int_0^2 x \, dx + a_1 \int_0^2 x^2 \, dx = \int_0^2 x^3 \, dx.$$

Performing the integration gives

$$2a_0 + 2a_1 = \frac{8}{3},$$
$$2a_0 + \frac{8}{3} a_1 = 4.$$

Solving the above system gives

$$a_0 = -\frac{2}{3}, \quad a_1 = 2.$$

Thus

$$p(x) = -\frac{2}{3} + 2x.$$

Moreover,

$$\int_0^2 (f(x) - p(x))^2 = \int_0^2 \left(\frac{2}{3} - 2x + x^2 \right)^2 \, dx = 0.$$

We now turn to a discussion of methods to solve the $n + 1$ normal equations (2.1.1.3). Noting that

$$\int_a^b x^{j+k} \, dx = \frac{1}{j+k+1} x^{j+k+1} \Big|_a^b = \frac{1}{j+k+1} \left(b^{j+k+1} - a^{j+k+1} \right),$$

observe that the solutions to the normal equations (2.1.1.3) are of the form

$$\sum_{k=0}^{n} \frac{c_k}{j+k+1} \left(b^{j+k+1} - a^{j+k+1} \right) = \int_a^b x^k f(x) \, dx. \tag{2.1.1.4}$$

The matrix of the $(n+1) \times (n+1)$ linear system obtained by (2.1.1.4) is known as the **Hilbert matrix**

$$\begin{bmatrix} b-a & \frac{1}{2}(b^2 - a^2) & \dots & \frac{1}{n}(b^n - a^n) \\ \frac{1}{2}(b^2 - a^2) & b-a & \dots & \frac{1}{n-1}(b^{n-1} - a^{n-1}) \\ \vdots & \ddots & \ddots & \vdots \\ \frac{1}{n}(b^n - a^n) & \frac{1}{n-1}(b^{n-1} - a^{n-1}) & \dots & b-a \end{bmatrix}.$$

A few problems arise:

(1) The Hilbert matrix is notorious for roundoff error difficulties;

(2) The system (2.1.1.4) is dense (not sparse) and does not have an easily calculated numerical solution;
(3) There is no indication of how to use the calculation of p_n to obtain p_{n+1}, that is, the work done to find p_n does not lessen the amount of work required to find p_{n+1}.

Thus we are motivated to use a different polynomial basis for Π^n than

$$\mathcal{B} := \{1, x, x^2, \ldots, x^n\}.$$

Definition 2.1.1.2 (Linear Independence)**.** *The set of functions $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is said to be* **linearly independent** *on the interval $[a, b]$ if whenever*

$$\sum_{k=0}^{n} c_k \phi_k(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0$$

for all $x \in [a, b]$, we have $c_j = 0$ for all $j = 0, 1, \ldots, n$.

Definition 2.1.1.3 (Linear Dependence)**.** *The set of functions $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is* **linearly dependent** *if it is not linearly independent.*

Example 2.1.1.4. *We show that the set $\{1, x, x^2, \ldots, x^n\}$ is linearly independent on $[a, b]$.*
Let $c_j \in \mathbb{R}$, $j = 0, 1, \ldots, n$ be such that

$$p(x) := \sum_{k=0}^{n} c_k x^k = c_0 + c_1 x + \cdots + c_n x^n = 0$$

for all $x \in [a, b]$. Note that p is a polynomial of degree at most n such that

$$p(x) = 0$$

for all $x \in [a, b]$. Since $[a, b]$ is uncountable, p must vanish identically, so that

$$p(x) \equiv 0$$

on $[a, b]$. From this it follows that p is the zero polynomial, so that $c_j = 0$ for each $j = 0, 1, \ldots, n$.

We get the following generalization of the above example.

Theorem 2.1.1.5. *Suppose that, for each $j = 0, 1, \ldots, n$, $\phi_j(x)$ is a polynomial of degree precisely j. Then $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is linearly independent on any interval $[a, b]$.*

Proof. Let $\{\phi_0, \phi_1, \ldots, \phi_n\}$ be a set of functions that satisfy the assumptions of (2.1.1.5) and let $c_j \in \mathbb{R}$, $j = 0, 1, \ldots, n$ be such that

$$p(x) := c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0$$

for all $x \in [a, b]$. Since $\{1, x, \ldots, x^n\}$ is linearly independent on $[a, b]$, there exist coefficients $\beta_0, \beta_1, \ldots, \beta_n \in \mathbb{R}$ such that

$$p(x) = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_n x^n.$$

Thus the polynomial p vanishes identically on $[a, b]$, so that $\beta_j = 0$ for all $j = 0, 1, \ldots, n$. In particular, $\beta_n = 0$. But $c_n\phi_n(x)$ is the only term in p that contains x^n, so we must have $c_n = \beta_n = 0$. Thus

$$p(x) = \sum_{k=0}^{n} c_k\phi_k(x) = \sum_{k=0}^{n-1} c_k\phi_k(x).$$

Continuing in such fashion, it may be shown that the remaining constants $c_{n-1}, c_{n-2}, \ldots, c_0$ are all zero, from which it follows that $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is linearly independent on $[a, b]$. □

We get the following result from linear algebra.

Theorem 2.1.1.6. *Suppose that $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is a collection of linearly independent polynomials in Π^n. Then any polynomial $p \in \Pi^n$ can be written uniquely as a linear combination of $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$.*

Proof. By the invertible matrix theorem (1.1.5.1), it suffices to show uniqueness.

Let $p \in \Pi^n$ and suppose that there exist real coefficients α_j, β_j, $j = 0, 1, \ldots, n$, such that

$$p(x) = \sum_{k=0}^{n} \alpha_k \phi_k(x) \quad \text{and} \quad p(x) = \sum_{k=0}^{n} \beta_k \phi_k(x)$$

for all $x \in [a, b]$. Then

$$\left(\sum_{k=0}^{n} \alpha_k \phi_k(x) \right) - \left(\sum_{k=0}^{n} \beta_k \phi_k(x) \right) = \sum_{k=0}^{n} (\alpha_k - \beta_k) \phi_k(x) = 0.$$

Since the set $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is linearly independent, we must have $\alpha_j - \beta_j = 0$ for every $j = 0, 1, \ldots, n$, so that $\alpha_j = \beta_j$ for all $j = 0, 1, \ldots, n$.

It follows that any $p \in \Pi^n$ can be written uniquely as a linear combination of $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$. □

We introduce the concepts of a weight function and orthogonality.

Definition 2.1.1.7 (Weight Function)**.** *An integrable function w is called a* **weight function** *on the interval $[a, b]$ if $w(x) \geq 0$ for all $x \in [a, b]$ and w does not vanish identically on any subinterval of $[a, b]$, that is, the zero set*

$$\mathcal{Z}(w) := \{x \in [a, b] : w(x) = 0\}$$

has measure zero.

The purpose of the weight function w is to assign varying degrees of importance to approximations on certain portions of the interval $[a, b]$. For instance, the weight function

$$w(x) := \frac{1}{\sqrt{1 - x^2}}$$

places less emphasis on the center of the interval $[-1, 1]$ and more emphasis when $|x|$ is near 1.

We revisit the least squares problem with the addition of a weight function w. Let $\{\phi_0, \phi_1, \ldots, \phi_n\}$ be a set of linearly independent functions on $[a, b]$ and let w be a weight function on the interval $[a, b]$. Given $f \in C[a, b]$, recall that we want a linear combination

$$p(x) := \sum_{k=0}^{n} c_k \phi_k(x)$$

to minimize the $L^2[a, b]$ error

$$E(c_0, c_1, \ldots, c_n) = \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} c_k \phi_k(x) \right]^2 dx.$$

For each $j = 0, 1, \ldots, n$ we have

$$\frac{\partial E(c_0, c_1, \ldots, c_n)}{\partial c_j} = 2 \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} c_k \phi_k(x) \right] \phi_j(x) \; dx = 0.$$

Thus the normal equations (2.1.1.3) become

$$\int_a^b w(x) f(x) \phi_j(x) \; dx = \sum_{k=0}^{n} c_k \int_a^b w(x) \phi_k(x) \phi_j(x) \; dx, \quad j = 0, 1, \ldots, n. \tag{2.1.1.5}$$

If we can choose the set of functions $\{\phi_0, \phi_1, \ldots, \phi_n\}$ such that

$$\int_a^b w(x) \phi_k(x) \phi_j(x) \; dx = \begin{cases} 0, & k \neq j, \\ \alpha_j, & k = j, \end{cases} \tag{2.1.1.6}$$

then the normal equations (2.1.1.5) reduce to the remarkably simple system

$$\begin{aligned} \int_a^b w(x) f(x) \phi_j(x) \; dx &= \sum_{k=0}^{n} c_k \int_a^b w(x) \phi_k(x) \phi_j(x) \; dx \\ &= c_j \int_a^b w(x) [\phi_j(x)]^2 \; dx \\ &= c_j \alpha_j \end{aligned}$$

for each $j = 0, 1, \ldots, n$. From (2.1.1.5) and (2.1.1.6) we can solve for each c_j, $j = 0, 1, \ldots, n$ easily to find

$$c_j = \frac{1}{\alpha_j} \int_a^b w(x) f(x) \phi_j(x) \; dx, \quad j = 0, 1, \ldots, n. \tag{2.1.1.7}$$

We note here that the polynomials $\{\phi_0, \phi_1, \ldots, \phi_n\}$ chosen such that (2.1.1.6) holds are said to satisfy an **orthogonality condition**, and we can see that this greatly simplifies the least squares approximation problem.

Definition 2.1.1.8 (Orthogonal Set of Functions)**.** *The set of functions $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is said to be $w-$***orthogonal** *on the interval $[a, b]$ with respect to the weight functions w if*

$$\int_a^b w(x) \phi_k(x) \phi_j(x) \; dx = \begin{cases} 0, & k \neq j, \\ \alpha_j > 0, & k = j. \end{cases}$$

Definition 2.1.1.9 (Orthonormal Set of Functions)**.** *Let $\{\phi_0, \phi_1, \ldots, \phi_n\}$ satisfy the conditions of* (2.1.1.8)*. If, in addition,*

$$\int_a^b w(x) [\phi_j(x)]^2 \; dx := \alpha_j = 1$$

*for each $j = 0, 1, \ldots, n$, the set $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is called $w-$***orthonormal** *on the interval $[a, b]$.*

We get the following theorem.

Theorem 2.1.1.10 (Construction of Least Squares Approximant)**.** *If $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is an orthogonal set of functions on $[a, b]$ with respect to the weight function w, then the least squares approximation of f with respect to w is*

$$p(x) = \sum_{k=0}^{n} c_j \phi_j(x),$$

where, for each $j = 0, 1, \ldots, n$,

$$c_j := \frac{\int_a^b w(x) f(x) \phi_j(x)\ dx}{\int_a^b w(x)[\phi_j(x)]^2\ dx} = \frac{1}{\alpha_j} \int_a^b w(x) f(x) \phi_j(x)\ dx.$$

Proof. Let $p \in \Pi^n$ be the least squares approximant of f,

$$p(x) := \sum_{k=0}^{n} c_k \phi_k(x),$$

where c_j and $\phi_j(x)$ are defined as in the statement of (2.1.1.10), $j = 0, 1, \ldots, n$. If $q \in \Pi^n$ is any other polynomial, then

$$\int_a^b w(x)(f(x) - q(x))^2\ dx \tag{2.1.1.8}$$

$$= \int_a^b w(x)((f(x) - p(x)) + (p(x) - q(x)))^2\ dx$$

$$= \int_a^b w(x)(f(x) - p(x))^2\ dx + 2\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx + \int_a^b w(x)(p(x) - q(x))^2\ dx$$

$$\geq \int_a^b w(x)(f(x) - p(x))^2\ dx + 2\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx. \tag{2.1.1.9}$$

We pass to the consideration of

$$\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx.$$

First note that $p - q \in \Pi^n$, so there exist real coefficients a_j, $j = 0, 1, \ldots, n$, such that

$$p(x) - q(x) = \sum_{k=0}^{n} a_k \phi_k(x).$$

Observe

$$\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx$$

$$= \int_a^b w(x)(f(x) - p(x)) \left(\sum_{k=0}^{n} a_k \phi_k(x) \right)\ dx$$

$$= \sum_{k=0}^{n} a_k \int_a^b w(x)f(x)\phi_k(x)\ dx - \sum_{k=0}^{n} a_k \int_a^b w(x)p(x)\phi_k(x)\ dx$$
$$= \sum_{k=0}^{n} a_k \int_a^b w(x)f(x)\phi_k(x)\ dx - \sum_{k=0}^{n} a_k \int_a^b w(x)\left(\sum_{l=0}^{n} c_l\phi_l(x)\right)\phi_k(x)\ dx$$
$$= \sum_{k=0}^{n} a_k \int_a^b w(x)f(x)\phi_k(x)\ dx - \sum_{k=0}^{n} a_k \sum_{l=0}^{n} c_l \int_a^b w(x)\phi_l(x)\phi_k(x)\ dx.$$

By the orthogonality of the set $\{\phi_0, \phi_1, \ldots, \phi_n\}$ and the assumption that

$$c_j := \frac{1}{\alpha_j}\int_a^b w(x)f(x)\phi_j(x)\ dx, \quad j = 0, 1, \ldots, n,$$

we have

$$\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx = \sum_{k=0}^{n} a_k(c_k\alpha_k) - \sum_{k=0}^{n} a_k(c_k\alpha_k) = 0.$$

Returning to (2.1.1.8) and (2.1.1.9), we have thus

$$\int_a^b w(x)(f(x) - q(x))^2\ dx \geq \int_a^b w(x)(f(x) - p(x))^2\ dx,$$

which completes the proof. □

We also have the following characterization of the least squares approximant.

Theorem 2.1.1.11. *Let $f \in C[a, b]$ and let w be a weight function on $[a, b]$. Then p is the least squares approximation to f in Π^n if and only if*

$$\int_a^b w(x)(f(x) - p(x))q(x)\ dx = 0$$

for all $q \in \Pi^n$.

Proof. First, suppose that

$$\int_a^b w(x)(f(x) - p(x))q(x)\ dx = 0$$

holds for all $q \in \Pi^n$. Then

$$\int_a^b w(x)(f(x) - q(x))^2\ dx = \int_a^b w(x)((f(x) - p(x)) + (p(x) - q(x)))^2\ dx$$
$$= \int_a^b w(x)(f(x) - p(x))^2\ dx + 2\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx +$$
$$\int_a^b w(x)(f(x) - q(x))^2\ dx$$
$$\geq \int_a^b w(x)(f(x) - p(x))^2\ dx + 2\int_a^b w(x)(f(x) - p(x))(p(x) - q(x))\ dx.$$

But since $p-q \in \Pi^n$, the assumptions imply that the second term on the RHS is zero. Thus

$$\int_a^b w(x)(f(x)-q(x))^2\ dx \geq \int_a^b w(x)(f(x)-p(x))^2\ dx,$$

which proves that p is the least squares approximation to f.

Now for the converse assume that there exists $\overline{q} \in \Pi^n$ such that

$$\int_a^b w(x)(f(x)-p(x))\overline{q}(x)\ dx = \alpha \neq 0.$$

Then clearly

$$\beta := \int_a^b w(x)(\overline{q}(x))^2\ dx > 0.$$

Put $\lambda := \frac{\alpha}{\beta} \neq 0$. Then

$$\begin{aligned}
&\int_a^b w(x)(f(x)-p(x)-\lambda\overline{q}(x))^2\ dx \\
&\qquad = \int_a^b w(x)(f(x)-p(x))^2\ dx - 2\int_a^b w(x)(f(x)-p(x))\lambda\overline{q}(x)\ dx + \\
&\qquad\quad \int_a^b w(x)(\lambda\overline{q}(x))^2\ dx \\
&\qquad = \int_a^b w(x)(f(x)-p(x))^2\ dx - 2\lambda\alpha + \lambda^2\beta \\
&\qquad = \int_a^b w(x)(f(x)-p(x))^2\ dx - \lambda^2\beta,
\end{aligned}$$

by definition of λ. But since $\lambda^2\beta > 0$, we conclude that

$$\int_a^b w(x)(f(x)-p(x)-\lambda\overline{q}(x))^2 < \int_a^b w(x)(f(x)-p(x))^2\ dx,$$

and since $p+\lambda\overline{q} \in \Pi^n$, this implies that p is not the least squares approximation to f, a contradiction.

This completes the proof. □

We next give a formula for the construction of $w-$orthogonal polynomials. It will be helpful to first give the precise definition of the inner product on $L^2[a,b]$.

Definition 2.1.1.12 (L^2- Inner Product)**.** *Let $f,g \in L^2[a,b]$ and let w be a weight function on $[a,b]$. We define the weighted* **inner product** *$\langle f,g\rangle$ of f and g by*

$$\langle f,g\rangle := \int_a^b w(x)f(x)g(x)\ dx.$$

Theorem 2.1.1.13 (Construction of $w-$Orthogonal Polynomials)**.** *There exist polynomials $\phi_n \in \Pi^n$, $n = 0, 1, \ldots,$ such that*

$$\langle \phi_k, \phi_j \rangle = \int_a^b w(x)\phi_k(x)\phi_j(x)\ dx = 0, \quad k \neq j.$$

These polynomials are uniquely defined by the recursion

$$\phi_0(x) := 1,$$

$$\phi_1(x) := x - B_1, \quad B_1 := \frac{\langle x\phi_0, \phi_0 \rangle}{\langle \phi_0, \phi_0 \rangle} = \frac{\int_a^b xw(x)\ dx}{\int_a^b w(x)\ dx},$$

and, when $k \geq 2$,

$$\phi_k(x) := (x - B_k)\phi_{k-1}(x) - C_k^2\phi_{k-2}(x),$$

where

$$B_k := \frac{\langle x\phi_{k-1}, \phi_{k-1} \rangle}{\langle \phi_{k-1}, \phi_{k-1} \rangle} = \frac{\int_a^b xw(x)[\phi_{k-1}(x)]^2\ dx}{\int_a^b w(x)[\phi_{k-1}(x)]^2\ dx},$$

$$C_k^2 := \frac{\langle \phi_{k-1}, \phi_{k-1} \rangle}{\langle \phi_{k-2}, \phi_{k-2} \rangle} = \frac{\int_a^b w(x)[\phi_{k-1}(x)]^2\ dx}{\int_a^b w(x)[\phi_{k-2}(x)]^2\ dx}.$$

Proof. The proof follows from the Gram–Schmidt orthogonalization process and induction on n. □

The last result from this section is the following useful corollary.

Corollary 2.1.1.14. *Let $\{\phi_0, \phi_1, \ldots, \phi_n\}$ be the $w-$orthogonal set of functions given in* (2.1.1.13)*. Then $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is linearly independent on $[a, b]$ and*

$$\int_a^b w(x)\phi_n(x)Q_k(x)\ dx = 0$$

for any polynomial $Q_k(x)$ of degree $k < n$.

Proof. First, note by the recursion (2.1.1.13) that each ϕ_k, $k = 0, 1, \ldots, n$, is a polynomial of degree precisely k. Since a set of polynomials with such a property is linearly independent (2.1.1.5), it follows immediately that $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is a linearly independent set.

Now let $Q_k \in \Pi^n$ be a polynomial of degree $k < n$. Since $\{\phi_0, \phi_1, \ldots, \phi_n\}$ forms a basis for Π^n, there exist real coefficients a_j, $j = 0, 1, \ldots, k$ such that

$$Q_k(x) = \sum_{j=0}^{k} a_k\phi_k(x).$$

Since $\langle \phi_n, \phi_j \rangle = 0$ for each $j = 0, 1, \ldots, k$, we have

$$\begin{aligned}\int_a^b w(x)\phi_n(x)Q_k(x)\ dx &= \int_a^b w(x)\phi_n(x)\left(\sum_{j=0}^{k} a_j\phi_j(x)\right)\ dx\\ &= \sum_{j=0}^{k} a_j \int_a^b w(x)\phi_n(x)\phi_j(x)\ dx\\ &= 0.\end{aligned}$$

This completes the proof. □

Example 2.1.1.15. *The* **Legendre polynomials** *are orthogonal on* $[-1,1]$ *with respect to the weight function* $w(x)\equiv 1$. *Using the formulas given in* (2.1.1.13), *the first three Legendre polynomials are*

$$\begin{gathered}\phi_0(x) := 1,\\ B_1 := \frac{\int_{-1}^{1} x\ dx}{\int_{-1}^{1}\ dx},\\ \phi_1(x) := x - B_1 = x - 0 = x,\\ B_2 := \frac{\int_{-1}^{1} x^3\ dx}{\int_a^b x^2\ dx} = 0,\\ C_2^2 := \frac{\int_{-1}^{1} x^2\ dx}{\int_{-1}^{1}\ dx} = \frac{\frac{1}{3}x|_{-1}^{1}}{x|_{-1}^{1}} = \frac{2/3}{2} = \frac{1}{3},\\ \phi_2(x) := (x-B_2)\phi_1(x) - C_2^2\phi_0(x) = x\phi_1(x) - \frac{1}{3} = x^2 - \frac{1}{3}.\end{gathered}$$

Example 2.1.1.16. *Recalling the first example from this section, given* $f = x^2$, *we use an orthogonal basis given by* (2.1.1.13) *to calculate the polynomial* $p\in\Pi^1$ *that satisfies*

$$\int_0^2 (f(x)-p(x))^2\ dx \le \int_0^2 (f(x)-q(x))^2\ dx$$

for all $q\in\Pi^1$. *The orthogonal polynomials are*

$$\begin{gathered}\phi_0(x) = 1,\\ \phi_1(x) = x - \frac{\int_0^2 x\ dx}{\int_0^2\ dx} = x - \frac{\frac{1}{2}x^2|_0^2}{x|_0^2} = x - \frac{2}{2} = x - 1.\end{gathered}$$

Now by (2.1.1.10) *we find*

$$c_0 = \frac{\int_0^2 x^2\ dx}{\int_0^2\ dx} = \frac{\frac{1}{3}x^3|_0^2}{x|_0^2} = \frac{8/3}{2} = \frac{4}{3},$$

and

$$c_1 = \frac{\int_0^2 x^2(x-1)\ dx}{\int_0^2 (x-1)^2\ dx} = \frac{\int_0^2 x^3 - x^2\ dx}{\frac{1}{3}(x-1)^3|_0^2} = \frac{\frac{1}{4}x^4 - \frac{1}{3}x^3|_0^2}{2/3} = \frac{4-\frac{8}{3}}{2/3} = 2.$$

Hence, the least squares approximation p *to* f *is given by*

$$p(x) := c_0\phi_0(x) + c_1\phi_1(x) = \frac{4}{3} + 2(x-1) = \frac{4}{3} + 2x - 2 = -\frac{2}{3} + 2x.$$

Note that this least squares approximant p is precisely the approximant p found in the first example.

2.1.2. *Chebyshev Polynomials and Economization of Power Series.* The Chebyshev polynomials $\{T_j(x)\}_{j=0}^n$ form a $w-$orthogonal basis for Π^n on the interval $[-1,1]$ with respect to the weight function $w(x) := \frac{1}{\sqrt{1-x^2}}$. They may be constructed using (2.1.1.13); we construct them using an alternative method here.

For $x \in [-1,1]$, we define the Chebyshev polynomials by

$$T_n(x) := \cos(n \arccos(x)), \quad n \geq 0. \tag{2.1.2.1}$$

First note that

$$T_0(x) = \cos(0) = 1, \quad T_1(x) = \cos(\arccos(x)) = x.$$

We introduce the notation $\theta := \arccos(x)$. Then we have

$$T_n(\theta(x)) = T_n(\theta) = \cos(n\theta), \quad \theta \in [0,\pi].$$

We obtain a recurrence relation by noting that

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

Adding gives

$$T_{n+1}(\theta) = 2\cos(n\theta)\cos(\theta) - T_{n-1}(\theta).$$

Since $\theta = \arccos(x)$, evidently $x = \cos(\theta)$, so that

$$T_{n+1}(x) = 2x\cos(n\arccos(x)) - T_{n-1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Hence the recurrence relation for the construction of $\{T_n(x)\}$ is as follows:

$$\begin{aligned} T_0(x) &:= 1, \\ T_1(x) &:= x, \\ T_n(x) &:= 2xT_{n-1}(x) - T_{n-2}(x), \quad n \geq 2. \end{aligned} \tag{2.1.2.2}$$

Note from (2.1.2) that for all $n \geq 0$, $T_n(x)$ is a polynomial of degree n, and for $n \geq 1$, $T_n(x)$ has leading coefficient 2^{n-1}.

Example 2.1.2.1. *The first five Chebyshev polynomials are*

$$T_0(x) = 1,$$
$$T_1(x) = x,$$
$$T_2(x) = 2xT_1(x) - T_0(x) = 2x(x) - 1 = 2x^2 - 1,$$
$$T_3(x) = 2xT_2(x) - T_1(x) = 2x(2x^2-1) - x = 4x^3 - 3x,$$
$$T_4(x) = 2xT_3(x) - T_2(x) = 2x(4x^3-3x) - (2x^2-1) = 8x^4 - 8x^2 + 1.$$

We now show that $\{T_n(x)\}$ is $w-$orthogonal with respect to the weight function

$$w(x) := \frac{1}{\sqrt{1-x^2}}$$

on $[-1,1]$. Consider

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \int_{-1}^{1} \frac{\cos(n\arccos(x))\cos(m\arccos(x))}{\sqrt{1-x^2}}\, dx.$$

Reintroducing $\theta := \arccos(x)$, we have $d\theta = -\frac{1}{\sqrt{1-x^2}}dx$, and

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = -\int_{\pi}^{0} \cos(n\theta)\cos(m\theta)\, d\theta = \int_{0}^{\pi} \cos(n\theta)\cos(m\theta)\, d\theta.$$

Since

$$\cos(n\theta)\cos(m\theta) = \frac{1}{2}[\cos((n+m)\theta) + \cos((n-m)\theta)],$$

we have

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \frac{1}{2}\int_{0}^{\pi} \cos((n+m)\theta)\, d\theta + \frac{1}{2}\int_{0}^{\pi} \cos((n-m)\theta)\, d\theta.$$

If $n \neq m$,

$$\begin{aligned}\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx &= \frac{1}{2}\left[\frac{1}{n+m}\sin((n+m)\theta)\right]_0^{\pi} + \frac{1}{2}\left[\frac{1}{n-m}\sin((n-m)\theta)\right]_0^{\pi} \\ &= \left[\frac{1}{2(n+m)}\sin((n+m)\theta)\right]_0^{\pi} + \left[\frac{1}{2(n-m)}\sin((n-m)\theta)\right]_0^{\pi} \\ &= 0,\end{aligned}$$

since n and m are integers.

If $n = m$,

$$\begin{aligned}\int_{-1}^{1} \frac{[T_n(x)]^2}{\sqrt{1-x^2}}\, dx &= \frac{1}{2}\int_{0}^{\pi} \cos(2n\theta)\, d\theta + \frac{1}{2}\int_{0}^{\pi} d\theta \\ &= \frac{1}{2}\left[\frac{1}{2n}\sin(2n\theta)\right]_0^{\pi} + \frac{1}{2}\theta\Big|_0^{\pi} \\ &= \left[\frac{1}{4n}\sin(2n\theta)\right]_0^{\pi} + \frac{\pi}{2} \\ &= \frac{\pi}{2},\end{aligned}$$

for $n \geq 1$. Note if $n = 0$, then

$$\int_{-1}^{1} \frac{[T_0(x)]^2}{\sqrt{1-x^2}}\, dx = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} = \pi.$$

The Chebyshev polynomials are frequently used to minimize approximation error. We first give an important result regarding the zeros of the Chebyshev polynomials and their first derivatives.

Theorem 2.1.2.2 (Zeros and Extreme Values of the Chebyshev Polynomials)**.** *The Chebyshev polynomial $T_n(x)$ of degree $n \geq 1$ has n simple zeros in $[-1,1]$ at*

$$x_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, 2, \ldots, n.$$

Moreover, $T_n(x)$ assumes its absolute extreme values in $[-1,1]$ at

$$z_k = \cos\left(\frac{k\pi}{n}\right), \quad k = 0, 1, \ldots, n,$$

with

$$T_n(z_k) = (-1)^k, \quad k = 0, 1, \ldots, n.$$

Proof. By definition of T_n and x_k, $k = 1, 2, \ldots, n$,

$$\begin{aligned} T_n(x_k) &= \cos(n \arccos(x_k)) \\ &= \cos\left(n \arccos\left(\cos\left(\frac{2k-1}{2n}\pi\right)\right)\right) \\ &= \cos\left(n\left(\frac{2k-1}{2n}\pi\right)\right) \\ &= \cos\left(\frac{2k-1}{2}\pi\right) \\ &= \cos\left(k\pi - \frac{\pi}{2}\right) \\ &= 0, \end{aligned}$$

since k is an integer. Since the x_k are distinct, note that these are distinct zeros. Furthermore

$$\begin{aligned} T_n'(x_k) &= \frac{n \sin\left(n \arccos\left(\cos\left(\frac{2k-1}{2n}\pi\right)\right)\right)}{\sqrt{1 - \left[\cos\left(\frac{2k-1}{2n}\pi\right)\right]^2}} \\ &= \frac{n \sin\left(n\left(\frac{2k-1}{2n}\right)\right)}{\sqrt{1 - \left[\cos\left(\frac{k}{n}\pi - \frac{\pi}{2n}\right)\right]^2}} \\ &= \frac{n \sin\left(k\pi - \frac{\pi}{2}\right)}{\sqrt{1 - \left[\cos\left(\frac{k}{n}\pi - \frac{\pi}{2n}\right)\right]^2}} \neq 0, \end{aligned}$$

so that the x_k are simple zeros, $k = 1, 2, \ldots, n$.

Also observe that

$$\begin{aligned} T_n'(z_k) &= \frac{n \sin\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right)}{\sqrt{1 - \left[\cos\left(\frac{k\pi}{n}\right)\right]^2}} = \frac{n \sin\left(n\left(\frac{k\pi}{n}\right)\right)}{\sqrt{\sin^2\left(\frac{k\pi}{n}\right)}} \\ &= \frac{n \sin(k\pi)}{\sin\left(\frac{k\pi}{n}\right)} = 0, \end{aligned}$$

for $k = 1, 2, \ldots, n-1$. Since $T_n' \in \Pi^{n-1}$, all of the $n-1$ zeros of T_n' occur at these points z_k. Including the endpoints $z_0 := -1$ and $z_n := 1$,

$$\begin{aligned} T_n(z_k) &= \cos\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right) \\ &= \cos\left(n\left(\frac{k\pi}{n}\right)\right) \\ &= \cos(k\pi) \\ &= (-1)^k, \quad k = 0, 1, \ldots, n. \end{aligned}$$

Hence, $|T_n| \leq 1$ on $[-1, 1]$ and has $n+1$ extreme values on $[-1, 1]$. This completes the proof. □

We introduce the notion of a monic polynomial.

Definition 2.1.2.3 (Monic Polynomial)**.** *A* **monic polynomial** *is a polynomial in which the leading coefficient is equal to* 1*, that is,*

$$p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0,$$

for some coefficients $c_0, c_1, \ldots, c_{n-1} \in \mathbb{R}$.

We will denote by $\widetilde{\Pi}^n$ the set of all monic polynomials of degree exactly n. More precisely,

$$\widetilde{\Pi}^n := \left\{ p \in \Pi^n : p(x) := x^n + \sum_{k=0}^{n-1} c_k x^k, c_j \in \mathbb{R}, j = 0, 1, \ldots, n-1 \right\}.$$

The monic Chebyshev polynomials $\widetilde{T}_n(x)$ are derived from the Chebyshev polynomials $T_n(x)$ by dividing by the leading coefficient 2^{n-1}. We obtain

$$\widetilde{T}_0(x), \quad \text{and} \quad \widetilde{T}_n(x) = \frac{1}{2^{n-1}} T_n(x), \quad n = 1, 2, \ldots. \tag{2.1.2.3}$$

We also get the recurrence

$$\begin{aligned} \widetilde{T}_0(x) &= 1, \\ \widetilde{T}_1(x) &= x \\ \widetilde{T}_2(x) &= x\widetilde{T}_1(x) - \frac{1}{2}\widetilde{T}_0(x) = x^2 - \frac{1}{2}, \\ &\vdots \\ \widetilde{T}_n(x) &= x\widetilde{T}_{n-1}(x) - \frac{1}{4}\widetilde{T}_{n-2}(x), \quad n \geq 3, . \end{aligned} \tag{2.1.2.4}$$

Noting that $\widetilde{T}_n(x)$ is just a multiple of $T_n(x)$, (2.1.2.2) implies that the zeros of $\widetilde{T}_n(x)$, $n \geq 1$, also occur at

$$x_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 0, 1, \ldots, n,$$

and the extreme values of $\widetilde{T}_n(x)$, $n \geq 1$, occur at

$$z_k = \cos\left(\frac{k\pi}{n}\right), \quad k = 0, 1, \ldots, n,$$

where

$$\widetilde{T}_n(z_k) = \frac{(-1)^k}{2^{n-1}}, \quad k = 0, 1, \ldots, n,$$

by the construction.

From this construction we get an important minimization property of the monic Chebyshev polynomials $\widetilde{T}_n(x)$ that distinguishes $\{\widetilde{T}_j(x)\}_{j=0}^n$ from other sets of polynomials in $\widetilde{\Pi}^n$.

Theorem 2.1.2.4 (Minimization Property of Monic Chebyshev Polynomials)**.** *The monic Chebyshev polynomials $\widetilde{T}_n(x)$, $n \geq 1$, have the property*

$$\frac{1}{2^{n-1}} = \max_{x\in[-1,1]} \left|\widetilde{T}_n(x)\right| \leq \max_{x\in[-1,1]} |p_n(x)|$$

for all $p_n \in \widetilde{\Pi}^n$. Moreover, equality occurs only if $p_n = \widetilde{T}_n$. In the case $n = 0$, we have

$$1 = \max_{x\in[-1,1]} \left|\widetilde{T}_0(x)\right| = \widetilde{T}_0(x).$$

Proof. If $n = 0$, then $\widetilde{\Pi}^0 = \{p(x) \equiv 1\}$ and $\widetilde{T}_0(x) \equiv 1$, which establishes the result for this case.

Now let $n \geq 1$. Suppose that $p_n \in \widetilde{\Pi}^n$ and

$$\max_{x\in[-1,1]} |p_n(x)| \leq \frac{1}{2^{n-1}} = \max_{x\in[-1,1]} \left|\widetilde{T}_n(x)\right|.$$

Define the difference polynomial $Q := \widetilde{T}_n - p_n$. Since both $\widetilde{T}_n, p_n \in \widetilde{\Pi}^n$, they are both monic polynomials of degree precisely n, so that Q is a polynomial of degree at most $n-1$, $Q \in \Pi^{n-1}$. At the $n+1$ extreme points z_k, $k = 0, 1, \ldots, n$ of $\widetilde{T}_n$, we have

$$Q(z_k) = \widetilde{T}_n(z_k) - p_n(z_k) = \frac{(-1)^k}{2^{n-1}} - p_n(z_k).$$

Since $|p_n(z_k)| \leq \frac{1}{2^{n-1}}$, $k = 0, 1, \ldots, n$ by the assumption, we have for k even

$$Q(z_k) \geq 0$$

and for k odd

$$Q(z_k) \leq 0.$$

Since Q is continuous, the intermediate value theorem implies that for each $j = 0, 1, \ldots, n-1$, there exists $t \in [z_j, z_{j+1}]$ such that $Q(t) = 0$. Thus Q has n zeros in $[-1, 1]$, and since Q has degree at most $n-1$, Q must vanish identically,

$$Q(x) \equiv 0.$$

This implies

$$p_n \equiv \widetilde{T}_n,$$

which completes the proof. □

We immediately get the following corollary.

