MA637 - Numerical Analysis and Computing Lecture Notes

Abhinav Jha
Indian Institute of Technology, Gandhinagar
Winter Semester 2024-2025

Preface

These notes are designed to provide a structured and comprehensive understanding of the course content. They will cover key topics, concepts, and computational techniques that are fundamental to numerical analysis. Please note that this is the first iteration (Version 1.0.0) of the notes and hence there is a chance that some of the content is incorrect. If you find some flaws, please email me at `abhinav.jha@iitgn.ac.in`.

Time. Space. Reality. It's more than a linear path. It's a prism of endless possibility,
where a single choice can branch out into infinite realities,
creating alternate worlds from the ones you know.
I am The Watcher. I am your guide through these vast new realities.
Follow me, and ponder the question... What if?

—Watcher

Contents

Chapter 1

Interpolation

Interpolation has various definitions depending on the search engine. For example, *Wikipedia* states,

"Interpolation is a type of estimation, a method of constructing (finding) new data points based on the range of a discrete set of known data points."

Blackphoto says,

"It is a technique used by digital scanners, cameras, and printers to increase the size of an image in pixels by averaging the colour and brightness values of surrounding pixels."

One can see such an example in image processing. A rather famous (or infamous) example is the *Ecce Homo* painting (see Fig. 1 (left)). This is a fresco painting painted in 1930 by the Spanish painter Elías García Martínez depicting Jesus Christ. With wear and tear, the painting got degraded, and in 2012, an 81-year-old lady, Cecilia Giménez "tried" to restore it (see Fig. 1 (right)); as we can see, it is not very good, and hence it was named *Ecce Mono*. We can get much better results with modern image processing techniques (which inherently use a form of interpolation).

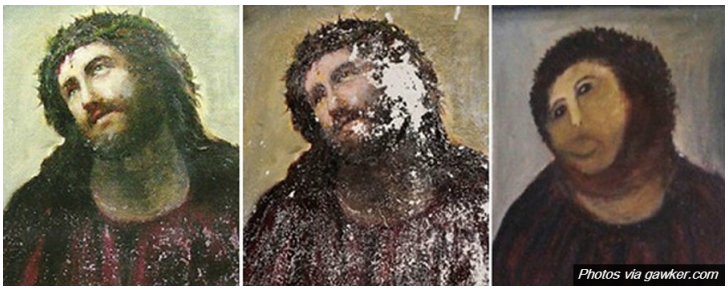

Figure 1.1: Elías García Martínez, Ecce Homo: The leftmost photograph, taken in 2010, shows some initial flaking of the paintwork. The central photograph was taken in July 2012, just a month before the attempted restoration, showing the extent of damage and deterioration. The rightmost photograph documents the artwork following Giménez's efforts to repair it.

In interpolation, we try to approximate general functions by a "simple" class of functions. In analysis, a powerful result connects the continuous functions and polynomial approximation:

the Weierstrass Approximation theorem given by Karl Weierstrass.

Figure 1.2: Karl Weierstrass: 31 October 1815-19 February 1897

Theorem 1.1. *[1, Theorem 5.4.14]* **(Weierstrass Approximation Theorem)** *Let $f \in \mathcal{C}[a,b]$. Then for each $\varepsilon > 0$ there exists a polynomial $p(x)$ with the property that*

$$|f(x) - p(x)| < \varepsilon \qquad \text{for all} \quad x \in [a,b].$$

This theorem is important because polynomials have excellent differentiation and integration properties as their derivatives and integrals are polynomials. Another interpretation of Theorem 1 is that given a continuous function on a closed and bounded interval, there exists a polynomial, i.e. , as "close" to the given function as desired.

But in analysis, there exists one more kind of polynomial approximation, and that is the Taylor's theorem

Theorem 1.2. *[1, Theorem 6.4.1]* **(Taylor's Theorem)** *Suppose $f \in \mathcal{C}^n[a,b]$ and $f^{(n+1)}$ exists on $[a,b]$ and $x_0 \in [a,b]$. For every $x \in [a,b]$ there exists a number $\xi(x) \in [x_0, x]$ with*

$$f(x) = P_n(x) + R_n(x),$$

where $P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$ and $R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}$.

Figure 1.3: Brook Taylor: 18 August 1685-29 December 1731

There are two issues here:

1. We need to know the higher derivatives of $f(x)$.

2. This is a *local* approximation, i.e., the approximation is excellent near x_0 but we need certain global approximation. For example, if we do the Taylor series expansion for $\exp(x)$ around zero, then it becomes worse as we move away from zero (see Fig. 1.4).

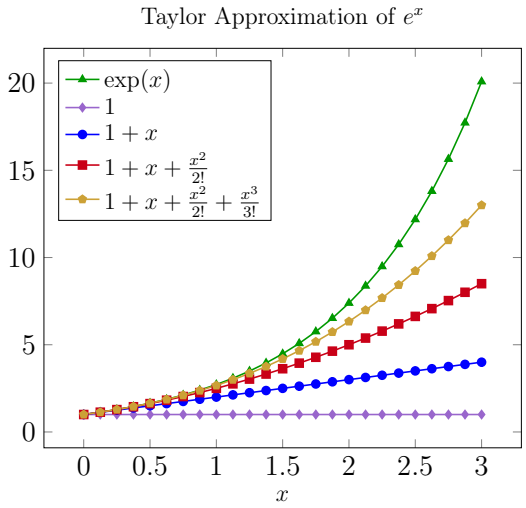

Figure 1.4: Taylor polynomials for exponential function approximated at $x = 0$.

However, it should be noted that the Taylor theorem is still a powerful result whose main purpose is the derivation of numerical techniques and error estimation.

1.1 Polynomial Interpolation

Suppose we have a finite set of data points f_i associated with parameters x_i. We want to depict these data points as a function $f(x)$ with the property that $f(x_i) = f_i$. This is clearly not well-defined since there are many such functions. But if we restrict to finite-dimensional spaces (such as polynomials), then we can define such functions, or to be more precise, the process is well-defined.

We first start with the idea of polynomial interpolation. Polynomials representing an unknown functional dependence of the discrete set of data points are called *interpolants*. The main problem that we want to tackle with interpolation is:

Problem: Given a set of $(n+1)$ data points say $\{(x_i, f_i)\}_{i=0}^{n}$ find a polynomial $p_n(x)$ of degree n satisfying

$$p_n^{\mathbb{V}}(x_i) = f_i \qquad \text{for all} \quad i = 0, 1, \ldots, n.$$

Now the general form of a polynomial $p_n^{\mathbb{V}}(x)$ is given by

$$p_n^{\mathbb{V}}(x) = \sum_{i=0}^{n} c_i x^{n-i} := c_0 x^n + c_1 x^{n-1} + \cdots + c_n,$$

for coefficients $c_i \in \mathbb{R}$. Since each polynomial of degree n can be determined by $(n+1)$ coefficients, we can re-write the above problem as solving the following system of equations:

$$c_0 x_i^n + c_1 x_i^{n-1} + \cdots + c_{n-1} x_i + c_n = f_i \qquad i = 0, 1, 2, \ldots, n,$$

or in the matrix form

$$\mathbf{V}\mathbf{c} = \mathbf{f} \tag{1.1}$$

where

$$\mathbf{V} = \begin{bmatrix} x_0^n & x_0^{n-1} & \dots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \dots & x_1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_n^n & x_n^{n-1} & \dots & x_n & 1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}, \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix}.$$

This system has a unique solution if $\mathbf{V}$ is invertible [5, Theorem 3.10], which is equivalent to saying that $\det(\mathbf{V}) \neq 0$. This matrix $\mathbf{V}$ is called as the *Vandermonde matrix* and it's determinant is given by

$$\det(\mathbf{V}) = \prod_{i=0}^{n-1} \prod_{j=i+1}^{n} (x_i - x_j).$$

Figure 1.5: Alexandre-Théophile Vandermonde: 28 February 1735 – 1 January 1796

This determinant is non-zero if we have distinct points. Hence, from now on, we assume we have $(n+1)$ distinct points.

The algorithm for using the polynomial interpolation using Vandermonde matrix for finding solution at a given point x_{eval} is given in Algorithm 1.

Note that we have introduced the notation $p_n^{\mathbb{V}}(x)$ only to denote the polynomial $p_n(x)$ computed using the Vandermonde matrix.

1.1.1 Drawbacks

Even though Eq. (1.1) has a perfect mathematical solution, computationally, it is not that good. The reason being Vandermonde matrices are ill-conditioned[1] (we will do conditioning of a system in the following chapters). The matrix $\mathbf{V}$ has a large condition number leading to inaccurate solutions.

To understand why the Vandermonde matrix is ill-conditioned for large n, we can plot x^k for $0 \leq k \leq n$ in $[0, 1]$ (see Fig. 1.6). Even though x^k are distinct for larger k, they tend to look the same. As a result, it is harder to identify projections of a particular polynomial $p_n^{\mathbb{V}}(x)$ into the nearly collinear basis of monomials x^k for large k.

[1] **Ill-Conditioned System**: In numerical analysis, the condition number of a function quantifies the extent to which the output can change in response to small variations in the input. It measures a function's sensitivity to input changes or errors, indicating how much an input error can propagate into the output. A problem with a low condition number is said to be *well-conditioned* while a problem with a high condition number is said to be *ill-conditioned*.

Algorithm 1 Vandermonde Interpolation

Given: Data sets $\{(x_i, f_i)\}_{i=0}^{n}$, Evaluation point x_{eval}.
Find: Interpolated polynomial $p_n^{\mathbb{V}}(x_{\text{eval}})$.