Corollary 2.1.2.5. *For any $p_n \in \widetilde{\Pi}^n$, where $n \geq 1$,*

$$\max_{x\in[-1,1]} |p_n(x)| \geq \frac{1}{2^{n-1}}.$$

In the case $n = 0$,

$$\max_{x\in[-1,1]} |p_n(x)| = p_n(x) = 1.$$

We now show how the Chebyshev polynomials can be used to minimize the error in polynomial interpolation. Recall the error formula (1.1.4.2)

$$f(\overline{x}) - p(\overline{x}) = \frac{\omega(\overline{x}) f^{n+1}(\xi)}{(n+1)!},$$

where

$$\omega(x) := \prod_{j=0}^{n}(x - x_j).$$

Since n is prescribed and there is generally no control over ξ, we choose to minimize ω. Noting that $\omega \in \widetilde{\Pi}^{n+1}$, we have just shown that the minimum infinity norm of ω on $[-1, 1]$ is obtained when $\omega \equiv \widetilde{T}_{n+1}$.

From (2.1.2.4) and the above observations comes the following important theorem regarding the minimization of the error formula for the error in polynomial interpolation (1.1.4.2).

Theorem 2.1.2.6 (Error in Polynomial Interpolation at Chebyshev Zeros)**.** *Suppose that $p \in \Pi^n$ is the unique interpolating polynomial of degree n of the function f with support abscissas at the Chebyshev zeros x_k, $k = 1, 2, \ldots, n + 1$, of $T_{n+1}(x)$. Then there exists a number $\xi \in [-1, 1]$ such that*

$$\max_{x\in[-1,1]} |f(x) - p(x)| \leq \frac{1}{2^n(n+1)!}|f^{(n+1)}(\xi)|.$$

Proof. Let $p \in \Pi^n$ satisfy the hypotheses of (2.1.2.6). Recall from the formula for error in polynomial interpolation (1.1.4.2) that there exists a number $\xi \in [-1, 1]$ with

$$|f(x) - p(x)| = \frac{\omega(x) f^{(n+1)}(\xi)}{(n+1)!},$$

where

$$\omega(x) := \prod_{k=0}^{n}(x - x_k),$$

provided that $x \in [-1, 1]$. But since each x_k is the $k-$th zero of $\widetilde{T}_{n+1}$, $k = 1, 2, \ldots, n+1$, we have that ω coincides with $\widetilde{T}_{n+1}$, $\omega \equiv \widetilde{T}_{n+1}$. Hence it follows by (2.1.2.4) that

$$\begin{aligned} |f(x) - p(x)| &= \left| \frac{\widetilde{T}_{n+1}(x) f^{(n+1)}(\xi)}{(n+1)!} \right| \leq \frac{1}{(n+1)} |f^{(n+1)}(\xi)| \max_{x\in[-1,1]} |\widetilde{T}_{n+1}(x)| \\ &= \frac{1}{2^n(n+1)!}|f^{(n+1)}(\xi)|, \end{aligned}$$

which completes the proof. □

Example 2.1.2.7. *Let $T_2(x)$ be the standard Chebyshev polynomial on the domain $[-1,1]$. Given $f(x) = (1+x\sqrt{2})^2$, define $p(x)$ to be the first-order polynomial that interpolates f at the roots of T_2. We calculate p.*

Recall

$$T_0(x) := 1,$$
$$T_1(x) := x,$$
$$T_2(x) := 2xT_1(x) - T_0(x) = 2x^2 - 1.$$

Thus the roots of T_2 are $x = \pm\frac{1}{2}$.

We have

	$k=0$	$k=1$
$x_0 := -\frac{1}{2}$	$f[x_0] = 3 - 2\sqrt{2}$	
		$f[x_0, x_1] = 4\sqrt{2}$
$x_1 := \frac{1}{2}$	$f[x_1] = 3 + 2\sqrt{2}$	

Hence

$$p(x) = 3 - 2\sqrt{2} + 4\sqrt{2}\left(x + \frac{1}{2}\right).$$

Theorem (2.1.2.6) gives us about the tightest upper bound on the error we can achieve without further knowledge of the function f. Note that we cannot always choose the nodes in this fashion, however, and that choosing support abscissas at the Chebyshev zeros does not guarantee that the interpolating polynomial p is the *best* approximation of f.

We now discuss methods to generalize the Chebyshev polynomials $\{T_n(x)\}$ to an arbitrary interval $[a,b]$. We can generalize the Chebyshev polynomials to the interval $[a,b]$ by applying an affine mapping

$$\widetilde{x} = \frac{1}{2}[(a+b) + (b-a)x]$$

for numbers $x \in [-1,1]$. That is, the numbers $x \in [-1,1]$ map to the numbers $\widetilde{x} \in [a,b]$. We get the following theorem.

Theorem 2.1.2.8. *The Chebyshev zeros can be generalized from the interval $[-1,1]$ to the interval $[a,b]$ by applying an affine mapping. In general, $T_n(x)$ has the following n zeros on the closed interval $[a,b]$:*

$$\widetilde{x}_k = \frac{1}{2}\left[(a+b) + (b-a)\cos\left(\frac{2k-1}{2n}\pi\right)\right], \quad k = 1, 2, \ldots, n.$$

Proof. Denote the zeros of $T_n(x)$ on the interval $[-1,1]$ by

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

Recall that an affine mapping $\widetilde{x} : [-1,1] \to [a,b]$ has the form

$$\widetilde{x}(x) = \lambda x + \beta$$

for all $x \in [-1,1]$. To map the endpoints of $[-1,1]$ to the endpoints of $[a,b]$, we define this affine mapping $\widetilde{x}$ by

$$\widetilde{x}(-1) := a = -\lambda + \beta,$$
$$\widetilde{x}(1) := b = \lambda + \beta.$$

Solving for λ and β gives

$$\lambda = \frac{b-a}{2}, \quad \beta = \frac{a+b}{2},$$

so that $\widetilde{x}$ is given by

$$\widetilde{x}(x) = \frac{1}{2}\left[(a+b) + (b-a)x\right].$$

Furthermore, we have

$$\widetilde{x}_k = \widetilde{x}(x_k) = \frac{1}{2}\left[(a+b) + (b-a)\cos\left(\frac{2k-1}{2n}\pi\right)\right]$$

for each $k = 1, 2, \ldots, n$, which completes the proof. □

We get the following analog to (2.1.2.4).

Theorem 2.1.2.9 (Minimization on Arbitrary Interval)**.** *Let $p \in \Pi^n$ on the interval $[a, b]$ be such that*

$$p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

for real coefficients c_j, $j = 0, 1, \ldots, n$. Then

$$\max_{x\in[a,b]} |p(x)| = \max_{x\in[a,b]} |c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n| \geq |c_n| \frac{(b-a)^n}{2^{2n-1}}.$$

Proof. Let $p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$. Set

$$q(x) := p\left(a + \frac{b-a}{2}(x+1)\right).$$

Noting that q is simply p on the interval $[-1, 1]$, we have that

$$\max_{x\in[a,b]} |p(x)| = \max_{x\in[-1,1]} |q(x)|.$$

The leading coefficient on q is

$$c_n \frac{(b-a)^n}{2^n}.$$

Without loss of generality, we assume that p is of degree n, so that evidently $c_n \neq 0$. It follows that

$$\begin{aligned}
\max_{x\in[a,b]} |p(x)| &= \max_{x\in[-1,1]} |q(x)| \\
&= \max_{x\in[-1,1]} \left| c_n \frac{(b-a)^n}{2^n} x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0 \right| \\
&= \left(|c_n| \frac{(b-a)^n}{2^n} \right) \max_{x\in[-1,1]} \left| x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \right| \\
&\geq |c_n| \frac{(b-a)^n}{2^n} \left(\frac{1}{2^{n-1}} \right) \\
&= |c_n| \frac{(b-a)^n}{2^{2n-1}}.
\end{aligned}$$

Moreover, equality occurs only if

$$\frac{2^n}{c_n (b-a)^n} p(x) = T_n(x).$$

This completes the proof. □

We arrive at the following interpolation minimization property for the interval $[a, b]$.

Theorem 2.1.2.10. *Suppose that $p \in \Pi^n$ is the unique interpolating polynomial of the function f with support abscissas at the Chebyshev zeros x_k, $k = 1, 2, \ldots, n+1$ of $T_{n+1}(x)$ generalized to the interval $[a, b]$. Then there exists a number $\xi \in [a, b]$ such that*

$$\max_{x\in[a,b]} |f(x) - p(x)| \leq \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!} |f^{(n+1)}(\xi)|.$$

Proof. The proof follows immediately from (2.1.2.6) and (2.1.2.9). □

The last application of the Chebyshev polynomials is in reducing the degree of an approximating polynomial with minimal increase in error.

Consider approximating an arbitrary polynomial of degree n

$$p_n(x) := c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

on the interval $[-1, 1]$ with a polynomial of degree at most $n-1$. We want to find a polynomial $p_{n-1} \in \Pi^{n-1}$ so that the quantity

$$\max_{x\in[-1,1]} |p_n(x) - p_{n-1}(x)|$$

is minimized.

Note that

$$\frac{1}{c_n}(p_n(x) - p_{n-1})$$

is a monic polynomial of degree n. By the minimization property of the monic Chebyshev polynomials (2.1.2.4), we have

$$\max_{x\in[-1,1]} \left| \frac{1}{c_n}(p_n(x) - p_{n-1}(x)) \right| \geq \frac{1}{2^{n-1}}.$$

Equality occurs precisely when

$$\frac{1}{c_n}(p_n(x) - p_{n-1}(x)) = \widetilde{T}_n(x).$$

Rearranging, we see that we should choose

$$p_{n-1}(x) = p_n(x) - c_n \widetilde{T}_n(x).$$

Then with this choice of p_{n-1} we have

$$\max_{x\in[-1,1]} |p_n(x) - p_{n-1}(x)| = |c_n| \max_{x\in[-1,1]} |\widetilde{T}_n(x)| = \frac{|c_n|}{2^{n-1}}.$$

2.2. **Uniform Approximation.**

2.2.1. *Best Uniform Approximation.* Let $(V, \|\cdot\|)$ be a normed linear space and let W be a subspace of V. The essence of the approximation problem is as follows: given a vector $v \in V$, find a vector $w \in W$ such that the distance from w to v is minimized, that is, find $w^* \in W$ such that

$$\|v - w^*\| \leq \|v - w\| \quad \text{for all } w \in W.$$

We call such a w^* the best approximation to v out of W under $\|\cdot\|$.

We get the following theorem.

Theorem 2.2.1.1 (Existence and Uniqueness of Best Approximation)**.** *Let $(V, \|\cdot\|)$ be a normed linear space and W a finite-dimensional subspace of V. Then, for all $v \in V$, there exists a unique $w^* \in W$ such that*

$$\|v - w^*\| \leq \|v - w\|$$

for all $w \in W$.

Proof. The proof is attributed to Tonelli. □

The space $(V, \|\cdot\|)$ considered in this section is the space $(C[a,b], \|\cdot\|_\infty)$ of continuous functions on $[a,b]$ under the infinity (uniform, supremum) norm

$$\|f\|_\infty := \sup_{x\in[a,b]} |f(x)|, \quad \text{for } f \in C[a,b].$$

Note that, letting W be the space Π^n, (2.2.1.1) immediately gives the following result.

Theorem 2.2.1.2. *Let $f \in C[a,b]$. Then there exists a unique p^* in Π^n such that*

$$\max_{x\in[a,b]} |f(x) - p^*(x)| \leq \min_{p\in\Pi^n} \max_{x\in[a,b]} |f(x) - p(x)|.$$

Also note (2.2.1.2) is equivalent to the Stone–Weierstrass approximation theorem. That is, given any $f \in C[a,b]$, there exists a sequence $\{p_n\}_{n=1}^\infty$ of polynomials of degree $n = 1, 2, \ldots$, converging uniformly to f on $[a,b]$. The remainder of this section characterizes the best uniform approximation to such an f out of Π^n.

Definition 2.2.1.3 (Error Function)**.** *Let $f \in C[a,b]$ and let $p^* \in \Pi^n$ be the best uniform approximation to f out of Π^n. We define the* **error function** *by*

$$E_n(f; [a,b]) := E_n(f) := \|f - p^*\|_\infty = \max_{x\in[a,b]} |f(x) - p^*(x)|.$$

Lemma 2.2.1.4. *Let $f \in C[a,b]$. Then*

$$E_0(f) \geq E_1(f) \geq E_2(f) \geq \ldots$$

and, moreover,

$$\lim_{n\to+\infty} E_n(f) = 0.$$

Proof. Note first that the inequalities

$$E_0(f) \geq E_1(f) \geq E_2(f) \geq \ldots$$

follow immediately from the nesting of the polynomial spaces

$$\Pi^0 \subset \Pi^1 \subset \Pi^2 \subset \ldots.$$

The fact that

$$\lim_{n\to+\infty} E_n(f) = 0$$

follows from the Stone–Weierstrass approximation theorem, which states that we may approximate f uniformly by polynomials to any desired tolerance. □

Let $p^* \in \Pi^n$ be the best uniform approximation of $f \in C[a,b]$. Define the signed error

$$e(x) := f(x) - p^*(x)$$

and note $\|e(x)\|_\infty = E_n(f)$. We get the following preliminary result.

Lemma 2.2.1.5. *Let $f \in C[a,b]$ and let $p^* \in \Pi^n$ be the best uniform approximation of f. Then there exist at least two distinct points $x_1, x_2 \in [a,b]$ such that*

$$|e(x_1)| = E_n(f) = |e(x_2)|$$

and

$$e(x_1) = -e(x_2).$$

Proof. The signed error function $e(x)$ is continuous and bounded by its extreme values at $y = \pm E_n(f)$ by definition. Moreover, by definition of $E_n(f)$, $e(x)$ has at least one extreme value at $\pm E_n(f)$, say, without loss of generality, that there is $x_1 \in [a,b]$ such that $e(x_1) = E_n(f)$.

By contradiction, suppose that $e(x) > -E_n(f)$ throughout $[a,b]$. Define

$$\min_{x\in[a,b]} e(x) := m > -E_n(f)$$

and

$$c := \frac{E_n(f) + m}{2} > 0.$$

Since c is a constant, note $p := p^* + c \in \Pi^n$. Then $f(x) - p(x) = f(x) - (p^*(x) + c) = f(x) - p^*(x) - c = e(x) - c$, and

$$-(E_n(f) - c) = -((2c - m) - c) = m - c \le e(x) - c \le E_n(f) - c.$$

But since $e(x) - c = f(x) - p(x)$, we have evidently

$$\|f - p\|_\infty = E_n(f) - c,$$

a contradiction to the assumption that p^* is the best uniform approximation of f. Thus there must exist a point $x_2 \in [a,b]$ such that $e(x_2) = -E_n(f)$. This proves the result. □

As it turns out, for a best approximation $p \in \Pi^n$, the signed error $e(x)$ oscillates and must touch the lines $y = \pm E_n(f)$ alternately $n+2$ times. This in fact characterizes the best uniform approximation and gives the following important theorem.

Theorem 2.2.1.6 (Chebyshev Equioscillation Theorem)**.** *Let $f \in C[a,b]$. Then a polynomial $p^* \in \Pi^n$ is the best uniform approximation to f out of Π^n on $[a,b]$ if and only if there exists an alternating set of points $x_j, j = 1, 2, \ldots, n+2$,*

$$a \leq x_1 < x_2 < \cdots < x_{n+2} \leq b$$

in $[a,b]$ such that $e(x)$ assumes its extreme values with alternating signs

$$e(x_j) = \pm E_n(f), \quad j = 1, 2, \ldots, n+2,$$

and

$$e(x_j) = -e(x_{j+1}), \quad j = 1, 2, \ldots, n+1.$$

Before proving the theorem, note that it is biconditional. That is, only one polynomial $p \in \Pi^n$ may have the equioscillation property as described in the statement of the theorem. This means that a polynomial $p \in \Pi^n$ that has such an equioscillation property is a sufficient condition to conclude that p the best uniform approximation to f out of Π^n.

Proof. ($\Longleftarrow$) Suppose that $\{x_j\}_{j=1}^{n+2}$ forms an alternating set for the signed error $e(x) = f(x) - p^*(x)$. We show that p^* is the best uniform approximation to f out of Π^n on $[a,b]$.

By contradiction, suppose not. Then there exists $p \in \Pi^n$ such that

$$\|f - p\|_\infty < \|f - p^*\|_\infty.$$

In particular, since $\{x_j\}_{j=1}^{n+2}$ forms an alternating set,

$$|f(x_j) - p(x_j)| < \|f - p^*\|_\infty = |f(x_j) - p^*(x_j)|, \quad j = 1, 2, \ldots, n+2.$$

Then the difference

$$[f(x_j) - p^*(x_j)] - [f(x_j) - p(x_j)] = p(x_j) - p^*(x_j)$$

changes signs at x_j for each $j = 1, 2, \ldots, n+2$. Since $p - p* \in \Pi^n$, clearly the polynomial $p - p^*$ is continuous, so that the intermediate value theorem implies that $p - p^*$ has a zero in each subinterval $[x_j, x_{j+1}]$, $j = 1, 2, \ldots, n+1$. Thus $p - p^*$ must vanish identically, so that

$$p - p^* \equiv 0,$$

and, moreover,

$$p \equiv p^*,$$

a contradiction to the hypotheses.

This proves the converse.

($\Longrightarrow$) Note that, since $e(x)$ is continuous on the closed interval $[a,b]$, $e(x)$ is uniformly continuous on $[a,b]$. Put $\epsilon := \frac{E_n(f)}{2}$ and select $\delta >$ such that

$$|e(x_1) - e(x_2)| < \epsilon$$

for any $x_1, x_2 \in [a,b]$ such that $|x_1 - x_2| < \delta$. Let a partition

$$\Delta := \{a = z_0 < z_1 < \cdots < z_N = b\}$$

be such that $\max_{j=0,1,\ldots,N-1} |z_{j+1} - z_j| < \delta$.

Note by (2.2.1.5) that there exists at least one subinterval $[z_j, z_{j+1}]$ such that $e(x) = 2\epsilon$ and at least one subinterval $[z_j, z_{j+1}]$ such that $e(x) = -2\epsilon$. Denote by I_j, $j = 1, 2, \ldots, m$ the subintervals $[z_j, z_{j+1}]$ such that $e(x)$ achieves its extreme values $\pm 2\epsilon$. Also note that either $e(x) > \epsilon$ or $e(x) < \epsilon$ throughout each I_j, $j = 1, 2, \ldots, m$.

Define $\sigma^j(e) := \operatorname{sgn}(e(x))$ for each $x \in I_j$, $j = 1, 2, \ldots, m$. We wish to show that there are at least $n+1$ sign changes in the sequence $\sigma^1(e), \sigma^2(e), \ldots, \sigma^m(e)$. By contradiction, suppose that there are less than $n+1$ sign changes. We show that there is $p \in \Pi^n$ with $p \neq p^*$ such that

$$\|f - p\|_\infty = \max_{x\in[a,b]} |f(x) - p(x)| < E_n(f).$$

Appealing again to (2.2.1.5), there is at least one sign change in $\sigma^1(e), \sigma^2(e), \ldots, \sigma^m(e)$. Thus we may group the subintervals I_j, $j = 1, 2, \ldots, m$ by common sign. Put

$$\begin{aligned}
G_1 &:= \sigma^1(e) = \sigma^2(e) = \cdots = \sigma^{j_1}(e) \implies \{I_1, I_2, \ldots, I_{j_1}\},\\
G_2 &:= \sigma^{j_1+1}(e) = \sigma^{j_1+2}(e) = \cdots = \sigma^{j_2}(e) \implies \{I_{j_1+1}, I_{j_1+2}, \ldots, I_{j_2}\},\\
&\vdots\\
G_k &:= \sigma^{j_{k-1}+1} = \sigma^{j_{k-1}+2} = \cdots = \sigma^{j_k} \implies \{I_{j_{k-1}+1}, I_{j_{k-1}+2}, \ldots, I_{j_k}\}
\end{aligned}$$

with $j_k = m$. Each subset G_j, $j = 1, 2, \ldots, k$ contains at least one element, and we have $k-1$ sign changes. For a contradiction, assume that $k < n+2$, so that there are $k-1 < n+1$ changes of sign. Since $\sigma^{j_i} \neq \sigma^{j_i+1}$ for $i = 1, 2, \ldots, k$, it is clear that the closed subintervals $I_{j_i} \neq I_{j_{i+1}}$ are disjoint. We can therefore choose points $t_1, t_2, \ldots, t_{k-1}$ with the property that $t_i > x$ for all $x \in I_{j_i}$ and $t_i < x$ for all $x \in I_{j_{i+1}}$, for $i = 1, 2, \ldots, k-1$. Form the polynomial

$$q(x) := \prod_{i=1}^{k-1}(t_i - x) = (t_1 - x)(t_2 - x)\ldots(t_{k-1} - x).$$

Since $k - 1 \leq n$, we have evidently that $q \in \Pi^{k-1} \subseteq \Pi^n$. Also note that q vanishes only at x_i, $i = 1, 2, \ldots, k-1$, and is nonzero elsewhere, so that q has constant sign on each I_j, $j = 1, 2, \ldots, m$, and thus each group G_i, $i = 1, 2, \ldots, k$. Moreover, q has the property that

$\operatorname{sgn}(q)$	+	−	+	...
Group	1	2	3	...

Therefore either $\operatorname{sgn}(q) = \operatorname{sgn}(e)$ or $\operatorname{sgn}(q) = -\operatorname{sgn}(e)$ for all I_j simultaneously. Define

$$\ell(x) := \begin{cases} q(x), & \operatorname{sgn}(q) = \operatorname{sgn}(e) \text{ throughout } I_1,\\ -q(x), & \text{otherwise.}\end{cases}$$

Then $\operatorname{sgn}(\ell) = \operatorname{sgn}(e)$ on each I_j, $j = 1, 2, \ldots, m$.

Next, put

$$S := \overline{[a, b]] \setminus \left(\bigcup_{j=1}^{m} I_j\right)}$$

and define

$$E_n' := \max_{x\in S} |e(x)|.$$

Then $E_n' < E_n$. Construct

$$p(x) := p^*(x) + \lambda\ell(x),$$

where λ is such that

$$0 < \lambda < \frac{1}{2\max_{x\in[a,b]} |\ell(x)|}(E_n - E_n').$$

We now show that $\max_{x\in[a,b]} |f-p| < 2\epsilon$ for a contradiction. On any interval I_j with $e(x) > 0$, then $\ell(x) > 0$ by the construction, and we have

$$\begin{aligned} 0 < \lambda\ell(x) &< \frac{\ell(x)}{2\max_{x\in[a,b]}|\ell(x)|}(E_n - E_n') \\ &\leq \frac{E_n - E_n'}{2} \leq \frac{E_n}{2} < e(x) \end{aligned}$$

on I_j, $j = 1, 2, \ldots, m$. That is, $e(x) - \lambda\ell(x) > 0$ throughout I_j. Thus

$$\begin{aligned} \|f-p\|_\infty &= \|f - (p^* + \lambda\ell)\|_\infty \\ &= \|e - \lambda\ell\|_\infty \\ &= e - \lambda\ell \\ &\leq E_n(f) - \lambda \min_{x\in I_j} \ell(x) \\ &< E_n(f), \end{aligned}$$

since $\lambda > 0$ and ℓ does not vanish on I_j, $j = 1, 2, \ldots, m$. A similar argument applies in the case that $e(x) < 0$ on I_j.

It only remains to show that $\|f-p\|_\infty < \|f-p^*\|_\infty$ on

$$S = \overline{[a,b] \setminus \left(\bigcup_{j=1}^{m} I_j\right)}.$$

Throughout S, we have

$$\begin{aligned} \|f-p\|_\infty &= \max_{x\in S} |e - \lambda\ell| \\ &\leq \max_{x\in S} |e| + \ell \max_{x\in S} |\ell(x)| \\ &< E_n' + \frac{\max_{x\in S}|\ell(x)|}{2\max_{x\in[a,b]}|\ell(x)|}(E_n - E_n') \\ &\leq E_n' + \frac{1}{2}(E_n - E_n') \\ &< E_n, \end{aligned}$$

since $E_n' < E_n$. Hence, we have shown that $\|f-p\|_\infty < \|f-p^*\|_\infty$, a contradiction to the assumption that p^* is the best uniform approximation to f out of Π^n on $[a,b]$.

This completes the proof of the theorem. □

The next theorem establishes uniqueness of the best uniform approximation.

Theorem 2.2.1.7 (Uniqueness of Best Uniform Approximation)**.** *If $p^* \in \Pi^n$ is a best uniform approximation to $f \in C[a,b]$ out of Π^n on $[a,b]$, then p^* is unique. More precisely, if $p \in \Pi^n$ and $p \neq p^*$, then*

$$\|f-p\|_\infty > \|f-p^*\|_\infty.$$

Proof. Suppose that p and p^* are both best uniform approximations to $f \in C[a,b]$ out of Π^n, so that

$$\|f-p\|_\infty = \|f-p^*\|_\infty = E_n(f).$$

Then

$$q := \frac{p^* + p}{2}$$

is also a best uniform approximation to f, for

$$\|f - q\|_\infty = \left\| \frac{1}{2}(f - p^*) + \frac{1}{2}(f - p) \right\|_\infty \leq \frac{1}{2}\|f - p^*\|_\infty + \frac{1}{2}\|f - p\|_\infty = E_n(f),$$

and equality holds since p, p^* are both best uniform approximations.

By (2.2.1.6), there exists an alternating set $\{x_1, x_2, \ldots, x_{n+2}\}$ for $f - q$. Thus for some integer $l = 0, 1$, we have

$$f(x_j) - q(x_j) = \frac{f(x_j) - p^*(x_j)}{2} + \frac{f(x_j) - p(x_j)}{2} = (-1)^{l+j} E_n(f), \quad j = 1, 2, \ldots, n+2. \tag{2.2.1.1}$$

Since

$$\frac{1}{2}\|f - p^*\|_\infty = \frac{1}{2}E_n(f) \quad \text{and} \quad \frac{1}{2}\|f - p\|_\infty = \frac{1}{2}E_n(f),$$

(2.2.1.1) can hold only if

$$f(x_j) - p^*(x_j) = f(x_j) - p(x_j) = (-1)^{l+j} E_n(f), \quad j = 1, 2, \ldots, n+2.$$

Thus we see that

$$p^*(x_j) = p(x_j), \quad j = 1, 2, \ldots, n+2,$$

which implies that

$$p \equiv p^*.$$

This completes the proof. □

Example 2.2.1.8. *Let $f(x) = \frac{2}{1+x}$ and $p(x) = \frac{11}{6} - x$, for $x \in [0, 1]$. We show that p is not the best uniform approximant to f out of $\Pi^1[0, 1]$.*

Observe that

$$\begin{aligned} e(x) = (f - p)(x) = f(x) - p(x) &= \frac{2}{1+x} - \frac{11 - 6x}{6} \\ &= \frac{12 - (11 - 6x)(1 + x)}{6(1 + x)} \\ &= \frac{12 - (11 + 5x - 6x^2)}{6(1 + x)} \\ &= \frac{6x^2 - 5x + 1}{6(1 + x)} \\ &= \frac{(1 - 3x)(1 - 2x)}{6 + 6x}. \end{aligned}$$

However, we note

$$\max_{x \in [0,1]} |e(x)| = \frac{1}{6},$$

and $e(0) = e(1) = \frac{1}{6}$. Moreover, $x = 0, 1$ are the only values for which this extreme value is achieved. We conclude that p is not the best uniform approximation to f.

3. Numerical Quadrature

In this section, we wish to calculate the definite integral of a real–valued function $f(x)$ on the interval $[a, b]$:

$$\int_a^b f(x)\ dx.$$

Recall that for some simple integrands $f(x)$, the indefinite integral

$$\int^t f(x)\ dx = F(t), \quad F'(x) = f(x),$$

can be obtained in closed form. It then follows from the fundamental theorem of calculus that

$$\int_a^b f(x)\ dx = F(b) - F(a).$$

As a general rule, however, definite integrals are often computed using discretization methods which approximate the integral by finite sums corresponding to a partition of the interval $[a, b]$. This process is known as numerical quadrature.

3.1. The Integration Formulas of Newton and Cotes.

3.1.1. *Newton–Cotes Formulas.* We first give the definition of a *quadrature rule.*

Definition 3.1.1.1 (Quadrature Rule, Quadrature Weights)**.** *A* **quadrature rule** *is a method that approximates the definite integral* $\int_a^b f(x)\ dx$ *by*

$$\sum_{j=0}^{n} \alpha_j f(x_j) \approx \int_a^b f(x)\ dx.$$

Moreover, the numbers α_j, $j = 0, 1, \ldots, n$ *are called the* **quadrature weights**, *whose values may depend only on the choice of* n, *but not on* a, b, *or* f.

In this section, we consider the definite integral

$$\int_a^b f(x)\ dx.$$

We obtain the integration formulas of Newton and Cotes if the integrand $f(x)$ is replaced by an interpolating polynomial $p(x)$ and then take $\int_a^b p(x)\ dx$ as an approximation for $\int_a^b f(x)\ dx$.

For the Newton–Cotes formulas, we must have a uniform partition of the interval $[a, b]$,

$$x_j := a + jh, \quad j = 0, 1, \ldots, n,$$

of step length $h := \frac{b-a}{n}$, for $n > 0$. Let $p_n \in \Pi^n$ be the interpolating polynomial of degree n or less with

$$p_n(x_j) = f(x_j) =: f_j, \quad j = 0, 1, \ldots, n.$$

By Lagrange's interpolation formula (1.1.1.3),

$$p_n(x) = \sum_{j=0}^{n} f_j L_j(x)$$

$$= \sum_{j=0}^{n} f_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{x - x_k}{x_j - x_k}$$
$$= \sum_{j=0}^{n} f_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{x - (a+kh)}{(a+jh) - (a+kh)}.$$

Now let the variable t be such that $x = a + th$. Then

$$p_n(x) = \sum_{j=0}^{n} f_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{(a+th) - (a+kh)}{(a+jh) - (a+kh)}$$
$$= \sum_{j=0}^{n} f_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{th - kh}{jh - kh}$$
$$= \sum_{j=0}^{n} f_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{t - k}{j - k}.$$

Define

$$\varphi_j(t) := L_j(x) = \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{t - k}{j - k}.$$

Integration by substitution $x = a + th$ gives

$$\int_a^b f(x)\ dx \approx \int_a^b p_n(x)\ dx = \int_a^b \sum_{j=0}^{n} f_j L_j(x)\ dx$$
$$= h \sum_{j=0}^{n} f_j \int_0^n \varphi_j(t)\ dt$$
$$= h \sum_{j=0}^{n} f_j \alpha_j,$$

where the *quadrature weights* α_j, $j = 0, 1, \ldots, n$ are such that

$$\alpha_j := \int_0^n \varphi_j(t)\ dt$$

and α_j depends only on n for each $j = 0, 1, \ldots, n$.

For any natural number n, the *Newton–Cotes formulas*

$$\int_a^b p_n(x)\ dx = h \sum_{j=0}^{n} f_j \alpha_j = h \sum_{j=0}^{n} f_j \int_0^n \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{t - k}{j - k}\ dt, \quad f_j := f(a + jh), \quad h := \frac{b-a}{n}$$

provide approximate values for $\int_a^b f(x)\ dx$.

Theorem 3.1.1.2. *If f is a polynomial of degree k, then selecting $n \geq k$ in the Newton–Cotes formulas gives*

$$\int_a^b f(x)\ dx = \int_a^b p_n(x)\ dx.$$

That is, quadrature is exact on Π^n.

Proof. The proposition follows immediately from the uniqueness of polynomial interpolation (1.1.1.1), noting that choosing $n \geq k$ gives

$$p_n \equiv f$$

on $[a, b]$. □

Theorem 3.1.1.3. *Let α_j, $j = 0, 1, \ldots, n$ be the weights for a Newton–Cotes quadrature rule. Then*

$$\sum_{j=0}^{n} \alpha_j = n.$$

Proof. By (3.1.1.2), choosing any integer $n > 0$ implies that the Newton–Cotes formulas integrate $f(x) \equiv 1$ exactly. Moreover, in the case $f(x) \equiv 1$, we have evidently that $f_j = 1$ for each $j = 0, 1, \ldots, n$. Thus

$$\begin{aligned} b - a = \int_a^b dx = \int_a^b f(x)\ dx \\ &= h \sum_{j=0}^{n} f_j \alpha_j \\ &= h \sum_{j=0}^{n} \alpha_j. \end{aligned}$$

That is,

$$b - a = h \sum_{j=0}^{n} \alpha_j.$$

Recalling that $h = \frac{b-a}{n}$, we have

$$\sum_{j=0}^{n} \alpha_j = n.$$

□

If s is a common denominator for the weights α_j so that the numbers

$$\sigma_j := s\alpha_j, \quad j = 0, 1, \ldots, n$$

are integers, then the Newton–Cotes formulas may be written

$$\begin{aligned} \int_a^b f(x)\ dx \approx \int_a^b p_n(x)\ dx = h \sum_{j=0}^{n} f_j \alpha_j \\ &= \frac{h}{s} \sum_{j=0}^{n} f_j \sigma_j = \frac{b-a}{ns} \sum_{j=0}^{n} f_j \sigma_j. \end{aligned}$$

Example 3.1.1.4 (Trapezoid Rule)**.** *The* **trapezoid rule** *is obtained by the Newton–Cotes formulas in the case $n = 1$. Then $h = b - a$, so that $f_0 = a$ and $f_1 = b$. Observe that*

$$\alpha_0 = \int_0^1 \frac{t-1}{0-1}\, dt = \int_0^1 1 - t\, dt = t - \frac{1}{2}t^2\Big|_0^1 = 1 - \frac{1}{2} = \frac{1}{2}$$

and

$$\alpha_1 = \int_0^1 \frac{t-0}{1-0}\, dt = \int_0^1 t\, dt = \frac{1}{2}t^2\Big|_0^1 = \frac{1}{2}.$$

Hence we have the approximation

$$\int_a^b f(x)\, dx \approx \int_a^b p_1(x)\, dx = (b-a)\sum_{j=0}^{1} f_j\alpha_j = \frac{b-a}{2}(f(a) + f(b)).$$

This is the **trapezoid rule**.

Example 3.1.1.5 (Simpson's Rule)**. Simpson's Rule** *is obtained by the Newton–Cotes formulas in the case $n = 2$. Then $h = \frac{b-a}{2}$, so that $f_0 = a$, $f_1 = \frac{a+b}{2}$, and $f_2 = b$. Observe further*

$$\alpha_0 = \int_0^2 \frac{(t-1)(t-2)}{(0-1)(0-2)}\, dt = \frac{1}{2}\int_0^2 t^2 - 3t + 2\, dt = \frac{1}{2}\left[\frac{1}{3}t^3 - \frac{3}{2}t^2 + 2t\right]_0^2 = \frac{1}{2}\left[\frac{8}{3} - 6 + 4\right] = \frac{1}{3},$$

$$\alpha_1 = \int_0^2 \frac{(t-0)(t-2)}{(1-0)(1-2)}\, dt = -\int_0^2 t^2 - 2t\, dt = -\left[\frac{1}{3}t^3 - t^2\right]_0^2 = -\left[\frac{8}{3} - 4\right] = \frac{4}{3},$$

and

$$\alpha_2 = \int_0^2 \frac{(t-0)(t-1)}{(2-0)(2-1)}\, dt = \frac{1}{2}\int_0^2 t^2 - t\, dt = \frac{1}{2}\left[\frac{1}{3}t^3 - \frac{1}{2}t^2\right]_0^2 = \frac{1}{2}\left[\frac{8}{3} - 2\right] = \frac{1}{3}.$$

Hence we have the approximation

$$\begin{aligned}\int_a^b f(x)\, dx \approx \int_a^b p_2(x)\, dx &= \frac{b-a}{2}\sum_{j=0}^{2} f_j\alpha_j \\ &= \frac{b-a}{2}\left[\frac{1}{3}f(a) + \frac{4}{3}f\left(\frac{a+b}{2}\right) + \frac{1}{3}f(b)\right],\end{aligned}$$

or

$$\int_a^b f(x)\, dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

This is **Simpson's rule**.