Step 1: Compute Vandermonde Matrix
Initialize an empty Vandermonde matrix $\mathbf{V}$ of size $(n+1) \times (n+1)$
for $i = 0$ **to** n **do**
  for $j = 0$ **to** n **do**
    $\mathbf{V}_{i,j} = x_i^{(n-j)}$
  end for
end for

Step 2: Solve the System of Linear Equations
Solve the system $\mathbf{V} \cdot \mathbf{c} = \mathbf{f}$ to get coefficient vector $\mathbf{c}$

Step 3: Evaluate the Vandermonde Polynomial $p_n^{\mathbb{V}}(x)$ at x_{eval}
Initialize $p_n^{\mathbb{V}}(x_{\text{eval}}) = 0$
for $i = 0$ **to** n **do**
  $p_n^{\mathbb{V}}(x_{\text{eval}}) = p_n^{\mathbb{V}}(x_{\text{eval}}) + c_i \cdot x_{\text{eval}}^{(n-i)}$
end for

---

return $p_n^{\mathbb{V}}(x_{\text{eval}})$

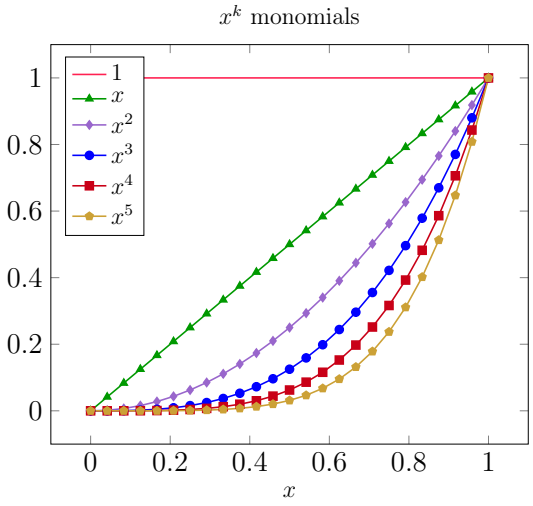

Figure 1.6: Monomial basis x^k for $k = 0, 1, \ldots, 5$.

1.2 Lagrange Interpolation

After examining how unstable polynomial interpolation is, we need to develop more stable methods. One of the most known methods is the Lagrange interpolation. The formula was first published by Waring in 1779, rediscovered by Euler in 1783, and published by Lagrange in 1795 (Jeffreys & Jeffreys, 1988).

Let us start with a basic example of two points (x_0, f_0) and (x_1, f_1), then we define

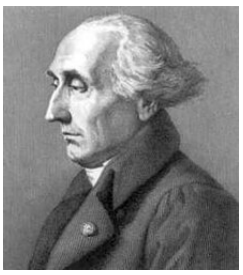

Figure 1.7: Joseph-Louis Lagrange: 25 January 1736-10 April 1813

functions:

$$\mathbb{L}_0(x) = \frac{x - x_1}{x_0 - x_1} \qquad \text{and} \qquad \mathbb{L}_1(x) = \frac{x - x_0}{x_1 - x_0}. \tag{1.2}$$

Then, a linear interpolating polynomial passing through the above points is given by:

$$p_1(x) = \mathbb{L}_0(x)f_0 + \mathbb{L}_1(x)f_1 = \frac{x - x_1}{x_0 - x_1}f_0 + \frac{x - x_0}{x_1 - x_0}f_1,$$

as $\mathbb{L}_0(x_0) = 1; \mathbb{L}_0(x_1) = 0; \mathbb{L}_1(x_0) = 0;$ and $\mathbb{L}_1(x_1) = 1$, we have $p_1(x_0) = f_0$ and $p_1(x_1) = f_1$. This polynomial is called the *Lagrange linear interpolating polynomial*. In fact, this is a unique polynomial. Fig. 1.8 shows $\mathbb{L}_0(x)$ and $\mathbb{L}_1(x)$ for $x_0 = 0$ and $x_1 = 1$.

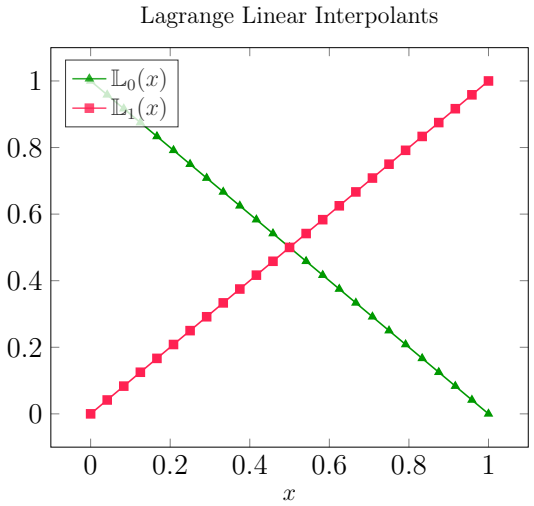

Figure 1.8: Lagrange linear interpolating polynomials for $x_i = \{0, 1\}$.

What happens if we generalize this concept, i.e., we have $\{(x_i, f_i)\}_{i=0}^n$? In this case we first need to construct for each $i = 0, 1, \ldots, n$ a function $\mathbb{L}_{n,i}(x)$ with the property that

$$\mathbb{L}_{n,i}(x_k) = \delta_{ik} \qquad \text{for} \quad k = 0, \ldots, n.$$

Based on Eq. (1.2) the general form should look like:

$$\mathbb{L}_{n,i}(x) = \prod_{j=0, j\neq i}^{n} \left(\frac{x - x_j}{x_i - x_j} \right).$$

Then, we can define the polynomial as

$$p_n^{\mathbb{L}}(x) = \sum_{i=0}^{n} f_i \mathbb{L}_{n,i}(x), \tag{1.3}$$

where we have used the notation $p_n^{\mathbb{L}}(x)$ to denote the interpolating polynomial obtained by the Lagrange interpolation. If the degree of the polynomial is clear we can write $\mathbb{L}_{n,i}(x)$ as $\mathbb{L}_i(x)$. We call $\mathbb{L}_{n,i}(x)$ as the n^{th} *Lagrange interpolating polynomial* (see Fig. 1.9). One can compute the Lagrange interpolating polynomial using algorithm 2.

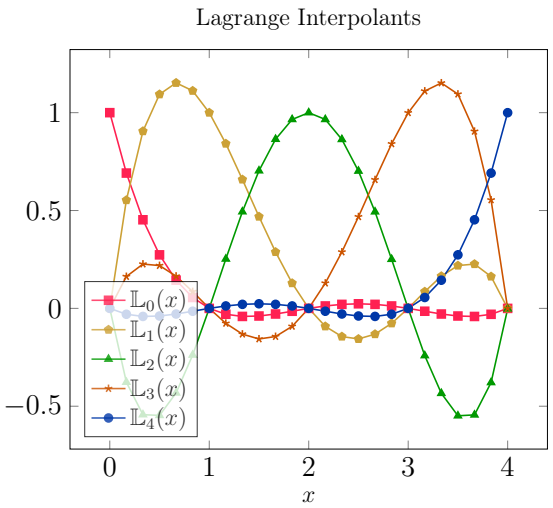

Figure 1.9: Lagrange interpolating polynomials defined over $x_i = 0, 1, 2, 3, 4$.

We have certain remarks for the Lagrange interpolation:

1. We note that in Eq. (1.3) $p_n^{\mathbb{L}}(x)$ maps the linear space $\mathbb{R}^{n+1}$ to the space of polynomials $\mathbb{P}_n$ which is a linear map.

2. We can extend the Lagrange interpolant to any continuous function $f(x)$ by

$$p_n^{\mathbb{L}} f(x) = \sum_{i=0}^{n} f(x_i)\mathbb{L}_i(x).$$

3. The operator $p_n^{\mathbb{L}}(x)$ is a projection, i.e., $p_n^{\mathbb{L}} q = q$ for all $q \in \mathbb{P}_n$.

Now we present a theorem that tells us about the error obtained using Lagrange interpolation.

Theorem 1.3. *Suppose* $\{x_0, x_1, \ldots, x_n\}$ *are distinct numbers in the interval* $[a, b]$ *and* $f \in \mathcal{C}^{n+1}[a, b]$. *Then for each* $x \in [a, b]$ *there exists a number* $\xi(x) \in (a, b)$ *with*

$$f(x) = p_n^{\mathbb{L}}(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{i=0}^{n} (x - x_i), \tag{1.4}$$

where $p_n^{\mathbb{L}}(x)$ *is given by Eq.* (1.3).

Proof. Note that if $x = x_k$ then $f(x_k) = p_n^{\mathbb{L}}(x_k)$ for any $k = 0, 1, \ldots, n$. Hence Eq. (1.4) is trivial for any $\xi(x) \in (a, b)$.

Suppose $x \neq x_k$ for any $k = 0, 1, \ldots, n$ then define a function g for t in $[a, b]$ as

$$g(t) = f(t) - p_n^{\mathbb{L}}(t) - \left[f(x) - p_n^{\mathbb{L}}(x)\right] \prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)}.$$

Since $f \in \mathcal{C}^{n+1}[a, b]$ and $p_n^{\mathbb{L}} \in \mathcal{C}^{\infty}[a, b]$ we have $g \in \mathcal{C}^{n+1}[a, b]$.

Theorem 1.4. *[3, Theorem 1.10]* **(Generalized Rolle's Theorem)**

Suppose $f \in \mathcal{C}[a, b]$ is n-times differentiable on (a, b). If $f(x) = 0$ at $(n+1)$ distinct points $a \leq x_0 < x_1 < \cdots < x_n \leq b$ then there exists a number $c \in (x_0, x_n)\,(\subset (a, b))$ such that $f^{(n)}(c) = 0$.

For $t = x_k$ for any k, we have

$$g(x_k) = f(x_k) - p_n^{\mathbb{L}}(x_k) = 0.$$

Moreover $g(x) = 0$. Thus $g \in \mathcal{C}^{n+1}[a, b]$ with $(n + 2)$ distinct zeros. By Generalized Rolle's theorem 1.2 there exists a $\xi \in (a, b)$ for which $g^{(n+1)}(\xi) = 0$. So,

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - p_n^{\mathbb{L}(n+1)}(\xi) - \left[f(x) - p_n^{\mathbb{L}}(x)\right] \frac{d^{n+1}}{dt^{n+1}} \left[\prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)}\right]_{t=\xi}. \tag{1.5}$$

Now, $p_n^{\mathbb{L}}(x)$ is a polynomial of degree at most n. Hence, $p_n^{\mathbb{L}(n+1)}(x) = 0$. Also, $\prod_{i=0}^{n} \frac{(t-x_i)}{(x-x_i)}$ is a polynomial of degree $(n + 1)$ with leading coeffecient being $\frac{1}{\prod_{i=0}^{n}(x-x_i)}$. Hence,

$$\frac{d^{n+1}}{dt^{n+1}} \left(\prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)}\right) = \frac{(n + 1)!}{\prod_{i=0}^{n}(x - x_i)}.$$

Hence, Eq. (1.5) becomes

$$f^{(n+1)}(\xi) - \left[f(x) - p_n^{\mathbb{L}}(x)\right] \frac{(n + 1)!}{\prod_{i=0}^{n}(x - x_i)} = 0 \qquad \Rightarrow \qquad f(x) = p_n^{\mathbb{L}}(x) + \frac{f^{(n+1)}(\xi)}{(n + 1)!} \prod_{i=0}^{n}(x - x_i).$$

□

Note that this error term is similar to Taylor's theorem, but it has information on all the points instead of the error being concentrated along one point.

1.2.1 Drawbacks

Lagrange interpolant suffers from certain drawbacks. The first one is regarding its *computational complexity*[2]. For the evaluation of an unknown point x, we will check the computational

[2] **Computational Complexity**: Computational complexity measures how hard it is for a computer to solve a problem as the size of the problem increases. It tells us how much time and resources are needed to find a solution.

Algorithm 2 Lagrange Interpolation

Given: Data sets $\{(x_i, f_i)\}_{i=0}^{n}$, Evaluation point x_{eval}.
Find: Interpolated polynomial $p_n^{\mathbb{L}}(x_{\text{eval}})$.
Step 1: Compute Lagrange Basis Polynomials $\mathbb{L}_i(x)$
for $i = 0$ **to** n **do**
  $\mathbb{L}_i(x_{\text{eval}}) = 1$
  for $j = 0$ **to** n **do**
    if $j \neq i$ **then**
      $\mathbb{L}_i(x_{\text{eval}}) = \mathbb{L}_i(x_{\text{eval}}) \times \frac{x_{\text{eval}} - x_j}{x_i - x_j}$
    end if
  end for
end for

Step 2: Compute Lagrange Polynomial $p_n^{\mathbb{L}}(x)$ **at** x_{eval}
Initialize $p_n^{\mathbb{L}}(x_{\text{eval}}) = 0$
for $i = 0$ **to** n **do**
  $p_n^{\mathbb{L}}(x_{\text{eval}}) = p_n^{\mathbb{L}}(x_{\text{eval}}) + f_i \times \mathbb{L}_i(x_{\text{eval}})$
end for

---

return $p_n^{\mathbb{L}}(x_{\text{eval}})$

complexity. An individual Lagrange interpolating polynomial of degree n looks like

$$\mathbb{L}_i(x) = \prod_{j=0, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)},$$

and then $p_n^{\mathbb{L}}(x) = f_0 + \mathbb{L}_0(x) + f_1\mathbb{L}_1(x) + \cdots + f_n\mathbb{L}_n(x)$. For the computation of each $\mathbb{L}_i(x)$ we need $\mathcal{O}(n)$ multiplications. As we have $(n+1)$ points, we need $\mathcal{O}\left(n^2 + n\right)$ operations. The final operation for computing of $p_n^{\mathbb{L}}(x)$ is of multiplication and addition and hence a total of $\mathcal{O}(n)$ operations. Therefore, in totality, we need $\mathcal{O}(n^2)$ operations, which is not very nice as, generally, we prefer to have linear ($\mathcal{O}(n)$) complexity.

Apart from the above drawback, another major drawback is that if we want to add a new point, say (x_{n+1}, f_{n+1}), then we need to perform new computations from scratch.

But there are advantages as well; for example, the computation of $\{\mathbb{L}_i(x)\}_{i=0}^{n}$ is independent of $f(x_k)$. Another one is that it does not depend on the arrangement of nodes.

1.2.2 Runge Phenomena

In 1901, Carl David Tolmé Runge observed that while approximating

$$f(x) = \frac{1}{1 + 25x^2}, \quad x \in [-1, 1],$$

using polynomial approximation, there are large errors at the endpoints of the interval while using equally spaced points (see Fig. 1.11). This is what is called as the *Runge phenomena* and the above function is called the *Runge function*.

Figure 1.10: Carl David Tolmé Runge: 30 August 1856-3 January 1927

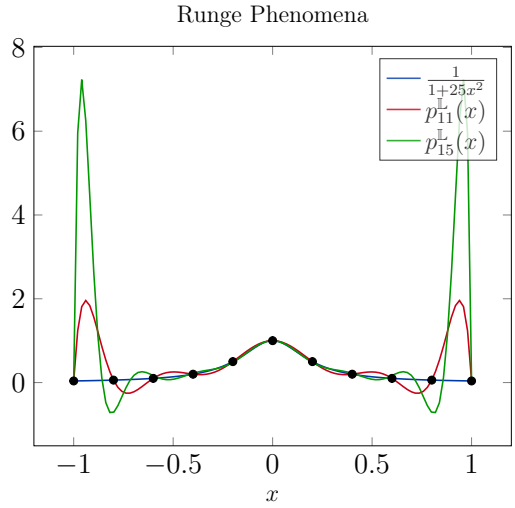

Figure 1.11: Runge phenomena for the function $1/(1+25x^2)$. $p_{11}^{\mathbb{L}}(x)$ refers to an approximation computed using 11 points (the dots refer to $\{x_i\}_{i=0}^{10}$), $p_{15}^{\mathbb{L}}(x)$ refers to approximation using 15 points.

Let us look at the interpolation error and try to understand this phenomenon. In Theorem 1.3 it was seen that

$$f(x) - p_n^{\mathbb{L}}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \quad \text{for} \quad \xi \in (-1, 1).$$

Thus, we have

$$\max_{-1 \le x \le 1} |f(x) - p_n^{\mathbb{L}}(x)| \le \max_{-1 \le x \le 1} \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} \right| \max_{-1 \le x \le 1} \prod_{i=0}^{n} |x - x_i|.$$

Now, it can be shown (although not very easily) that $\max_{-1 \le x \le 1} \prod_{i=0}^{n} |x - x_i| \le h^{n+1} n!$ where $h = 2/n$ and we suppose that the $(n+1)^{\text{th}}$ derivative of $f(x)$ can be bounded by M_{n+1} which in turn can be bounded by $5^{n+1}(n+1)!$ (see this PDF). Hence in total

$$\lim_{n \to \infty} \left(\max_{-1 \le x \le 1} |f(x) - p_n^{\mathbb{L}}(x)| \right) \le \lim_{n \to \infty} \left(\left(\frac{10}{n} \right)^{n+1} n! \right) = \infty.$$

To mitigate this problem, one idea is to use a non-uniform grid with points accumulated at the endpoints. If one is interested, I suggest this excellent review paper by Berrut and Trefethen [2].

1.3 Newton Divided Difference Interpolation

We noticed in Sec. 1.2 that Lagrange interpolation suffers from $\mathcal{O}(n^2)$ evaluation computational complexity. Now, we have another interpolating method that overcomes this and is referred to as the *Newton Divided Differences*.

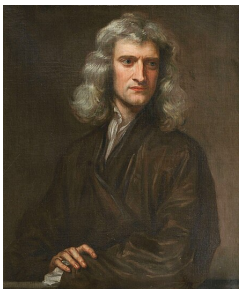

Figure 1.12: Isaac Newton: 4 January 1643-31 March 1727

Let $p_n(x)$ be a polynomial interpolating the data points $\{(x_i, f_i)\}_{i=0}^n$. Another way of expressing such a polynomial is

$$p_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n \prod_{i=0}^{n-1}(x - x_i), \tag{1.6}$$

for appropriate constants $\{a_i\}_{i=0}^n$. Now, the question is, how do we determine these coefficients? At $x = x_0$ we have $p_n(x_0) = f_0$. Hence, $y_0 = a_0$. Similarly at $x = x_1$, $p_n(x_1) = f_1$ which implies

$$a_1 = \frac{f_1 - a_0}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0}.$$

Now, we can continue in this manner and compute each a_i. For this, we introduce the divided difference (DD) notation. The zeroth divided difference of a function $f(x)$ with respect to x_i is denoted by $f[x_i] = f(x_i) = f_i$. For the rest, we define them in a recursive way.

- 1^{st} DD of $f(x)$ with respect to x_i and x_{i+1} is

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}. \tag{1.7}$$

- 2^{nd} DD of $f(x)$ with respect to $x_i,\ x_{i+1}$ and x_{i+2} is

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}. \tag{1.8}$$

- k^{th} DD of $f(x)$ with respect to $x_i,\ x_{i+1}, \ldots, x_{i+k}$ is

$$f[x_i, x_{i+1}, \ldots, x_{i+k-1}, x_{i+k}] = \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}. \tag{1.9}$$

- n^{th} DD of $f(x)$ with respect to $x_0,\ x_1, \ldots, x_n$ is

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}. \tag{1.10}$$

Hence, we can rewrite the polynomial $p_n(x)$ defined in Eq. (1.6) as

$$p_n^{\mathbb{N}}(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0)(x - x_1)\ldots(x - x_{k-1}). \tag{1.11}$$

We have introduced the notation $p_n^{\mathbb{N}}(x)$ to identify the Newton DD polynomial.

For simplicity, let us look at the DD table we obtain for 4 points (see Table 1.1)

$\boldsymbol{x}$	$\boldsymbol{f(x) = 0^{\text{th}}}$ DD	$\mathbf{1}^{\text{st}}$ DD	$\mathbf{2}^{\text{nd}}$ DD	$\mathbf{3}^{\text{rd}}$ DD
x_0	$f[x_0]$			
		$f[x_0, x_1] = \frac{f[x_1]-f[x_0]}{x_1-x_0}$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2] = \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}$	
		$f[x_1, x_2] = \frac{f[x_2]-f[x_1]}{x_2-x_1}$		$f[x_0, x_1, x_2, x_3] = \frac{f[x_1,x_2,x_3]-f[x_0,x_1,x_2]}{x_3-x_0}$
x_2	$f[x_2]$		$f[x_1, x_2, x_3] = \frac{f[x_2,x_3]-f[x_1,x_2]}{x_3-x_1}$	
		$f[x_2, x_3] = \frac{f[x_3]-f[x_2]}{x_3-x_2}$		
x_3	$f[x_3]$			

Table 1.1: Divided difference table for four points x_0, x_1, x_2, x_3.