Example 3.1.1.6. *We discuss more rigorous methods for deriving the error in Newton–Cotes quadrature in the following section, but observe, for the trapezoid rule, that by* (1.1.4.2) *we have*

$$(f - p_1)(x) = \frac{f^{(2)}(\xi)}{2}(x-a)(x-b)$$

for some $\xi \in [a, b]$. Letting $I(f) = \int_a^b f(x)\, dx$ and $\tilde{I}_1(f)$ the trapezoidal approximation, we have

$$I(f) - \tilde{I}_1(f) = \int_a^b \frac{f''(\xi)}{2}(x-a)(x-b)\, dx = \frac{1}{2}\int_a^b f''(\xi)(x-a)(x-b)\, dx.$$

Since $(x-a)(x-b)$ does not change sign on $[a,b]$, we have by the weighted mean value theorem for integrals that

$$\begin{aligned} I(f) - \tilde{I}_1(f) &= \frac{f''(\xi)}{2} \int_a^b x^2 - (a+b)x + ab \; dx \\ &= \frac{f''(\xi)}{2} \left[\frac{1}{3}x^3 - \frac{a+b}{2}x^2 + abx \right]_a^b \\ &= \frac{f''(\xi)}{2} \left[-\frac{(b-a)^3}{6} \right] \\ &= -\frac{h^3}{12} f''(\xi). \end{aligned}$$

Example 3.1.1.7. *Similarly to the previous example, we see that for Simpson's rule,* (1.1.4.2) *implies that*

$$(f - p_2)(x) = \frac{f'''(\xi)}{6}(x-a)\left(x - \frac{a+b}{2}\right)(x-b)$$

for some $\xi \in [a,b]$. Thus

$$\begin{aligned} I(f) - \tilde{I}_2(f) &= \frac{1}{6} \int_a^b f'''(\xi)(x-a)\left(x - \frac{a+b}{2}\right)(x-b) \; dx \\ &= \frac{1}{6} \int_a^{\frac{a+b}{2}} f'''(\xi)(x-a)\left(x - \frac{a+b}{2}\right)(x-b) \; dx + \\ &\quad \frac{1}{6} \int_{\frac{a+b}{2}}^b f'''(\xi)(x-a)\left(x - \frac{a+b}{2}\right)(x-b) \; dx. \end{aligned}$$

By the weighted mean value theorem for integrals,

$$\begin{aligned} I(f) - \tilde{I}_2(f) &= \frac{f'''(\xi_1)}{6} \int_a^{\frac{a+b}{2}} (x-a)\left(x - \frac{a+b}{2}\right)(x-b) \; dx + \\ &\quad \frac{f'''(\xi_2)}{6} \int_{\frac{a+b}{2}}^b (x-a)\left(x - \frac{a+b}{2}\right)(x-b) \; dx \\ &= \frac{f'''(\xi_1)}{6}\left(\frac{(b-a)^4}{64}\right) + \frac{f'''(\xi_2)}{6}\left(-\frac{(b-a)^4}{64}\right) \\ &= \frac{h^4}{24}(f'''(\xi_1) - f'''(\xi_2)), \end{aligned}$$

for some $\xi_1 \in \left(a, \frac{a+b}{2}\right)$ and $\xi_2 \in \left(\frac{a+b}{2}, b\right)$. If we assume that $f \in C^4[a,b]$, then the mean value theorem implies that there exists $\xi \in (a,b)$ such that

$$f^{(4)}(\xi) = \frac{f'''(\xi_1) - f'''(x_2)}{\xi_1 - \xi_2}.$$

Thus

$$I(f) - \tilde{I}_2(f) = \frac{h^4(\xi_1 - \xi_2)}{24} f^{(4)}(\xi).$$

We show in the next section that a higher–order error term for Simpson's rule may be derived.

The Newton–Cotes formulas are generally not applied to the entire interval $[a, b]$, since the resulting quadrature error may be very large. Instead, we apply Newton–Cotes to a partition of $[a, b]$ and then approximate the full integral by taking the sum of approximations to the subintegrals. The resulting quadrature rule is called a *composite* rule.

Example 3.1.1.8 (Composite Trapezoid Rule)**.** *Let*

$$x_j := a + jh, \quad j = 0, 1, \ldots, n, \quad h := \frac{b-a}{n}$$

be a partition of the interval $[a, b]$. The trapezoid rule provides the approximation

$$\int_{x_j}^{x_{j+1}} f(x)\ dx \approx \frac{h}{2}[f(x_j) + f(x_{j+1})] =: \tilde{I}_j(f)$$

on each subinterval $[x_j, x_{j+1}]$, $j = 0, 1, \ldots, n-1$. For the entire interval $[a, b]$, we have

$$\begin{aligned}\int_a^b f(x)\ dx &= \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f(x)\ dx \\ &\approx \frac{h}{2} \sum_{j=0}^{n-1} [f(x_j) + f(x_{j+1})] \\ &= \frac{h}{2}[f(a) + 2f(a+h) + 2f(a+2h) + \cdots + 2f(b-h) + f(b)].\end{aligned}$$

This is the **composite trapezoid rule**.

Example 3.1.1.9 (Total Error for Composite Trapezoid Rule)**.** *In the next section, we will show that in each subinterval $[x_j, x_{j+1}]$, $j = 0, 1, \ldots, n-1$ there exists $\xi_j \in (x_j, x_{j+1})$ with*

$$I_j(f) - \tilde{I}_j(f) = \frac{h^3}{12} f''(\xi_j)$$

for $f \in C^2[a, b]$. Summing these individual error terms gives

$$\frac{h^3}{12} \sum_{j=0}^{n-1} f''(\xi_j) = \frac{h^2}{12} \left(\frac{b-a}{n} \right) \sum_{j=0}^{n-1} f''(\xi_j).$$

Since

$$\min_{j=0,1,\ldots,n-1} f''(x_j) \leq \frac{1}{n} \sum_{j=0}^{n-1} f''(\xi_j) \leq \max_{j=0,1,\ldots,n-1} f''(\xi_j)$$

and f'' is continuous on $[a, b]$, the intermediate value theorem implies that there exists $\xi \in (a, b)$ such that

$$f''(\xi) = \frac{1}{n} \sum_{j=0}^{n-1} f''(\xi_j).$$

Hence the total error for the composite trapezoid rule is

$$I(f) - \tilde{I}(f) = \frac{b-a}{12} h^2 f''(\xi),$$

for some $\xi \in (a, b)$.

Example 3.1.1.10 (Composite Simpson's Rule)**.** *We take n to be even. Then we may apply Simpson's rule to each subinterval*

$$[x_{2j}, x_{2j+2}], \quad j = 0, 1, \ldots, \frac{n}{2} - 1.$$

Simpson's rule provides the approximation

$$\int_{x_{2j}}^{x_{2j+2}} f(x)\ dx \approx \frac{h}{3}[f(x_{2j}) + 4f(x_{2j+1}) + f(x_{2j+2})] =: \tilde{I}_{2j}(f)$$

on each subinterval $[x_{2j}, x_{2j+2}]$. Then for the entire interval $[a, b]$, we have

$$\begin{aligned}
\int_a^b f(x)\ dx &= \sum_{j=0}^{\frac{n}{2}-1} \int_{x_{2j}}^{x_{2j+2}} f(x)\ dx \\
&\approx \frac{h}{3} \sum_{j=0}^{\frac{n}{2}-1} [f(x_{2j}) + 4f(x_{2j+1}) + f(x_{2j+2})] \\
&= \frac{h}{3} \left[f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \ldots + \right. \\
&\quad \left. 4f(b-3h) + 2f(b-2h) + 4f(b-h) + f(b) \right],
\end{aligned}$$

where the step size h is

$$h = \frac{1}{2}(x_{2j+2} - x_{2j}) = \frac{b-a}{n}.$$

This is the **composite Simpson's rule***.*

Example 3.1.1.11 (Total Error in Composite Simpson's Rule)**.** *In each subinterval $[x_{2j}, x_{2j+2}]$, there exists $\xi_{2j} \in (x_{2j}, x_{2j+2})$ with*

$$I_{2j}(f) - \tilde{I}_{2j}(f) = \frac{h^5}{90} f^{(4)}(\xi_{2j}),$$

for $f \in C^4[a, b]$. Summing these individual error terms gives

$$\frac{h^5}{90} \sum_{j=0}^{\frac{n}{2}-1} f^{(4)}(\xi_{2j}) = \frac{h^4}{90} \left(\frac{b-a}{n} \right) \sum_{j=0}^{\frac{n}{2}-1} f^{(4)}(\xi_{2j}).$$

Proceeding similarly to the error for the composite trapezoid rule, since

$$\min j = 0, 1, \ldots, \frac{n}{2} - 1 f^{(4)}(\xi_{2j}) \leq \frac{1}{n} \sum_{j=0}^{\frac{n}{2}-1} f^{(4)}(\xi_{2j}) \leq \max_{j=0,1,\ldots,\frac{n}{2}-1} f^{(4)}(\xi_{2j})$$

and $f^{(4)}$ is continuous on $[a, b]$ by the assumption, it follows by the intermediate value theorem that there exists $\xi \in (a, b)$ such that

$$f^{(4)}(\xi) = \frac{1}{n} \sum_{j=0}^{\frac{n}{2}-1} f^{(4)}(\xi_{2j}).$$

Hence the total error for the composite Simpson's rule is

$$I(f) - \tilde{I}(f) = \frac{b-a}{90} h^4 f^{(4)}(\xi),$$

for some $\xi \in (a, b)$.

Additional quadrature rules may be derived using Hermite interpolating polynomials of the integrand f.

Example 3.1.1.12 (Hermite Cubic Quadrature)**.** *Let* $p \in \Pi^3$ *be such that*

$$\begin{aligned} p(a) &= f(a), \quad p(b) = f(b), \\ p'(a) &= f'(a), \quad p'(b) = f'(b). \end{aligned}$$

In the special case $a = 0$, $b = 1$, *observe*

x_j	f			
$t_0 := 0$	$f(0)$			
		$f'(0)$		
$t_1 := 0$	$f(0)$		$f(1) - f(0) - f'(0)$	
		$f(1) - f(0)$		$f'(1) - 2f(1) + 2f(0) + f'(0)$
$t_2 := 1$	$f(1)$		$f'(1) - f(1) + f(0)$	
		$f'(1)$		
$t_3 := 1$	$f(1)$			

Thus

$$p(x) = f(0) + f'(0)x + [f(1) - f(0) - f'(0)]x^2 + [f'(1) - 2f(1) + 2f(0) + f'(0)]x^2(x - 1).$$

Integration gives

$$\begin{aligned} \int_0^1 p(x)\, dx &= \Big[f(0)x + \frac{1}{2}f'(0)x^2 + \frac{1}{3}(f(1) - f(0) - f'(0))x^3 + \\ &\qquad \frac{1}{4}(f'(1) - 2f(1) + 2f(0) + f'(0))x^4 - \frac{1}{3}(f'(1) + 2f(0) - 2f(1) + f'(0))x^3 \Big]_0^1 \\ &= f(0) + \frac{1}{2}f'(0) + \frac{1}{3}f(1) - \frac{1}{3}f(0) - \frac{1}{3}f'(0) + \frac{1}{4}f'(1) - \frac{1}{2}f(1) + \frac{1}{2}f(0) + \\ &\qquad \frac{1}{4}f'(0) - \frac{1}{3}f'(1) - \frac{2}{3}f(0) + \frac{2}{3}f(1) - \frac{1}{3}f'(0) \\ &= \frac{1}{2}f(0) + \frac{1}{2}f(1) + \frac{1}{12}f'(0) - \frac{1}{12}f'(1) \\ &= \frac{1}{2}[f(0) + f(1)] + \frac{1}{12}[f'(0) - f'(1)]. \end{aligned}$$

From this expression we can generalize to any interval $[a, b]$ *by introducing*

$$t := a + x(b - a).$$

Note that $dt = (b - a)dx$. *Then*

$$\begin{aligned} \int_a^b p(x)`dx &= \frac{b-a}{2}[f(a) + f(b)] + \frac{b-a}{12}\left[\frac{d}{dx}f(t(x))|_{t=0} - \frac{d}{dx}f(t(x))|_{t=1}\right] \\ &= \frac{b-a}{2}[f(a) + f(b)] + \frac{(b-a)^2}{12}[f'(a) - f'(b)], \end{aligned}$$

by the chain rule.

Example 3.1.1.13 (Composite Hermite Cubic Quadrature)**.** *Let*

$$x_j := a + jh, \quad j = 0, 1, \ldots, n, \quad h := \frac{b-a}{n}$$

be a partition of $[a, b]$*. On each subinterval* $[x_j, x_{j+1}]$*,* $j = 0, 1, \ldots, n$*, Hermite cubic quadrature provides the approximation*

$$\int_{x_j}^{x_{j+1}} f(x)\ dx \approx \frac{h}{2}[f(x_j) + f(x_{j+1})] + \frac{h^2}{12}[f'(x_j) - f'(x_{j+1})] =: \tilde{I}_j(f).$$

Summing these n *terms over the entire interval* $[a, b]$ *gives*

$$\begin{aligned}
\int_a^b f(x)\ dx &= \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f(x)\ dx \\
&\approx h \sum_{0}^{n} \left[\frac{1}{2} \left(f(x_j) + f(x_{j+1}) \right) + \frac{h}{12} \left(f'(x_j) - f'(x_{j+1}) \right) \right] \\
&= \frac{h}{2} \left[\sum_{j=0}^{n} (f(x_j) + f(x_{j+1})) \right] + \frac{h^2}{12}(f'(a) - f'(b)) \\
&= \frac{h}{2} \left[f(a) + 2f(a+h) + 2f(a+2h) + \cdots + 2f(b-h) + f(b) \right] + \\
&\quad \frac{h^2}{12}[f'(a) - f'(b)],
\end{aligned}$$

since $\sum_{j=0}^{n-1}[f'(x_j) - f'(x_{j+1})]$ *telescopes. This is the* **composite Hermite cubic quadrature rule**.

Example 3.1.1.14. *Similar methods to the trapezoid and Simpson's rules in the previous examples may be used to show that the total error for composite Hermite cubic quadrature is*

$$I(f) - \tilde{I}(f) = -\frac{b-a}{720} h^4 f^{(4)}(\xi)$$

for some $\xi \in (a, b)$*, provided that* $f \in C^4[a, b]$*.*

In comparison to the trapezoid rule, note that the composite Hermite cubic quadrature has improved the order of the method by 2 with minimal effort, namely, the computation of the two derivatives $f'(a)$ and $f'(b)$. Moreover, if these two boundary derivatives are known to agree, for instance, if f is periodic, then the trapezoid rule itself is a method of order 4.

This discussion prompts the following definition.

Definition 3.1.1.15 (Superconvergence)**.** *A* **superconvergent** *method is a method that converges faster than generally expected.*

3.2. Peano's Error Representation.

3.2.1. *Peano's Error Representation.* The quadrature rules considered so far are of the form

$$\tilde{I}(f) := \sum_{k=0}^{m_0} a_{k_0} f(x_{k_0}) + \sum_{k=0}^{m_1} a_{k_1} f'(x_{k_1}) + \cdots + \sum_{k=0}^{m_n} a_{k_n} f^{(n)}(x_{k_n}).$$

The quadrature error

$$R(f) := \tilde{I}(f) - I(f) = \tilde{I}(f) - \int_a^b f(x)\ dx$$

is a linear operator

$$R(\alpha f + \beta g) = \alpha R(f) + \beta R(g)$$

on some normed linear function space V, where $\alpha, \beta \in \mathbb{R}$ and $f, g \in V$. For instance, we may have $V = C^n[a,b]$ or $V = \Pi^n[a,b]$. The following integral representation of the quadrature error $R(f)$ is attributed to Peano.

Theorem 3.2.1.1 (Peano's Error Representation)**.** *Suppose that $R(p) = 0$ holds for all polynomials $p \in \Pi^n$, that is, every polynomial of degree at most n is integrated exactly by $\tilde{I}(f)$. Then for all functions $f \in C^{(n+1)}[a,b]$,*

$$R(f) = \int_a^b f^{(n+1)}(t)K(t)\ dt,$$

where

$$K(t) := \frac{1}{n!}R_x[(x-t)_+^n], \quad (x-t)_+^n := \begin{cases}(x-t)^n, & x \geq t,\\ 0, & x < t,\end{cases}$$

and

$$R_x[(x-t)_+^n]$$

denotes the error of $(x-t)_+^n$ considered as a function of x.

Definition 3.2.1.2 (Peano Kernel)**.** *The function*

$$K(t) := \frac{1}{n!}R_x[(x-t)_+^n]$$

is called the **Peano kernel** *of the operator R.*

Proof. We first consider the Taylor series expansion of $f(x)$ about $x = a$,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + r_n(x),$$

where r_n is taken to be the integral remainder

$$r_n(x) = \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\ dt = \frac{1}{n!}\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt.$$

Applying the linear operator R to $f(x)$ then gives

$$R(f) = R(r_n) = \frac{1}{n!}R_x\left(\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right),$$

since all terms preceding r_n in the Taylor series expansion belong to Π^n.

In order to prove the theorem, we must show that we may safely interchange the operator R_x with integration. We first show that

$$\frac{d^k}{dx^k}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right] = \int_a^b f^{(n+1)}(t)\left[\frac{d^k}{dx^k}[(x-t)_+^n]\right]dt$$

for $k = 1, 2, \ldots, n$. For $k < n$ this follows immediately from the fact that $(x-t)_+^n$ is $n-1$ times continuously differentiable. For $k = n-1$, in particular, we have

$$\frac{d^{n-1}}{dx^{n-1}}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right] = \int_a^b f^{(n+1)}(t)\frac{d^{n+1}}{dx^{n+1}}[(x-t)_+^n]dt,$$

and thus

$$\begin{aligned}\frac{d^{n+1}}{dx^{n+1}}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right] &= n!\int_a^b f^{(n+1)}(t)(x-t)_+\ dt \\ &= n!\int_a^x f^{(n+1)}(t)(x-t)\ dt.\end{aligned}$$

Then, by the fundamental theorem of calculus,

$$\begin{aligned}\frac{d^n}{dx^n}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right] &= \frac{d}{dx}\left[\frac{d^{n-1}}{dx^{n-1}}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right]\right] \\ &= \frac{d}{dx}\left[n!\int_a^x f^{(n+1)}(t)(x-t)\ dt\right] \\ &= n!\int_a^x f^{(n+1)}(t)\ dt + n!f^{(n+1)}(x)(x-x) \\ &= n!\int_a^x f^{(n+1)}(t)\ dt \\ &= n!\int_a^b f^{(n+1)}(t)\frac{d^n}{dx^n}\left[(x-t)_+^n\right]dt.\end{aligned}$$

This proves that the differential operators

$$\frac{d^k}{dx^k}, \quad k = 1, 2, \ldots, n,$$

commute with integration. Since $\tilde{I}(f)$ is a linear combination of these differential operators, it also commutes with integration. Particularly, observe that

$$\begin{aligned}&R_x\left(\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right) \\ &\quad= \tilde{I}\left(\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right) - \int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\ dx \\ &\quad= \sum_{k=0}^{m_0} a_{k_0}\int_a^b f^{(n+1)}(t)(x_k-t)_+^n\ dt + \cdots + \sum_{k=0}^{m_n} a_{k_n}\int_a^b f^{(n+1)}(t)\left[\frac{d^n}{dx^n}\left[(x-t)_+^n\right]_{x=x_k}\right]dt - \\ &\qquad\int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\ dx \\ &\quad= \sum_{k=0}^{m_0} a_{k_0}\int_a^b f^{(n+1)}(t)(x_k-t)_+^n\ dt + \cdots + \sum_{k=0}^{m_n} a_{k_n}\left(\frac{d^n}{dx^n}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\right]_{x=x_k}\right) - \\ &\qquad\int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\ dx.\end{aligned}$$

Since

$$\int_a^b f^{(n+1)}(t)R_x[(x-t)_+^n]dt = \sum_{k=0}^{m_0} a_{k_0}\int_a^b f^{(n+1)}(t)(x_k-t)_+^n\ dt + \cdots + \sum_{k=0}^{m_n} a_{k_n}\int_a^b f^{(n+1)}(t)\left[\frac{d^n}{dx^n}(x-t)_+^n\right]_{x=x_k} dt - \int_a^b f^{(n+1)}(t)\int_a^b (x-t)_+^n\ dx\ dt,$$

it only remains to show that

$$\int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\ dx = \int_a^b f^{(n+1)}(t)\int_a^b (x-t)_+^n\ dx\ dt.$$

By Fubini's Theorem, it follows

$$\begin{aligned}\int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dt\ dx &= \int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\ dx\ dt\\ &= \int_a^b f^{(n+1)}(t)\int_a^b (x-t)_+^n\ dx\ dt.\end{aligned}$$

This shows that the entire operator R_x commutes with integration.

It follows that

$$R(f) = \frac{1}{n!}\int_a^b f^{(n+1)}(t)R_x[(x-t)_+^n]\ dt = \int_a^b f^{(n+1)}(t)K(t)\ dt.$$

This proves the theorem. □

Example 3.2.1.3 (Trapezoid Rule)**.** *We find the Peano kernel for the trapezoid rule on the interval* $[0,1]$*. Recall that*

$$\tilde{I}(f) = \frac{1}{2}[f(0)+f(1)].$$

Clearly $\tilde{I}(f)$ *is exact on* Π^1*, so we may apply* (3.2.1.1) *with* $n=1$*. The Peano kernel* $K(t)$ *becomes*

$$\begin{aligned}K(t) &= \frac{1}{1!}R_x[(x-t)_+] = \tilde{I}[(x-t)_+] - I[(x-t)_+]\\ &= \frac{1}{2}(0-t)_+ + \frac{1}{2}(1-t)_+ - \int_0^1 (x-t)_+\ dx.\end{aligned}$$

By definition of $(x-t)_+$*, we have for* $t\in[0,1]$ *that*

$$(0-t)_+ = 0, \quad (1-t)_+ = 1-t$$

and

$$\begin{aligned}\int_0^1 (x-t)_+\ dx &= \int_t^1 (x-t)\ dx\\ &= \frac{1}{2}x^2 - tx\Big|_{x=t}^1\\ &= \frac{1}{2} - t - \frac{1}{2}t^2 + t^2\\ &= \frac{1}{2}t^2 - t + \frac{1}{2}\end{aligned}$$

$$= \frac{1}{2}(1-t)^2.$$

Hence the Peano kernel for the trapezoid rule on the interval $[0,1]$ *is*

$$\begin{aligned} K(t) &= \frac{1}{2}(1-t) - \frac{1}{2}(1-t)^2 \\ &= \frac{1}{2}(1-t)(1-(1-t)) \\ &= \frac{1}{2}(1-t)t, \\ &= \frac{1}{2}t - \frac{1}{2}t^2, \quad 0 \le t \le 1. \end{aligned}$$

Example 3.2.1.4 (Simpson's Rule)**.** *We find the Peano kernel for Simpson's rule on the interval* $[-1,1]$*. Recall that*

$$\tilde{I}(f) = \frac{1}{3}[f(-1) + 4f(0) + f(1)].$$

Clearly $\tilde{I}(f)$ *is exact on* Π^2*; we show that* $\tilde{I}(f)$ *is exact on* Π^3*.*

Consider integrating a polynomial p *of degree three,* $p \in \Pi^3$*. Let* $q \in \Pi^2$ *be such that*

$$q(-1) = p(-1), \quad q(0) = p(0), \quad q(1) = p(1).$$

Define the polynomial

$$S := p - q \in \Pi^3.$$

Then S *vanishes at the points* $x = -1, 0, 1$*. Since* $S \in \Pi^3$ *and has the three roots* $-1, 0, 1$*,* S *is evidently of the form*

$$S(x) = ax(x+1)(x-1) = a(x^2+x)(x-1) = a(x^3 - x^2 + x^2 - x) = ax^3 - ax.$$

Since $q \in \Pi^2$*,* $R(q) = 0$*. Thus*

$$\begin{aligned} R(p) = R(p) - R(q) = R(p-q) = R(S) &= \tilde{I}(S) - I(S) \\ &= \frac{1}{3}[S(-1) + 4S(0) + S(1)] - \int_{-1}^{1} S(x)\, dx \\ &= \int_{-1}^{1} ax^3 - ax\, dx = 0. \end{aligned}$$

Thus we may apply (3.2.1.1) *with* $n = 3$*. The Peano kernel becomes*

$$\begin{aligned} K(t) &= \frac{1}{3!} R_x[(x-t)_+^3] \\ &= \frac{1}{18}\left[(-1-t)_+^3 + 4(0-t)_+^3 + (1-t)_+^3\right] - \frac{1}{6}\int_{-1}^{1} (x-t)_+^3\, dx. \end{aligned}$$

By definition of $(x-t)_+^3$*, we have for* $t \in [-1,1]$ *that*

$$(-1-t)_+^3 = 0, \quad (1-t)_+^3 = (1-t)^3,$$

$$(0-t)_+^3 = -t_+^3 = \begin{cases} 0, & t \ge 0, \\ -t^3, & t < 0 \end{cases},$$

and

$$\begin{aligned}\int_{-1}^{1}(x-t)_+^3\ dx &= \int_t^1 (x-t)^3 \\ &= \frac{1}{4}(x-t)^4\Big|_{x=t}^1 \\ &= \frac{1}{4}(1-t)^4.\end{aligned}$$

Thus the Peano kernel for Simpson's rule for $t \in [0,1]$ *is*

$$\begin{aligned}K(t) &= \frac{1}{18}(1-t)^3 - \frac{1}{24}(1-t)^4 \\ &= \frac{1}{72}(1-t)^3(4-3(1-t)) \\ &= \frac{1}{72}(1-t)^3(1+3t).\end{aligned}$$

Likewise, the Peano kernel for $t \in [-1,0]$ *is*

$$\begin{aligned}K(t) &= \frac{1}{18}(-4t^3+(1-t)^3) - \frac{1}{24}(1-t)^4 \\ &= \frac{1}{72}(1-t)^3(1+3t) - \frac{2}{9}t^3 \\ &= \frac{1}{72}(1+t)^3(1-3t).\end{aligned}$$

Thus we see that on $[-1,1]$ *we have*

$$K(t) = \begin{cases} \frac{1}{72}(1-t)^3(1+3t), & 0 \le t \le 1, \\ K(-t), & -1 \le t \le 0. \end{cases}$$

To show how Peano's error formula is commonly used, we recall the following result from calculus.

Theorem 3.2.1.5 (Weighted Mean Value Theorem for Integrals)**.** *If* f *is continuous on the interval* $[a,b]$ *and* g *is an integrable function that does not change sign on* $[a,b]$*, then there exists a number* $\xi \in (a,b)$ *such that*

$$\int_a^b f(x)g(x)\ dx = f(\xi)\int_a^b g(x)\ dx.$$

For a large class of quadrature rules, the Peano kernel $K(t)$ has constant sign on $[a,b]$. In particular, the Peano kernels for every Newton–Cotes quadrature rule have constant sign on their respective intervals of integration. In this situation, the weighted mean value theorem for integral calculus (3.2.1.5) gives

$$R(f) = \int_a^b f^{(n+1)}(t)K(t)\ dt = f^{(n+1)}(\xi)\int_a^b K(t)\ dt \tag{3.2.1.1}$$

for some $\xi \in (a,b)$, provided that $f \in C^{n+1}[a,b]$. Moreover, since $K(t)$ does not depend on f, we may determine the integral by applying the operator R, for instance, to the polynomial

$p(x) = x^{n+1}$. This gives

$$R(x^{n+1}) = (n+1)! \int_a^b K(t)\ dt.$$

Rearranging this expression, we obtain

$$\int_a^b K(t)\ dt = \frac{R(x^{n+1})}{(n+1)!},$$

and inserting for the integral in (3.2.1.1) gives

$$R(f) = f^{(n+1)}(\xi) \int_a^b K(t)\ dt = \frac{R(x^{n+1})}{(n+1)!} f^{(n+1)}(\xi),$$

for some $\xi \in (a, b)$.

That is, once we determine that the Peano kernel $K(t)$ has constant sign on $[a, b]$, we no longer need it to determine the quadrature error $R(f)$. In summary,

Theorem 3.2.1.6 (Quadrature Error Formula)**.** *Suppose that $R(p) = 0$ holds for all polynomials $p \in \Pi^n$, that is, every polynomial of degree at most n is integrated exactly by $\tilde{I}(f)$. If the Peano kernel*

$$K(t) := \frac{1}{n!} R_x[(x-t)_+^n]$$

has constant sign on $[a, b]$, then for all functions $f \in C^{n+1}[a, b]$ there exists a number $\xi \in (a, b)$ such that

$$R(f) = \tilde{I}(f) - I(f) = \frac{R(x^{n+1})}{(n+1)!} f^{(n+1)}(\xi).$$

Example 3.2.1.7 (Error for Trapezoid Rule)**.** *We find the error for the trapezoid rule on the interval $[0, 1]$. Recall that the Peano kernel in this situation is*

$$K(t) = \frac{1}{2} t(1-t), \quad 0 \le t \le 1,$$

which is nonnegative throughout $[0, 1]$. We find

$$\begin{aligned} R(x^2) = \tilde{I}(x^2) - I(x^2) &= \frac{1}{2}[0+1] - \int_0^1 x^2\ dx \\ &= \frac{1}{2} - \frac{1}{3} x^3 \Big|_0^1 \\ &= \frac{1}{2} - \frac{1}{3} = \frac{1}{6}. \end{aligned}$$

Hence, for $f \in C^2[0, 1]$, there exists $\xi \in (0, 1)$ such that

$$R(f) = \frac{R(x^2)}{2!} f''(\xi) = \frac{1}{12} f''(\xi).$$

Note that, if we assume that the Peano kernel has constant sign throughout the arbitrary interval $[a, b]$, we have thus

$$R(x^2) = \tilde{I}(x^2) - I(x^2) = \frac{b-a}{2} \left[a^2 + b^2\right] - \int_a^b x^2\ dx$$

$$= \frac{b-a}{2}\left[a^2+b^2\right] - \frac{1}{3}x^3\Big|_a^b$$
$$= \frac{b-a}{2}\left[a^2+b^2\right] - \frac{1}{3}\left[b^3-a^3\right]$$
$$= \frac{(b-a)^3}{6} = \frac{h^3}{6},$$

for step size $h := b-a$. *Hence if* $f \in C^2[a,b]$, *then there is* $\xi \in (a,b)$ *such that*

$$R(f) = \frac{R(x^2)}{2!}f''(\xi) = \frac{h^3}{12}f''(\xi),$$

which was assumed in the previous section.

Example 3.2.1.8 (Error for Simpson's Rule)**.** *We find the error for Simpson's rule on the interval* $[-1,1]$. *Recall that the Peano kernel was found to be*

$$K(t) = \begin{cases} \frac{1}{72}(1-t)^3(1+3t), & 0 \le t \le 1, \\ K(-t), & -1 \le t \le 0, \end{cases}$$

which is nonnegative throughout $[-1,1]$. *We find*

$$R(x^4) = \tilde{I}(x^4) - I(x^4) = \frac{1}{3}[1+0+1] - \int_{-1}^{1} x^4\,dx$$
$$= \frac{2}{3} - \frac{1}{5}x^5\Big|_{-1}^{1}$$
$$= \frac{2}{3} - \frac{2}{5} = \frac{4}{15}.$$

Thus for any $f \in C^4[-1,1]$ *there is* $\xi \in (-1,1)$ *such that*

$$R(f) = \frac{R(x^4)}{4!}f^{(4)}(\xi) = \frac{1}{90}f^{(4)}(\xi).$$

Assuming that the Peano kernel has constant sign throughout $[a,b]$, *we have*

$$R(x^4) = \tilde{I}(x^4) - I(x^4) = \frac{b-a}{6}\left[a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right] - \int_a^b x^4\,dx$$
$$= \frac{b-a}{6}\left[a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right] - \frac{1}{5}x^5\Big|_a^b$$
$$= \frac{b-a}{6}\left[a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right] - \left[\frac{b^5}{5} - \frac{a^5}{5}\right]$$
$$= \frac{(b-a)^5}{120} = \frac{4}{15}h^5,$$

for step size $h := \frac{b-a}{2}$. *Hence if* $f \in C^4[a,b]$, *then there is* $\xi \in (a,b)$ *such that*

$$R(f) = \frac{R(x^4)}{4!}f^{(4)}(\xi) = \frac{h^5}{90}f^{(4)}(\xi).$$

Note that this is a higher–order error term than was derived in the previous section.

In general, the Newton–Cotes formulas of degree n integrate without error on Π^n if n is odd and Π^{n+1} if n is even. The Peano kernels for the Newton–Cotes formulas are of constant sign, and we have for the $n-$th degree Newton–Cotes formula

$$R_n(f) = \begin{cases} \dfrac{R_n(x^{n+1})}{(n+1)!} f^{(n+1)}(\xi), & n \text{ is odd}, \\ \dfrac{R_n(x^{n+2})}{(n+2)!} f^{(n+2)}(\xi), & n \text{ is even}. \end{cases}$$

Example 3.2.1.9 (Hermite Cubic Quadrature)**.** *Lastly, we derive the error induced by Hermite cubic quadrature. Recall*

$$\tilde{I}(f) = \frac{h}{2}(f(a) + f(b)) + \frac{h^2}{12}(f'(a) - f'(b)), \quad h := b - a,$$

which clearly integrates exactly polynomials $p \in \Pi^3$. For $n = 3$, we obtain the following Peano kernel $K(t)$:

$$\begin{aligned} K(t) &= \frac{1}{6} R_x[(x-t)_+^3] \\ &= \frac{1}{6}\left[\frac{h}{2}((a-t)_+^3 + (b-t)_+^3) + \frac{h^2}{4}((a-t)_+^2 - (b-t)_+^2) - \int_a^b (x-t)_+^3 \, dx\right] \\ &= \frac{1}{6}\left[\frac{h}{2}(b-t)^3 - \frac{h^2}{4}(b-t)^2 - \int_t^b (x-t)^3 \, dx\right] \\ &= \frac{1}{6}\left[\frac{h}{2}(b-t)^3 - \frac{h^2}{4}(b-t)^2 - \frac{1}{4}(x-t)^4\right] \\ &= -\frac{1}{24}(b-t)^2(b-h-t)^2 = -\frac{1}{24}(b-t)^2(a-t)^2, \end{aligned}$$

which is clearly nonpositive throughout $[a, b]$, so that we may apply (3.2.1.6)*. We find*

$$\begin{aligned} R(x^4) = \tilde{I}(x^4) - I(x^4) &= \frac{b-a}{2}(a^4 + b^4) + \frac{(b-a)^2}{12}(4a^3 - 4b^3) - \int_a^b x^4 \, dx \\ &= \frac{(b-a)}{2}(a^4 + b^4) + \frac{(b-a)^2}{3}(a^3 - b^3) - \frac{1}{5}\Big|_a^b \\ &= \frac{(b-a)}{2}(a^4 + b^4) + \frac{(b-a)^2}{3}(a^3 - b^3) - \left(\frac{b^5}{5} - \frac{a^5}{5}\right) \\ &= -\frac{(b-a)^5}{30} = -\frac{h^5}{30}, \end{aligned}$$

for step size $h := b - a$. Hence if $f \in C^4[a, b]$, then there exists $\xi \in (a, b)$ such that

$$R(f) = \frac{R(x^4)}{4!} f^{(4)}(\xi) = -\frac{h^5}{720} f^{(4)}(\xi),$$

which was assumed in the previous section.

3.3. **Gaussian Integration Methods.**

3.3.1. *Gaussian Quadrature.* We recall the definition of a *weight function.*

Definition 3.3.1.1 (Weight Function)**.** *A* **weight function** *on the interval* $[a, b]$ *is a function* $w(x)$ *that satisfies the following properties:*

(1) $w(x) \geq 0$ *is measurable on* $[a, b]$,

(2) All moments $\mu_k := \int_a^b x^k w(x)\ dx$, $k = 0, 1, \ldots$, *exist and are finite,*

(3) $\int_a^b w(x)\ dx > 0$.

We note that the conditions for a weight function $w(x)$ are met if $w(x)$ is positive and continuous on an interval with finite measure.

In this section we consider integrals of the form

$$I(f) := \int_a^b w(x) f(x)\ dx,$$

where $w(x)$ is a given nonnegative weight function on $[a, b]$. We again examine quadrature rules of the type

$$\tilde{I}(f) = \sum_{j=0}^{n} w_j f(x_j).$$

For Newton–Cotes rules, the abscissas were required to form a uniform partition of the interval $[a, b]$. Here, we try to choose the nodes x_j and weights w_j so as to maximize the order of the quadrature method. This leads to a class of quadrature rules known as the *Gaussian quadrature* formulas.

We will define

$$\bar{\Pi}^n := \{p : p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n\}$$

to be the set of all normed polynomials of degree n.

We recall the following definitions from the section regarding approximation theory, particularly, see the section on least squares function approximation.

Definition 3.3.1.2 (L^2 Inner Product)**.** *Let* $f, g \in L^2[a, b]$. *We define the* **weighted inner product** $\langle f, g \rangle$ *of* f *and* g *by*

$$\langle f, g \rangle := \int_a^b w(x) f(x) g(x)\ dx.$$

Definition 3.3.1.3 (L^2 Norm)**.** *Let* $f \in L^2[a, b]$. *We define the* **norm** $\|f\|$ *of* f *on* $L^2[a, b]$ *by*

$$\|f\| := \sqrt{\langle f, f \rangle} = \sqrt{\int_a^b w(x)[f(x)]^2\ dx}.$$

Definition 3.3.1.4 ($w-$Orthogonal)**.** *The functions* $f, g \in L^2[a, b]$ *are said to be* $w-$**orthogonal** *on* $[a, b]$ *if*

$$\langle f, g \rangle = 0.$$

We also recall the following result regarding the construction of $w-$orthogonal polynomials from (2.1.1.13).

Theorem 3.3.1.5 (Construction of $w-$Orthogonal Polynomials)**.** *There exist polynomials $\phi_n \in \bar{\Pi}^n$, $n = 0, 1, \ldots,$ such that*

$$\langle \phi_i, \phi_j \rangle = 0, \quad \text{for } i \neq j.$$

These polynomials ϕ_n are uniquely defined by the recursion

$$\phi_0(x) := 1,$$

$$\phi_1(x) := x - B_1, \quad B_1 := \frac{\int_a^b xw(x)\ dx}{\int_a^b w(x)\ dx},$$

and, when $k \geq 2$,

$$\phi_k(x) := (x - B_k)\phi_{k-1}(x) - C_k^2 \phi_{k-2}(x),$$

where

$$B_k := \frac{\langle x\phi_{k-1}, \phi_{k-1} \rangle}{\langle \phi_{k-1}, \phi_{k-1} \rangle} = \frac{\int_a^b xw(x)[\phi_{k-1}(x)]^2\ dx}{\int_a^b w(x)[\phi +_{k-1}(x)]^2\ dx},$$

$$C_k^2 := \frac{\langle \phi_{k-1}, \phi_{k-1} \rangle}{\langle \phi_{k-2}, \phi_{k-2} \rangle} = \frac{\int_a^b w(x)[\phi_{k-1}(x)]^2\ dx}{\int_a^b w(x)[\phi +_{k-2}(x)]^2\ dx}.$$

Corollary 3.3.1.6. *If $\{\phi_0, \phi_1, \ldots, \phi_n\}$ are the $w-$orthogonal polynomials as given by* (2.1.1.13)*, then for any $p \in \Pi^{n-1}$, we have*

$$\langle p, \phi_n \rangle = 0.$$

Proof. This corollary is equivalent to (2.1.1.14). □

We arrive at the following result regarding the roots of the $n-th$ orthogonal polynomial ϕ_n.

Theorem 3.3.1.7 (Roots of ϕ_n)**.** *The roots x_i, $i = 1, 2, \ldots, n$ of $\phi_n(x)$ are real and simple. Moreover, each x_i lies in the open interval (a, b).*

Proof. First note that $\phi_0 \equiv 1$ has no roots. Thus, assume that $n \geq 1$, for otherwise the theorem follows vacuously.

If $\phi_n(x) > 0$ for all $x \in (a, b)$, then

$$\int_a^b w(x)\phi_n(x)\phi_0(x)\ dx = \int_a^b w(x)\phi_n(x)\ dx > 0,$$

but, by the orthogonality condition, $\int_a^b w(x)\phi_n(x)\phi_0(x)\ dx = 0$. Thus $\phi_n(x)$ changes sign at least once on (a, b).

Let

$$a < z_1 < z_2 < \cdots < z_k < b$$

be the distinct real roots of ϕ_n with odd multiplicity. Define the polynomial

$$q(x) := \prod_{j=1}^{k} (x - z_j).$$

Then clearly $q \in \Pi^k$. Moreover, the polynomial $\phi_n(x)q(x)$ does not change sign on (a, b). It follows that

$$\int_a^b w(x)\phi_n(x)q(x)\ dx \neq 0.$$

Since $k \leq n$, we have by the orthogonality condition that $k = n$, for otherwise, $\int_a^b w(x)\phi_n(x)q(x)$ would equal zero. Hence, $\phi_n(x)$ has n zeros at z_k, $k = 1, 2, \ldots, n$, where each z_k lies in the interval (a, b).

This proves the theorem. □

Theorem 3.3.1.8. *The $n \times n$ matrix*

$$A := \begin{bmatrix} \phi_0(t_1) & \ldots & \phi_0(t_1) \\ \vdots & \ddots & \vdots \\ \phi_{n-1}(t_n) & \ldots & \phi_{n-1}(t_n) \end{bmatrix}$$

is nonsingular for mutually distinct arguments t_j, $j = 1, 2, \ldots, n$.

Proof. By contradiction, suppose that A is singular. Then there exists a vector

$$c := [c_0, c_1, \ldots, c_n]^\top \neq 0$$

such that $c^\top A = 0$. The polynomial

$$q(x) := \sum_{j=0}^{n-1} c_j\phi_j(x)$$

has the n distinct roots $t_1, t_2, \ldots, t_n$. That is,

$$q(t_k) = \sum_{j=0}^{n-1} c_j\phi_j(t_k) = 0, \quad k = 1, 2, \ldots, n.$$

But since each ϕ_j, $j = 0, 1, \ldots, n-1$ is a polynomial of degree precisely j, $q \in \Pi^{n-1}$ and has n distinct roots and thus vanishes identically,

$$q \equiv 0.$$

Since the polynomials $\{\phi_0, \phi_1, \ldots, \phi_{n-1}\}$ are linearly independent, $q(x) \equiv 0$ implies that $c = 0$, a contradiction to the assumption.

This completes the proof. □

The Theorem (3.3.1.8), together with the invertible matrix theorem (1.1.5.1), shows that the interpolation problem of finding a function of the form

$$p(x) = \sum_{j=0}^{n-1} c_j\phi_j(x)$$

is always solvable, with $p(t_j) = f_j$, $j = 1, 2, \ldots, n$. The condition that the arguments t_j, $j = 1, 2, \ldots, n$ are mutually distinct is known as the *Haar condition*. Any sequence of functions $f_0, f_1, \ldots,$ that satisfy this Haar condition is said to form a *Chebyshev system*. In particular, Theorem (3.3.1.8) states that sequences of $w-$orthogonal polynomials $\phi_0, \phi_1, \ldots,$ for instance, those $w-$orthogonal polynomials constructed via (2.1.1.13) form Chebyshev systems.

We arrive at the main result of this section.

Theorem 3.3.1.9 (Characterization of Nodes and Weights)**.**

(1) Let $x_1, x_2, \ldots, x_n$ be the roots of the $n-$th $w-$orthogonal polynomial $\phi_n(x)$, and let $w_1, w_2, \ldots, w_n$ be the solution of the nonsingular system of equations

$$\sum_{i=1}^{n} \phi_k(x_i) w_i = \begin{cases} \langle \phi_0, \phi_0 \rangle, & \text{if } k = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3.1.1}$$

Then $w_i > 0$ for each $i = 1, 2, \ldots, n$, and

$$\int_a^b w(x)p(x)\ dx = \sum_{i=1}^{n} w_i p(x_i) \tag{3.3.1.2}$$

holds for all polynomials $p \in \Pi^{2n-1}$.

(2) Conversely, if the numbers x_i, w_i, $i = 1, 2, \ldots, n$ are such that (3.3.1.2) holds for all $p \in \Pi^{2n-1}$, then the x_i, $i = 1, 2, \ldots, n$ are the roots of the $n-$th $w-$orthogonal polynomial ϕ_n and the weights w_i, $i = 1, 2, \ldots, n$ satisfy (3.3.1.1).

(3) It is not possible to find numbers x_i, w_i, $i = 1, 2, \ldots, n$ such that (3.3.1.2) holds for all polynomials $p \in \Pi^{2n-1}$.

Proof. Since the roots x_i, $i = 1, 2, \ldots, n$ of ϕ_n are real mutually distinct arguments in (a, b) (3.3.1.7), the matrix

$$A := \begin{bmatrix} \phi_0(x_1) & \ldots & \phi_0(x_n) \\ \vdots & \ddots & \vdots \\ \phi_{n-1}(x_1) & \ldots & \phi_{n-1}(x_n) \end{bmatrix}$$

is nonsingular (3.3.1.8). Thus the system (3.3.1.1) has a unique solution.

Let $p \in \Pi^{2n-1}$ be arbitrary. By polynomial division, we may write

$$p(x) := \phi_n(x)q(x) + r(x),$$

for some $q, r \in \Pi^{n-1}$. since $\{\phi_0, \phi_1, \ldots, \phi_{n-1}\}$ forms a basis for Π^{n-1}, there exist unique coefficients $r_i, q_i \in \mathbb{R}$, $i = 0, 1, \ldots, n-1$ such that

$$q(x) = \sum_{i=0}^{n-1} q_i \phi_i(x), \quad r(x) = \sum_{i=0}^{n-1} r_i \phi_i(x).$$

It follows

$$\begin{aligned} \int_a^b w(x)p(x)\ dx &= \int_a^b w(x)[\phi_n(x)q(x) + r(x)]\ dx \\ &= \int_a^b w(x)\phi_n(x)q(x)\ dx + \int_a^b w(x)r(x)\ dx \\ &= \sum_{i=0}^{n-1} q_i \int_a^b w(x)\phi_i(x)\phi_n(x)\ dx + \sum_{i=0}^{n-1} r_i \int_a^b w(x)\phi_i(x)\ dx \\ &= \sum_{i=0}^{n-1} r_i \int_a^b w(x)\phi_i(x)\phi_0(x)\ dx \end{aligned}$$

$$= r_0 \int_a^b w(x) \left[\phi_0(x)\right]^2 \, dx.$$

On the other hand, since $\phi_n(x_i) = 0$, $i = 1, 2, \ldots, n$, we have by (3.3.1.1) that

$$\begin{aligned}
\sum_{i=1}^{n} w_i p(x_i) &= \sum_{i=1}^{n} w_i [\phi_n(x_i) q(x_i) + r(x_i)] \\
&= \sum_{i=1}^{n} w_i q(x_i) \phi_n(x_i) + \sum_{i=1}^{n} w_i r(x_i) \\
&= \sum_{i=1}^{n} w_i \left[\sum_{k=0}^{n-1} r_k \phi_k(x_i) \right] \\
&= \sum_{k=0}^{n-1} r_k \sum_{i=1}^{n} w_i \phi_k(x_i) \\
&= r_0 \int_a^b w(x) \left[\phi_0(x)\right]^2 \, dx,
\end{aligned}$$

which proves (3.3.1.2).

We now show that $w_i > 0$ for each $i = 1, 2, \ldots, n$. Define the polynomials $p_j(x)$, $j = 1, 2, \ldots, n$ as follows:

$$p_j(x) := \prod_{\substack{k=1 \\ k \neq j}}^{n} (x - x_k)^2 \in \Pi^{2n-2}.$$

Since $p_j \geq 0$ on $[a, b]$ and clearly does not vanish identically, applying (3.3.1.2), it follows

$$\begin{aligned}
0 &< \int_a^b w(x) p_j(x) \, dx \\
&= \sum_{i=1}^{n} w_i p_j(x_i) \\
&= w_j \prod_{\substack{k=1 \\ k \neq j}}^{n} (x - x_k)^2.
\end{aligned}$$

Noting that the product is strictly positive, we have $w_j > 0$ for each $j = 1, 2, \ldots, n$.

This completes the proof of (3.3.1.9)[1].

We now show (3.3.1.9)[2]. By contradiction, suppose that there exist numbers x_i, w_i, $i = 1, 2, \ldots, n$ such that (3.3.1.2) holds for all polynomials $p \in \Pi^{2n}$. Put

$$\tilde{p}(x) := \prod_{j=1}^{n} (x - x_j)^2 \in \Pi^{2n}.$$

Then since $\tilde{p} \geq 0$ on $[a, b]$ and does not vanish identically, it follows from (3.3.1.2) that

$$0 < \int_a^b w(x) \tilde{p}(x) \, dx$$

$$\begin{aligned}
&= \sum_{i=1}^{n} w_i \tilde{p}(x_i) \\
&= \sum_{i=1}^{n} w_i \prod_{j=1}^{n} (x_i - x_j) \\
&= 0,
\end{aligned}$$

which implies that $0 < \int_a^b w(x)\tilde{p}(x)\ dx = 0$, which is clearly absurd. This proves (3.3.1.9)[3].

Lastly, we show (3.3.1.9)[2]. Suppose that the numbers x_i, w_i, $i = 1, 2, \ldots, n$ are such that (3.3.1.2) holds for all $p \in \Pi^{2n-1}$. Note that the abscissas x_i, $i = 1, 2, \ldots, n$ must be distinct, for otherwise, we may reformulate terms and sum to obtain a quadrature rule that is exact on Π^{2n-1} with less than n points, a contradiction to (3.3.1.9)[3].

Recall that if (3.3.1.2) holds and the x_i, $i = 1, 2, \ldots, n$ are distinct, then $w_i > 0$ for each $i = 1, 2, \ldots, n$.

We apply (3.3.1.2) to each ϕ_j, $j = 0, 1, \ldots, n-1$ to find

$$\begin{aligned}
\sum_{i=1}^{n} w_i \phi_j(x_i) &= \int_a^b w(x)\phi_j(x)\ dx \\
&= \int_a^b w(x)\phi_j(x)\phi_0(x)\ dx \\
&= \begin{cases} \int_a^b w(x)\left[\phi_0(x)\right]^2\ dx, & j = 0, \\ 0, \quad \text{otherwise.} \end{cases}
\end{aligned}$$

This proves (3.3.1.1).

It remains to show that $\phi_n(x_i) = 0$ for each x_i, $i = 1, 2, \ldots, n$. Since $\phi_j\phi_n \in \Pi^{2n-1}$ for $j = 0, 1, \ldots, n-1$, (3.3.1.2) gives

$$\begin{aligned}
\sum_{i=1}^{n} w_i \phi_n(x_i)\phi_j(x_i) &= \int_a^b w(x)\phi_n(x)\phi_j(x)\ dx \\
&= 0.
\end{aligned}$$

That is, the vector

$$c := [w_1\phi_n(x_1), w_2\phi_n(x_2), \ldots, w_n\phi_n(x_n)]^\top$$

solves the homogeneous system $Ac = 0$, with

$$A := \begin{bmatrix} \phi_0(x_1) & \ldots & \phi_0(x_n) \\ \vdots & \ddots & \vdots \\ \phi_{n-1}(x_1) & \ldots & \phi_{n-1}(x_n). \end{bmatrix}$$

Since the abscissas x_i, $i = 1, 2, \ldots, n$ are distinct, A is nonsingular (3.3.1.8). Thus $c = 0$, so that $w_i\phi_n(x) = 0$ for each $i = 1, 2, \ldots, n$. Furthermore, since $w_i > 0$ for each $i = 1, 2, \ldots, n$, it follows that

$$\phi_n(x_i) = 0, \quad i = 1, 2, \ldots, n.$$

This completes the proof. □

Note that by Theorem (3.3.1.9), we have characterized the quantities x_i and w_i which enter the Gaussian quadrature rules for given weight functions $w(x)$, but it remains to discuss their

actual calculation. We assume that the coefficients B_j, C_j of the $w-$ orthogonal polynomial recursion (2.1.1.13) are given.

We consider the tridiagonal matrices

$$J_n := \begin{bmatrix} B_1 & C_2 & & & \\ C_2 & B_1 & C_2 & & \\ & C_3 & B_3 & \ddots & \\ & & \ddots & \ddots & C_n \\ & & & C_n & B_n \end{bmatrix} \tag{3.3.1.3}$$

as well as their principal submatrices

$$J_j := \begin{bmatrix} B_1 & C_2 & & & \\ C_2 & B_1 & C_2 & & \\ & C_3 & B_3 & \ddots & \\ & & \ddots & \ddots & C_j \\ & & & C_j & B_j \end{bmatrix},$$

where B_k, C_k, are such that

$$\phi_k(x) = (x - B_k)\phi_{k-1}(x) - C_k^2\phi_{k-2}(x),$$

that is,

$$B_k := \frac{\int_a^b xw(x)[\phi_{k-1}(x)]^2\ dx}{\int_a^b w(x)[\phi_{k-1}(x)]^2\ dx},$$
$$C_k := \frac{\int_a^b w(x)[\phi_{k-1}(x)]^2\ dx}{\int_a^b w(x)[\phi_{k-2}(x)]^2\ dx}.$$

We have the following result regarding the roots of $\phi_n(x)$ and the eigenvalues of J_n.

Theorem 3.3.1.10. *The roots x_i, $i = 1, 2, \ldots, n$ of the $n-$th $w-$orthogonal polynomial ϕ_n are the eigenvalues of the tridiagonal matrix J_n.*

Proof. We use induction. Recall that $\phi_0(x) \equiv 1$ has no roots and corresponds to the empty matrix J_0. Observe

$$\begin{aligned} \phi_i(x) &= (x - B_1)\phi_0(x) \\ &= x - B_1 \\ &= -\det(J_1 - x[1]). \end{aligned}$$

Let I_n denote the $n \times n$ identity matrix. Now

$$\begin{aligned} \phi_2(x) &= (x - B_2)\phi_1(x) - C_2^2\phi_0(x) \\ &= (x - B_2)(x - B_1) - C_2^2 \\ &= (B_2 - x)(B_1 - x) - C_2^2 \\ &= \det\left(\begin{bmatrix} B_1 - x & C_2 \\ C_2 & B_2 - x \end{bmatrix}\right) \\ &= \det(J_2 - I_2 x). \end{aligned}$$

This shows the base cases $j = 0, 1, 2$.

We show by induction on j that

$$\phi_j(x) = (-1)^j \det(J_j - I_j x).$$

Note

$$J_j - I_j x = \left[\begin{array}{ccccccc|c} B_1 - x & C_2 & & & & & & 0 \\ C_2 & B_2 - x & C_3 & & & & & \\ & C_3 & B_3 - x & \ddots & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & B_{j-2} - x & C_{j-1} & & \\ & & & & C_{j-1} & B_{j-1} - x & C_j & \\ 0 & & & & & C_j & B_j - x \end{array}\right]$$

so that

$$\begin{aligned} \det(J_j - I_j x) &= (B_j - x)\det(J_{j-1} - I_{j-1}x) - \\ &\qquad c_j \det\left(\begin{bmatrix} B_1 - x & C_2 & & & & \\ C_2 & B_2 - x & C_3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & C_{j-2} & B_{j-2} - x & C_{j-1} & \\ & & & 0 & C_j \end{bmatrix}\right) \\ &= (B_j - x)\det(J_{j-1} - I_{j-1}x) - C_j^2 \det(I_{j-2} - J_{j-2}x) \\ &= (B_j - x)(-1)^{j-1}\phi_{j-1}(x) - C_j^2(-1)^{j-2}\phi_{j-2}(x), \end{aligned}$$

by the induction hypothesis. Thus,

$$\begin{aligned} (-1)^j \det(I_j - J_j x) &= (-1)^{2j-1}(B_j - x)\phi_{j-1}(x) - (-1)^{2j-2}C_j^2\phi_{j-2}(x) \\ &= (-1)^{2j}(x - B_j)\phi_{j-1}(x) - C_j^2\phi_{j-2}(x) \\ &= (x - B_j)\phi_{j-1}(x) - C_j^2\phi_{j-2}(x) \\ &= \phi_j(x), \end{aligned}$$

by the recursion (2.1.1.13).

Hence,

$$\phi_n(x) = (-1)^n \det(J_n - I_n x) = 0$$

if and only if x is an eigenvalue of J_n. This proves the theorem. □

We note here that since the roots of ϕ_n are real and distinct (3.3.1.7), it follows immediately by (3.3.1.10) that the tridiagonal matrices J_n (3.3.1.3) have n real and distinct eigenvalues.

We present a few prerequisite definitions from linear algebra.

Definition 3.3.1.11 (Unitary Matrix)**.** *A square matrix U is said to be* **unitary** *if*

$$U^H U = I,$$

where $U^H := (\overline{U})^\top$, the conjugate transpose of U.

Definition 3.3.1.12 (Similar Matrices)**.** *Let A, B be square matrices. If there exists a nonsingular matrix T such that*

$$T^{-1}AT = B,$$

then we say that A is **similar** *to B, and write $A \sim B$.*

Definition 3.3.1.13 (Unitarily Similar)**.** *Let A be similar to B,*

$$T^{-1}AT = B.$$

If T is a unitary matrix, then we say that A and B are **unitarily similar***.*

We recall the following important result from linear algebra.

Lemma 3.3.1.14. *If A and B are similar matrices, then the eigenvalues of A are precisely the eigenvalues of B.*

Proof. Let A and B be similar matrices. Then there exists T nonsingular such that

$$T^{-1}AT = B.$$

Recalling that $\det(AB) = \det(A)\det(B)$, it follows for all $\lambda \in \mathbb{C}$ that

$$\begin{aligned}
\det(B - \lambda I) &= \det(T^{-1}AT - \lambda I) \\
&= \det(T^{-1}AT - \lambda T^{-1}T) \\
&= \det(T^{-1}(AT - \lambda T)) \\
&= \det(T^{-1}(A - \lambda I)T) \\
&= \det(T^{-1})\det(A - \lambda I)\det(T) \\
&= \det(T^{-1}T)\det(A - \lambda I) \\
&= \det(A - \lambda I).
\end{aligned}$$

□

Theorem 3.3.1.15 (Schur Normal Form)**.** *For every $n \times n$ matrix A, there exists a unitary $n \times n$ matrix U such that*

$$U^H AU = \begin{bmatrix} \lambda_1 & * & \dots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda_n \end{bmatrix},$$

where λ_j, $j = 1, 2, \dots, n$ are the eigenvalues of A.

We call the form of the matrix $U^H AU$ in (3.3.1.15) the *Schur Normal Form* (also *Schur Canonical Form*).

Proof. We use induction.

For the case $n = 1$, we choose $U := [1]$ and we're done.

For the induction hypothesis, assume that (3.3.1.15) holds for matrices up to size $(n-1) \times (n-1)$ for some integer $n > 1$. Let A be $n \times n$. Further, let λ_1 be an eigenvalue for A with associated eigenvector $\vec{x}_1 \neq \vec{0}$. We may rescale $\vec{x}_1$ as needed, so that $\vec{x}_1^H \vec{x}_1 = 1$.

We apply the Gram–Schmidt orthogonalization process to generate vectors $\vec{x}_2, \vec{x}_3, \dots, \vec{x}_n$ that form an orthonormal basis for $\mathbb{C}^n$. Then the matrix

$$X := [\vec{x}_1 \ \vec{x}_2 \ \dots \ \vec{x}_n]$$

is an $n \times n$ unitary matrix, for $X^H X$ has entries

$$\vec{x}_i^H \vec{x}_j = \begin{cases} 1, & i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Denote by

$$\vec{e}_j := [0, \ldots, \underbrace{0, 1, 0}_{j-\text{th entry}}, \ldots, 0]^\top, \quad j = 1, 2, \ldots, n$$

the standard basis vectors for $\mathbb{C}^n$. Further, given a matrix B, denote by $\vec{B}_j$ the $j-$th column of B. Note that evidently $B\vec{e}_j = \vec{B}_j$.

Now the first column of $X^H AX$ is

$$(X^H AX)\vec{e}_1 = X^H A(X\vec{e}_1) = X^H A\vec{x}_1 = X^H \lambda_1 \vec{x}_1 = \lambda_1 X^H \vec{x}_1 = \lambda_1 \vec{e}_1,$$

since $\vec{x}_1$ is an eigenvector of A and X is unitary. It follows

$$X^H AX = \left[\begin{array}{c|c} \lambda_1 & \vec{a}^H \\ \hline 0 & \\ \vdots & A_1 \\ 0 & \end{array}\right],$$

where A_1 is an $(n-1) \times (n-1)$ matrix and $\vec{a} \in \mathbb{C}^{n-1}$. By the induction hypothesis, there exists an $(n-1) \times (n-1)$ unitary matrix U_1 such that

$$U_1^H A_1 U_1 = \begin{bmatrix} \lambda_2 & * & \ldots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix}.$$

Define the $n \times n$ matrix

$$U := X \left[\begin{array}{c|ccc} 1 & 0 & \ldots & 0 \\ \hline 0 & & & \\ \vdots & & U_1 & \\ 0 & & & \end{array}\right].$$

Then U is unitary, for

$$\begin{aligned} U^H U &= \left[\begin{array}{c|ccc} 1 & 0 & \ldots & 0 \\ \hline 0 & & & \\ \vdots & & U_1^H & \\ 0 & & & \end{array}\right] X^H X \left[\begin{array}{c|ccc} 1 & 0 & \ldots & 0 \\ \hline 0 & & & \\ \vdots & & U_1 & \\ 0 & & & \end{array}\right] \\ &= \left[\begin{array}{c|ccc} 1 & 0 & \ldots & 0 \\ \hline 0 & & & \\ \vdots & & U_1^H & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} 1 & 0 & \ldots & 0 \\ \hline 0 & & & \\ \vdots & & U_1 & \\ 0 & & & \end{array}\right] \\ &= \left[\begin{array}{c|ccc} 1 & 0 & \ldots & 0 \\ \hline 0 & & & \\ \vdots & & U_1^H U_1 & \\ 0 & & & \end{array}\right] = I_n, \end{aligned}$$

since X and U_1 are both unitary. Moreover,

$$
\begin{aligned}
U^H A U &= \left[\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & U_1^H & \\ 0 & & & \end{array}\right] X^H A X \left[\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & U_1 & \\ 0 & & & \end{array}\right] \\
&= \left[\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & U_1^H & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} \lambda_1 & & \vec{a}^H & \\ \hline 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & U_1 & \\ 0 & & & \end{array}\right] \\
&= \left[\begin{array}{c|ccc} \lambda_1 & & \vec{a}^H & \\ \hline 0 & & & \\ \vdots & & U_1^H A_1 & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & U_1 & \\ 0 & & & \end{array}\right] \\
&= \left[\begin{array}{c|ccc} \lambda_1 & & \vec{a}^H U_1 & \\ \hline 0 & & & \\ \vdots & & U_1^H A_1 U_1 & \\ 0 & & & \end{array}\right] \\
&= \left[\begin{array}{c|ccc} \lambda_1 & \multicolumn{3}{c}{\vec{a}^H U_1} \\ \hline 0 & \lambda_2 & * & \dots & * \\ \vdots & 0 & \lambda_3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & * \\ 0 & 0 & \dots & 0 & \lambda_n \end{array}\right] \\
&= \begin{bmatrix} \lambda_1 & * & \dots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda_n \end{bmatrix},
\end{aligned}
$$

which completes the proof. □

Definition 3.3.1.16 (Hermitian Matrix)**.** *A square matrix A is said to be* **Hermitian** *if*

$$A^H = A.$$

Theorem 3.3.1.17. *For every $n \times n$ Hermitian matrix $A = A^H$, there exists a unitary matrix*

$$U = [\vec{x}_1 \ \vec{x}_2 \ \dots \ \vec{x}_n]$$

with $U^H A U = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$. Moreover,

(1) The eigenvalues λ_j, $j = 1, 2, \dots, n$ of A are all real–valued;
(2) $A\vec{x}_j = \lambda_j$, that is, the columns of U are the eigenvectors of A;
(3) Since U is unitary,

$$\vec{x}_i^H \vec{x}_j = \begin{cases} 1, & i = j, \\ 0, & otherwise. \end{cases}$$

Proof. Let A be an $n \times n$ Hermitian matrix. By (3.3.1.15), there exists an $n \times n$ unitary matrix U such that

$$U^H AU = \begin{bmatrix} \lambda_1 & * & \dots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}.$$

But since A is Hermitian,

$$(U^H AU)^H = U^H A^H (U^H)^H = U^H A^H U = U^H AU,$$

so that

$$\begin{aligned} (U^H AU)^H &= \begin{bmatrix} \lambda_1 & * & \dots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}^H \\ &= \begin{bmatrix} \bar{\lambda}_1 & 0 & \dots & 0 \\ * & \bar{\lambda}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \dots & * & \bar{\lambda}_n \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & * & \dots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}, \end{aligned}$$

from which it follows that all off–diagonal entries are zeros. It also immediately follows $\lambda_j = \bar{\lambda}_j$ for each $j = 1, 2, \dots, n$, so that λ_j, $j = 1, 2, \dots, n$ are real–valued.

Finally,

$$U^H AU = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix},$$

so that

$$AU = (U^H)^{-1} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix} = U \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}.$$

Hence,

$$[A\vec{x}_1 \ A\vec{x}_2 \ \dots \ A\vec{x}_n] = [\lambda_1 \vec{x}_1 \ \lambda_2 \vec{x}_2 \ \dots \ \lambda_n \vec{x}_n],$$

and thus it follows $A\vec{x}_j = \lambda_j \vec{x}_j$, $j = 1, 2, \dots, n$. □

Note by definition that if A is a symmetric ($A = A^\top$) real–valued matrix, then consequently $A = A^H$, so that A is Hermitian. Also recall that the tridiagonal matrices J_n

(3.3.1.3) are symmetric and real–valued. Thus by Theorems (3.3.1.10) and (3.3.1.17) there exists a unitary matrix

$$U = [\vec{u}_1 \ \vec{u}_2 \ \dots \ \vec{u}_n]$$

with $U^H J_n U = \text{diag}(x_1, x_2, \dots, x_n)$, where the x_j, $j = 1, 2, \dots, n$ are the roots of the $n-$th $w-$orthogonal polynomial ϕ_n, and consequently are real and distinct. The eigenvectors $\{\vec{u}_j\}_{j=1}^n$ are orthogonal here, and we rescale these to obtain the quadrature weights w_i.

Theorem 3.3.1.18 (Weights for Gaussian Quadrature)**.** *Let $\vec{u}^{(i)} := [\vec{u}_1^{(i)}, \vec{u}_2^{(i)}, \dots, \vec{u}_n^{(i)}]^\top$ be an eigenvector of J_n for the eigenvalue x_i, $i = 1, 2, \dots, n$. Suppose that $\vec{u}^{(i)}$ is scaled in such a way that*

$$\left\|\vec{u}^{(i)}\right\|_2^2 = \left(\vec{u}^{(i)}\right)^H \vec{u}^{(i)} = \sum_{k=1}^{n} \left(\vec{u}_k^{(i)}\right)^2 = \int_a^b w(x)\ dx.$$

Then the weights w_i, $i = 1, 2, \dots, n$ of the $n-$point Gaussian quadrature rule are given by

$$w_i = (\vec{u}_1^{(i)})^2, \quad i = 1, 2, \dots, n.$$

Proof. We first verify that the vector

$$\tilde{u}^{(i)} := [\rho_0\phi_0(x_i) \ \rho_1\phi_1(x_i) \ \dots \ \rho_{n-1}\phi_{n-1}(x_i)]^\top$$

where

$$\rho_j := \frac{1}{C_1 C_2 \dots C_j C_{j+1}},$$

$j = 0, 1, \dots, n-1$ is an eigenvector of J_n for the eigenvalue x_i. By the recursion for ϕ_j, the first row of $J_n\tilde{u}^{(i)}$ is, for any x,

$$\begin{aligned} B_1\rho_0\phi_0(x) + C_2\rho_1\phi_1(x) &= \frac{B_1}{C_1}\phi_0(x) + \frac{C_2}{C_1C_2}\phi_1(x) \\ &= B_1 + \phi_1(x) \\ &= x \\ &= x\rho_0\phi_0(x). \end{aligned}$$

Similarly, for $j = 2, \dots, n-1$, the $j-$th entry of $J_n\tilde{u}^{(i)}$ is

$$\begin{aligned} &C_j\rho_{j-2}\phi_{j-2}(x) + B_j\rho_{j-1}\phi_{j-1}(x) + C_{j+1}\rho_j\phi_j(x) \\ &\quad = C_j(C_j\rho_{j-1})\phi_{j-2}(x) + B_j\rho_{j-1}\phi_{j-1}(x) + \rho_{j-1}\phi_j(x) \\ &\quad = \rho_{j-1}[C_j^2\phi_{j-2}(x) + B_j\phi_{j-1}(x) + \phi_j(x)] \\ &\quad = \rho_{j-1}[C_j^2\phi_{j-2}(x) + B_j\phi_{j-1}(x) + ((x - B_j)\phi_{j-1}(x) - C_j^2\phi_{j-2}(x))] \\ &\quad = x\rho_{j-1}\phi_{j-1}(x). \end{aligned}$$

Finally, the last entry of $J_n\tilde{u}^{(i)}$ is

$$\rho_{n-1}[C_n^2\phi_{n-2}(x) + B_n\phi_{n-1}(x)] = x\rho_{n-1}\phi_{n-1}(x) - \rho_{n-1}\phi_n(x),$$

so that

$$x_i\rho_{n-1}\phi_{n-1}(x_i) - \rho_{n-1}\phi_n(x_i) = x_i\rho_{n-1}\phi_{n-1}(x_i).$$

Thus $J_n\tilde{u}^{(i)} = x_i\tilde{u}^{(i)}$ for each $i = 1, 2, \dots, n$. This shows that $\tilde{u}^{(i)}$ is an eigenvector of J_n.