The Lagrange interpolating polynomial has a polynomial basis as $\mathbb{L}_{n,i}(x)$, we can consider the Newton DD as another method with a basis defined by $\omega_i(x) = \prod_{k=0}^{i-1}(x - x_k)$ for $i \geq 1$ and $\omega_0(x) = 1$ (see Fig. 1.13).

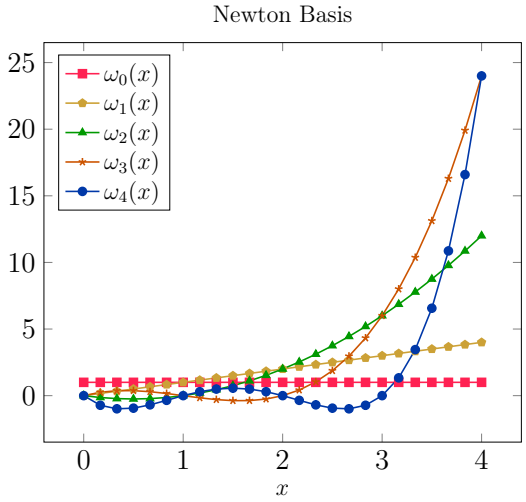

Figure 1.13: Newton basis polynomials defined over $x_i = 0, 1, 2, 3, 4$.

Now, we try to establish a relation between the DD and the derivatives of f. First, we recall the mean value theorem

Theorem 1.5. *[1, Theorem 6.2.4]* **(Mean Value Theorem)** *If $f \in C[a, b]$ and $f(x)$ is differentiable in (a, b) then there exists a $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Now, if we apply the MVT on the interval $[x_i, x_{i+1}]$ then there exists a $\xi \in (x_i, x_{i+1})$ such that

$$f'(\xi) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = f[x_i, x_{i+1}].$$

In fact, we can generalize this concept.

Theorem 1.6. *Suppose that $f \in \mathcal{C}^n[a, b]$ and $x_0, x_1, \ldots, x_n$ are distinct numbers in $[a, b]$. Then there exists a $\xi \in (a, b)$ with*

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

Proof. Let $g(x) = f(x) - p_n^{\mathbb{N}}(x)$. Since, $f(x_i) = p_n^{\mathbb{N}}(x_i)$ at $i = 0, 1, \ldots, n$. Then $g(x)$ has $(n+1)$ distinct zeros in $[a, b]$. So by generlized Rolle's theorem 1.2 there exists a $\xi \in (a, b)$ with $g^{(n)}(\xi) = 0$, so

$$0 = f^{(n)}(\xi) - p_n^{\mathbb{N}(n)}(\xi).$$

Since, $p_n^{\mathbb{N}}(x)$ is polynomial of degree n with leading coefficient $f[x_0, x_1, \ldots, x_n]$, we have

$$p_n^{\mathbb{N}(n)}(x) = n! f[x_0, x_1, \ldots, x_n] \qquad \text{for all } \; x.$$

Hence,

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

□

Next, we give a result which gives an explicit representation of the Newton DD formula.

Theorem 1.7. *For distinct points $x_0, \ldots, x_n$, the n^{th} coefficient of the Newton interpolation satisfies*

$$f[x_0, x_1, \ldots, x_n] = \sum_{k=0}^{n} \frac{f(x_k)}{\prod_{i \neq k}(x_k - x_i)},$$

where $f[x_0] = f(x_0)$ in the case $n = 0$.

Proof. Using the representation Eq. (1.11) of the interpolant the n^{th} derivative of $p_n^{\mathbb{N}}(x)$ is given by to $f[x_0, \ldots, x_n]\omega_n^{(n)}(x)$ where $\omega_n(x) = \prod_{i=0}^{n-1}(x - x_i)$.

Now, the polynomial $p_n^{\mathbb{N}}(x)$ is just another representation of $p_n^{\mathbb{L}}(x)$. Hence their n^{th} derivatives must match.

For the Lagrange polynomial the n^{th} derivative is $\sum_{k=0}^{n} f(x_k)\mathbb{L}_k^{(n)}(x)$ (see Eq. (1.3)). Hence,

$$f[x_0, x_1, \ldots, x_n]\omega_n^{(n)}(x) = \sum_{k=0}^{n} f(x_k)\mathbb{L}_k^{(n)}(x).$$

Now, the k^{th} Lagrange interpolating polynomial is given by

$$\mathbb{L}_k(x) = \prod_{j \neq k} \left(\frac{x - x_j}{x_k - x_j} \right),$$

and it's n^{th} derivative $\mathbb{L}_k^{(n)}(x) = \frac{n!}{\prod_{j \neq k}(x_k - x_j)}$ since x^n is the leading term. Since the leading term in $\omega_n(x)$ is x^n,we get $\omega_n^{(n)}(x) = n!$. Cancelling out these factorial term we get the expression. □

The algorithm for computing the Newton DD interpolation is provided in Algorithm 3.

Algorithm 3 Newton Interpolation

Given: Data sets $\{(x_i, f_i)\}_{i=0}^{n}$, Evaluation point x_{eval}.
Find: Interpolated polynomial $p_n^{\mathbb{N}}(x_{\text{eval}})$.

Step 1: Construct Divided Difference Table
Initialize DD as a zero matrix of size $(n+1) \times (n+1)$.
for $i = 0$ **to** n **do**
  $\text{DD}_{i,0} = f_i$
end for
for $j = 1$ **to** n **do**
  for $i = 0$ **to** $n - j$ **do**
    Compute:

$$\text{DD}_{i,j} = \frac{\text{DD}_{i+1,j-1} - \text{DD}_{i,j-1}}{x_{i+j} - x_i}.$$

  end for
end for

Step 2: Evaluate Newton Polynomial $p_n^{\mathbb{N}}(x)$ at x_{eval}
Initialize $p_n^{\mathbb{N}}(x_{\text{eval}}) = \text{DD}_{0,0}$.
for $k = 1$ **to** n **do**
  Initialize $\omega = 1$.
  for $j = 0$ **to** $k - 1$ **do**
    $\omega = \omega \times (x_{\text{eval}} - x_j)$
  end for
  Update the interpolated value:

$$p_n^{\mathbb{N}}(x_{\text{eval}}) = p_n^{\mathbb{N}}(x_{\text{eval}}) + \text{DD}_{0,k} \cdot \omega.$$

end for

---

return $p_n^{\mathbb{N}}(x_{\text{eval}})$.

1.3.1 Computational Complexity

We can rewrite the Newton interpolation as

$$\begin{aligned} p_n^{\mathbb{N}}(x) &= a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots + a_n(x-x_0)(x-x_1)\ldots(x-x_{n-1}) \\ &= a_0 + (x-x_0)\left[a_1 + (x-x_1)\left\{a_2 + \cdots + (x-x_{n-2})\left\{a_{n-1} + a_n(x-x_{n-1})\right\}\right\}\right]. \end{aligned}$$

We notice that each term requires one multiplication and one addition for evaluation, and we have n points. Hence, we require $2n$ operations, which is of $\mathcal{O}(n)$, whereas for Lagrange, we have $\mathcal{O}(n^2)$.

Another advantage of the Newton interpolation over Lagrange interpolation is that it is easy to update the DD table whenever we have a new data set as it does not require new computation only a modification of the DD table.

1.3.2 Forward Difference Formula

Suppose we have an equal spacing of points; then we can rewrite Newton's formula in a better way. Let $h = x_{i+1} - x_i$ for all $i = 0, 1, \ldots, n-1$ and $x = x_0 + sh$. Then we can rewrite Eq. (1.11) as

$$\begin{aligned} p_n^{\mathbb{N}}(x) &= f[x_0] + (x-x_0)f[x_0,x_1] + (x-x_0)(x-x_1)f[x_0,x_1,x_2] + \ldots \\ &\quad + (x-x_0)\ldots(x-x_{n-1})f[x_0,x_1,\ldots,x_n] \\ &= f[x_0] + shf[x_0,x_1] + s(s-1)h^2 f[x_0,x_1,x_2] + \ldots \\ &\quad + s(s-1)\ldots(s-n+1)h^n f[x_0,x_1,\ldots,x_n] \\ &= f[x_0] + \sum_{k=1}^{n} s(s-1)\ldots(s-k+1)h^k f[x_0,x_1,\ldots,x_k]. \end{aligned}$$

Using the binomial coefficient notation

$$^s\mathrm{C}_k = \frac{s(s-1)\ldots(s-k+1)}{k!},$$

we can express

$$p_n^{\mathbb{N}}(x) = p_n^{\mathbb{N}}(x_0 + sh) = f[x_0] + \sum_{k=1}^{n} {}^s\mathrm{C}_k k! h^k f[x_0,x_1,\ldots,x_k].$$

Let us use the Δ notation for forward difference, i.e, $\Delta f(x_0) = f(x_1) - f(x_0)$. Similarly for higher differences we use the notation $\Delta^2 f(x_0) = \Delta f(x_1) - \Delta f(x_0)$, then we can rewrite the divided differences as

$$\begin{aligned} f[x_0,x_1] &= \frac{f[x_1]-f[x_0]}{x_1-x_0} = \frac{\Delta f(x_0)}{h} \\ f[x_0,x_1,x_2] &= \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0} = \frac{1}{2h}\frac{\Delta f(x_1)-\Delta f(x_0)}{h} = \frac{\Delta^2 f(x_0)}{2!h^2} \\ &\vdots \\ f[x_0,x_1,\ldots,x_n] &= \frac{\Delta^n f(x_0)}{n!h^n}. \end{aligned}$$

Hence,

$$p_n^{\mathbb{N}}(x) = f(x_0) + \sum_{k=1}^{n} {}^s\mathrm{C}_k \Delta^k f(x_0).$$

1.4 Hermite Interpolation

Definition 1.8. Let $\{x_0, x_1, \ldots, x_n\}$ be $(n+1)$ distinct points in $[a, b]$ and for $i = 0, 1, \ldots, n$ let m_i be a non-negative integer. Suppose that $f \in \mathcal{C}^m[a, b]$ where $m = \max_{0 \le i \le n} m_i$. The *osculating polynomial* approximating $f(x)$ is the polynomial $p(x)$ of least degree such that

$$\frac{d^k p(x_i)}{dx^k} = \frac{d^k f(x_i)}{dx^k}, \quad \text{for} \quad i = 0, 1, \ldots, n \quad \text{and} \quad k = 0, 1, \ldots, m_i.$$

Not when $n = 0$ the osculating polynomial approximating f is the m_0^{th} Taylor polynomial for f at x_0. When $m_i = 0$ for all i then the osculating polynomial is the n^{th} Lagrange polynomial interpolating f at $x_0, x_1, \ldots, x_n$.