Recall from (3.3.1.1) that

$$\sum_{i=1}^{n} w_i \phi_k(x_i) = \begin{cases} \int_a^b w(x)\ dx, & k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Define $\vec{w} := [w_1, w_2, \ldots, w_n]^\top$ and

$$U := [\tilde{u}^{(1)} \mid \tilde{u}^{(2)} \mid \ldots \mid \tilde{u}^{(n)}] = \begin{bmatrix} \rho_0\phi_0(x_1) & \rho_0\phi_0(x_2) & \ldots & \rho_0\phi_0(x_n) \\ \rho_1\phi_1(x_1) & \rho_1\phi_1(x_2) & \ldots & \rho_1\phi_1(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1}\phi_{n-1}(x_1) & \rho_{n-1}\phi_{n-1}(x_2) & \ldots & \rho_{n-1}\phi_{n-1}(x_n) \end{bmatrix}.$$

For each $k = 1, 2, \ldots, n$, row k of U is given by

$$[\rho_{k-1}\phi(x_1)\ \rho_{k-1}\phi_{k-1}(x_2)\ \ldots\ \rho_{k-1}\phi_{k-1}(x_n)] = \rho_{k-1}[\phi_{k-1}(x_1)\ \phi_{k-1}(x_2)\ \ldots\ \phi_{k-1}(x_n)].$$

Thus, the $k-$th entry of $U\vec{w}$ is

$$(U\vec{w})_k = \rho_{k-1} \sum_{i=1}^{n} w_i \phi_{k-1}(x_i).$$

We now solve for the weights w_i, $i = 1, 2, \ldots, n$ by observing that

$$U\vec{w} = \left[\int_a^b w(x)\ dx, 0, \ldots, 0\right]^\top,$$

since $\rho_0 = 1$. Since the eigenvectors of J_n are orthogonal, we have

$$\begin{aligned} \left(\tilde{u}^{(i)}\right)^\top U\vec{w} &= \left(\tilde{u}^{(i)}\right)^\top \tilde{u}^{(i)} w_i \\ &= \left(\int_a^b w(x)\ dx\right) \left(\tilde{u}^{(i)}\right)^\top \vec{e}_1 \\ &= \left(\int_a^b w(x)\ dx\right) (\rho_0\phi_0(x_i)) \\ &= \int_a^b w(x)\ dx. \end{aligned}$$

By the hypothesis,

$$\left(\vec{u}^{(i)}\right)^\top \left(\vec{u}^{(i)}\right) = \int_a^b w(x)\ dx,$$

and we have already shown that $\vec{u}^{(i)}$ and $\tilde{u}^{(i)}$ are both eigenvectors of J_n with the associated eigenvalue x_i. Since there are n distinct eigenvalues, $\tilde{u}^{(i)}$ is a multiple of $\vec{u}^{(i)}$. the first entry of $\tilde{u}^{(i)}$ is $\rho_0\phi_0(x) = 1$, so that

$$\vec{u}^{(i)} = \vec{u}_1^{(i)} \tilde{u}^{(i)}.$$

Finally, observe that

$$\begin{aligned} w_i &= \frac{1}{(\tilde{u}^{(i)})^\top (\tilde{u}^{(i)})} \int_a^b w(x)\ dx \\ &= \frac{(\vec{u}^{(i)})^\top (\vec{u}^{(i)})}{(\tilde{u}^{(i)})^\top (\tilde{u}^{(i)})} \end{aligned}$$

$$= \left(\vec{u}_1^{(i)}\right)^2.$$

This completes the proof. □

Remark. *Suppose that $\{\vec{u}_j\}_{j=1}^n$ is a set of eigenvectors for J_n. Recall that we want to construct a set $\{\tilde{u}_j\}_{j=1}^n$ of eigenvectors such that*

$$\|\tilde{u}_j\|_2^2 = \int_a^b w(x)\ dx$$

for each $j = 1, 2, \ldots, n$. We may write

$$\tilde{u}_j = k_j \vec{u}_j$$

for some $k_j \in \mathbb{R}$, so that then

$$\|\tilde{u}_j\|_2^2 = \|k_j \vec{u}_j\|_2^2 = k_j^2 \|\vec{u}_j\|_2^2,$$

which implies that taking

$$k_j^2 := \frac{\int_a^b w(x)\ dx}{\|\vec{u}_j\|_2^2}$$

will scale the eigenvectors as needed.

Example 3.3.1.19. *We derive the abscissa x_i and weights w_i for $w(x) := 1$ on the interval $[-1, 1]$.*

Note that the first three orthogonal polynomials are

$$\phi_0(x) = 1,$$

$$\phi_1(x) = x - \frac{\int_{-1}^1 x\ dx}{\int_{-1}^1\ dx} = x,$$

$$\phi_2(x) = \left(x - \frac{\int_{-1}^1 x^3\ dx}{\int_{-1}^1 x^2\ dx}\right) x - \frac{\int_{-1}^1 x^2\ dx}{\int_{-1}^1\ dx} = x^2 - \frac{\left[\frac{1}{3}x^3\right]_{-1}^1}{x|_{-1}^1} = x^2 - \frac{1}{3} = (x-0)\phi_1(x) - \frac{1}{3}\phi_0(x).$$

Thus the zeros of ϕ_2 and thus the abscissas x_i, $i = 1, 2$ are

$$x_1 := -\frac{1}{\sqrt{3}}, \quad x_2 := \frac{1}{\sqrt{3}}.$$

Moreover, the matrix J_2 is given by

$$J_2 = \begin{bmatrix} B_1 & C_2 \\ C_2 & B_2 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & 0 \end{bmatrix}.$$

Finding the eigenvectors of J_2, we obtain

$$J_2 - I_2 x_1 = J_2 + \frac{1}{\sqrt{3}} I_2 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \implies \vec{u}_1 = k_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

and

$$J_2 - I_2 x_2 = J_2 - \frac{1}{\sqrt{3}} I_2 = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix} \implies \vec{u}_2 = k_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Noting that $\int_{-1}^{1} w(x)\ dx = \int_{-1}^{1} dx = 2$, *choosing* $k_1 = k_2 = 1$ *scales the eigenvectors as needed. Hence,*

$$w_1 = \left(\vec{u}_1^{(1)}\right)^2 = 1,$$
$$w_2 = \left(\vec{u}_1^{(2)}\right)^2 = 1.$$

The quadrature rule is

$$\tilde{I}(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Example 3.3.1.20. *We derive a quadrature rule* $\tilde{I}(f)$ *that will integrate*

$$I(f) = \int_{-1}^{1} x^2 f(x)\ dx$$

exactly whenever f *is a polynomial of degree* 2 *or less.*

The first three orthogonal polynomials are

$$\phi_0(x) = 1,$$
$$\phi_1(x) = x - \frac{\int_{-1}^{1} x^3\ dx}{\int_{-1}^{1} x^2\ dx} = x,$$
$$\phi_2(x) = \left(x - \frac{\int_{-1}^{1} x^5\ dx}{\int_{-1}^{1} x^4\ dx}\right) x - \frac{\int_{-1}^{1} x^4\ dx}{\int_{-1}^{1} x^2\ dx} = x^2 - \frac{2/5}{2/3} = x^2 - \frac{3}{5} = (x-0)\phi_1(x) - \frac{3}{5}\phi_0(x).$$

Thus the zeros of ϕ_2 *and therefore the quadrature abscissas* x_i, $i = 1, 2$ *are*

$$x_1 := -\sqrt{\frac{3}{5}}, \quad x_2 := \sqrt{\frac{3}{5}}.$$

Further, the matrix J_2 *is*

$$J_2 = \begin{bmatrix} B_1 & C_2 \\ C_2 & B_2 \end{bmatrix} = \begin{bmatrix} 0 & \sqrt{\frac{3}{5}} \\ \sqrt{\frac{3}{5}} & 0 \end{bmatrix}.$$

Finding the eigenvectors of J_2, *we obtain*

$$J_2 - I_2 x_1 = J_2 + \sqrt{\frac{3}{5}} I_2 = \begin{bmatrix} \sqrt{\frac{3}{5}} & \sqrt{\frac{3}{5}} \\ \sqrt{\frac{3}{5}} & \sqrt{\frac{3}{5}} \end{bmatrix} \implies \vec{u}_1 := k_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

and

$$J_2 - I_2 x_2 = J_2 - \sqrt{\frac{3}{5}} I_2 = \begin{bmatrix} -\sqrt{\frac{3}{5}} & \sqrt{\frac{3}{5}} \\ \sqrt{\frac{3}{5}} & -\sqrt{\frac{3}{5}} \end{bmatrix} \implies \vec{u}_2 := k_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Note that $\|u_j\|_2^2 = 2$ *for* $j = 1, 2$. *Moreover,*

$$\int_{-1}^{1} w(x)\ dx = \int_{-1}^{1} x^2\ dx = \left[\frac{1}{3}x^3\right] = \frac{2}{3}.$$

Put

$$k := \sqrt{\frac{2/3}{2}} = \sqrt{\frac{1}{3}}$$

and note $\tilde{u}_j := k\vec{u}_j$ *is the required basis. We obtain*

$$w_1 = w_2 = \left(\tilde{u}_1^{(1)}\right)^2 = \frac{1}{3}.$$

Hence, the quadrature rule is

$$\tilde{I}(f) = \frac{1}{3} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{1}{3} f\left(\sqrt{\frac{3}{5}}\right).$$

Theorem 3.3.1.21 (Error in Gaussian Quadrature)**.** *If* $f \in C^{2n}[a,b]$, *then*

$$\int_a^b w(x) f(x)\, dx - \sum_{i=1}^{n} w_i f(x_i) = \frac{f^{(2n)}(\xi)}{(2n!)} \int_a^b w(x)[\phi_n(x)]^2\, dx,$$

for some $\xi \in (a,b)$.

Proof. Let $p \in \Pi^{2n-1}$ be the unique Hermite interpolating polynomial satisfying

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad i = 1, 2, \ldots, n.$$

Since the Gaussian quadrature rule is exact on Π^{2n-1}, we have

$$\int_a^b w(x)p(x)\, dx = \sum_{i=1}^{n} w_i p(x_i) = \sum_{i=1}^{n} w_i f(x_i).$$

Therefore, the error term has the integral representation

$$\begin{aligned} I(f) - \tilde{I}(f) &= \int_a^b w(x)f(x)\, dx - \sum_{i=1}^{n} w_i f(x_i) \\ &= \int_a^b w(x)f(x)\, dx - \int_a^b w(x)p(x)\, dx \\ &= \int_a^b w(x)(f(x) - p(x))\, dx. \end{aligned}$$

Since the x_i, $i = 1, 2, \ldots, n$ are the roots of ϕ_n, it follows from the error in Hermite interpolation (1.1.5.7) that there exists $\zeta \in (a,b)$ such that

$$f(x) - p(x) = \frac{f^{(2n)}(\zeta)}{(2n)!} \prod_{i=1}^{n} (x - x_i)^2 = \frac{f^{(2n)}(\zeta)}{(2n)!} [\phi_n(x)]^2.$$

Next, the function

$$\frac{f^{(2n)}(\zeta(x))}{(2n)!} = \frac{f(x) - p(x)}{[\phi_n(x)]^2}$$

is continuous on $[a,b]$. Since $w \geq 0$ on $[a,b]$, by the weighted mean value theorem for integrals, it follows

$$I(f) - \tilde{I}(f) = \int_a^b w(x)(f(x) - p(x))\, dx$$

$$= \frac{1}{(2n)!} \int_a^b w(x) f^{(2n)}(\zeta) [\phi_n(x)]^2 \, dx$$
$$= \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b w(x) [\phi_n(x)]^2 \, dx$$

for some $\xi \in (a, b)$. This completes the proof. □

Example 3.3.1.22. *In the case*

$$f(x) = x^4,$$

we derive the explicit formula for the quadrature error $I(f) - \tilde{I}(f)$ from the first example.

Recall $n = 2$, $w(x) \equiv 1$, and $\phi_2(x) = x^2 - \frac{1}{3}$. Hence, applying (3.3.1.21)*, we have*

$$\begin{aligned} I(f) - \tilde{I}(f) &= \frac{4!}{4!} \int_{-1}^{1} \left(x^2 - \frac{1}{3} \right)^2 \, dx \\ &= \int_{-1}^{1} x^4 - \frac{2}{3} x^2 + \frac{1}{3} \, dx \\ &= \left[\frac{1}{5} x^5 - \frac{2}{9} x^3 + \frac{1}{3} x \right]_{-1}^{1} \\ &= \frac{2}{5} - \frac{4}{9} + \frac{2}{3} = \frac{28}{45}. \end{aligned}$$

4. Systems of Linear Equations

In this section we consider direct methods for solving systems of linear equations

$$\mathbf{A}\vec{x} = \vec{b}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{n1} \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

Here $\mathbf{A}$ is a given square $n \times n$ matrix and $\vec{b}$ a given vector. The direct methods discussed in this section produce the solution to the system $\mathbf{A}\vec{x} = \vec{b}$ in finitely many steps, assuming computations without roundoff errors.

This problem is closely related to that of computing the inverse $\mathbf{A}^{-1}$ of the matrix $\mathbf{A}$ provided that this inverse exists. For if $\mathbf{A}^{-1}$ is known, the solution $\vec{x}$ of $\mathbf{A}\vec{x} = \vec{b}$ can be obtained by matrix–vector multiplication,

$$\vec{x} = \mathbf{A}^{-1}\vec{b}.$$

Conversely, the $i-$th column $\bar{a}_i$ of $\mathbf{A}^{-1} = [\bar{a}_1, \dots, \bar{a}_n]^\top$ is the solution of the linear system $\mathbf{A}\vec{x} = \vec{e}_i$, where $\vec{e}_i = [0, \dots, 0, 1, 0, \dots, 0]^\top$ is the $i-$th unit vector.

4.1. Gaussian Elimination: The Triangular Decomposition of a Matrix.

4.1.1. *Gaussian Elimination.* We seek a solution to a system of linear equations

$$\mathbf{A}\vec{x} = \vec{b}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n. \end{bmatrix} \tag{4.1.1.1}$$

Here, $\mathbf{A}$ is a square $n \times n$ matrix and $\vec{b} \in \mathbb{R}^n$. The system (4.1.1.1) is transformed by rearrangements and linear combinations into a system of the form

$$\mathbf{R}\vec{x} = \vec{c}, \quad \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{nn} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & r_{nn} \end{bmatrix},$$

which has the same solution $\vec{x}$ as $\mathbf{A}\vec{x} = \vec{b}$. Here, $\mathbf{R}$ is an upper triangular matrix, so we can solve $\mathbf{R}\vec{x} = \vec{c}$ easily by back substitution

$$x_i := \frac{\left(c_i - \sum_{k=i+1}^{n} r_{ik}x_k\right)}{r_{ii}}, \quad i = n, n-1, \dots, 1.$$

In the first step of the algorithm we subtract a multiple of the first equation from all other equations so that the coefficients of x_1 vanish in these equations. Thus x_1 remains only in the first equation, which is possible only if $a_{11} \neq 0$, which can be achieved by swapping rows as necessary, so long as at least one $a_{i1} \neq 0$. The operations are carried out on the matrix

$$(\mathbf{A}, \vec{b}) = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{bmatrix}.$$

The first step of the Gaussian elimination process leads to a matrix $(\mathbf{A}', \vec{b}')$ of the form

$$(\mathbf{A}', \vec{b}') = \begin{bmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} & b'_1 \\ 0 & a'_{22} & \dots & a'_{2n} & b'_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a'_{n2} & \dots & a'_{nn} & b'_n \end{bmatrix}.$$

We may describe this step formally as follows:

Algorithm 4.1.1.1.

(1) Determine an element $a_{r1} \neq 0$ and proceed with (2). If no such r exists, then A is singular. Set $(\mathbf{A}', \vec{b}') = (\mathbf{A}, \vec{b})$ and stop.

(2) Interchange rows r and 1 of $(\mathbf{A}, \vec{b})$. The result is the matrix $(\bar{\mathbf{A}}, \bar{\vec{b}})$.

(3) For $i = 2, 3, \dots, n$, subtract the multiple

$$l_{i1} := \frac{\bar{a}_{i1}}{\bar{a}_{11}}$$

of row 1 from row i of the matrix $(\bar{\mathbf{A}}, \bar{\vec{b}})$. The desired matrix $(\mathbf{A}', \vec{b}')$ is obtained as the result.

The transition $(\mathbf{A}, \vec{b}) \to (\bar{\mathbf{A}}, \bar{\vec{b}}) \to (\mathbf{A}', \vec{b}')$ can be described by using matrix multiplications

$$(\bar{\mathbf{A}}, \bar{\vec{b}}) = P_1(\mathbf{A}, \vec{b}), \quad (\mathbf{A}', \vec{b}') = G_1(\bar{\mathbf{A}}, \bar{\vec{b}}) = G_1 P_1(\mathbf{A}, \vec{b}), \tag{4.1.1.2}$$

where P_1 is a permutation matrix

$$P_1 := \begin{bmatrix} 0 & & & 0 & 1 & 0 & & 0 \\ & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & 1 & & & & \\ 1 & & & & 0 & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ 0 & & & & & & & 1 \end{bmatrix}$$

and G_1 is a lower triangular matrix

$$G_1 := \begin{bmatrix} 1 & 0 & \dots & 0 \\ -l_{21} & 1 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ -l_{n_1} & 0 & \dots & 1 \end{bmatrix}.$$

Matrices such as G_1 that differ in at most one column from an identity matrix are called *Frobenius matrices*. Both matrices P_1 and G_1 are nonsingular:

$$P_1^{-1} = P_1, \quad G_1^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & 0 & \dots & 1 \end{bmatrix}.$$

Thus, the equation systems $\mathbf{A}\vec{x} = \vec{b}$ and $\mathbf{A}'\vec{x} = \vec{b}'$ have the same solution $\vec{x}$:

$$\mathbf{A}\vec{x} = \vec{b} \implies \mathbf{A}'\vec{x} = G_1 P_1 \mathbf{A}\vec{x} = G_1 P_1 \vec{b} = \vec{b}',$$
$$\mathbf{A}'\vec{x} = \vec{b}' \implies \mathbf{A}\vec{x} = P_1^{-1} G_1^{-1} \mathbf{A}'\vec{x} = P_1^{-1} G_1^{-1} \vec{b}' = \vec{b}.$$

The element $a_{r1} = \bar{a}_{11}$ determined in (4.1.1.1)[1] is called the *pivot element*, and step (1) is called *pivot selection*. Note that in pivot selection we may choose any $a_{r1} \neq 0$ as the pivot. For reasons of numerical stability, usually the choice

$$|a_{r1}| = \max_i |a_{i1}|$$

is made. It is assumed in making this choice that the orders of magnitudes of the elements of $\mathbf{A}$ are roughly equal (in this situation $\mathbf{A}$ is said to be *equilibrated*). This sort of pivot selection is called *partial pivot selection*.

We replace (1) and (2) in (4.1.1.1) as follows:

Algorithm 4.1.1.2.

(1) Determine r so that

$$|a_{r1}| = \max_i |a_{i1}|$$

and continue with (2) if $a_{r1} \neq 0$. Otherwise, $\mathbf{A}$ is singular; set $(\mathbf{A}', \vec{b}') = (\mathbf{A}, \vec{b})$, stop.

(2) Interchange rows 1 and r of $(\mathbf{A}, \vec{b})$. Let the resulting matrix be $(\bar{\mathbf{A}}, \bar{\vec{b}})$.

After the first elimination step, the resulting matrix has the form

$$(\mathbf{A}', \vec{b}') = \left[\begin{array}{c|c|c} a'_{11} & a'^{\top} & b'_1 \\ \hline 0 & \tilde{\mathbf{A}} & \tilde{b} \end{array}\right]$$

with an $(n-1)-$row matrix $\tilde{\mathbf{A}}$. The next elimination step consists of simply applying the same algorithm to the smaller matrix $(\tilde{\mathbf{A}}, \tilde{b})$. Carrying on in this fashion, a sequence of matrices

$$(\mathbf{A}, \vec{b}) := (\mathbf{A}^{(0)}, \vec{b}^{(0)}) \to (\mathbf{A}^{(1)}, \vec{b}^{(1)}) \to \cdots \to (\mathbf{A}^{(n-1)}, \vec{b}^{(n-1)}) =: (\mathbf{R}, \vec{c})$$

is obtained which begins with the given matrix $(\mathbf{A}, \vec{b})$ and ends with the desired matrix $(\mathbf{R}, \vec{c})$. In this sequence the $j-$th intermediate matrix has the form

$$(\mathbf{A}^{(j)}, \vec{b}^{(j)}) = \left[\begin{array}{ccc|ccc|c} * & \dots & * & * & \dots & * & * \\ & \ddots & \vdots & \vdots & & \vdots & \vdots \\ & & & & & & \\ 0 & & * & * & \ddots & * & * \\ \hline 0 & \dots & 0 & * & \dots & * & * \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & * & \dots & * & * \end{array}\right] = \left[\begin{array}{c|c|c} \mathbf{A}_{11}^{(j)} & \mathbf{A}_{12}^{(j)} & \vec{b}_1^{(j)} \\ \hline 0 & \mathbf{A}_{22}^{(j)} & \vec{b}_2^{(j)} \end{array}\right]$$

with a $j-$row upper triangular matrix $\mathbf{A}_{11}^{(j)}$. The matrix $(\mathbf{A}^{(j)}, \vec{b}^{(j)})$ is obtained from $(\mathbf{A}^{(j-1)}, \vec{b}^{(j-1)})$ by applying the elimination algorithm (4.1.1.1) on the $(n-j+1) \times (n-j+2)$ matrix $(\mathbf{A}_{22}^{(j-1)}, \vec{b}_2^{(j-1)})$. The elements of $\mathbf{A}_{11}^{(j)}$, $\mathbf{A}_{12}^{(j)}$, and $\vec{b}_1^{(j)}$ do not change from this step on, and thus they agree with the corresponding elements of $(\mathbf{R}, \vec{c})$. Moreover, the ensuing steps can be described using matrix multiplication. That is,

$$(\mathbf{A}^{(j)}, \vec{b}^{(j)}) = G_j P_j (\mathbf{A}^{(j-1)}, \vec{b}^{(j-1)}),$$

$$(\mathbf{R}, \vec{c}) = G_{n-1}P_{n-1}G_{n-2}P_{n-2}\dots G_1P_1(\mathbf{A}, \vec{b}),$$

with permutation matrices P_j and nonsingular Frobenius matrices G_j, $j = 1, 2, \dots, n-1$ of the form

$$G_j = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{j+1,j} & 1 & & \\ & & \vdots & \vdots & \ddots & \\ 0 & & -l_{n,j} & 0 & & 1 \end{bmatrix}.$$

In the $j-$th elimination step $(\mathbf{A}^{(j-1)}, \vec{b}^{(j-1)}) \to (\mathbf{A}^{(j)}, \vec{b}^{(j)})$, the elements below the diagonal in the $j-$th column vanish. For implementation of this algorithm on a computer, the locations which were once occupied by these elements may be used for the storage of the quantities l_{ij}, $i = j+1, j+2, \dots, n$, of G_j, that is, we work with a matrix of the form

$$T^{(j)} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1j} & r_{1,j+1} & \dots & r_{1n} & c_1 \\ \lambda_{21} & r_{22} & \dots & r_{2j} & r_{2,j+1} & & r_{2n} & c_2 \\ \lambda_{31} & \lambda_{32} & & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & r_{jj} & r_{j,j+1} & \dots & r_{j,n} & c_j \\ \vdots & \vdots & & \lambda_{j+1,j} & a^{(j)}_{j+1,j+1} & \dots & a^{(j)}_{j+1,n} & b^{(j)}_{j+1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{n1} & a^{(j)}_{n,j+1} & \dots & a^{(j)}_{n,n} & b^{(j)}_n \end{bmatrix}.$$

Here, the subdiagonal elements $\lambda_{k+1,k}, \lambda_{k+2,k}, \dots, \lambda_{nk}$ of the $k-$th column are a certain permutation of the elements $l_{k+1,k}, l_{k+2,k}, \dots, l_{nk}$ in G_k.

The $j-$th step $T^{(j-1)} \to T^{(j)}$, $j = 1, 2, \dots, n-1$ can be described as follows, where the elements of $T^{(j-1)}$ are denoted by t_{ik}, and those of $T^{(j)}$ by t'_{ik}, $i = 1, 2, \dots, n$, $k = 1, 2, \dots, n+1$:

Algorithm 4.1.1.3.

(1) Partial pivot selection: Determine r so that

$$|t_{rj}| = \max_{i \geq j} |t_{ij}|.$$

If $t_{rj} = 0$, set $T^{(j)} := T^{(j-1)}$; $\mathbf{A}$ is singular, stop. Otherwise, continue with (2).

(2) Interchange rows r and j of $T^{(j-1)}$, and denote the result by $\bar{T} = (\bar{t}_{ik})$.

(3) Replace

$$t'_{ij} := l_{ij} := \frac{\bar{t}_{ij}}{\bar{t}_{jj}}, \quad \text{for } i = j+1, j+2, \dots, n,$$
$$t'_{ik} := \bar{t}_{ik} - l_{ij}\bar{t}_{jk}, \quad \text{for } i = j+1, j+2, \dots, n \text{ and } k = j+1, j+2, \dots, n,$$
$$t'_{ik} := \bar{t}_{ik}, \quad \text{otherwise.}$$

We note that in (3) the elements $l_{j+1,j}, l_{j+2,j}, \dots, l_{nj}$ of G_j are store in their natural order as $t'_{j+1,j}, t'_{j+2,j}, \dots, t'_{nj}$. This order, however, may be changed in the subsequent elimination steps $T^{(k)} \to T^{(k+1)}$, $k \geq j$, because in (2) the rows of the entire matrix $T^{(k)}$ are rearranged.

This has the following effect: the lower triangular matrix L and the upper triangular matrix R,

$$L := \begin{bmatrix} 1 & 0 & \dots & 0 \\ t_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ t_{n1} & \dots & t_{n,n-1} & 1 \end{bmatrix}, \quad R := \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & & t_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & t_{nn} \end{bmatrix},$$

which are contained in the final matrix $T^{(n-1)} = (t_{ik})$, provide a triangular decomposition of the matrix $P\mathbf{A}$:

$$LR = P\mathbf{A}.$$

In this decomposition, P is the product of all of the permutations

$$P = P_{n-1}P_{n-2}\dots P_2P_1.$$

Example 4.1.1.4.

$$\begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix}.$$

$$\left[\begin{array}{ccc|c} 3^* & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{array}\right] \to \left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \frac{2}{3} & \frac{1}{3} & -1 & \frac{17}{3} \\ \frac{1}{3} & \frac{2}{3}^* & -1 & \frac{10}{3} \end{array}\right] \to \left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \frac{1}{3} & \frac{2}{3} & -1 & \frac{10}{3} \\ \frac{2}{3} & \frac{1}{3} & -1 & \frac{17}{3} \end{array}\right]$$
$$\to \left[\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \frac{1}{3} & \frac{2}{3} & -1 & \frac{10}{3} \\ \frac{2}{3} & \frac{1}{2} & -\frac{1}{2} & 4 \end{array}\right]$$

Thus the triangular equation system is

$$\begin{bmatrix} 3 & 1 & 6 \\ 0 & \frac{2}{3} & -1 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{10}{3} \\ 4 \end{bmatrix}.$$

Its solution is

$$\begin{aligned} x_3 &= -2(4) = -8, \\ x_2 &= \frac{3}{2}\left(\frac{10}{3} + x_3\right) = \frac{3}{2}\left(\frac{10}{3} - 8\right) = \frac{3}{2}\left(-\frac{14}{3}\right) = -7, \\ x_1 &= \frac{1}{3}(2 - 6x_3 - x_2) = \frac{1}{3}(2 - 6(-8) - (-7)) = \frac{1}{3}(57) = 19. \end{aligned}$$

Further

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad P\mathbf{A} = \begin{bmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix},$$

and the matrix $P\mathbf{A}$ has the triangular decomposition $P\mathbf{A} = LR$ with

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{2}{3} & \frac{1}{2} & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 3 & 1 & 6 \\ 0 & \frac{2}{3} & -1 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix}.$$

If the lower–upper triangular decomposition is known for a matrix $\mathbf{A}$, that is, if

$$P\mathbf{A} = LR$$

is known, then the equation system $\mathbf{A}\vec{x} = \vec{b}$ can be solved immediately with any right–hand side $\vec{b}$, for it follows that

$$P\mathbf{A}\vec{x} = LR\vec{x} = P\vec{b},$$

from which $\vec{x}$ can be found by solving both of the triangular systems

$$L(R\vec{x}) = L\vec{y} = P\vec{b} \quad \text{(using forward substitution)},$$
$$R\vec{x} = \vec{y} \quad \text{(using back substitution)}.$$

Thus, with the help fo the Gaussian elimination algorithm, it can be shown constructively that each square nonsingular matrix $\mathbf{A}$ has a triangular decomposition of the form

$$P\mathbf{A} = LR.$$

We also note that Gaussian elimination and direct triangular decomposition differ only in the ordering of operations. Both algorithms are, theoretically and numerically, entirely equivalent. In Gaussian elimination, the scalar products are formed only in pieces, with temporary storing of the intermediate results. Direct triangular decomposition forms each scalar product as a whole.

A second result pertains to the determinant of $\mathbf{A}$. Suppose that we are given a triangular decomposition

$$P\mathbf{A} = LR.$$

Note that $\det(P) = \pm 1$ and $\det(L) = 1$. Thus it follows

$$\begin{aligned}\det(\mathbf{A}) &= \pm\det(P\mathbf{A}) \\ &= \pm\det(LR) \\ &= \pm\det(L)\det(R) \\ &= \pm\det(R) \\ &= \pm\prod_{k=1}^{n} r_{kk}.\end{aligned}$$

Hence, we may get $\pm\det(A)$ by:

(1) Factor $P\mathbf{A} = LR$,
(2) Take $\prod_{k=1}^{n} r_{kk}$.

A further practical property of the method of triangular decomposition is that, for banded matrices with bandwidth m,

$$\mathbf{A} = \begin{bmatrix} * & \dots & * & 0 & \dots & 0 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ * & & \ddots & & \ddots & 0 \\ 0 & \ddots & & \ddots & & * \\ \vdots & \ddots & \ddots & & \ddots & \vdots \\ 0 & \dots & 0 & * & \dots & * \end{bmatrix}, \quad a_{ij} = 0, \text{ for } |i-j| \geq m,$$

the matrices L and R of the decomposition $P\mathbf{A} = LR$ are not full: R is a banded upper triangular matrix with bandwidth $2m-1$,

$$R = \begin{bmatrix} * & \dots & * & 0 & \dots & 0 \\ & \ddots & & \ddots & \ddots & \vdots \\ & & \ddots & & \ddots & 0 \\ & & & \ddots & & * \\ & & & & \ddots & \vdots \\ & & & & & * \end{bmatrix},$$

and in each column of L there are at most m elements different from zero. In contrast, the inverses $\mathbf{A}^{-1}$ of banded matrices are usually filled with nonzero entries. Thus, if $m << n$ ($\mathbf{A}$ is $n \times n$), using the triangular decomposition of $\mathbf{A}$ to solve $\mathbf{A}\vec{x} = \vec{b}$ results in a considerable savings in computation and storage over using $\mathbf{A}^{-1}$.

4.2. **The Gauss–Jordan Algorithm.**

4.2.1. *Gauss–Jordan Algorithm.* In the event that we want to find the inverse $\mathbf{A}^{-1}$ of a nonsingular matrix $\mathbf{A}$, we may use triangular decomposition or the Gauss–Jordan algorithm. Both methods require the same amount of work.

If the triangular decomposition $P\mathbf{A} = LR$ is known, then the $i-$th column $\bar{a}_i$ of $\mathbf{A}^{-1}$ is obtained as the solution of the system

$$LR\bar{a}_i = P\vec{e}_i,$$

where $\vec{e}_i$ is the $i-$th coordinate vector. The Gauss–Jordan method is obtained if we attempt to invert the mapping $\vec{x} \mapsto \mathbf{A}\vec{x} = \vec{y}$, $\vec{x}, \vec{y} \in \mathbb{R}^n$, determined by $\mathbf{A}$ in a systematic manner.

Consider the system $\mathbf{A}\vec{x} = \vec{y}$:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= y_1, \\ &\vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= y_n. \end{aligned}$$

In the first step of the Gauss–Jordan method, we switch x_1 for one of the variables y_r. To do this, an $a_{r1} \neq 0$ is found, for example by partial pivot selection

$$|a_{r1}| := \max_i |a_{i1}|,$$

and equations r and 1 are interchanged. In this way, a system

$$\begin{aligned} \bar{a}_{11}x_1 + \cdots + \bar{a}_{1n}x_n &= \bar{y}_1, \\ &\vdots \\ \bar{a}_{n1}x_1 + \cdots + \bar{a}_{nn}x_n &= \bar{y}_n \end{aligned} \tag{4.2.1.1}$$

is obtained in which the variables $\bar{y}_1, \dots, \bar{y}_n$ are some permutation of the variables $y_1, \dots, y_n$ and $\bar{a}_{11} = a_{r1}$, $\bar{y}_1 = y_r$ holds. Now $\bar{a}_{11} \neq 0$, for otherwise we would have $a_{i1} = 0$ for all $i = 1, 2, \dots, n$, which means $\mathbf{A}$ is singular, a contradiction. By solving the first equation of (4.2.1.1) for x_1 and substituting the result into the remaining equations, the system

$$\begin{aligned} a'_{11}\bar{y}_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n &= x_1,\\ a'_{21}\bar{y}_1 + a'_{22}x_2 + \cdots + a'_{2n}x_n &= \bar{y}_2,\\ &\vdots\\ a'_{n1}\bar{y}_1 + a'_{n2}x_2 + \cdots + a'_{nn}x_n &= \bar{y}_n \end{aligned}$$

is obtained with

$$a'_{11} := \frac{1}{\bar{a}_{11}}, \quad a'_{1k} := -\frac{\bar{a}_{1k}}{\bar{a}_{11}}, \quad a'_{i1} := \frac{\bar{a}_{i1}}{\bar{a}_{11}},$$
$$a'_{ik} := \bar{a}_{ik} - \frac{\bar{a}_{i1}\bar{a}_{1k}}{\bar{a}_{11}}, \quad \text{for } i,k = 2,3,\ldots,n.$$

In the next step, the variable x_1 is exchanged for one of the variables $\bar{y}_2,\ldots,\bar{y}_n$, then x_3 is exchanged for one of the remaining y variables, and so on. If the successive equation systems are represented by their matrices, then starting from $\mathbf{A}^{(0)} := \mathbf{A}$, a sequence

$$\mathbf{A}^{(0)} \to \mathbf{A}^{(1)} \to \cdots \to \mathbf{A}^{(n)}$$

is obtained. The matrix $\mathbf{A}^{(j)} = (a^{(j)}_{ik})$ stands for the matrix of a "mixed equation system" of the form

$$\begin{aligned} a^{(j)}_{11}\tilde{y}_1 + \cdots + a^{(j)}_{1j}\tilde{y}_j + a^{(j)}_{1,j+1}x_{j+1} + \cdots + a^{(j)}_{1n}x_n &= x_1,\\ &\vdots\\ a^{(j)}_{j1}\tilde{y}_1 + \cdots + a^{(j)}_{jj}\tilde{y}_j + a^{(j)}_{j,j+1}x_{j+1} + \cdots + a^{(j)}_{jn}x_n &= x_j,\\ a^{(j)}_{j+1,1}\tilde{y}_1 + \cdots + a^{(j)}_{j+1,j}\tilde{y}_j + a^{(j)}_{j+1,j+1}x_{j+1} + \cdots + a^{(j)}_{j+1,n}x_n &= \tilde{y}_{j+1},\\ &\vdots\\ a^{(j)}_{n1}\tilde{y}_1 + \cdots + a^{(j)}_{nj}\tilde{y}_j + a^{(j)}_{n,j+1}x_{j+1} + \cdots + a^{(j)}_{nn}x_n &= \tilde{y}_n. \end{aligned}$$

In this system $(\tilde{y}_1,\ldots,\tilde{y}_j,\tilde{y}_{j+1},\ldots,\tilde{y}_n)$ is a certain permutation of the original variables $(y_1,\ldots,y_n)$. In the transition $\mathbf{A}^{(j-1)} \to \mathbf{A}^{(j)}$ the variable x_j is swapped according to the rules given below. For simplicity, the elements of $\mathbf{A}^{(j-1)}$ are denoted by a_{ik}, and those of $\mathbf{A}^{(j)}$ by a'_{ik}.

Algorithm 4.2.1.1.

(1) Partial pivot selection: Determine r so that

$$|a_{rj}| = \max_{i\geq j} |a_{ij}|.$$

If $a_{rj} = 0$, $\mathbf{A}$ is singular, stop.

(2) Interchange rows r and j of $\mathbf{A}^{(j-1)}$, and call the result $\bar{\mathbf{A}} = (\bar{a}_{ik})$.

(3) Compute $\mathbf{A}^{(j)} = (a'_{ik})$ according to the formulas

$$a'_{jj} := \frac{1}{\bar{a}_{jj}},$$
$$a'_{jk} := -\frac{\bar{a}_{jk}}{\bar{a}_{jj}}, \quad a'_{ij} := \frac{\bar{a}_{ij}}{\bar{a}_{jj}} \quad \textit{for } i,k \neq j,$$

$$a'_{ik} := \bar{a}_{ik} - \frac{\bar{a}_{ij}\bar{a}_{jk}}{\bar{a}_{jj}}.$$

Note that

$$\mathbf{A}^{(n)}\hat{y} = \vec{x}, \quad \hat{y} = [\hat{y}_1, \ldots, \hat{y}_n]^\top,$$

where $\hat{y}_1, \ldots, \hat{y}_n$ is a certain permutation of the original variables $y_1, \ldots, y_n$, $\hat{y} = P\vec{y}$ which, since it corresponds to the interchange step (4.2.1.1)[2], can easily be determined. It follows

$$(\mathbf{A}^{(n)}P)\vec{y} = \vec{x},$$

and therefore, since $\mathbf{A}\vec{x} = \vec{y}$,

$$\mathbf{A}^{-1} = \mathbf{A}^{(n)}P.$$

Example 4.2.1.2.

$$\mathbf{A} := \mathbf{A}^{(0)} := \begin{bmatrix} 1^* & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{bmatrix} \to \mathbf{A}^{(1)} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1^* & 2 \\ 1 & 2 & 5 \end{bmatrix}$$
$$\to \mathbf{A}^{(2)} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 1 & -2 \\ -1 & 2 & 1^* \end{bmatrix} \to A^{(3)} = \begin{bmatrix} 3 & -3 & 1 \\ -3 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$
$$= \mathbf{A}^{-1}.$$

4.3. **The Choleski Decomposition.**

4.3.1. *The Choleski Decomposition.* The methods discussed thus far for solving equations can fail if no pivot selection is carried out, that is, if we restrict ourselves to taking the diagonal elements in order as pivots. However, there is an important class of matrices for which no pivot selection is necessary in computing triangular factors: the choice of each diagonal element in order always yields a nonzero pivot. Furthermore, it is numerically stable to use these pivots. We refer to the class of positive definite matrices.

Definition 4.3.1.1 (Positive Definite Matrix)**.** *A (possibly complex) $n \times n$ matrix $\mathbf{A}$ is said to be positive definite if it satisfies:*

(1) $\mathbf{A} = \mathbf{A}^H$, that is, $\mathbf{A}$ is Hermitian;
(2) $\vec{x}^H\mathbf{A}\vec{x} > 0$ for all $\vec{x} \in \mathbb{C}^n$, $\vec{x} \neq \vec{0}$.

We call a matrix $\mathbf{A} = \mathbf{A}^H$ *positive semidefinite* if $\vec{x}^H\mathbf{A}\vec{x} \geq 0$ holds for all $\vec{x} \in \mathbb{C}^n$.

Theorem 4.3.1.2. *For any positive definite matrix $\mathbf{A}$ the matrix $\mathbf{A}^{-1}$ exists and is positive definite. All principal submatrices of a positive definite matrix are also positive definite, and all principal minors of a positive definite matrix are positive.*

Proof. The inverse of a positive definite matrix $\mathbf{A}$ exists: if it were not so, then $\vec{x} \neq \vec{0}$ exists with $\mathbf{A}\vec{x} = \vec{0}$. Consequently

$$\vec{x}^H\mathbf{A}\vec{x} = \vec{x}^H\vec{0} = 0,$$

a contradiction to the assumption that $\mathbf{A}$ is positive definite.

Moreover, $\mathbf{A}^{-1}$ is positive definite: we have

$$(\mathbf{A}^{-1})^H = (\mathbf{A}^H)^{-1} = \mathbf{A}^{-1},$$

and if $\vec{y} \neq \vec{0}$ it follows

$$\vec{x} = \mathbf{A}^{-1}\vec{y} \neq \vec{0}.$$

Hence,

$$\vec{y}^H \mathbf{A}^{-1}\vec{y} = (\mathbf{A}\vec{x})^H \vec{x} = \vec{x}^H \mathbf{A}^H \vec{x} = \vec{x}^H \mathbf{A}\vec{x} > 0,$$

which shows that $\mathbf{A}^{-1}$ is indeed positive definite.

Every principal submatrix

$$\tilde{\mathbf{A}} = \begin{bmatrix} a_{i_1 i_1} & \dots & a_{i_1 i_k} \\ \vdots & & \vdots \\ a_{i_k i_1} & \dots & a_{i_k i_k} \end{bmatrix}$$

of a positive definite matrix $\mathbf{A}$ is also positive definite: clearly $\tilde{\mathbf{A}}^H = \tilde{\mathbf{A}}$. Moreover, every

$$\tilde{x} := [\tilde{x}_1, \dots, \tilde{x}_k]^\top \in \mathbb{C}^k, \quad \tilde{x} \neq \vec{0},$$

can be expanded to

$$\vec{x} := [x_1, \dots, x_n]^\top \in \mathbb{C}^n, \quad \vec{x} \neq \vec{0},$$

where

$$x_\mu := \begin{cases} \tilde{x}_j, & \mu = i_j,\ j = 1, 2, \dots, k, \\ 0, & \text{otherwise.} \end{cases}$$

From this construction it follows that

$$\tilde{x}^H \tilde{\mathbf{A}} \tilde{x} = \vec{x}^H \mathbf{A}\vec{x} > 0.$$

To complete the proof, it suffices to show that $\det(\mathbf{A}) > 0$ for any matrix $\mathbf{A}$ that is positive definite. We use induction.

Case $n = 1$ is trivial.

Now assume for the induction hypothesis that the theorem holds for positive definite matrices up to size $(n-1) \times (n-1)$, and let $\mathbf{A}$ be a positive definite $n \times n$ matrix. Then

$$\mathbf{A}^{-1} =: \begin{bmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{n1} & \dots & \alpha_{nn} \end{bmatrix}$$

is also positive definite, and consequently $\alpha_{11} = \vec{e}_1^H \mathbf{A}^{-1} \vec{e}_1 > 0$. Also, since

$$\mathbf{A} \begin{bmatrix} \alpha_{11} \\ \vdots \\ \alpha_{n1} \end{bmatrix} = \vec{e}_1,$$

we have by Cramer's Rule

$$\alpha_{11} = \frac{\det \left(\begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \dots & a_{nn} \end{bmatrix} \right)}{\det(\mathbf{A})}$$

$$= \frac{\det\left(\begin{bmatrix} a_{12} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \dots & a_{nn} \end{bmatrix}\right)}{\det(\mathbf{A})},$$

where

$$\det\left(\begin{bmatrix} a_{12} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \dots & a_{nn} \end{bmatrix}\right) > 0$$

by the induction hypothesis. Since $\alpha_{11} > 0$, we get $\det(\mathbf{A}) > 0$, which proves the theorem. □

Theorem 4.3.1.3. *For each $n \times n$ positive definite matrix $\mathbf{A}$ there is a unique $n \times n$ lower triangular matrix $\mathbf{L}$, $l_{ik} = 0$ for $k > i$, with $l_{ii} > 0$, $i = 1, 2, \dots, n$, satisfying $\mathbf{L}\mathbf{L}^H$. Moreover, if $\mathbf{A}$ is real, then so is $\mathbf{L}$.*

Note that $l_{ii} = 1$ if not required here. The matrix $\mathbf{L}$ is called the *Choleski factor* of $\mathbf{A}$, and $\mathbf{A} = \mathbf{L}\mathbf{L}^H$ its *Choleski decomposition.*

Proof. We use induction on n.

For case $n = 1$ the theorem is trivial: A positive definite 1×1 matrix $\mathbf{A} := (\alpha)$ is a positive number $\alpha > 0$, which can be written uniquely in the form

$$\alpha := l_{11}\bar{l}_{11}, \quad l_{11} := +\sqrt{\alpha}.$$

Now assume for the induction hypothesis that the theorem holds for positive definite matrices up to size $(n-1) \times (n-1)$. Let $\mathbf{A}$ be a positive definite $n \times n$ matrix. We may partition $\mathbf{A}$ into

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_{n-1} & \vec{b} \\ \hline \vec{b}^H & a_{nn} \end{array}\right],$$

where $\vec{b} \in \mathbb{C}^{n-1}$ and $\mathbf{A}_{n-1}$ is a positive definite $(n-1) \times (n-1)$ matrix by (4.3.1.2). By the induction hypothesis, there is a unique matrix $\mathbf{L}_{n-1}$ of size $(n-1) \times (n-1)$ satisfying

$$\mathbf{A}_{n-1} = \mathbf{L}_{n-1}\mathbf{L}_{n-1}^H, \quad l_{ik} = 0 \text{ for } k > i, \quad l_{ii} > 0.$$

We consider a matrix $\mathbf{L}$ of the form

$$\mathbf{L} = \left[\begin{array}{c|c} \mathbf{L}_{n-1} & 0 \\ \hline \vec{c}^H & \alpha \end{array}\right],$$

and try to determine $\vec{c} \in \mathbb{C}^{n-1}$, $\alpha > 0$ such that

$$\mathbf{L}\mathbf{U} = \left[\begin{array}{c|c} \mathbf{L}_{n-1} & 0 \\ \hline \vec{c}^H & \alpha \end{array}\right] \left[\begin{array}{c|c} \mathbf{L}_{n-1}^H & \vec{c} \\ \hline 0 & \alpha \end{array}\right] = \left[\begin{array}{c|c} \mathbf{L}_{n-1}\mathbf{L}_{n-1}^H & \mathbf{L}_{n-1}\vec{c} \\ \hline \vec{c}^H\mathbf{L}_{n-1}^H & \vec{c}^H\vec{c} + \alpha^2 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{n-1} & \vec{b} \\ \hline \vec{b}^H & a_{nn} \end{array}\right],$$

where

$$\mathbf{U} := \left[\begin{array}{c|c} \mathbf{L}_{n-1}^H & \vec{c} \\ \hline 0 & \alpha \end{array}\right].$$

This implies that we must have

$$\mathbf{L}_{n-1}\vec{c} = \vec{b},$$
$$\vec{c}^H\vec{c} + \alpha^2 = a_{nn}.$$

The first equation must have a unique solution $\vec{c} = \mathbf{L}_{n-1}^{-1}\vec{b}$, since $\mathbf{L}_{n-1}$, as a triangular matrix with strictly positive diagonal entries, has $\det(\mathbf{L}_{n-1}) > 0$. As for the second equation, define $\alpha := \sqrt{a_{nn} - \vec{c}^H\vec{c}}$. We need only show that $a_{nn} - \vec{c}^H\vec{c} > 0$ so that $\alpha \in \mathbb{R}$, $\alpha > 0$. But since $\vec{c}^H\vec{c} + \alpha^2 = a_{nn}$ holds in any case,

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) = \det(\mathbf{L}_{n-1})\det(\mathbf{L}_{n-1}^H)\alpha^2 > 0,$$

since $\mathbf{A}$ is positive definite by the hypothesis. But $\mathbf{A}_{n-1}$ is also positive definite (4.3.1.2) so that

$$\det(\mathbf{L}_{n-1})\det(\mathbf{L}_{n-1}^H) = \det(\mathbf{A}_{n-1}) > 0.$$

Hence, $\alpha^2 = a_{nn} - \vec{c}^H\vec{c} > 0$, as required.

This completes the proof. □

Check notes for how to compute entries l_{ij} of $\mathbf{L}$. We note here than an important implication of this computation of the entries l_{ij} is that

$$|l_{ij}| \leq \sqrt{a_{ii}}, \quad j = 1, 2, \ldots, k, \quad k = 1, 2, \ldots, n.$$

That is, the elements of $\mathbf{L}$ cannot grow too large.

4.4. **Error Bounds.**

4.4.1. *Error Bounds.* In general the solution $\vec{x} \in \mathbb{C}^n$ to the system $\mathbf{A}\vec{x} = \vec{b}$ is not computed exactly. We get

$$\tilde{x} = \vec{x} + \Delta\vec{x}, \quad \Delta\vec{x} \neq \vec{0}.$$

We want to discuss the error

$$\Delta\vec{x} := \tilde{x} - \vec{x}.$$

Definition 4.4.1.1 (Norm)**.** *A* **norm** *is a function $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ which assigns to each vector $\vec{x} \in \mathbb{C}^n$ a real value $\|\vec{x}\|$, which serves as a measure for the "size" of $\vec{x}$. A norm satisfies the following three properties:*

(1) $\|\vec{x}\| > 0$ for all $\vec{x} \in \mathbb{C}^n$, $\vec{x} \neq \vec{0}$ (positivity);
(2) $\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|$ for all $\alpha \in \mathbb{C}$, $\vec{x} \in \mathbb{C}^n$ (homogeneity);
(3) $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ for all $\vec{x}, \vec{y} \in \mathbb{C}^n$.

Theorem 4.4.1.2 (Reverse Triangle Inequality)**.** *For each norm $\|\cdot\|$ the inequality*

$$\|\vec{x} - \vec{y}\| \geq \left|\|\vec{x}\| - \|\vec{y}\|\right|,$$

for all $\vec{x}, \vec{y} \in \mathbb{C}^n$, holds.

Proof. Observe that

$$\|\vec{x}\| = \|(\vec{x} - \vec{y}) + \vec{y}\| \leq \|\vec{x} - \vec{y}\| + \|\vec{y}\|.$$

Consequently,

$$\|\vec{x}\| - \|\vec{y}\| \leq \|\vec{x} - \vec{y}\|.$$

Similarly

$$\|\vec{x} - \vec{y}\| = \|\vec{y} - \vec{x}\| \geq \|\vec{y}\| - \|\vec{x}\|,$$

which gives

$$\|\vec{x} - \vec{y}\| \geq \left|\|\vec{x}\| - \|\vec{y}\|\right|,$$

which proves the result. □

Example 4.4.1.3 (Vector Norms)**.** *Some common vector norms on $\mathbb{C}^n$ are as follows:*

(1) The Euclidean norm:

$$\|\vec{x}\|_2 := \sqrt{\vec{x}^H \vec{x}} = \left(\sum_{i=1}^{n} |x_i|^2 \right).$$

(2) The maximum (infinity) norm:

$$\|\vec{x}\|_\infty := \max_i |x_i|.$$

Theorem 4.4.1.4 (Norms are Uniformly Continuous)**.** *Each norm $\|\cdot\|$ in $\mathbb{C}^n$ is a uniformly continuous function with respect to the metric $\rho(\vec{x}, \vec{y}) := \max_i |x_i - y_i|$ on $\mathbb{C}^n$.*

Proof. Let $\vec{x} := [x_1, \ldots, x_n]^\top \in \mathbb{C}^n$ and $\vec{y} := [y_1, \ldots, y_n]^\top \in \mathbb{C}^n$. By the reverse triangle inequality,

$$|\|\vec{x} + \vec{y}\| - \|\vec{x}\|| \leq \|(\vec{x} + \vec{y}) - \vec{x}\| = \|\vec{y}\|.$$

Now $\vec{y} = \sum_{i=1}^n y_i \vec{e}_i$, where $\vec{e}_i$, $i = 1, 2, \ldots, n$ are the usual unit vectors. Therefore

$$\begin{aligned} \|\vec{y}\| &\leq \sum_{i=1}^{n} |y_i| \|\vec{e}_i\| \\ &\leq \max_i |y_i| \sum_{i=1}^{n} \|\vec{e}_i\| \\ &\leq M \max_i |y_i|, \end{aligned}$$

with $M := \sum_{i=1}^n \|\vec{e}_i\|$. Note that M depends only on $\|\cdot\|$ and n.

Fix $\epsilon > 0$. Then for all $\vec{y} \in \mathbb{C}^n$ satisfying

$$\|\vec{y}\|_\infty < \frac{\epsilon}{M},$$

it follows

$$|\|\vec{x} + \vec{y}\| - \|\vec{x}\|| \leq \|\vec{y}\| < M \left(\frac{\epsilon}{M} \right) = \epsilon.$$

Hence $\|\cdot\|$ is uniformly continuous. □

Theorem 4.4.1.5 (Equivalence of Norms)**.** *All norms on $\mathbb{C}^n$ are equivalent in the following sense: for each pair of norms $\|\cdot\|_a$, $\|\cdot\|_b$, there are positive constants m and M satisfying*

$$m\|\vec{x}\|_b \leq \|\vec{x}\|_a \leq M\|\vec{x}\|_b,$$

for all $\vec{x} \in \mathbb{C}^n$.

Proof. We show that any norm $\|\cdot\|$ is equivalent to the infinity norm,

$$m\|\vec{x}\|_\infty \leq \|\vec{x}\| \leq M\|\vec{x}\|_\infty,$$

so that any pair of norms $\|\cdot\|_a$, $\|\cdot\|_b$, are thus equivalent to $\|\cdot\|_\infty$.

By the homogeneity of vector norms, it suffices to consider only those vectors in the set

$$S := \{\vec{x} \in \mathbb{C}^n : \|\vec{x}\|_\infty = 1\},$$

which is compact in $\mathbb{C}^n$. Since S is compact and $\|\cdot\|$ is uniformly continuous (4.4.1.4), it follows from the extreme value theorem that

$$M := \max_{\vec{x}\in S} \|\vec{x}\| > 0, \quad m := \min_{\vec{x}\in S} \|\vec{x}\| > 0$$

exist. Thus, for all $\vec{y} \neq \vec{0}$, $\frac{\vec{y}}{\|\vec{y}\|_\infty} \in S$, and we have

$$m \leq \left\| \frac{\vec{y}}{\|\vec{y}\|_\infty} \right\| = \frac{1}{\|\vec{y}\|_\infty} \|\vec{y}\| \leq M,$$

which shows

$$m\|\vec{y}\|_\infty \leq \|\vec{y}\| \leq M\|\vec{y}\|_\infty.$$

This proves the theorem. □

For matrices $\mathbf{A} \in M(m,n)$ of fixed dimensions, norms $\|\mathbf{A}\|$ can be introduced. The properties

(1) $\|\mathbf{A}\| > 0$ for all $\mathbf{A} \neq \mathbf{0}$,
(2) $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$,
(3) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$

hold.

Definition 4.4.1.6 (Consistent Matrix Norm)**.** *The matrix norm $\|\cdot\|$ is said to be* **consistent** *with the vector norms $\|\cdot\|_a$ on $\mathbb{C}^n$ and $\|\cdot\|_b$ on $\mathbb{C}^m$ if*

$$\|\mathbf{A}\vec{x}\|_b \leq \|\mathbf{A}\|\|\vec{x}\|_a$$

for all $\vec{x} \in \mathbb{C}^n$ and $\mathbf{A} \in M(m,n)$.

Definition 4.4.1.7 (Submultiplicative Matrix Norm)**.** *A matrix norm $\|\cdot\|$ for square matrices $\mathbf{A} \in M(n,n)$ is called* **submultiplicative** *if*

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$$

for all $\mathbf{A}, \mathbf{B} \in M(n,n)$.

Note that choosing $B := I$ implies that $\|I\| \geq 1$ for submultiplicative matrix norms.

Example 4.4.1.8 (Matrix Norms)**.**

(1) Row-sum norm (also infinity norm):

$$\|\mathbf{A}\|_\infty := \max_i \left\{ \sum_{k=1}^{n} |a_{ik}| \right\}.$$

(2) Schur–Norm (also Frobenius norm):

$$\|\mathbf{A}\|_F := \left(\sum_{i,k=1}^{n} |a_{ik}|^2 \right)^{1/2}.$$

(3) Max norm:

$$\|\mathbf{A}\|_{max} := \max_{i,k} |a_{ik}|.$$

(4) Column-sum norm (also $1-$norm):

$$\|\mathbf{A}\|_1 := \max_k \left\{ \sum_{i=1}^{n} |a_{ik}| \right\}.$$

Norms (1), (2), and (4) are submultiplicative, (3) is not. Norm (2) is consistent with the Euclidean vector norm.

Definition 4.4.1.9 (Least Upper Bound Norm)**.** *Given a vector norm $\|\cdot\|$ on $\mathbb{C}^n$, a corresponding matrix norm for square matrices, the* **least upper bound norm** *or* **subordinate matrix norm***, can be defined by*

$$\mathrm{lub}(\mathbf{A}) := \max_{\vec{x}\neq 0} \frac{\|\mathbf{A}\vec{x}\|}{\|\vec{x}\|}.$$

Theorem 4.4.1.10. *Each subordinate matrix norm $\mathrm{lub}_v(\mathbf{A})$ is consistent with the vector norm $\|\cdot\|_v$ used to define it. Moreover, $\mathrm{lub}_v(\mathbf{A})$ is the smallest of all the matrix norms $\|\mathbf{A}\|$ which are consistent with the vector norm $\|\cdot\|_v$. Also, each subordinate matrix norm $\mathrm{lub}_v(\mathbf{A})$ is submultiplicative.*

Proof. The norm $\mathrm{lub}_v(\mathbf{A})$ is consistent with $\|\cdot\|_v$: Observe for all $\vec{x} \neq \vec{0}$ that

$$\|\mathbf{A}\vec{x}\|_v = \left\{\frac{\|\mathbf{A}\vec{x}\|_v}{\|\vec{x}\|_v}\right\}\|\vec{x}\|_v \leq \max_{\vec{x}\neq 0}\left\{\frac{\|\mathbf{A}\vec{x}\|_v}{\|\vec{x}\|_v}\right\}\|\vec{x}\|_v = \mathrm{lub}_v(\mathbf{A})\|\vec{x}\|_v.$$

The norm $\mathrm{lub}_v(\mathbf{A})$ is the smallest of all matrix norms $\|\mathbf{A}\|$ consistent with $\|\cdot\|_v$: Note that for a given matrix norm $\|\cdot\|$ and $\vec{x} \neq \vec{0}$, we have

$$\|\mathbf{A}\vec{x}\|_v \leq \|\mathbf{A}\|\|\vec{x}\|_v.$$

Thus for $\vec{x} \neq \vec{0}$,

$$\frac{\|\mathbf{A}\vec{x}\|_v}{\|\vec{x}\|_v} \leq \|\mathbf{A}\|$$

This is for all $\vec{x} \neq \vec{0}$, so that finally

$$\mathrm{lub}_v(\mathbf{A}) \leq \|\mathbf{A}\|.$$

Each subordinate matrix norm $\mathrm{lub}_v(\mathbf{A})$ is submultiplicative: Since $\mathrm{lub}_v(\mathbf{A})$ is the smallest matrix norm consistent with $\|\cdot\|_v$, we have evidently that

$$\mathrm{lub}_v(\mathbf{AB}) = \max_{\vec{x}\neq 0}\frac{\|\mathbf{AB}\vec{x}\|_v}{\|\vec{x}\|_v} \leq \max_{\vec{x}\neq 0}\mathrm{lub}_v(\mathbf{A})\frac{\|\mathbf{B}\vec{x}\|_v}{\|\vec{x}\|_v} = \mathrm{lub}_v(\mathbf{A})\mathrm{lub}_v(\mathbf{B}).$$

This completes the proof. □

Also note that

$$\mathrm{lub}(\mathbf{I}) = \max_{\vec{x}\neq 0}\frac{\|\mathbf{I}\vec{x}\|}{\|\vec{x}\|} = 1.$$

The consistency of $\mathrm{lub}(\mathbf{A})$ shows that $\mathrm{lub}(\mathbf{A})$ is the greatest magnification which a vector may attain under the mapping determined by $\mathbf{A}$. That is, it shows how much $\|\mathbf{A}\vec{x}\|$, the norm of an image point, can exceed $\|\vec{x}\|$, the norm of a source point.

Example 4.4.1.11. *For the maximum norm $\|\cdot\|_\infty = \max_v |\vec{x}|_v$, the subordinate matrix norm is the row–sum norm. Observe for any matrix $\mathbf{A} = (a_{ik})$ and $\vec{x} \neq \vec{0}$,*

$$\frac{\|\mathbf{A}\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = \frac{\max_i\left|\sum_{k=1}^{n} a_{ik}x_k\right|}{\max_k |x_k|}$$

$$\begin{aligned}
&\leq \frac{\max_i \sum_{k=1}^{n} |a_{ik}x_k|}{\max_k |x_k|} \\
&= \frac{\max_i \sum_{k=1}^{n} |a_{ik}||x_k|}{\max_k |x_k|} \\
&\leq \frac{\{\max_k |x_k|\} \max_i \sum_{k=1}^{n} |a_{ik}|}{\max_k |x_k|} \\
&= \|\mathbf{A}\|_\infty.
\end{aligned}$$

This shows $\mathrm{lub}_\infty(\mathbf{A}) \leq \|\mathbf{A}\|_\infty$.

To see that $\mathrm{lub}_\infty(\mathbf{A}) \geq \|\mathbf{A}\|_\infty$, *suppose that*

$$\|\mathbf{A}\|_\infty = \max_i \left\{ \sum_{k=1}^{n} |a_{ik}| \right\} = \sum_{k=1}^{n} |a_{i^*k}|,$$

that is, say i^* *is the maximum row. Define a vector*

$$\vec{x} := [x_1, x_2, \ldots, x_n]^\top$$

as follows:

$$x_k := \begin{cases} 1, & a_{i^*k} \geq 0, \\ -1, & \textit{otherwise.} \end{cases}$$

Clearly $\|\vec{x}\|_\infty = 1$. *Thus*

$$\begin{aligned}
\frac{\|\mathbf{A}\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = \|\mathbf{A}\vec{x}\|_\infty &= \max_i \left| \sum_{k=1}^{n} a_{ik}x_k \right| \\
&= \left| \sum_{k=1}^{1} a_{i^*k}x_k \right| \\
&= \sum_{k=1}^{n} a_{i^*k}x_k \\
&= \sum_{k=1}^{n} |a_{i^*k}| \\
&= \|\mathbf{A}\|_\infty.
\end{aligned}$$

It follows that $\mathrm{lub}_\infty(\mathbf{A}) \geq \|\mathbf{A}\|_\infty$.

Example 4.4.1.12. *For the Euclidean vector norm* $\|\cdot\|_2 = \sqrt{\vec{x}^H\vec{x}}$ *we have the subordinate matrix norm*

$$\begin{aligned}
\mathrm{lub}_2(\mathbf{A}) &= \max_{\vec{x}\neq 0} \frac{\|\mathbf{A}\vec{x}\|_2}{\|\vec{x}\|_2} \\
&= \max_{\vec{x}\neq 0} \frac{\sqrt{(\mathbf{A}\vec{x})^H(\mathbf{A}\vec{x})}}{\sqrt{\vec{x}^H\vec{x}}}
\end{aligned}$$

$$= \max_{\vec{x}\neq 0} \frac{\sqrt{\vec{x}^H \mathbf{A}^H \mathbf{A}\vec{x}}}{\sqrt{\vec{x}^H\vec{x}}}$$
$$= \sqrt{\lambda_{max}(\mathbf{A}^H\mathbf{A})},$$

where $\lambda_{max}(\mathbf{A}^H\mathbf{A})$ *denotes the largest eigenvalue of the matrix* $\mathbf{A}^H\mathbf{A}$.

In the following we assume that $\|\vec{x}\|$ is an arbitrary vector norm an $\|\mathbf{A}\|$ is a consistent submultiplicative matrix norm. Specifically, we can always take the subordinate norm $\text{lub}(\mathbf{A})$ as $\|\mathbf{A}\|$.

Theorem 4.4.1.13. *Let* $\vec{x}$ *be the solution to the system*

$$\mathbf{A}\vec{x} = \vec{b}$$

and suppose that $\tilde{x} := \vec{x} + \Delta\vec{x}$ *is an approximation to* $\vec{x}$ *such that*

$$\mathbf{A}(\vec{x} + \Delta\vec{x}) = \vec{b} + \Delta\vec{b}.$$

Then

$$\|\Delta\vec{x}\| \leq \|\mathbf{A}^{-1}\|\|\Delta\vec{b}\|.$$

Proof. Observe that

$$\begin{aligned}\mathbf{A}\tilde{x} &= \mathbf{A}(\vec{x} + \Delta\vec{x})\\ &= \mathbf{A}\vec{x} + \mathbf{A}\Delta\vec{x}\\ &= \vec{b} + \mathbf{A}\Delta\vec{x}\\ &= \vec{b} + \Delta\vec{b},\end{aligned}$$

where $\mathbf{A}\Delta\vec{x} = \Delta\vec{b}$. Presumably $\mathbf{A}$ is invertible, so that

$$\Delta\vec{x} = \mathbf{A}^{-1}\Delta\vec{b}.$$

It follows

$$\|\Delta\vec{x}\| \leq \|\mathbf{A}^{-1}\Delta\vec{b}\| \leq \|\mathbf{A}^{-1}\|\|\vec{b}\|.$$

□

Definition 4.4.1.14 (Condition Number)**.** *For a nonsingular square matrix* $\mathbf{A}$, *the* **condition number** $\kappa(\mathbf{A})$ *is defined by*

$$\kappa(\mathbf{A}) := \|\mathbf{A}\|\|\mathbf{A}^{-1}\|.$$

Theorem 4.4.1.15. *Let* $\vec{x}$ *be the solution to the system*

$$\mathbf{A}\vec{x} = \vec{b}$$

and assume that $\tilde{x} := \vec{x} + \Delta\vec{x}$ *is an approximation to* $\vec{x}$ *such that*

$$\mathbf{A}(\vec{x} + \Delta\vec{x}) = \vec{b} + \Delta\vec{b}.$$

Then

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \leq \kappa(\mathbf{A})\frac{\|\Delta\vec{b}\|}{\|\vec{b}\|}.$$

Proof. By (4.4.1.13), we have $\|\Delta\vec{x}\| \leq \|\mathbf{A}^{-1}\|\|\Delta\vec{b}\|$. Thus

$$\begin{aligned}\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} &\leq \frac{\|\mathbf{A}^{-1}\|\|\Delta\vec{b}\|}{\|\vec{x}\|} \\ &= \frac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\|\Delta\vec{b}\|}{\|\mathbf{A}\|\|\vec{x}\|} \\ &\leq \kappa(\mathbf{A})\frac{\|\Delta\vec{b}\|}{\|\mathbf{A}\vec{x}\|} \\ &= \kappa(\mathbf{A})\frac{\|\Delta\vec{b}\|}{\|\vec{b}\|}.\end{aligned}$$

□

For the case that $\kappa(\mathbf{A}) := \text{lub}(\mathbf{A})\text{lub}(\mathbf{A^{-1}})$, the condition of $\mathbf{A}$ is a measure of the sensitivity of th relative error in the solution to relative changes in the RHS. Moreover, since $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, $\kappa(\mathbf{A})$ satisfies

$$1 = \text{lub}(\mathbf{I}) = \text{lub}(\mathbf{A}\mathbf{A^{-1}}) \leq \text{lub}(\mathbf{A})\text{lub}(\mathbf{A^{-1}}) = \kappa(\mathbf{A}).$$

Note that this also holds for all submultiplicative matrix norms.

Definition 4.4.1.16 (Residual Operator)**.** *For a system $\mathbf{A}\vec{x} = \vec{b}$, we define a residual operator $\vec{r}(\cdot)$ by*

$$\vec{r}(\vec{y}) := \vec{b} - \mathbf{A}\vec{y}$$

for all $\vec{y} \in \mathbb{C}^n$.

For the true solution to $\mathbf{A}\vec{x} = \vec{b}$, we have evidently that $\vec{r}(\vec{y}) = \vec{0}$. Otherwise, $\|\vec{r}(\vec{x})\| > 0$.

Note that we can express the error in (4.4.1.13) is terms of the residual $\vec{r}(\cdot)$. To see this, let $\tilde{x} := \vec{x} + \Delta\vec{x}$ be an approximate solution to the system $\mathbf{A}\vec{x} = \vec{b}$ with residual

$$\vec{r}(\tilde{x}) := \vec{b} - \mathbf{A}\tilde{x} = \mathbf{A}(\vec{x} - \tilde{x}).$$

Then $\tilde{x}$ is the exact solution of

$$\mathbf{A}\tilde{x} = \vec{b} - \vec{r}(\tilde{x}),$$

so that $\vec{r}(\tilde{x}) = -\Delta\vec{b}$. Hence, it follows

$$\|\Delta\vec{x}\| \leq \|\mathbf{A}^{-1}\|\|\vec{r}(\tilde{x})\|.$$

To motivate the following result, set $\mathbf{B} := \mathbf{A} + \Delta\mathbf{A}$. If $\mathbf{A}^{-1}$ exists, then we may find $\mathbf{F}$ such that

$$\mathbf{AF} = \Delta\mathbf{A}.$$

Moreover, in this situation,

$$\mathbf{B} = \mathbf{A} + \mathbf{AF} = \mathbf{A}(\mathbf{I} + \mathbf{F}).$$

Then $\mathbf{A}^{-1}\mathbf{B} = \mathbf{I} + \mathbf{F}$.

Lemma 4.4.1.17. *If $\mathbf{F}$ is an $n \times n$ matrix with $\|\mathbf{F}\| < 1$, then $(\mathbf{I} + \mathbf{F})^{-1}$ exists and satisfies*

$$\|(\mathbf{I} + \mathbf{F})^{-1}\| \leq \frac{\|\mathbf{I}\|}{1 - \|\mathbf{F}\|}.$$

Proof. By the reverse triangle inequality,

$$\begin{aligned}\|(\mathbf{I}+\mathbf{F})\vec{x}\| &= \|\vec{x}+\mathbf{F}\vec{x}\| \geq |\|\vec{x}\| - \|\mathbf{F}\vec{x}\|| \geq \|\vec{x}\| - \|\mathbf{F}\vec{x}\| \\ &\geq \|\vec{x}\| - \|\mathbf{F}\|\|\vec{x}\| = (1-\|\mathbf{F}\|)\|\vec{x}\|\end{aligned}$$

holds for all $\vec{x}$. From $1-\|\mathbf{F}\| > 0$, it follows that $\|(\mathbf{I}+\mathbf{F})\vec{x}\| > 0$ for all $\vec{x} \neq \vec{0}$, that is, $(\mathbf{I}+\mathbf{F})\vec{x} = \vec{0}$ has only the trivial solution $\vec{x} = \vec{0}$, so that $\mathbf{I}+\mathbf{F}$ is nonsingular.

To prove the inequality, observe that

$$\begin{aligned}\|\mathbf{I}\| &= \|(\mathbf{I}+\mathbf{F})(\mathbf{I}+\mathbf{F})^{-1}\| \\ &= \|(\mathbf{I}+\mathbf{F})^{-1} + \mathbf{F}(\mathbf{I}+\mathbf{F})^{-1}\| \\ &\geq \big|\|(\mathbf{I}+\mathbf{F})^{-1}\| - \|\mathbf{F}\|\|(\mathbf{I}+\mathbf{F})^{-1}\|\big| \\ &\geq \|(\mathbf{I}+\mathbf{F})^{-1}\| - \|\mathbf{F}\|\|(\mathbf{I}+\mathbf{F})^{-1}\| \\ &= \big(\|(\mathbf{I}+\mathbf{F})^{-1}\|\big)\,(1-\|\mathbf{F}\|) > 0,\end{aligned}$$

from which we have

$$\|(\mathbf{I}+\mathbf{F})^{-1}\| \leq \frac{\|\mathbf{I}\|}{1-\|\mathbf{F}\|}.$$

This completes the proof. □

Before stating the next theorem, recall that matrix norms are submultiplicative for these results.