Hermite Polynomials

If $m_i = 1$ for all $i = 0, 1, \ldots, n$ then we get the Hermite polynomials. For a given function f these polynomials agree with f at $x_0, x_1, \ldots, x_n$. In addition they agree with there derivatives as well.

Figure 1.14: Charles Hermite: 24 December 1822- 14 January 1901

Theorem 1.9. *If $f \in \mathcal{C}^1[a,b]$ and $x_0, x_1, \ldots, x_n \in [a,b]$ are distinct, the unique polynomial of least degree agreeing with f and f' at $x_0, x_1, \ldots, x_n$ is the Hermite polynomial of degree at most $2n+1$ given by*

$$p^{\mathbb{H}}_{2n+1}(x) = \sum_{j=0}^{n} f(x_j)\mathbb{H}_{n,j}(x) + \sum_{j=0}^{n} f'(x_j)\hat{\mathbb{H}}_{n,j}(x),$$

where for $\mathbb{L}_{n,j}(x)$ denoting the j^{th} Lagrange coefficient polynomial of degree n we have

$$\mathbb{H}_{n,j}(x) = \left[1 - 2(x - x_j)\mathbb{L}_{n,j}(x_j)'\right]\mathbb{L}^2_{n,j}(x) \qquad \text{and} \qquad \hat{\mathbb{H}}_{n,j}(x) = (x - x_j)\mathbb{L}^2_{n,j}(x).$$

Moreover, if $f \in \mathcal{C}^{2n+2}[a,b]$ then

$$f(x) = p^{\mathbb{H}}_{2n+1}(x) + \frac{(x-x_0)^2(x-x_1)^2\ldots(x-x_n)^2}{(2n+2)!} f^{(2n+2)}\left(\xi(x)\right),$$

for some unknown $\xi(x) \in (a,b)$.

Proof. We know that $\mathbb{L}_{n,j}(x_i) = \delta_{ij}$. Hence when $i \neq j$ $\mathbb{H}_{n,j}(x_i) = 0$ and $\hat{\mathbb{H}}_{n,j}(x_i) = 0$ whereas for each i

$$\mathbb{H}_{n,i}(x_i) = \left[1 - 2(x_i - x_i)\mathbb{L}'_{n,i}(x_i)\right]\mathbb{L}^2_{n,i}(x_i) = 1 \qquad \text{and} \qquad \hat{\mathbb{H}}_{n,i}(x_i) = (x_i - x_i)\mathbb{L}^2_{n,i}(x_i) = 0.$$

Hence, we can say $\mathbb{H}_{n,j}(x_i) = \delta_{ij}$ and $\hat{\mathbb{H}}_{n,j}(x_i) = 0$ for all i, j. As a consequence

$$p^{\mathbb{H}}_{2n+1}(x_i) = \sum_{j=0}^{n} f(x_j)\mathbb{H}_{n,j}(x_i) + \sum_{j=0}^{n} f'(x_j)\hat{\mathbb{H}}_{n,j}(x_i) = f(x_i),$$

so $p^{\mathbb{H}}_{2n+1}$ agrees with f at $x_0, x_1, \ldots, x_n$.

Now we need to show that they match at the derivatives as well, i.e., $p^{\mathbb{H}'}_{2n+1}$ and f' match at x_i. We will tackle this be differentiating both the terms $\mathbb{H}_{n,j}$ and $\hat{\mathbb{H}}_{n,j}$.

The derivative of $\mathbb{H}_{n,j}(x)$ is given by

$$\begin{aligned}
\mathbb{H}'_{n,j}(x) &= \left[1 - 2(x - x_j)\mathbb{L}'_{n,j}(x_j)\right] 2\mathbb{L}_{n,j}(x)\mathbb{L}'_{n,j}(x) + \mathbb{L}^2_{n,j}(x)\left[-2\mathbb{L}'_{n,j}(x_j)\right] \\
&= 2\mathbb{L}_{n,j}(x)\left[\left\{1 - 2(x - x_j)\mathbb{L}'_{n,j}(x_j)\right\}\mathbb{L}'_{n,j}(x) - \mathbb{L}_{n,j}(x)\mathbb{L}'_{n,j}(x_j)\right].
\end{aligned}$$

As $\mathbb{L}_{n,j}(x_i) = \delta_{ij}$ we get that at $i \neq j$, $\mathbb{H}'_{n,j}(x_i) = 0$. At $i = j$ we have

$$\begin{aligned}
\mathbb{H}'_{n,i}(x_i) &= \left[1 - 2(x_i - x_i)\mathbb{L}'_{n,i}(x_i)\right] 2\mathbb{L}_{n,i}(x_i)\mathbb{L}'_{n,i}(x_i) + \mathbb{L}^2_{n,i}(x_i)\left[-2\mathbb{L}'_{n,i}(x_i)\right] \\
&= \left[2\mathbb{L}'_{n,i}(x_i) - 2\mathbb{L}'_{n,i}(x_i)\right] = 0.
\end{aligned}$$

Now for the second term the derivative is given by

$$\begin{aligned}
\hat{\mathbb{H}}'_{n,j}(x) &= (x - x_j)2\mathbb{L}_{n,j}(x)\mathbb{L}'_{n,j}(x) + \mathbb{L}^2_{n,j}(x) \\
&= \mathbb{L}_{n,j}(x)\left[2(x - x_j)\mathbb{L}'_{n,j}(x) + \mathbb{L}_{n,j}(x)\right].
\end{aligned}$$

At $x = x_i$ we have $\mathbb{L}_{n,j}(x_i) = \delta_{ij}$. Hence $\hat{\mathbb{H}}'_{n,j}(x_i) = 0$ if $i \neq j$ and at $i = j$

$$\begin{aligned}\hat{\mathbb{H}}'_{n,i}(x_i) &= (x_i - x_i)2\mathbb{L}_{n,i}(x_i)\mathbb{L}'_{n,i}(x_i) + \mathbb{L}^2_{n,i}(x_i) \\ &= 1.\end{aligned}$$

Hence $\hat{\mathbb{H}}'_{n,j}(x_i) = \delta_{ij}$. Therefore

$$p^{\mathbb{H}'}_{2n+1}(x_i) = \sum_{j=0}^{n} f(x_j)\mathbb{H}'_{n,j}(x_i) + \sum_{j=0}^{n} f'(x_j)\hat{\mathbb{H}}'_{n,j}(x_i) = f'(x_i).$$

Therefore $p^{\mathbb{H}}_{2n+1}$ agrees with f and $p^{\mathbb{H}'}_{2n+1}$ with f' at $x_0, x_1, \ldots, x_n$. So,we have existence of a polynomial that agrees with f and f' at $\{x_i\}_{i=0}^n$.

For the uniqueness we will use the method of contradiction. Suppose there exists another polynomial of least degree say $q(x)$ such that

$$q(x_i) = f(x_i) \qquad \text{and} \qquad q'(x_i) = f'(x_i) \qquad \forall \ \ i.$$

Now consider the polynomial $D(x) = p^{\mathbb{H}}_{2n+1}(x) - q(x)$ of degree at most $(2n+1)$. Obviously

$$D(x_i) = 0 \quad \text{and} \quad D'(x_i) = 0 \qquad \forall \ \ i$$

Hence x_i are distinct roots of multiplicity two. Therefore we have $2n+2$ roots, which is a only possible if $D(x) = 0$. Hence, we get $p^{\mathbb{H}}_{2n+1}(x) = q(x)$ leading to a contradiction.

For showing the error term we will use the same strategy as in theorem 1.3., if $x = x_i$ for some i then we can choose $\xi(x)$ arbitrary.

Suppose $x \neq x_i$ for any i, then define

$$g(t) = f(t) - p^{\mathbb{H}}_{2n+1}(t) - \left[f(x) - p^{\mathbb{H}}_{2n+1}(x)\right] \prod_{i=0}^{n} \frac{(t - x_i)^2}{(x - x_i)^2}.$$

Now $g(x) = 0$ and $g(x_i) = 0$ for all i. Hence $g(t)$ has distinct $n+2$ roots in $[a, b]$. Hence, by Rolle's theorem $g'(t)$ has $n+1$ distinct roots between $x_0, x_1, \ldots, x_n$ and x, say $\xi_0, \xi_1, \ldots, \xi_n$.

Now taking the derivative of $g(t)$ with respect to t we get

$$\begin{aligned}g'(t) &= f'(t) - p^{\mathbb{H}'}_{2n+1}(t) - \frac{\left[f(x) - p^{\mathbb{H}}_{2n+1}(x)\right]}{\prod_{i=0}^{n}(x - x_i)^2} \frac{d}{dt}\left(\prod_{i=0}^{n}(t - x_i)^2\right) \\ &= f'(t) - p^{\mathbb{H}'}_{2n+1}(t) - \frac{\left[f(x) - p^{\mathbb{H}}_{2n+1}(x)\right]}{\prod_{i=0}^{n}(x - x_i)^2} \frac{d}{dt}\left((t - x_0)^2(t - x_1)^2 \ldots (t - x_n)^2\right) \\ &= f'(t) - p^{\mathbb{H}'}_{2n+1}(t) - \frac{\left[f(x) - p^{\mathbb{H}}_{2n+1}(x)\right]}{\prod_{i=0}^{n}(x - x_i)^2} \Big(2(t - x_0)(t - x_1)^2 \ldots (t - x_n)^2 \\ &\quad + (t - x_0)^2 2(t - x_1) \ldots (t - x_n)^2 + \cdots + (t - x_0)^2(t - x_1)^2 \ldots 2(t - x_n)\Big) \\ &= f'(t) - p^{\mathbb{H}'}_{2n+1}(t) - 2\frac{\left[f(x) - p^{\mathbb{H}}_{2n+1}(x)\right]}{\prod_{i=0}^{n}(x - x_i)^2} \sum_{k=0}^{n}(t - x_k) \prod_{j=0, j\neq k}^{n} (t - x_j)^2\end{aligned}$$

At $t = x_i$ for any i we have $g'(x_i) = 0$ for $i = 0, 1, \ldots, n$. Hence, $g'(t)$ has $2n+2$ roots. Using the generalized Rolle's theorem on $g'(t)$ and then following the same pattern as in Theorem 1.3 we get the result. $\square$

Theorem 1.9 gives all the details about the Hermite polynomials but it is computationally expensive as we need to compute the Lagrange polynomials and it's derivatives.

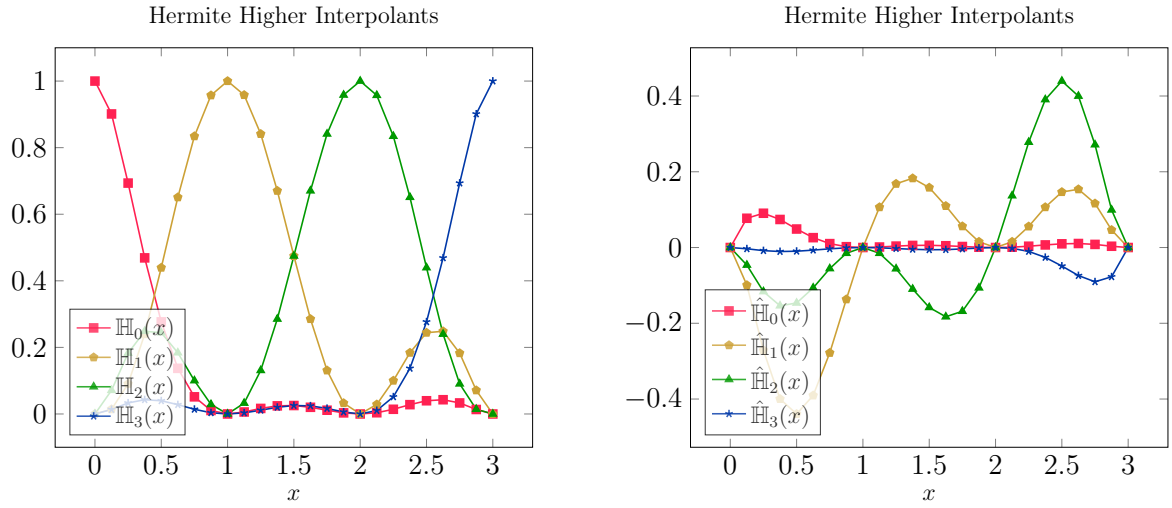

Figure 1.15: Hermite interpolating polynomials defined over $x_i = 0, 1, 2, 3$.

1.4.1 Hermite Polynomials using Divided Difference

For computing the Hermite polynomials using the Newton DD. We will use the relation between the n^{th} DD and the n^{th} derivative of $f(x)$ as in Theorem 1.6.

Suppose we have $(n+1)$ distinct points $\{x_i\}_{i=0}^{n}$, we define a new sequence $\{z_i\}_{i=0}^{2n+1}$ by

$$z_{2i} = z_{2i+1} = x_i \qquad \text{for each} \quad i = 0, 1, \ldots, n,$$

i.e., $z_0 = z_1 = x_0$, $z_2 = z_4 = x_1$, and so on. Then we can construct the DD table using these values. Since $z_{2i} = z_{2i+1}$ we cannot define $f[z_{2i}, z_{2i+1}]$. However from Theorem 1.6 we can make a reasonable substitution that

$$f[z_{2i}, z_{2i+1}] = f'(z_{2i}) = f'(x_i).$$

Hence we can use the derivative entries for the undefined DD.

The remaining entries of the DD are defined in the same manner and we get the Hermite polynomial as

$$p_{2n+1}^{\mathbb{H}}(x) = f[z_0] + \sum_{k=1}^{2n+1} f[z_0, z_1, \ldots, z_k](x - z_0)(x - z_1)\ldots(x - z_{k-1}).$$

For an example let us consider a data set of two points x_0 and x_1. Then the DD table is given by table 1.2.

The algorithm for Hermite interpolation is given in algorithm 4.

z	$f(z)$	1^{st} DD	2^{nd} DD	3^{rd} DD
$z_0 = x_0$	$f[z_0] = f(x_0)$			
		$f[z_0, z_1] = f'(x_0)$		
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2] = \frac{f[z_1, z_2] - f[z_0, z_1]}{z_2 - z_0}$	
		$f[z_1, z_2] = \frac{f[z_2] - f[z_1]}{z_2 - z_1}$		$f[z_0, z_1, z_2, z_3] = \frac{f[z_1, z_2, z_3] - f[z_0, z_1, z_2]}{z_3 - z_0}$
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3] = \frac{f[z_2, z_3] - f[z_1, z_2]}{z_3 - z_1}$	
		$f[z_2, z_3] = f'(x_1)$		
$z_3 = x_1$	$f[z_3] = f(x_1)$			

Table 1.2: Divided difference table for two points and the Hermite polynomial

Algorithm 4 Hermite Interpolation

Given: Data sets $\{(x_i, f_i, f_i')\}_{i=0}^{n}$, Evaluation point x_{eval}.
Find: Interpolated polynomial $p_{2n+1}^{\mathbb{H}}(x_{\text{eval}})$.

Step 1: Create z_i and $f(z_i)$ arrays
Construct $\{z_i\}_{i=0}^{2n+1}$ and $\{f(z_i)\}_{i=0}^{2n+1}$
for $i = 0$ **to** n **do**
$z_{2i} = z_{2i+1} = x_i, \qquad f(z_{2i}) = f(z_{2i+1}) = f_i$
end for

Step 2: Construct Divided Difference Table
Initialize DD as a zero matrix of size $(2n+2) \times (2n+2)$.
for $i = 0$ **to** $2n+1$ **do**
$\text{DD}_{i,0} = f(z_i)$
end for
for $j = 1$ **to** $2n+1$ **do**
for $i = 0$ **to** $2n+1-j$ **do**
if $j = 1$ **and** $i\%2 = 0$ **then**
$\text{DD}_{i,j} = f'_{\frac{i}{2}}$
else
Compute:

$$\text{DD}_{i,j} = \frac{\text{DD}_{i+1,j-1} - \text{DD}_{i,j-1}}{z_{i+j} - z_i}.$$

end if
end for
end for

Step 3: Evaluate Hermite Polynomial $p_{2n+1}^{\mathbb{H}}(x)$ at x_{eval}
Initialize $p_{2n+1}^{\mathbb{H}}(x) = \text{DD}_{0,0}$.
for $k = 1$ **to** $2n+1$ **do**
Initialize $\omega = 1$.
for $j = 0$ **to** $k-1$ **do**
$\omega = \omega \times (x_{\text{eval}} - z_j)$
end for
Update the interpolated value:

$$p_{2n+1}^{\mathbb{H}}(x_{\text{eval}}) = p_{2n+1}^{\mathbb{H}}(x_{\text{eval}}) + \text{DD}_{0,k} \cdot \omega.$$

end for

---

return $p_{2n+1}^{\mathbb{H}}(x_{\text{eval}})$.

1.5 Spline Interpolation

Both Lagrange and Newton interpolation methods suffer from the Runge phenomenon, where oscillations occur at the edges of the interval, especially with high-degree polynomials. This issue arises because these methods rely on a single global polynomial, meaning that every data point influences the entire approximation. This "global approximation" can lead to poor performance for non-uniform or large datasets.

An alternative approach is to divide the interval into smaller sub-intervals and use piecewise polynomial approximation. This strategy, known as local interpolation, reduces the influence of distant data points, resulting in more stable and accurate approximations.

Given a set of points $\{(x_i, f_i)\}_{i=0}^{n}$ we can use piecewise-linear interpolation that consists of joining set of data points using straight lines (see Fig. 1.16). An immediate disadvantage of such an interpolation is that the approximating polynomial is not differentiable at the nodal points which geometrically mean the function is not "smooth".

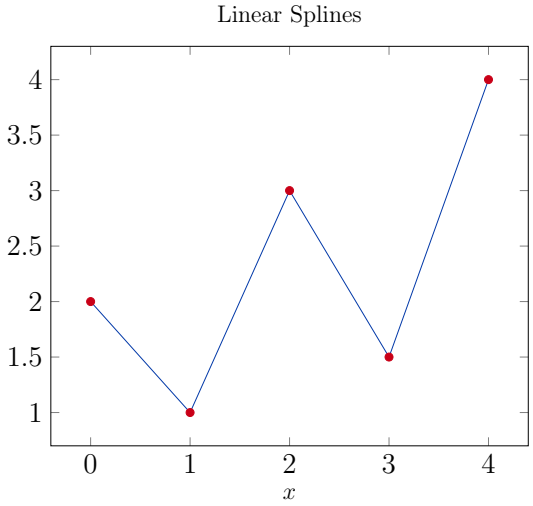

Figure 1.16: Linear spline defined over $x_i = 0, 1, 2, 3, 4$.

To address the limitations of linear interpolation, we use *splines*, which are piecewise polynomials of higher degree. The term "spline" was introduced by Isaac Jacob Schoenberg in the 1930s, inspired by drafting tools called "flat splines". These tools were used to draw smooth curves on paper before the advent of computer-aided design. A spline curve behaves like a flexible beam, ensuring continuity in both slope and curvature.

1.5.1 Cubic Splines

The most common piecewise polynomial approximation uses the cubic polynomials between each successive pair of nodes and is called *cubic spline interpolation* (see Fig. 1.18) .