Theorem 4.4.1.18. *Let* $\mathbf{A}$ *be a nonsingular* $n \times n$ *matrix,* $\mathbf{B} = \mathbf{A}(\mathbf{I}+\mathbf{F})$, $\|\mathbf{F}\| < 1$, *and* $\vec{x}$ *and* $\Delta\vec{x}$ *be defined by* $\mathbf{A}\vec{x} = \vec{b}$, $\mathbf{B}(\vec{x}+\Delta\vec{x}) = \vec{b}$. *It follows that*

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \leq \frac{\|\mathbf{F}\|}{1-\|\mathbf{F}\|}\|\mathbf{I}\|$$

as well as

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \leq \frac{\theta}{1-\theta}\|\mathbf{I}\|,$$

where

$$\theta := \kappa(\mathbf{A})\frac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|},$$

provided that $\theta < 1$.

Recall that we defined

$$\mathbf{B}-\mathbf{A} = \mathbf{A}\mathbf{F} = \Delta\mathbf{A}.$$

Thus

$$\frac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|} = \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}$$

gives a relative error for $\mathbf{A}$.

Proof. The matrix $\mathbf{B}^{-1}$ exists by (4.4.1.17), and

$$\begin{aligned}\Delta\vec{x} &= \mathbf{B}^{-1}\vec{b} - \vec{x} = \mathbf{B}^{-1}\vec{b} - \mathbf{A}^{-1}\vec{b} = (\mathbf{B}^{-1}-\mathbf{A}^{-1})\vec{b} \\ &= \mathbf{B}^{-1}(\mathbf{A}-\mathbf{B})\mathbf{A}^{-1}\vec{b},\end{aligned}$$

where $\vec{x} = \mathbf{A}^{-1}\vec{b}$. Furthermore,

$$\begin{aligned}\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} &= \frac{\|\mathbf{B}^{-1}(\mathbf{A}-\mathbf{B})\mathbf{A}^{-1}\vec{b}\|}{\|\mathbf{A}^{-1}\vec{b}\|}\\ &\le \frac{\|\mathbf{B}^{-1}(\mathbf{A}-\mathbf{B})\|\|\mathbf{A}^{-1}\vec{b}\|}{\|\mathbf{A}^{-1}\vec{b}\|}\\ &= \|\mathbf{B}^{-1}(\mathbf{A}-\mathbf{B})\|.\end{aligned}$$

Now $\mathbf{B}^{-1} = (\mathbf{I}+\mathbf{F})^{-1}\mathbf{A}^{-1}$ and $\mathbf{A}-\mathbf{B} = -\mathbf{A}\mathbf{F}$, so that by (4.4.1.17) we have

$$\begin{aligned}\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} &\le \|-(\mathbf{I}+\mathbf{F})^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{F}\|\\ &= \|(\mathbf{I}+\mathbf{F})^{-1}\mathbf{F}\|\\ &\le \|(\mathbf{I}+\mathbf{F})^{-1}\|\|\mathbf{F}\|\\ &\le \frac{\|\mathbf{F}\|}{1-\|\mathbf{F}\|}\|\mathbf{I}\|.\end{aligned}$$

Moreover, since $\mathbf{F} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})$ and

$$\|\mathbf{F}\| = \|\mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\| \le \|\mathbf{A}^{-1}\|\|\mathbf{B}-\mathbf{A}\| = \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|} = \kappa(\mathbf{A})\frac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|},$$

it follows

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \le \frac{\kappa(\mathbf{A})\dfrac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|}}{1-\kappa(\mathbf{A})\dfrac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|}}\|\mathbf{I}\|,$$

which completes the proof. □

According to (4.4.1.18), $\kappa(\mathbf{A})$ also measures the sensitivity of the solution $\vec{x}$ of $\mathbf{A}\vec{x} = \vec{b}$ to perturbations of the matrix $\mathbf{A}$.

If we put

$$\mathbf{C} := (\mathbf{I}+\mathbf{F})^{-1} = \mathbf{B}^{-1}\mathbf{A}, \quad \mathbf{F} = \mathbf{A}^{-1}\mathbf{B}-\mathbf{I},$$

it then follows from (4.4.1.17) that

$$\|\mathbf{B}^{-1}\mathbf{A}\| \le \frac{\|\mathbf{I}\|}{1-\|\mathbf{I}-\mathbf{A}^{-1}\mathbf{B}\|}.$$

If further we assume that $\mathbf{A} = \mathbf{B}(\mathbf{I}+\tilde{\mathbf{F}})$ for some $\tilde{\mathbf{F}}$ with $\|\tilde{\mathbf{F}}\| < 1$, then interchanging $\mathbf{A}$ and $\mathbf{B}$ and noting that $\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^{-1}$ gives

$$\|\mathbf{A}^{-1}\| \le \|\mathbf{A}^{-1}\mathbf{B}\|\|\mathbf{B}\| \le \frac{\|\mathbf{I}\|}{1-\|\mathbf{I}-\mathbf{B}^{-1}\mathbf{A}\|}\|\mathbf{B}^{-1}\|.$$

In particular, the residual estimate

$$\|\Delta\vec{x}\| \le \|\mathbf{A}^{-1}\|\|\vec{r}(\tilde{x})\|$$

leads us to the bound

$$\|\tilde{x}-\vec{x}\| \le \frac{\|\mathbf{I}\|\|\mathbf{B}^{-1}\|}{1-\|\mathbf{I}-\mathbf{B}^{-1}\mathbf{A}\|}\|\vec{r}(\tilde{x})\|,$$

where $\vec{r}(\tilde{x}) = \vec{b} - \mathbf{A}\tilde{x}$, and where $\mathbf{B}$ is an approximate inverse to $\mathbf{A}$ with $\|\mathbf{I} - \mathbf{B}^{-1}\mathbf{A}\| < 1$.

The estimates up to this point give bounds on the error $\Delta\vec{x} := \tilde{x} - \vec{x}$, but the evaluation of the bounds requires at least an approximate knowledge of $\mathbf{A}^{-1}$. The estimates given next do not require any knowledge of $\mathbf{A}^{-1}$.

In general, the given data $\mathbf{A}_0, \vec{b}_0$ of an equation system $\mathbf{A}_0\vec{x} = \vec{b}_0$ are inexact, being tainted, for example, by measurement errors $\Delta\mathbf{A}$, $\Delta\vec{b}$. Thus, it is reasonable to accept an approximate solution $\tilde{x}$ to the system as "correct" if $\tilde{x}$ is the exact solution to a "neighboring" system

$$\mathbf{A}\tilde{x} = \vec{b},$$

with

$$\mathbf{A} \in \mathbb{A} := \{\mathbf{A} : |\mathbf{A} - \mathbf{A}_0| \leq \Delta\mathbf{A}\},$$
$$\vec{b} \in \mathbb{B} := \{\vec{b} : |\vec{b} - \vec{b}_0| \leq \Delta\vec{b}\}.$$

The notation used here is

$$|\mathbf{A}| = (|a_{ik}|), \quad \text{where } \mathbf{A} = (a_{ik}),$$
$$|\vec{b}| = (|b_1|, |b_2|, \ldots, |b_n|)^\top, \quad \text{where } \vec{b} = (b_1, b_2, \ldots, b_n)^\top,$$

and the relation $\leq$ between vectors and matrices is to be understood as holding componentwise.

Theorem 4.4.1.19. *Let $\Delta\mathbf{A} > 0$ and $\Delta\vec{b} > 0$. Associated with any approximate solution $\tilde{x}$ of the system $\mathbf{A}_0\vec{x} = \vec{b}_0$, there is a matrix $\mathbf{A} \in \mathbb{A}$ and vector $\vec{b} \in \mathbb{B}$ satisfying*

$$\mathbf{A}\tilde{x} = \vec{b}$$

if and only if

$$|\vec{r}(\tilde{x})| \leq \Delta\mathbf{A}|\tilde{x}| + \Delta\vec{b},$$

where $\vec{r}(\tilde{x}) := \vec{b}_0 - \mathbf{A}_0\tilde{x}$ is the residual of $\tilde{x}$.

Proof. ($\Longrightarrow$) Assume that $\mathbf{A}\tilde{x} = \vec{b}$ for some $\mathbf{A} \in \mathbb{A}$ and $\vec{b} \in \mathbb{B}$. Then

$$\mathbf{A} = \mathbf{A}_0 + \delta\mathbf{A}, \quad \vec{b} = \vec{b}_0 + \delta\vec{b},$$

where

$$|\delta\mathbf{A}| \leq \Delta\mathbf{A} \quad \text{and} \quad |\delta\vec{b}| \leq \Delta\vec{b}.$$

It follows

$$\begin{aligned}
|\vec{r}(\tilde{x})| &= |\vec{b}_0 - \mathbf{A}_0\tilde{x}| \\
&= |(\vec{b} - \delta\vec{b}) - (\mathbf{A} - \delta\mathbf{A})\tilde{x}| \\
&= |\vec{b} - \delta\vec{b} - \mathbf{A}\tilde{x} + \delta\mathbf{A}\tilde{x}| \\
&= |\delta\mathbf{A}\tilde{x} - \delta\vec{b}| \\
&\leq |\delta\mathbf{A}||\tilde{x}| + |\delta\vec{b}| \\
&\leq \Delta\mathbf{A}|\tilde{x}| + \Delta\vec{b},
\end{aligned}$$

which completes the proof for this case.

($\Longleftarrow$) For the converse, suppose that

$$|\vec{r}(\tilde{x})| \leq \Delta\mathbf{A}|\tilde{x}| + \Delta\vec{b}.$$

We introduce the following notation:

(1) $\tilde{x} =: (x_1, x_2, \ldots, x_n)^\top$,
(2) $\vec{b}_0 =: (b_1, b_2, \ldots, b_n)^\top$,
(3) $\vec{r} := \vec{r}(\tilde{x}) = (r_1, r_2, \ldots, r_n)^\top$,
(4) $\vec{s} := \Delta\vec{b} + \Delta\mathbf{A}|\tilde{x}| \geq 0, \quad \vec{s} =: (s_1, s_2, \ldots, s_n)^\top$.

We construct $\delta\vec{b}$, $\delta\mathbf{A}$ as follows. For $i = 1, 2, \ldots, n$, if $s_i = 0$, then set $(\delta\vec{b})_i = 0$ and set $(\delta\mathbf{A})_{ij} = 0$ for all $j = 1, 2, \ldots, n$. Otherwise, $s_i > 0$. In this case, put

$$(\delta\mathbf{A})_{ij} = \frac{r_i(\Delta\mathbf{A})_{ij}\mathrm{sgn}\,(x_i)}{s_i}$$

for $j = 1, 2, \ldots, n$, and

$$(\delta\vec{b})_i = \frac{-r_i(\Delta\vec{b})_i}{s_i}.$$

Since $|\vec{r}(\tilde{x})| \leq \Delta\mathbf{A}|\tilde{x}| + \Delta\vec{b} = \vec{s}$,

$$\frac{|r_i|}{|s_i|} \leq 1$$

when $s_i > 0$.

Now take $\mathbf{A} := \mathbf{A}_0 + \delta\mathbf{A}$, $\vec{b} := \vec{b}_0 + \delta\vec{b}$. Note by the construction that $|\delta\mathbf{A}| \leq \Delta\mathbf{A}$, $|\delta\vec{b}| \leq \Delta\vec{b}$, which implies that $\mathbf{A} \in \mathbb{A}$, $\vec{b} \in \mathbb{B}$.

We verify that $\vec{b} = \mathbf{A}\tilde{x}$. For any $i = 1, 2, \ldots, n$, there are two cases.

If $s_i = 0$, then

$$|\vec{r}(\tilde{x})| \leq \Delta\mathbf{A}|\tilde{x}| + \Delta\vec{b} = \vec{s}$$

implies that

$$r_i = 0 = (\vec{b}_0 - \mathbf{A}_0\tilde{x})_i.$$

Furthermore, $(\delta\vec{b})_i = 0$ and $(\delta\mathbf{A})_{ij} = 0$ for $j = 1, 2, \ldots, n$, so that $b_i = (\vec{b})_i$ and $(\mathbf{A})_{ij} = (\mathbf{A}_0)_{ij}$. Thus

$$(\vec{b} - \mathbf{A}\tilde{x})_i = (\vec{b}_0 - \mathbf{A}_0\tilde{x})_i = 0.$$

Now consider the case that $s_i > 0$. We may write

$$\begin{aligned}
(\vec{b}_0 - \mathbf{A}_0\tilde{x})_i = r_i &= \frac{s_i}{s_i} r_i \\
&= \frac{[(\Delta\vec{b})_i + \sum_{j=1}^n (\Delta\mathbf{A})_{ij}|x_j|]r_i}{s_i} \\
&= \frac{(\Delta\vec{b})_i r_i}{s_i} + \sum_{j=1}^n \left[(\Delta\mathbf{A})_{ij}\frac{r_i}{s_i}\mathrm{sgn}\,(x_j)\right] x_j \\
&= -(\delta\vec{b})_i + \sum_{j=1}^n (\delta\mathbf{A})_{ij} x_j \\
&= (\delta\mathbf{A}\tilde{x} - \delta\vec{b})_i,
\end{aligned}$$

so that evidently $(\vec{b}_0 + \delta\vec{b})_i = ((\mathbf{A}_0 + \delta\mathbf{A})\tilde{x})_i$. Hence,

$$(\vec{b})_i = (\mathbf{A}\tilde{x})_i.$$

This completes the proof. □

The criterion expressed in Theorem (4.4.1.19) allows us to draw conclusions about the fitness of a solution from the smallness of its residual. For instance, if all components of $\mathbf{A}_0$ and $\vec{b}_0$ have the same relative accuracy ϵ,

$$\Delta\mathbf{A} = \epsilon|\mathbf{A}_0|, \quad \Delta\vec{b} = \epsilon|\vec{b}_0|,$$

then Theorem (4.4.1.19) is satisfied if

$$|\vec{r}(\tilde{x})| = |\mathbf{A}_0\tilde{x} - \vec{b}_0| \leq \Delta\mathbf{A}\tilde{x} + \Delta\vec{b} = \epsilon(|\vec{b}_0| + |\mathbf{A}_0||\tilde{x}|).$$

From this inequality, the smallest ϵ can be computed for which a given $\tilde{x}$ can still be accepted as a usable solution.

4.5. **Orthogonalization Techniques of Householder and Gram–Schmidt.**

4.5.1. *Orthogonalization Techniques of Householder and Gram–Schmidt.* Recall that the methods discussed thus far for solving

$$\mathbf{A}\vec{x} = \vec{b}$$

have consisted of multiplying $\mathbf{A}\vec{x}$ by approximate matrices $P^{(j)}$, $j = 1, 2, \ldots, n$, so that the system obtained by the outcome

$$\mathbf{A}^{(n)}\vec{x} = \vec{b}^{(n)}$$

may be solved directly. The sensitivity of $\vec{x}$ to changes in the arrays of the intermediate systems

$$\mathbf{A}^{(j)}\vec{x} = \vec{b}^{(j)}, \quad [\mathbf{A}^{(j)}, \vec{b}^{()}] = p^{(j)}[\mathbf{A}^{(j-1)}, \vec{b}^{(j-1)}],$$

is given by

$$\kappa(\mathbf{A}^{(\mathbf{j})}) = \text{lub}(\mathbf{A}^{(\mathbf{j})})\text{lub}((\mathbf{A}^{(\mathbf{j})})^{-\mathbf{1}}).$$

Denote the roundoff error incurred in the $j-$th step of this process by $\epsilon^{(j)}$:

$$\epsilon^{(j)} := [\mathbf{A}^{(j-1)}, \vec{b}^{(j-1)}] \to [\mathbf{A}^{(j)}, \vec{b}^{(j)}].$$

These roundoff errors are amplified by the factors $\kappa(\mathbf{A}^{(\mathbf{j})})$ in their effect on $\vec{x}$, and we have

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \dot{\leq} \sum_{j=0}^{n-1} \epsilon^{(j)}\kappa(\mathbf{A}^{(\mathbf{j})}).$$

If there exists $\mathbf{A}^{(j)}$ with

$$\kappa(\mathbf{A}^{(\mathbf{j})}) >> \kappa(\mathbf{A}^{(\mathbf{0})}),$$

then evidently the sequence of computations is not numerically stable. That is, $\epsilon^{(j)}$ has a stronger influence that the initial error $\epsilon^{(0)}$. Our goal is to choose $P^{(j)}$ so that

$$\kappa(\mathbf{A}^{(\mathbf{j-1})}) \geq \kappa(\mathbf{A}^{(\mathbf{j})}).$$

Lemma 4.5.1.1. *Let $\mathbf{U} \in M(n, n)$ be unitary. Then*

(1) $\|\mathbf{U}\vec{x}\|_2 = \|\vec{x}\|_2$ for all $\vec{x} \in \mathbb{C}^n$;
(2) $\text{lub}_2(\mathbf{U}) = \text{lub}_2(\mathbf{U}^{\mathbf{H}}) = 1$.

Proof. Let $\mathbf{U} \in M(n,n)$ be unitary. Then

$$\|\mathbf{U}\vec{x}\|_2^2 = (\mathbf{U}\vec{x})^H(\mathbf{U}\vec{x}) = \vec{x}^H\mathbf{U}^H\mathbf{U}\vec{x} = \vec{x}^H\vec{x} = \|\vec{x}\|_2^2,$$

which proves (1).

For (2), observe that

$$\mathrm{lub}_2(\mathbf{U}) = \max_{x\neq 0}\frac{\|\mathbf{U}\vec{x}\|_2}{\|\vec{x}\|_2} = \max_{\vec{x}\neq 0}\frac{\|\vec{x}\|_2}{\|\vec{x}\|_2} = 1.$$

Since $\mathbf{U}^H$ is also unitary, this completes the proof. □

Lemma 4.5.1.2. *Let $\mathbf{A} \in M(n,n)$ and let $\mathbf{U} \in M(n,n)$ be unitary. Then*

$$\kappa(\mathbf{A}) = \kappa(\mathbf{UA}),$$

where $\kappa(\cdot)$ denotes the condition number of $\cdot$ with respect to the subordinate matrix norm $\mathrm{lub}_2(\cdot)$. induced by the vector $2-$norm.

Proof. Since $\mathrm{lub}_2(\cdot)$ is submultiplicative, we have

$$\begin{aligned}\mathrm{lub}_2(\mathbf{A}) &= \mathrm{lub}_2(\mathbf{U^H UA})\\ &= \mathrm{lub}_2(\mathbf{U^H})\mathrm{lub}_2(\mathbf{UA})\\ &\leq \mathrm{lub}_2(\mathbf{UA})\\ &\leq \mathrm{lub}_2(\mathbf{U})\mathrm{lub}_2(\mathbf{A})\\ &= \mathrm{lub}_2(\mathbf{A}).\end{aligned}$$

Thus

$$\mathrm{lub}_2(\mathbf{A}) = \mathrm{lub}_2(\mathbf{UA}).$$

Analogously,

$$\mathrm{lub}_2((\mathbf{UA})^{-\mathbf{1}}) = \mathrm{lub}_{\cdot}(\mathbf{A}^{-\mathbf{1}})$$

Hence,

$$\kappa(\mathbf{A}) = \mathrm{lub}_2(\mathbf{A})\mathrm{lub}_2(\mathbf{A}^{-\mathbf{1}}) = \mathrm{lub}_2(\mathbf{UA})\mathrm{lub}_2((\mathbf{UA})^{-\mathbf{1}}) = \kappa(\mathbf{UA}).$$

□

In this section we choose the transformation matrices $P^{(j)}$ to be unitary. From this property it follows that the condition numbers associated with the systems

$$\mathbf{A}^{(j)}\vec{x} = \vec{b}^{(j)}$$

do not change. Furthermore, the matrices $P^{(j)}$ should be chosen so that the $\mathbf{A}^{(j)}$ become simpler, in this case, to reduce $\mathbf{A}$ to upper triangular form.

Definition 4.5.1.3 (Householder Matrix)**.** *A* **Householder Matrix** *P is a matrix*

$$P := \mathbf{I} - 2\vec{w}\vec{w}^H,$$

with $\vec{w}^H\vec{w} = 1$, $\vec{w} \in \mathbb{C}^n$.

We get the following important properties for Householder matrices P.

Theorem 4.5.1.4 (Properties of Householder Matrices)**.** *Let P be a Householder matrix. Then P satisfies the following properties:*

(1) $P = P^H$ (P is Hermitian);
(2) $P^H P = \mathbf{I}$ (P is unitary);
(3) $P^2 = \mathbf{I}$ (P is involutory).

Proof. P is Hermitian: Observe

$$\begin{aligned} P^H &= (\mathbf{I} - 2\vec{w}\vec{w}^H)^H \\ &= \mathbf{I}^H - 2(\vec{w}\vec{w}^H)^H \\ &= \mathbf{I} - 2(\vec{w}^H)^H\vec{w}^H \\ &= \mathbf{I} - 2\vec{w}\vec{w}^H = P. \end{aligned}$$

P is unitary and involutory: We have

$$\begin{aligned} P^H P = P^2 &= (\mathbf{I} - 2\vec{w}\vec{w}^H)(\mathbf{I} - 2\vec{w}\vec{w}^H) \\ &= \mathbf{I}^2 - 2\mathbf{I}\vec{w}\vec{w}^H - 2\mathbf{I}\vec{w}\vec{w}^H + 4(\vec{w}\vec{w}^H)(\vec{w}\vec{w}^H) \\ &= \mathbf{I} - 4\vec{w}\vec{w}^H + 4\vec{w}(\vec{w}^H\vec{w})\vec{w}^H \\ &= \mathbf{I} - 4\vec{w}\vec{w}^H + 4\vec{w}\vec{w}^H \\ &= \mathbf{I}. \end{aligned}$$

This completes the proof. □

Geometrically, the map $\vec{x} \mapsto y =: P\vec{x} = \vec{x} - 2(\vec{w}^H\vec{x})\vec{w}$ describes a reflection of $\vec{w}$ with respect to the plane $\{\vec{z} : \vec{w}^H\vec{z} = 0\}$ and $\vec{x}$ and $\vec{y}$ then satisfy:

(1) $\vec{y}^H\vec{y} = (P\vec{x})^H(P\vec{x}) = \vec{x}^H P^H\vec{x} = \vec{x}^H\vec{x}$;
(2) $\vec{x}^H\vec{y} = \vec{x}^H P\vec{x} = (\vec{x}^H P\vec{x})^H$,

and $\vec{x}^H\vec{y}$ is real.

The remainder of this section deals with the construction of the Householder matrix P.

We wish to determine a vector $\vec{w}$, and thereby P, so that a given vector

$$\vec{x} := [x_1, x_2, \ldots, x_n]^\top$$

is transformed into a multiple of the first coordinate vector $\vec{e}_1$:

$$k\vec{e}_1 = P\vec{x},$$

for then we may use P to eliminate every entry below the diagonal in a column of a given matrix. Note that

$$\|\vec{x}\|_2^2 = \vec{x}^H\vec{x} = \vec{x}^H P^H P\vec{x} = (P\vec{x})^H P\vec{x} = (k\vec{e}_1)^H k\vec{e}_1 = |k|^2\vec{e}_1^H\vec{e}_1 = |k|^2.$$

Also,

$$k\vec{x}^H\vec{e}_1 = \vec{x}^H(k\vec{e}_1) = \vec{x}^H(P\vec{x}) = \vec{x}^H P^H\vec{x} = (\vec{x}^H P\vec{x})^H \in \mathbb{R},$$

which implies that $k\vec{x}^H\vec{e}_1$ is real.

Put $x_1 =: e^{i\alpha}|x_1|$. Then

$$k\vec{x}^H\vec{e}_1 = k\bar{x}_1 = k|x_1|e^{-i\alpha} \in \mathbb{R},$$

so that

$$k = \pm|k|e^{i\alpha}.$$

Recalling that

$$\|\vec{x}\|_2^2 = \vec{x}^H\vec{x} = |k|^2,$$

we have

$$k = \pm\|\vec{x}\|_2 e^{i\alpha}.$$

Now note

$$e^{i\alpha} = \frac{x_1}{|x_1|}$$

by definition. Thus we arrive at

$$k = \pm\frac{x_1}{|x_1|}\|\vec{x}\|_2.$$

In the case $x_1 = 0$, we put $k = \pm\|\vec{x}\|_2$.

We now define

$$\vec{w} := \frac{\vec{x} - k\vec{e_1}}{\|\vec{x} - k\vec{e_1}\|_2}.$$

We verify that this choice of $\vec{w}$ suffices. Clearly $\|\vec{w}\|_2 = 1$. Also,

$$\begin{aligned}
\frac{2\vec{w}^H\vec{x}}{\|\vec{x} - k\vec{e_1}\|_2} &= \frac{2(\vec{x}^H - (k\vec{e_1})^H)\vec{x}}{\|\vec{x} - k\vec{e_1}\|_2^2} \\
&= \frac{2\vec{x}^H\vec{x} - 2\vec{e}_1^H\bar{k}\vec{x}}{\|\vec{x} - k\vec{e_1}\|_2^2} \\
&= 2\frac{\|\vec{x}\|_2^2 - \bar{k}x_1}{\|\vec{x} - k\vec{e_1}\|_2^2} \\
&= 2\frac{\|\vec{x}\|_2^2 \mp |x_1|\|\vec{x}\|_2^2}{\vec{x}^H\vec{x} - \bar{k}x_1 - k\bar{x}_1 + |k|^2} \\
&= 2\frac{\|\vec{x}\|_2^2 \mp |x_1|\|\vec{x}\|_2^2}{2\|\vec{x}\|_2^2 \mp |x_1|\|\vec{x}\|_2^2} \\
&= 1,
\end{aligned}$$

since

$$\bar{k}x_1 = \pm\frac{\bar{x}_1}{|x_1|}\|\vec{x}\|_2(x_1) = \pm|x_1|\|\vec{x}\|_2 = k\bar{x}_1.$$

Moreover,

$$\begin{aligned}
2(\vec{w}^H\vec{x})\vec{w} &= 2\vec{w}^H\vec{x}\frac{\vec{x} - k\vec{e_1}}{\|\vec{x} - k\vec{e_1}\|_2} \\
&= (\|\vec{x} - k\vec{e_1}\|_2)\frac{\vec{x} - k\vec{e_1}}{\|\vec{x} - k\vec{e_1}\|_2} \\
&= \vec{x} - k\vec{e_1}.
\end{aligned}$$

Hence,

$$P\vec{x} = \vec{x} - 2(\vec{w}^H\vec{x})\vec{w} = k\vec{e_1}.$$

We now turn to a consideration of the roundoff error induced by this process. Observe that

$$\|\vec{x} - k\vec{e_1}\|_2^2 = |x_1 - k|^2 + |x_2|^2 + \cdots + |x_n|^2$$

$$
\begin{aligned}
&= \left| x_1 \mp \frac{x_1}{|x_1|}\|\vec{x}\|_2 \right|^2 + |x_2|^2 + \cdots + |x_n|^2 \\
&= |x_1|^2 \mp 2|x_1|\|\vec{x}\|_2 + \|\vec{x}\|_2^2 + |x_2|^2 + \cdots + |x_n|^2 \\
&= (|x_1| \mp \|\vec{x}\|_2)^2 + |x_2|^2 + \cdots + |x_n|^2.
\end{aligned}
$$

In order to avoid roundoff-related cancellation error in the computation of $|x_1| \mp \|\vec{x}\|_2$, we choose the sign in the definition of k to be negative:

$$k := -\frac{x_1}{|x_1|}\|\vec{x}\|_2.$$

In this case,

$$|x_1 - k|^2 = \left| x_1 + \frac{x_1}{|x_1|}\|\vec{x}\|_2 \right|^2 = |x_1|^2 + 2|x_1|\|\vec{x}\|_2^2 + \|\vec{x}\|_2^2,$$

from which it follows that

$$\|\vec{x} - k\vec{e}_1\|_2^2 = 2\|\vec{x}\|_2^2 + 2|x_1|\|\vec{x}\|_2,$$

which can only be near zero when $\|\vec{x}\|_2 \approx 0$, in which case we would have a nearly singular matrix P.

Furthermore,

$$
\begin{aligned}
P &= \mathbf{I} - 2\vec{w}\vec{w}^H \\
&= \mathbf{I} - 2\frac{(\vec{x} - k\vec{e}_1)(\vec{x} - k\vec{e}_1)^H}{\|\vec{x} - k\vec{e}_1\|_2^2} \\
&= \mathbf{I} - 2\frac{(\vec{x} - k\vec{e}_1)(\vec{x} - k\vec{e}_1)^H}{2\|\vec{x}\|_2^2 + 2|x_1|\|\vec{x}\|_2} \\
&= \mathbf{I} - \frac{(\vec{x} - k\vec{e}_1)(\vec{x} - k\vec{e}_1)^H}{\|\vec{x}\|_2^2 + |x_1|\|\vec{x}\|_2},
\end{aligned}
$$

which we can now write as

$$P = \mathbf{I} - \beta\vec{u}\vec{u}^H,$$

with

$$\vec{u} := \vec{x} - k\vec{e}_1 := \begin{bmatrix} x_1 + \frac{x_1}{|x_1|}\|\vec{x}\|_2 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \beta := \frac{1}{\|\vec{x}\|_2^2 + |x_1|\|\vec{x}\|_2}.$$

An $n \times n$ matrix $\mathbf{A} = \mathbf{A}^{(0)}$ can be reduced step by step using these unitary Householder matrices $P^{(j)}$,

$$\mathbf{A}^{(j)} = P^{(j)}\mathbf{A}^{(j-1)},$$

into an upper triangular matrix

$$P^{(n-1)} \ldots P^{(1)}\mathbf{A}^{(0)} = \mathbf{A}^{(n-1)} = R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix}.$$

To do this, the $n \times n$ unitary matrix $P^{(1)}$ is determined so that

$$P^{(1)} a_1^{(0)} = k\vec{e}_1,$$

where $a_1^{(0)}$ denotes the first column of $\mathbf{A}^{(0)}$. From this step we obtain

$$\mathbf{A}^{(1)} = P^{(1)}\mathbf{A}^{(0)} = \begin{bmatrix} k & * & \dots & * \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & & \vdots & \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}.$$

If the matrix $\mathbf{A}^{(j-1)}$ obtained after $j-1$ steps has the form

$$\mathbf{A}^{(j-1)} = \left[\begin{array}{ccc|ccc} * & \dots & * & * & \dots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & * & * & \dots & * \\ \hline & & & a_{jj}^{(j-1)} & \dots & a_{jn}^{(j-1)} \\ & 0 & & \vdots & & \vdots \\ & & & a_{nj}^{(j-1)} & \dots & a_{nn}^{(j-1)} \end{array}\right] =: \left[\begin{array}{c|c} \mathbf{D} & \mathbf{B} \\ \hline 0 & \tilde{\mathbf{A}}^{(j-1)} \end{array}\right],$$

then we determine the $(n-j+1) \times (n-j+1)$ unitary matrix $\tilde{P}^{(j)}$ so that

$$\tilde{P}^{(j)}\vec{a}_j^{(j-1)} = \tilde{P}^{(j)} \begin{bmatrix} a_{jj}^{(j-1)} \\ \vdots \\ a_{nj}^{(j-1)} \end{bmatrix} = k \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{C}^{n-j+1}.$$

Using $\tilde{P}^{(j)}$ the desired $n \times n$ unitary matrix is constructed as

$$P^{(j)} := \left[\begin{array}{c|c} \mathbf{I}_{j-1} & \mathbf{0} \\ \hline \mathbf{0} & \tilde{P}^{(j)} \end{array}\right].$$

After forming $\mathbf{A}^{(j)} = P^{(j)}\mathbf{A}^{(j-1)}$, the elements $a_{ij}^{(j)}$ for $i > j$ are annihilated, and the rows above the horizontal line remain untouched. In this way an upper triangular matrix

$$R := \mathbf{A}^{(n-1)}$$

is obtained after $n-1$ steps.

On a computer, the transformation of a matrix by

$$\tilde{P}^{(j)} = \mathbf{I} - \beta_j \vec{u}_j \vec{u}_j^H$$

is carried out as follows:

$$\tilde{P}^{(j)}\tilde{\mathbf{A}}^{(j-1)} = \tilde{\mathbf{A}}^{(j-1)} - \vec{u}_j(\beta_j \vec{u}_j^H \tilde{\mathbf{A}}^{(j-1)}) = \tilde{\mathbf{A}}^{(j-1)} - \vec{u}_j \vec{v}_j^H,$$

with $\vec{v}_j^H := \beta_j \vec{u}_j^H \tilde{\mathbf{A}}^{(j-1)}$, that is, the vector $\vec{v}_j$ is computed first, and then $\tilde{\mathbf{A}}^{(j-1)}$ is modified as indicated.