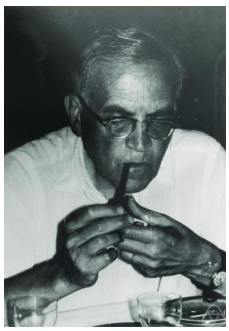

Figure 1.17: Isaac Jacob Schoenberg: 21 April 1903-21 February 1990

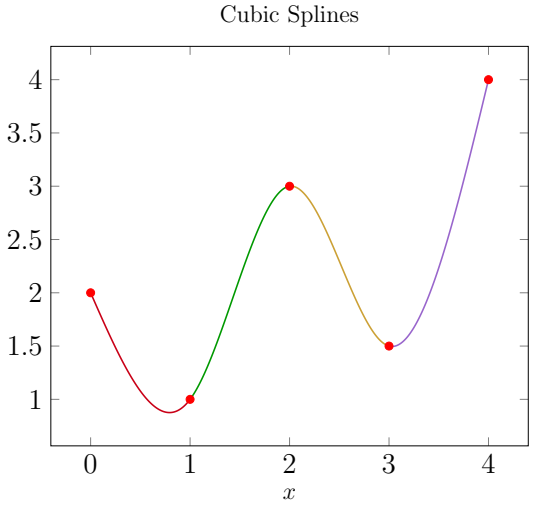

Figure 1.18: Cubic spline defined over $x_i = 0, 1, 2, 3, 4$.

Definition 1.10. Given a function f defined on $[a, b]$ and a set of nodes $a = x_0 < x_1 < \cdots < x_n = b$ (called *knots*), a cubic spline interpolant $p^{\mathbb{S}}$ for f is a function that satisfies the following conditions:

- **a)** $p^{\mathbb{S}}(x)$ is a cubic polynomial, whose restriction on the interval $[x_j, x_j + 1]$ is denoted by $p_j^{\mathbb{S}}(x)$ for each $j = 0, 1, \ldots, n - 1$.
- **b)** $p_j^{\mathbb{S}}(x_j) = f(x_j)$ and $p_j^{\mathbb{S}}(x_{j+1}) = f(x_{j+1})$ for $j = 0, 1, \ldots, n - 1$.
- **c)** $p_{j+1}^{\mathbb{S}}(x_{j+1}) = p_j^{\mathbb{S}}(x_{j+1})$ for $j = 0, 1, \ldots, n - 2$ (implied by **b)**.
- **d)** $p_{j+1}^{\mathbb{S}'}(x_{j+1}) = p_j^{\mathbb{S}'}(x_{j+1})$ for $j = 0, 1, \ldots, n - 2$.
- **e)** $p_{j+1}^{\mathbb{S}''}(x_{j+1}) = p_j^{\mathbb{S}''}(x_{j+1})$ for $j = 0, 1, \ldots, n - 2$.
- **f)** One of the following sets of boundary conditions is satisfied:
  - **i)** $p^{\mathbb{S}''}(x_0) = p^{\mathbb{S}''}(x_n) = 0$ (*natural* (or free) boundary).
  - **ii)** $p^{\mathbb{S}'}(x_0) = f'(x_0)$ and $p^{\mathbb{S}'}(x_n) = f'(x_n)$ (*clamped* boundary).

When the free boundary condition occurs the spline is called *natural spline* . In general clamped boundary conditions lead to more accurate results but it includes the information about the derivative of the function which is not easily available.

Notice that we have n intervals and on each interval we have 4 unknowns. Hence we have a total of $4n$ unknowns.

Construction of Cubic Splines

Let $[a, b]$ be an interval. We divide this interval into n subintervals denoted $[x_j, x_{j+1}]$ for any $j = 0, 1, 2, \ldots, n-1$, then for each subinterval we define a cubic polynomial as

$$p_j^{\mathbb{S}}(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3.$$

Since, $p_j^{\mathbb{S}}(x) = f(x_j)$ we get $a_j = f(x_j)$.

Now from condition **c)** we have

$$\begin{aligned} p_{j+1}^{\mathbb{S}}(x_{j+1}) &= p_j^{\mathbb{S}}(x_{j+1}) \\ a_{j+1} &= a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3, \end{aligned}$$

for $j = 0, 1, 2, \ldots, n-2$. Let us denote $x_{j+1} - x_j$ by h_j for $j = 1, 2, \ldots, n-1$. If we define $a_n = f(x_n)$, then we get the relation

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3, \qquad \text{for } \; j = 0, 1, \ldots, n-1. \tag{1.12}$$

We also note that

$$p_j^{\mathbb{S}'}(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2.$$

Substituting $x = x_j$, we get $p_j^{\mathbb{S}'}(x_j) = b_j$ for each $j = 0, 1, \ldots, n-2$. Defining $b_n = p^{\mathbb{S}'}(x_n)$ we get the relation

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2 \qquad \text{for } \; j = 0, 1, \ldots, n-1. \tag{1.13}$$

Now $p_j^{\mathbb{S}''}(x) = 2c_j + 6d_j(x - x_j)$ and hence $p_j^{\mathbb{S}''}(x_j) = 2c_j$. Defining $p^{\mathbb{S}''}(x_n) = 2c_n$, from condition **e)** we get

$$2c_{j+1} = 2c_j + 6d_j h_j \qquad \text{for } \; j = 0, 1, \ldots, n-1. \tag{1.14}$$

Solving for d_j in Eq. (1.14) we get $d_j = \frac{c_{j+1} - c_j}{3h_j}$ and substituting this back in Eq. (1.12) and Eq. (1.13) we get

$$\begin{aligned} a_{j+1} &= a_j + b_j h_j + c_j h_j^2 + \frac{(c_{j+1} - c_j)}{3h_j} h_j^3 \\ &= a_j + b_j h_j + \frac{(2c_j + c_{j+1})}{3} h_j^2. \end{aligned} \tag{1.15}$$

$$b_{j+1} = b_j + 2c_j h_j + h_j(c_{j+1} - c_j). \tag{1.16}$$

From Eq. (1.15) we get for b_j

$$b_j = \frac{a_{j+1} - a_j}{h_j} - \frac{h_j}{3}(2c_j + c_{j+1}). \tag{1.17}$$

Substituting Eq. (1.17) into Eq. (1.16) we get

$$\begin{aligned} \frac{a_{j+2} - a_{j+1}}{h_{j+1}} - \frac{h_{j+1}}{3}(2c_{j+1} + c_{j+2}) &= \frac{a_{j+1} - a_j}{h_j} - \frac{h_j}{3}(2c_j + c_{j+1}) + h_j(c_j + c_{j+1}) \\ \frac{a_{j+2} - a_{j+1}}{h_{j+1}} - \frac{a_{j+1} - a_j}{h_j} &= \frac{c_j h_j}{3} + \frac{2c_{j+1}h_j}{3} + \frac{2c_{j+1}h_{j+1}}{3} + \frac{c_{j+2}h_{j+1}}{3} \\ c_j h_j + 2c_{j+1}(h_j + h_{j+1}) + c_{j+2}h_{j+1} &= \frac{3(a_{j+2} - a_{j+1})}{h_{j+1}} - \frac{3(a_{j+1} - a_j)}{h_j}, \end{aligned}$$

for $j = 0, 1, 2, \ldots, n-2$.

For simplicity we do the shifting of the index by 1. Hence, finally we get the system of equation as

$$c_{j-1}h_{j-1} + 2c_j(h_{j-1} + h_j) + c_{j+1}h_j = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}), \tag{1.18}$$

for $j = 1, 2, \ldots, n-1$. The system of equations given by Eq. (1.18) involves the unknown $\{c_i\}_{i=0}^{n}$ as the values of $\{h_j\}_{j=0}^{n}$ and $\{a_j\}_{j=0}^{n}$ are known. Hence, if we can compute $\{c_j\}$ then we can compute $\{b_j\}_{j=0}^{n-1}$ using Eq. (1.17) and $\{d_j\}_{j=0}^{n-1}$ from Eq. (1.14). Hence after these computations we can compute $\{p_j^{\mathbb{S}}(x)\}_{j=0}^{n-1}$. So if Eq. (1.18) has an unique solution then we are done.

Theorem 1.11. *If f is defined at $a = x_0 < x_1 < \cdots < x_n = b$, then f has a unique natural spline interpolant $p^{\mathbb{S}}(x)$ on the nodes $x_0, x_1, \ldots, x_n$, i.e., a spline interpolant that satisfies the natural boundary condition $p^{\mathbb{S}''}(a) = p^{\mathbb{S}''}(b) = 0$.*

Proof. Let us consider $p_0^{\mathbb{S}}(x)$, which is given by $p_0^{\mathbb{S}}(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 + d_0(x - x_0)^3$. Now, $p_0^{\mathbb{S}''}(x) = 2c_0 + 6d_0(x - x_0)$ and at $x = x_0$, we have $p_0^{\mathbb{S}''}(x_0) = 2c_0 = 0$ which implies $c_0 = 0$. Similarly $c_n = 0$.

Let us look at Eq. (1.18)

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}),$$

for $j = 1, 2, \ldots, n-1$. If we substitute for each j we get a system of equation

$$\mathbf{Sc} = \mathbf{v},$$

where

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & \ldots & 0 & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \ldots & \ldots & 0 & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \ldots & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \ldots & \ldots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{v} = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

We use the following theorem to show that the matrix $\mathbf{S}$ is invertible.

Theorem 1.12. (Strictly Diagonal Dominant Matrix)

A strictly diagonally dominant matrix $\mathbf{A}$ *is nonsingular.*

We notice that our matrix $\mathbf{S}$ is strictly diagonally dominant[3] and hence it is invertible, which leads to an unique solution. □

The algorithm for natural spline interpolation can be found in Algorithm 5 We have a similar result for the clamped spline interpolation.

Theorem 1.13. *If* f *is defined at* $a = x_0 < x_1 < \cdots < x_n = b$ *and differentiable at* a *and* b *then* f *has a unique clamped spline interpolant* $p^{\mathbb{S}}(x)$ *on the nodes* $x_0, x_1, \ldots, x_n$*,i.e., a spline interpolant that satisfies the clamped boundary condition* $p^{\mathbb{S}'}(a) = f'(a)$ *and* $p^{\mathbb{S}'}(b) = f'(b)$.

Now we present a result regarding the error bound of the spline interpolation but we will not delve into it's proof as the proof requires a lot of technicalities from Numerical Analysis which is out of scope of this lecture.