The Householder reduction of an $n \times n$ matrix into triangular form requires about $2n^3/3$ operations. In this process one usually stores the data β_j and $\vec{u}_j$ so that the $n \times n$ unitary matrix

$$P = P^{(n-1)} \dots P^{(1)}$$

consisting of Householder matrices $P^{(j)}$, $j = 1, 2, \ldots, n-1$, can be inverted:

$$\begin{aligned} P^{-1} &= (P^{(n-1)} \ldots P^{(1)})^{-1} \\ &= (P^{(1)})^{-1} \ldots (P^{(n-1)})^{-1} \\ &= (P^{(1)})^{H} \ldots (P^{(n-1)})^{H} \\ &= (P^{(n-1)} \ldots P^{(1)})^{H} \\ &= P^{H}. \end{aligned}$$

Hence,

$$P\mathbf{A} = R,$$

or

$$\mathbf{A} = P^H R = QR, \quad Q := P^{-1} = P^H.$$

This is known as a *$QR-$decomposition* of the matrix $\mathbf{A}$ into a product of Q unitary and R upper triangular,

$$\mathbf{A} = QR.$$

Example 4.5.1.5. *We apply Householder transformations to the 3×3 matrix*

$$\mathbf{A} = \begin{bmatrix} 12 & 10 & 4 \\ 10 & 8 & -5 \\ 4 & -5 & 3 \end{bmatrix}$$

to produce a $QR-$decomposition,

$$\mathbf{A} = QR.$$

For the first step we have

$$\vec{a}_1^{(0)} = [12, 10, 4]^\top.$$

Hence we take

$$k = -\frac{12}{|12|}(2\sqrt{65}) = -2\sqrt{65}.$$

We take

$$\vec{w} = \frac{1}{30.8381} \begin{bmatrix} 12 + 2\sqrt{65} \\ 10 \\ 4 \end{bmatrix} = \begin{bmatrix} 0.9339 \\ 0.3320 \\ 0.1328 \end{bmatrix}.$$

Thus,

$$\begin{aligned} P^{(1)} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} 0.9339 \\ 0.3320 \\ 0.1328 \end{bmatrix} \cdot [0.9339, 0.3320, 0.1328] \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} 0.8721 & 0.3101 & 0.1240 \\ 0.3101 & 0.1103 & 0.0441 \\ 0.1240 & 0.0441 & 0.0176 \end{bmatrix} \\ &= \begin{bmatrix} -0.7442 & -0.6202 & -0.2481 \\ -0.6202 & 0.7795 & -0.0882 \\ -0.2481 & -0.0882 & 0.9647 \end{bmatrix}. \end{aligned}$$

Thus

$$\mathbf{A}^{(1)} = P^{(1)}\mathbf{A} = \begin{bmatrix} -0.7442 & -0.6202 & -0.2481 \\ -0.6202 & 0.7795 & -0.0882 \\ -0.2481 & -0.0882 & 0.9647 \end{bmatrix} \begin{bmatrix} 12 & 10 & 4 \\ 10 & 8 & -5 \\ 4 & -5 & 3 \end{bmatrix}$$
$$= \begin{bmatrix} -16.1245 & -11.1631 & -0.6202 \\ 0 & 0.4752 & -6.6428 \\ 0 & -8.0099 & 2.3429 \end{bmatrix}.$$

For the second step we have

$$\vec{a}_2^{(1)} = [0.4752, -8.0099]^\top.$$

Hence we take

$$k = -\frac{0.4752}{|0.4752|}(8.0240) = -8.0240.$$

We take

$$\vec{w} = \frac{1}{11.6788}\begin{bmatrix} 0.4752 + 8.0240 \\ -8.0099 \end{bmatrix} = \begin{bmatrix} 0.7277 \\ -0.6858 \end{bmatrix}.$$

Thus,

$$\tilde{P}^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2\begin{bmatrix} 0.7277 \\ -0.6858 \end{bmatrix} \cdot [0.7277, -0.6858]$$
$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2\begin{bmatrix} 0.5296 & -0.4991 \\ -0.4991 & 0.4704 \end{bmatrix}$$
$$= \begin{bmatrix} -0.0592 & 0.9982 \\ 0.9982 & 0.0592 \end{bmatrix}.$$

So

$$P^{(2)} := \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.0592 & 0.9982 \\ 0 & 0.9982 & 0.0592 \end{bmatrix},$$

and

$$R = \mathbf{A}^{(2)} = P^{(2)}\mathbf{A}^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.0592 & 0.9982 \\ 0 & 0.9982 & 0.0592 \end{bmatrix} \begin{bmatrix} -16.1245 & -11.1631 & -0.6202 \\ 0 & 0.4752 & -6.6428 \\ 0 & -8.0099 & 2.3429 \end{bmatrix}$$
$$= \begin{bmatrix} -16.1245 & -11.1631 & -0.6202 \\ 0 & -8.0240 & 2.7319 \\ 0 & 0 & -6.4921 \end{bmatrix}.$$

Moreover,

$$P = P^{(2)}P^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.0592 & 0.9982 \\ 0 & 0.9982 & 0.0592 \end{bmatrix} \begin{bmatrix} -0.7442 & -0.6202 & -0.2481 \\ -0.6202 & 0.7795 & -0.0882 \\ -0.2481 & -0.0882 & 0.9647 \end{bmatrix}$$
$$= \begin{bmatrix} -0.7442 & -0.6242 & -0.2481 \\ -0.2109 & -0.1342 & 0.9682 \\ -0.6338 & 0.7729 & -0.0309 \end{bmatrix}.$$

Thus we take

$$Q = P^H = \begin{bmatrix} -0.7442 & -0.2109 & -0.6338 \\ -0.6242 & -0.1342 & 0.7729 \\ -0.2481 & 0.9682 & -0.0309 \end{bmatrix},$$

and we see that

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 12 & 10 & 4 \\ 10 & 8 & -5 \\ 4 & -5 & 3 \end{bmatrix} \\ &= \begin{bmatrix} -0.7442 & -0.2109 & -0.6338 \\ -0.6242 & -0.1342 & 0.7729 \\ -0.2481 & 0.9682 & -0.0309 \end{bmatrix} \begin{bmatrix} -16.1245 & -11.1631 & -0.6202 \\ 0 & -8.0240 & 2.7319 \\ 0 & 0 & -6.4921 \end{bmatrix} \\ &= QR. \end{aligned}$$

Example 4.5.1.6. *We apply Householder transformations to the* 2×2 *matrix*

$$\mathbf{A} = \begin{bmatrix} 3 & 5 \\ 4 & -12 \end{bmatrix}$$

to produce a $QR-$*decomposition,*

$$\mathbf{A} = QR.$$

Note that we have $\vec{a}_1 = [3, 4]^\top$. *Hence we take*

$$k = -\frac{3}{|3|}(5) = -5.$$

We take

$$\vec{w} = \frac{1}{4\sqrt{5}} \begin{bmatrix} 3+5 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}.$$

Thus,

$$\begin{aligned} P &= \mathbf{I} - 2\vec{w}\vec{w}^H \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix} \cdot \begin{bmatrix} \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} \frac{4}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{bmatrix}. \end{aligned}$$

We find

$$\begin{aligned} R = P\mathbf{A} &= \begin{bmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} 3 & 5 \\ 4 & -12 \end{bmatrix} \\ &= \begin{bmatrix} -5 & \frac{33}{5} \\ 0 & -\frac{56}{5} \end{bmatrix}. \end{aligned}$$

Also,

$$Q = P^H = \begin{bmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{bmatrix},$$

and we see that

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} 3 & 5 \\ 4 & -12 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} -5 & \frac{33}{5} \\ 0 & -\frac{56}{5} \end{bmatrix} \\ &= QR.\end{aligned}$$

Note that the columns of Q are orthonormal. To solve $\mathbf{A}\vec{x} = \vec{b}$, we see

$$\mathbf{A}\vec{x} = QR\vec{x} = \vec{b} \implies R\vec{x} = Q^{-1}\vec{b} = P\vec{b},$$

which means one matrix–vector product and a back–substitution solve.

Also recall that Householder matrices help reduce roundoff error effects as compared to Gaussian elimination.

4.6. **Data Fitting.**

4.6.1. *The Data Fitting Problem.* In many applications we are concerned with determining the values of certain constants

$$x_1, x_2, \ldots, x_n.$$

Often it is difficult to measure the x_i, $i = 1, 2, \ldots, n$, directly. In such cases another more easily measurable quantity y is sampled, which depends in some way on the x_i and on further controllable experimental conditions z :

$$y = f(z; x_1, x_2, \ldots, x_n).$$

In order to determine the x_i, experiments are carried out under m different conditions $z_1, z_2, \ldots, z_m$ and the corresponding results

$$y_k = f(z_k; x_1, x_2, \ldots, x_n), \quad k = 1, 2, \ldots, m, \tag{4.6.1.1}$$

are measured. In general, at least n experiments, $m \geq n$, must be carried out so that the x_i can be uniquely determined. If $m > n$, however, then the equations (4.6.1.1) form an overdetermined system for the unknown parameters $x_1, x_2, \ldots, x_n$, which does not usually have a solution because the observed quantities y_i are perturbed by measurement errors. Consequently, instead of finding an exact solution to (4.6.1.1), the problem becomes one of finding the "best possible solution." Such a solution to (4.6.1.1) is taken to mean a set of values for the unknown parameters for which the expression

$$\sum_{k=1}^{m} (y_k - f_k(x_1, x_2, \ldots, x_n))^2$$

is minimized, where $f_k(\cdot) := f(z_k; \cdot)$.

Put $r_k := y_k - f_k(\cdot)$ for each $k = 1, 2, \ldots, m$. The r_k are called *residuals*, and the data fitting problem becomes one of minimizing

$$\sum_{k=1}^{m} \|r_k\|_2^2,$$

which in turn minimizes

$$\sum_{k=1}^{m} \|r_k\|_2,$$

for $\|r_k\|_2 \geq 0$.

If the functions $f_k(x_1, x_2, \ldots, x_n)$ have continuous partial derivatives in all of the variables x_i, $i = 1, 2, \ldots, n$, then we may give a necessary conditions for $\vec{x} := (x_1, x_2, \ldots, x_n)^\top$ to minimize the data fitting problem:

$$\frac{\partial}{\partial x_i} \sum_{k=1}^{m} (y_k - f_k(x_1, x_2, \ldots, x_n))^2 = 0, \quad i = 1, 2, \ldots, n.$$

These are called the *normal equations* for $\vec{x}$.

An important special case is the linear least squares problem, where all of the functions $f_k(x_1, x_2, \ldots, x_n)$ are linear in the parameters x_i. In this case there is an $m \times n$ matrix $\mathbf{A}$ with

$$\mathbf{A}\vec{x} = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{bmatrix}.$$

Putting $\vec{y} := (y_1, y_2, \ldots, y_m)^\top$, the normal equations reduce to a linear system

$$\begin{aligned} \nabla_x((\vec{y} - \mathbf{A}\vec{x})^\top(\vec{y} - \mathbf{A}\vec{x})) &= \nabla_x((\vec{y}^\top - (\mathbf{A}\vec{x})^\top)(\vec{y} - \mathbf{A}\vec{x})) \\ &= \nabla_x(\vec{y}^\top\vec{y} - \vec{y}^\top\mathbf{A}\vec{x} - (\mathbf{A}\vec{x})^\top\vec{y} + (\mathbf{A}\vec{x})^\top\mathbf{A}\vec{x}) \\ &= 2\mathbf{A}^\top\mathbf{A}\vec{x} - 2\mathbf{A}^\top\vec{y} = 0, \end{aligned}$$

or equivalently,

$$\begin{aligned} \frac{\partial}{\partial x_i} \sum_{k=1}^{m} (y_k - f_k(x_1, x_2, \ldots, x_n))^2 &= 2 \sum_{k=1}^{m} \left[\left(y_k - \sum_{j=1}^{m} a_{kj} x_j \right) (-a_{k_i}) \right] \\ &= 0. \end{aligned}$$

In any case, we get that

$$\mathbf{A}^\top\mathbf{A}\vec{x} = \mathbf{A}^\top\vec{y}$$

are the normal equations.

4.6.2. *Linear Least Squares: The Normal Equations.* In the following $\|\cdot\|$ will always denote the Euclidean norm

$$\|\vec{x}\| := \|\vec{x}\|_2 = \sqrt{\vec{x}^H\vec{x}}.$$

Let a real $m \times n$ matrix $\mathbf{A}$ and a vector $\vec{y} \in \mathbb{R}^m$ be given, and let

$$\|\vec{y} - \mathbf{A}\vec{x}\|_2^2 = (\vec{y} - \mathbf{A}\vec{x})^\top(\vec{y} - \mathbf{A}\vec{x}) \tag{4.6.2.1}$$

be minimized as a function of $\vec{x}$. We want to show that $\vec{x} \in \mathbb{R}^n$ is a solution to the normal equations

$$\mathbf{A}^\top\mathbf{A}\vec{x} = \mathbf{A}^\top\vec{y}$$

if and only if $\vec{x}$ is also a minimum point of (4.6.2.1).

Theorem 4.6.2.1. *The linear least squares problem*

$$\min_{\vec{x}\in\mathbb{R}^n} \|\vec{y} - \mathbf{A}\vec{x}\|$$

has at least one minimum point $\vec{x}_0$. If $\vec{x}_1$ is another minimum point, then $\mathbf{A}\vec{x}_0 = \mathbf{A}\vec{x}_1$. The residual $\vec{r} := \vec{y} - \mathbf{A}\vec{x}_0$ is uniquely determined and satisfies the equation

$$\mathbf{A}^\top \vec{r} = \vec{0}.$$

Every minimum point $\vec{x}_0$ is also a solution of the normal equations

$$\mathbf{A}^\top \mathbf{A}\vec{x} = \mathbf{A}^\top \vec{y}$$

and conversely.

Proof. We first show that every minimum point $\vec{x}_0$ is a solution to the normal equations and conversely. Let $R \subseteq \mathbb{R}^m$ be the linear subspace (column space) of the matrix $\mathbf{A}$:

$$R(\mathbf{A}) := \{\mathbf{A}\vec{x} : \vec{x} \in \mathbb{R}^n\},$$

which is spanned by the columns of $\mathbf{A}$, and let $R^\perp$ be its orthogonal complement

$$\begin{aligned} R^\perp &:= \{\vec{r} : \vec{r}^\top \vec{z} = \vec{0} \text{ for all } \vec{z} \in R(\mathbf{A})\} \\ &= \{\vec{r} : \vec{r}^\top \mathbf{A} = \vec{0}\}. \end{aligned}$$

Note that $\mathbb{R}^m = R \oplus R^\perp$. The vector $\vec{y} \in \mathbb{R}^m$ can be written uniquely in the form

$$\vec{y} = \vec{s} + \vec{r}, \quad \vec{s} \in R, \quad \vec{r} \in R^\perp,$$

and there is at least one $\vec{x}_0$ with

$$\mathbf{A}\vec{x}_0 = \vec{s}.$$

Now $\mathbf{A}^\top \vec{r} = (\vec{r}^\top \mathbf{A})^\top = \vec{0}^\top = \vec{0}$, so $\vec{x}_0$ satisfies

$$\mathbf{A}^\top \vec{y} = \mathbf{A}^\top(\vec{s} + \vec{r}) = \mathbf{A}^\top \vec{s} + \mathbf{A}^\top \vec{r} = \mathbf{A}^\top \vec{s} = \mathbf{A}^\top \mathbf{A}\vec{x}_0,$$

that is, $\vec{x}_0$ is a solution of the normal equations. In other words, the normal equations are solvable. Conversely, each solution $\vec{x}_1$ of the normal equations corresponds to a representation

$$\begin{gathered} \vec{y} = \vec{s} + \vec{r}, \quad \vec{s} := \mathbf{A}\vec{x}_1, \quad \vec{r} := \vec{y} - \mathbf{A}\vec{x}_1, \\ \vec{s} \in R, \quad \vec{r} \in R^\perp. \end{gathered}$$

Since this representation is unique, it follows that

$$\mathbf{A}\vec{x}_0 = \mathbf{A}\vec{x}_1$$

for all solutions $\vec{x}_0$, $\vec{x}_1$ of the normal equations.

We now show that each solution $\vec{x}_0$ of the normal equations is a minimum point for the problem

$$\min_{\vec{x}\in\mathbb{R}^n} \|\vec{y} - \mathbf{A}\vec{x}\|.$$

To see this, let $\vec{x} \in \mathbb{R}^n$ be arbitrary, and set

$$\vec{z} := \mathbf{A}\vec{x} - \mathbf{A}\vec{x}_0, \quad \vec{r} := \vec{y} - \mathbf{A}\vec{x}_0.$$

Now $\vec{z} \in R^\perp$, so that $\vec{r}^\top \vec{z} = \vec{0}$. Thus,

$$\begin{aligned} \|\vec{y} - \mathbf{A}\vec{z}\|^2 &\le \|(\vec{y} - \mathbf{A}\vec{x}_0) + (\mathbf{A}\vec{x}_0 - \mathbf{A}\vec{x})\|^2 \\ &= \|\vec{r} - \vec{z}\|^2 \end{aligned}$$

$$
\begin{aligned}
&= (\vec{r} - \vec{z})^\top (\vec{r} - \vec{z}) \\
&= (\vec{r}^\top - \vec{z}^\top)(\vec{r} - \vec{z}) \\
&= \vec{r}^\top \vec{r} - \vec{r}^\top \vec{z} - \vec{z}^\top \vec{r} - \vec{z}^\top \vec{z} \\
&= \|\vec{r}\|^2 + \|\vec{z}\|^2 - \vec{r}^\top \vec{z} - (\vec{r}^\top \vec{z})^\top \\
&= \|\vec{r}\|^2 + \|\vec{z}\|^2 \geq \|\vec{r}\|^2 \\
&= \|\vec{y} - \mathbf{A}\vec{x}_0\|^2,
\end{aligned}
$$

that is, $\vec{x}_0$ is a minimum point of the problem

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{y} - \mathbf{A}\vec{x}\|.$$

Since we have shown that the normal equations are solvable, we have shown the existence of a solution to the linear least squares problem.

This completes the proof. □

If the columns of $\mathbf{A}$ are linearly independent, that is, if $\vec{x} \neq \vec{0}$ implies $\mathbf{A}\vec{x} \neq \vec{0}$, then the matrix $\mathbf{A}^\top \mathbf{A}$ is positive definite, and thus, nonsingular. If this were not the case, then there would exist $\vec{x} \neq \vec{0}$ satisfying $\mathbf{A}^\top \mathbf{A}\vec{x} = \vec{0}$, from which

$$0 = \vec{x}^\top (\mathbf{A}^\top \mathbf{A}\vec{x}) = (\mathbf{A}\vec{x})^\top (\mathbf{A}\vec{x}) = \|\mathbf{A}\vec{x}\|^2$$

would yield a contradiction, for $\mathbf{A}\vec{x} \neq \vec{0}$. Therefore the normal equations

$$\mathbf{A}^\top \mathbf{A}\vec{x} = \mathbf{A}^\top \vec{y}$$

have a unique solution

$$\vec{x} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \vec{y},$$

which may be computed using for instance the Choleski factorization of $\mathbf{A}^\top \mathbf{A}$.

4.6.3. *The Use of Orthogonalization in Solving Linear Least–Squares Problems.* The linear least–squares problem of determining an $\vec{x} \in \mathbb{R}^n$ that minimizes

$$\|\vec{y} - \mathbf{A}\vec{x}\|, \quad \mathbf{A} \in \mathrm{Mat}(m, n), \quad m \geq n,$$

can be solved using the orthogonalization techniques ($QR-$factorization, Gram–Schmidt) discussed in the previous section. Let the matrix $\mathbf{A} =: \mathbf{A}^{(0)}$ and the vector $\vec{y} =: \vec{y}^{(0)}$ be transformed by a sequence of Householder transformations $P^{(i)}$, $\mathbf{A}^{(i)} = P^{(i)} \mathbf{A}^{(i-1)}$, $\vec{y} = P^{(i)} \vec{y}^{(i-1)}$, $i = 2, \ldots, n$. The final matrix $\mathbf{A}^{(n)}$ has the form

$$\mathbf{A}^{(n)} = \begin{bmatrix} R \\ \mathbf{0} \end{bmatrix}, \quad \text{with} \quad R =: \begin{bmatrix} r_{11} & \cdots & r_{nn} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix}, \tag{4.6.3.1}$$

since $m \geq n$ (here $\mathbf{R}$ is $n \times n$ and $\mathbf{0}$ is $(m-n) \times (m-n)$.). Let the vector $\vec{h} := \vec{y}^{(n)}$ be partitioned correspondingly:

$$\vec{h} = \begin{bmatrix} \vec{h}_1 \\ \vec{h}_2 \end{bmatrix}, \quad \vec{h}_1 \in \mathbb{R}^n, \quad \vec{h}_2 \in \mathbb{R}^{m-n}. \tag{4.6.3.2}$$

Note that the matrix $P = P^{(n)} \dots P^{(1)}$ is a product of unitary matrices and thus is unitary itself:

$$P^H P = \left(P^{(1)}\right)^H \dots \left(P^{(n)}\right)^H P^{(n)} \dots P^{(1)} = \mathbf{I},$$

and satisfies

$$\mathbf{A}^{(n)} = P\mathbf{A}, \quad \vec{h} = P\vec{y}.$$

Recall that unitary matrices $\mathbf{U}$ leave the Euclidean norm $\|\vec{x}\|_2$ of a vector $\vec{x}$ invariant:

$$\|\mathbf{U}\vec{x}\|_2^2 = (\mathbf{U}\vec{x})^H(\mathbf{U}\vec{x}) = \vec{x}^H\mathbf{U}^H\mathbf{U}\vec{x} = \vec{x}^H\vec{x} = \|\vec{x}\|_2.$$

Thus,

$$\|\vec{y} - \mathbf{A}\vec{x}\| = \|P(\vec{y} - \mathbf{A}\vec{x})\| = \|\vec{y}^{(n)} - \mathbf{A}^{(n)}\vec{x}\|.$$

However, from (4.6.3.1) and (4.6.3.2), the vector $\vec{y}^{(n)} - \mathbf{A}^{(n)}\vec{x}$ has the structure

$$\vec{y}^{(n)} - \mathbf{A}^{(n)}\vec{x} = \begin{bmatrix} \vec{h}_1 \\ \vec{h}_2 \end{bmatrix} - \begin{bmatrix} R \\ \mathbf{0} \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{h}_1 - R\vec{x} \\ \vec{h}_2 \end{bmatrix}.$$

Hence, $\|\vec{y} - \mathbf{A}\vec{x}\|$ is minimized if $\vec{x}$ is chosen so that

$$\vec{h}_1 = R\vec{x}.$$

The matrix R, being upper triangular, is nonsingular if and only if the columns of $\mathbf{A}$ are linearly independent ($\mathbf{A}$ has full rank), for then in this situation R has nonzero diagonal entries and thus a nonzero determinant. Furthermore, $\mathbf{A}\vec{z} = \vec{0}$ for a vector $\vec{z} \in \mathbb{R}^n$ if and only if

$$P\mathbf{A}\vec{z} = \vec{0},$$

and since $P\mathbf{A} = R$, $P\mathbf{A}\vec{z} = \vec{0}$ if and only if

$$R\vec{z} = \vec{0}.$$

If we assume that the columns of $\mathbf{A}$ are linearly independent, then

$$\vec{h}_1 = R\vec{x},$$

which is a triangular system, can be solved uniquely for $\vec{x}$ (specifically, $\vec{x} = R^{-1}\vec{h}_1$.). This $\vec{x}$ is, moreover, the unique minimum point for the given least–squares problem. (Note that if the columns of $\mathbf{A}$ are linearly dependent, then, although the value of $\min_{\vec{x}\in\mathbb{R}^n} \|\vec{y} - \mathbf{A}\vec{x}\|$ is uniquely determined, there are many minimum points $\vec{x}$.)

In the case that $\vec{h}_1 = R\vec{x}$, then the size of the residual is seen to be

$$\|\vec{y} - \mathbf{A}\vec{x}\| = \|\vec{y}^{(n)} - \mathbf{A}^{(n)}\vec{x}\| = \|\vec{h}_2\|.$$

Lastly, note that instead of using Householder matrices, the Gram–Schmidt technique (with reorthogonalization) can be used to obtain the solution.

4.6.4. *The Pseudoinverse of a Matrix.* For any arbitrary (complex) $m \times n$ matrix $\mathbf{A}$ there is an $n \times m$ matrix $\mathbf{A}^+$, called the *pseudoinverse* (or *Moore-Penrose inverse*) of $\mathbf{A}$. It is associated with $\mathbf{A}$ in a natural fashion and agrees with the inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$ in the case $m = n$ and $\mathbf{A}$ is nonsingular.

Denote by $R(\mathbf{A})$ the range space of $\mathbf{A}$ and $N(\mathbf{A})$ the null space of $\mathbf{A}$,

$$R(\mathbf{A}) := \{\mathbf{A}\vec{x} \in \mathbb{C}^m : \vec{x} \in \mathbb{C}^n\},$$
$$N(\mathbf{A}) := \{\vec{x} \in \mathbb{C}^n : \mathbf{A}\vec{x} = \vec{0}\},$$

together with their orthogonal complement spaces $R(\mathbf{A})^\perp \in \mathbb{C}^m$, $N(\mathbf{A})^\perp \in \mathbb{C}^n$. Further, let P be the $n \times n$ matrix which projects $\mathbb{C}^n$ onto $N(\mathbf{A})^\perp$, and let $\bar{P}$ be the $m \times m$ matrix which projects $\mathbb{C}^m$ onto $R(\mathbf{A})$:

$$P\vec{x} = \vec{0} \iff x \in N(\mathbf{A}), \quad P = P^H = P^2,$$
$$\bar{P}\vec{y} = \vec{y} \iff \vec{y} \in R(\mathbf{A}), \quad \bar{P} = \bar{P}^H = \bar{P}^2.$$

For each $\vec{y} \in R(\mathbf{A})$ there is a uniquely determined $\vec{x}_1 \in N(\mathbf{A})^\perp$ satisfying $\mathbf{A}\vec{x}_1 = \vec{y}$, that is, there is a well–defined mapping $f : R(\mathbf{A}) \to \mathbb{C}^n$ with

$$\mathbf{A}f(\vec{y}) = \vec{y}, \quad f(\vec{y}) \in N(\mathbf{A})^\perp \quad \text{for all } \vec{y} \in R(\mathbf{A}).$$

For, given $\vec{y} \in R(\mathbf{A})$, there is an $\vec{x}$ which satisfies $\vec{y} = \mathbf{A}\vec{x}$. Hence,

$$\vec{y} = \mathbf{A}(P\vec{x} + (\mathbf{I} - P)\vec{x}) = \mathbf{A}P\vec{x} = \mathbf{A}\vec{x}_1,$$

where $\vec{x}_1 := P\vec{x} \in N(\mathbf{A})^\perp$, since $(\mathbf{I} - P)\vec{x} \in N(\mathbf{A})$. Furthermore, if $\vec{x}_1, \vec{x}_2 \in N(\mathbf{A})^\perp$, $\mathbf{A}\vec{x}_1 = \mathbf{A}\vec{x}_2 = \vec{y}$, it follows that

$$\vec{x}_1 - \vec{x}_2 \in N(\mathbf{A}) \cap N(\mathbf{A})^\perp = \{\vec{0}\},$$

which implies that $\vec{x}_1 = \vec{x}_2$. Note f is linear.

The composite mapping $f \circ \bar{P} : \vec{y} \in \mathbb{C}^m \to f(\bar{P}(\vec{y})) \in \mathbb{C}^n$ is well–defined and linear, since $\bar{P}\vec{y} \in R(\mathbf{A})$. Hence, it is represented by an $n \times m$ matrix, which is precisely $\mathbf{A}^+$, the pseudoinverse of $\mathbf{A}$: $\mathbf{A}^+\vec{y} = f(\bar{P}(\vec{y}))$ for all $\vec{y} \in \mathbb{C}^m$.

We get the following properties for the pseudoinverse $\mathbf{A}^+$.

Theorem 4.6.4.1 (Properties of the Pseudoinverse)**.** *Let $\mathbf{A}$ be an $m \times n$ matrix. The pseudoinverse $\mathbf{A}^+$ is an $n \times m$ matrix satisfying:*

(1) $\mathbf{A}^+\mathbf{A} = P$ is the orthogonal projector $P : \mathbb{C}^n \to N(\mathbf{A})^\perp$ and $\mathbf{A}\mathbf{A}^+ = \bar{P}$ is the orthogonal projector $\bar{P} : \mathbb{C}^m \to R(\mathbf{A})$.

(2) The following formulas hold:

(a) $\mathbf{A}^+\mathbf{A} = (\mathbf{A}^+\mathbf{A})^H$;

(b) $\mathbf{A}\mathbf{A}^+ = (\mathbf{A}\mathbf{A}^+)^H$;

(c) $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$;

(d) $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$.

Proof. According to the definition of $\mathbf{A}^+$,

$$\mathbf{A}^+\mathbf{A}\vec{x} = f(\bar{P}(\mathbf{A}\vec{x})) = f(\mathbf{A}\vec{x}) = P\vec{x}$$

for all $\vec{x}$, so that $\mathbf{A}^+\mathbf{A} = P$. Since $P^H = P$, part (a) is satisfied.

Furthermore, from the definition of f,

$$\mathbf{A}\mathbf{A}^+ = \mathbf{A}(f(\bar{P}\vec{y})) = \bar{P}\vec{y}$$

for all $\vec{y} \in \mathbb{C}^m$. Thus, $\mathbf{A}\mathbf{A}^+ = \bar{P} = \bar{P}^H$. Since $\bar{P}^H = \bar{P}$, part (b) follows as well.

Finally, for all $\vec{x} \in \mathbb{C}^n$,

$$(\mathbf{A}^+)\mathbf{A}\vec{x} = \bar{P}\mathbf{A}\vec{x} = \mathbf{A}\vec{x}$$

according to the definition of P, and for all $\vec{y} \in \mathbb{C}^m$,

$$\mathbf{A}^+(\mathbf{A}\mathbf{A}^+)\vec{y} = \mathbf{A}^+\bar{P}\vec{y} = f(\bar{P}^2\vec{y}) = f(\bar{P}\vec{y}) = \mathbf{A}^+\vec{y}.$$

Hence, (c) and (d) hold.

This completes the proof. □

The properties (2a–d) of (4.6.4.1) uniquely characterize $\mathbf{A}^+$.

Theorem 4.6.4.2. *If Z is a matrix satisfying*

(1) $Z\mathbf{A} = (Z\mathbf{A})^H$;
(2) $\mathbf{A}Z = (\mathbf{A}Z)^H$;
(3) $\mathbf{A}Z\mathbf{A} = \mathbf{A}$;
(4) $Z\mathbf{A}Z = Z$;

then $Z = \mathbf{A}^+$.

Proof. We have the following chain of equalities:

$$\begin{aligned}
Z &= Z\mathbf{A}Z \\
&= Z(\mathbf{A}\mathbf{A}^+\mathbf{A})\mathbf{A}^+(\mathbf{A}\mathbf{A}^+\mathbf{A})Z \\
&= (Z\mathbf{A})^H(\mathbf{A}^+\mathbf{A})^H\mathbf{A}^+(\mathbf{A}\mathbf{A}^+)^H(\mathbf{A}Z)^H \\
&= \mathbf{A}^H Z^H \mathbf{A}^H \mathbf{A}^{+H} \mathbf{A}^+ \mathbf{A}^{+H} \mathbf{A}^H Z^H \mathbf{A}^H \\
&= (\mathbf{A}^H Z^H \mathbf{A}^H)\mathbf{A}^{+H} \mathbf{A}^+ \mathbf{A}^{+H} (\mathbf{A}^H Z^H \mathbf{A}^H) \\
&= (\mathbf{A}Z\mathbf{A})^H \mathbf{A}^{+H} \mathbf{A}^+ \mathbf{A}^{+H} (\mathbf{A}Z\mathbf{A})^H \\
&= \mathbf{A}^H \mathbf{A}^{+H} \mathbf{A}^+ \mathbf{A}^{+H} \mathbf{A}^H \\
&= (\mathbf{A}^+\mathbf{A})^H \mathbf{A}^+ (\mathbf{A}\mathbf{A}^+)^H \\
&= \mathbf{A}^+\mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{A}^+ \\
&= \mathbf{A}^+.
\end{aligned}$$

This completes the proof. □

We also have the following.

Corollary 4.6.4.3. *For all matrices* $\mathbf{A}$,

$$\mathbf{A}^{++} = \mathbf{A}$$

and

$$(\mathbf{A}^+)^H = (\mathbf{A}^H)^+.$$

Proof. This holds because $Z := \mathbf{A}$ (respectively $Z := (\mathbf{A}^+)^H$) has the properties of $(\mathbf{A}^+)^+$ (respectively $(\mathbf{A}^H)^+$) in (4.6.4.2). □

The pseudoinverse is often used to give an elegant representation of the solution to the least–squares problem

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{y} - \mathbf{A}\vec{x}\|_2.$$

Theorem 4.6.4.4 (Solution to Least–Squares Problem Using Pseudoinverse)**.** *The vector $\bar{x} := \mathbf{A}^+\vec{y}$ satisfies:*
(1) $\|\mathbf{A}\vec{x} - \vec{y}\|_2 \geq \|\mathbf{A}\bar{x} - \vec{y}\|_2$ for all $\vec{x} \in \mathbb{C}^n$;
(2) $\|\mathbf{A}\vec{x} - \vec{y}\|_2 = \|\mathbf{A}\bar{x} - \vec{y}\|_2$ and $\vec{x} \neq \bar{x}$ imply $\|\vec{x}\|_2 > \|\bar{x}\|_2$.

In other words, $\bar{x} := \mathbf{A}^+\vec{y}$ is the minimum point of the least–squares problem that has the smallest Euclidean norm, in the case that the least squares problem does not have a unique minimum point.

Proof. From (4.6.4.1), $\mathbf{A}\mathbf{A}^+$ is the orthogonal projector on $R(\mathbf{A})$. Thus, for all $\vec{x} \in \mathbb{C}^n$, it follows that

$$\mathbf{A}\vec{x} - \vec{y} = \vec{u} - \vec{v},$$
$$u := \mathbf{A}(\vec{x} - \mathbf{A}^+\vec{y}) \in R(\mathbf{A}), \quad \vec{v} := (\mathbf{I} - \mathbf{A}\mathbf{A}^+)\vec{y} = \vec{y} - \mathbf{A}\bar{x} \in R(\mathbf{A})^\perp.$$

Consequently, for all $\vec{x} \in \mathbb{C}^n$,

$$\begin{aligned}\|\mathbf{A}\vec{x} - \vec{y}\|_2^2 &= \|\vec{u} - \vec{v}\|_2^2 \\ &= \vec{u}^H\vec{u} - \vec{u}^H\vec{v} - \vec{v}^H\vec{u} + \vec{v}^H\vec{v} \\ &= \|\vec{u}\|_2^2 + \|\vec{v}\|_2^2 \\ &\geq \|\vec{v}\|_2^2 \\ &= \|\mathbf{A}\bar{x} - \vec{y}\|_2^2,\end{aligned}$$

and $\|\mathbf{A}\vec{x} - \vec{y}\|_2$ holds precisely if

$$\mathbf{A}\vec{x} = \mathbf{A}\mathbf{A}^+\vec{y}.$$

Now, $\mathbf{A}^+\mathbf{A}$ is the orthogonal projector on $N(\mathbf{A})^\perp$. Therefore, for all $\vec{x}$ such that $\mathbf{A}\vec{x} = \mathbf{A}\mathbf{A}^+\vec{y}$,

$$\vec{x} = \vec{u}_1 + \vec{v}_1,$$
$$\vec{u}_1 := \mathbf{A}^+\mathbf{A}\vec{x} = \mathbf{A}^+\mathbf{A}\mathbf{A}^+\vec{y} = \mathbf{A}^+\vec{y} = \bar{x} \in N(\mathbf{A})^\perp,$$
$$\vec{v}_1 := \vec{x} - \vec{u}_1 = \vec{x} - \bar{x} \in N(\mathbf{A}).$$

From this observation it follows that

$$\begin{aligned}\|\vec{x}\|_2^2 &= \|\vec{u}_1 + \vec{v}_1\|_2^2 \\ &= \vec{u}^H\vec{u} + \vec{u}^H\vec{v} + \vec{v}^H\vec{u} + \vec{v}^H\vec{v} \\ &= \|\vec{u}\|_2^2 + \|\vec{v}\|_2^2 \\ &\geq \|\vec{u}\|_2^2 \\ &= \|\bar{x}\|_2^2\end{aligned}$$

for all $\vec{x} \in \mathbb{C}^n$ satisfying $\vec{x} - \bar{x} \neq \vec{0}$ and $\|\mathbf{A}\vec{x} - \vec{y}\|_2 = \|\mathbf{A}\bar{x} - \vec{y}\|_2$. □

If the $m \times n$ matrix $\mathbf{A}$ with $m \geq n$ has maximal rank, that is, $\text{rank}(\mathbf{A}) = n$ (which occurs if and only if the columns of $\mathbf{A}$ are linearly independent), then there is an explicit formula for $\mathbf{A}^+$: It is easily verified that the matrix

$$Z := (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$$

has all properties given in (4.6.4.2) characterizing the pseudoinverse $\mathbf{A}^+$ so that

$$\mathbf{A}^+ = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H.$$

By means of the $QR-$decomposition of $\mathbf{A}$, $\mathbf{A} = QR$, this formula for $\mathbf{A}^+$ is equivalent to

$$\begin{aligned}\mathbf{A}^+ &= ((QR)^H(QR))^{-1}(QR)^H \\ &= (R^HQ^HQR)^{-1}R^HQ^H \\ &= (R^HR)^{-1}R^HQ^H \\ &= R^{-1}(R^H)^{-1}R^HQ^H \\ &= R^{-1}Q^H.\end{aligned}$$

This allows a numerically more stable computation of the pseudoinverse $\mathbf{A}^+ = R^{-1}Q^H$.

If $m < n$ and $\operatorname{rank}(\mathbf{A}) = m$ then because of $(\mathbf{A}^+)^H = (\mathbf{A}^H)^+$, the pseudoinverse $\mathbf{A}^+$ is given by

$$\mathbf{A}^+ = Q(R^H)^{-1},$$

if the matrix $\mathbf{A}^H$ has the $QR-$decomposition $\mathbf{A}^H = QR$.

For general $m \times n$ matrices $\mathbf{A}$ of arbitrary rank, the pseudoinverse $\mathbf{A}^+$ can be computed by means of the *singular value decomposition* of $\mathbf{A}$.

References

1. J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, third ed., Texts in Applied Mathematics, vol. 12, Springer-Verlag, New York, 2002, Translated from the German by R. Bartels, W. Gautschi and C. Witzgall. MR 1923481