Theorem 1.14. *Let* $f \in \mathcal{C}^4[a,b]$ *with* $M = \max_{a\le x\le b} |f^{(4)}(x)|$*.If* $p^{\mathbb{S}}(x)$ *is the unique clamped cubic spline interpolant to* f *with respect to the nodes* $a = x_0 < x_1 < \cdots < x_n = b$ *then for all* $x \in [a,b]$

$$|f(x) - p^{\mathbb{S}}(x)| \le \frac{5M}{384} \max_{0\le j\le n-1} (x_{j+1} - x_j)^4 .$$

A fourth order error bound also exist for the case of natural boundary splines, but they are more difficult to express. An alternative to the natural boundary condition is the *not-a-knot* condition, it states that $p^{\mathbb{S}'''}(x)$ has to be continuous at x_1 and x_{n-1}.

1.5.2 B-Splines

So far, we have focused on a specific type of spline function called cubic splines. A natural question arises: can we generalize this to splines of other degrees? The answer to this is yes. A generalization of the cubic splines is the basis splines or B-splines.

Let $\{x_i\}_{i=0}^n$ be the data points (or knots); then we define the zeroth degree B-spline as

$$B_{j,0}(x) = \begin{cases} 1, & x \in [x_j, x_{j+1}), \\ 0, & \text{else}, \end{cases}$$

[3] **Strictly Diagonally Dominant Matrix**: A matrix $\mathbf{A} = \{a_{ij}\}_{i=1,j=1}^n$ is said to be strictly diagonally dominant if

$$|a_{ii}| > \sum_{j\ne i} |a_{ij}| \qquad \forall i.$$

Algorithm 5 Cubic Natural Spline Interpolation

Given: Data sets $\{(x_i, f_i)\}_{i=0}^{n}$, Evaluation point x_{eval}.
Find: Interpolated polynomial $p^{\mathbb{S}}(x_{\text{eval}})$.

Step 1: Compute h_i arrays
Construct $\{h_i\}_{i=0}^{n-1}$,
for $i = 0$ **to** $n-1$ **do**
  $h_i = x_{i+1} - x_i$
end for

Step 2: Construct S and v
Initialize $\mathbf{S}$ as a zero matrix of size $(n+1) \times (n+1)$ and $\mathbf{v}$ as $(n+1)$.
for $i = 0$ **to** n **do**
  if $i = 0$ or $i = n$ **then**
    $\mathbf{S}_{ii} = 1$ and $\mathbf{v}_i = 0$
  else
    $\mathbf{S}_{ii} = 2(h_{i-1} + h_i)$
    $\mathbf{S}_{i,i-1} = h_{i-1}$
    $\mathbf{S}_{i,i+1} = h_i$
    $\mathbf{v}_i = \frac{3}{h_i}(f_{i+1} - f_i) - \frac{3}{h_{i-1}}(f_i - f_{i-1})$
  end if
end for

Step 3: Solve Sc = v
Solve the system $\mathbf{Sc} = \mathbf{v}$ to get coefficient vector $\mathbf{c}$.

Step 4: Locate x_{eval}
$\text{loc} = 0$
for $i = 0$ **to** n **do**
  if $x_{\text{eval}} \leq x_i$ **then**
    $\text{loc} = i - 1$
    break
  end if
end for

Step 5: Compute b_{loc} and d_{loc}
$b_{\text{loc}} = \frac{f_{\text{loc}+1} - f_{\text{loc}}}{h_{\text{loc}}} - \frac{h_{\text{loc}}}{3}(2\mathbf{c}_{\text{loc}} + \mathbf{c}_{\text{loc}+1})$
$d_{\text{loc}} = \frac{\mathbf{c}_{\text{loc}+1} - \mathbf{c}_{\text{loc}}}{3h_{\text{loc}}}$

Step 6: Evaluate Spline Polynomial $p^{\mathbb{S}}(x)$ at x_{eval}
$p^{\mathbb{S}}(x_{\text{eval}}) = f_{\text{loc}} + b_{\text{loc}}(x_{\text{eval}} - x_{\text{loc}}) + \mathbf{c}_{\text{loc}}(x_{\text{eval}} - x_{\text{loc}})^2 + d_{\text{loc}}(x_{\text{eval}} - x_{\text{loc}})^3$

---

return $p^{\mathbb{S}}(x_{\text{eval}})$

for $j = 0, 1, \ldots, n-1$. In Fig. 1.19 we have the zero degree spline $B_{0,0}(x)$ for $x_j = 0$. Higher-

degree splines are constructed recursively using lower-degree splines as follows:

$$B_{j,k}(x) = \frac{x - x_j}{x_{j+k} - x_j} B_{j,k-1}(x) + \frac{x_{j+k+1} - x}{x_{j+k+1} - x_{j+1}} B_{j+1,k-1}(x) \qquad k \geq 1.$$

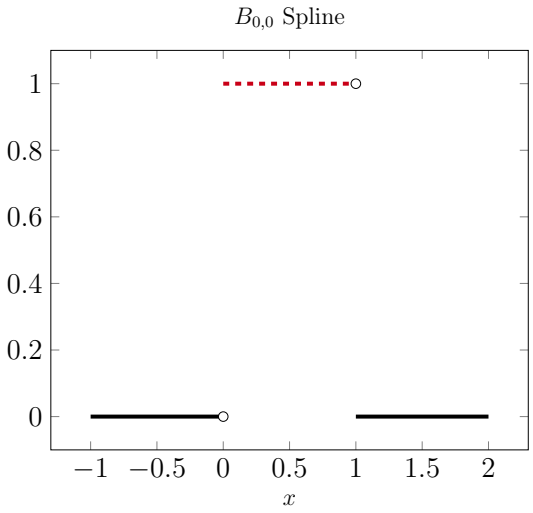

Figure 1.19: Zeroth degree B-spline $B_{0,0}(x)$ for $x_i = 0$.

Although not obvious, one can see that $B_{j,k}(x)$ has one more continuous derivative than $B_{j,k-1}(x)$. Thus while $B_{j,0}(x)$ is discontinuous, $B_{j,1}(x)$ is continuous, $B_{j,2}(x) \in \mathcal{C}^1(\mathbb{R})$, and $B_{j,3}(x) \in \mathcal{C}^2(\mathbb{R})$.

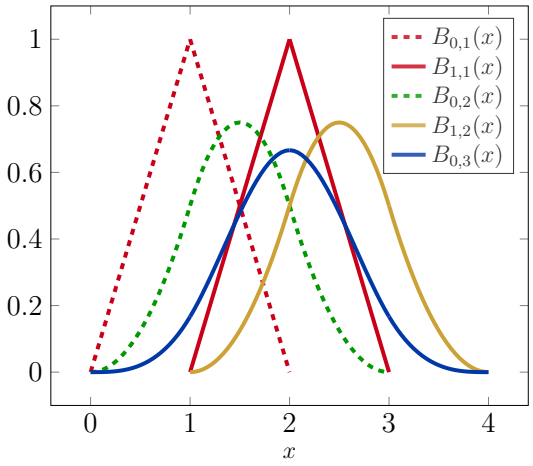

Figure 1.20: Higher degree B-spline polynomial for $x_i = 0$.

As the degree of the B-splines increases, they become more smooth, but the support of $B_{j,k}(x)$ also increases. Based on these results, we can make the following observations:

1. $B_{j,k}(x) \in \mathcal{C}^{k-1}(\mathbb{R})$ (Continuity).
2. $B_{j,k}(x) = 0$ if $x \notin (x_j, x_{j+k+1})$ (Compact Support) [4].

[4] **Compact Support**: Let $f : X \to \mathbb{R}$ be a real-valued function whose domain is an arbitrary set X. The support of f written as $\text{supp}(f)$, is the set of points in X where f is non-zero, i.e., $\text{supp}(f) = \{x \in X : f(x) \neq 0\}$. If $\text{supp}(f)$ is a compact set, then the support is referred to as compact support

3. $B_{j,k}(x) > 0$ for $x \in (x_j, x_{j+k+1})$ (Positivity).

Note: We can notice from Fig. 1.20 that as the degree of the function increases, the support of the function increases as well. Hence, we might get points outside $[x_0, x_n]$. To develop the method, we include additional points beyond the original domain as follows:

$$\cdots < x_{-2} < x_{-1} < x_0 < x_1 < \cdots < x_n < x_{n+1} < \ldots.$$

Let $p_k^{\mathbb{S}}(x)$ denote the spline of piecewise polynomial in $\mathbb{P}_k$. Then we have the following two conditions:

1. $p_k^{\mathbb{S}}(x_i) = f_i$ for $i = 0, 1, \ldots, n$.
2. $p_k^{\mathbb{S}} \in \mathcal{C}^{k-1}[x_0, x_n]$ for $k \geq 1$.

Notice that we have an abuse of notation here. In the previous section we used $p_j^{\mathbb{S}}$ to denote the restriction to $[x_j, x_{j+1}]$, whereas here $p_k^{\mathbb{S}}$ denote a spline of degree k.

Let $c_{j,k}$ denote the unknown coefficients, then

$$p_k^{\mathbb{S}}(x) = \sum_j c_{j,k} B_{j,k}(x).$$

Now the question remains on what values of j the summation applies. For the greatest flexibility, we take j for which

$$B_{j,k}(x) \neq 0 \quad \text{for some } x \in [x_0, x_n].$$

Now, for $k \geq 1$,$B_{j,k}(x)$ has support of (x_j, x_{j+k+1}) and hence

$$p_k^{\mathbb{S}}(x) = \sum_{j=-k}^{n-1} c_{j,k} B_{j,k}(x), \quad k > 1.$$

The inclusion of negative indices for j arises due to the support of the B-spline at boundary knots, particularly at x_0. For the B-spline $B_{j,k}(x)$ to contribute at x_0, its support must include x_0. Since the support of $B_{j,k}(x)$ spans from x_j to x_{j+k+1}, and the last point of this support is x_1 when considering x_0, we require $j + k + 1 = 1$. Solving for j, this gives $j = -k$, which explains the inclusion of "ghost points" $x_{-1}, x_{-2}, \ldots, x_{-k}$ in the extended knot sequence. At the other boundary, x_n, the support extends back to x_{n-1}, ensuring that $B_{j,k}(x)$ contributes only within the domain of the spline. To include all valid intervals in the original knot sequence, the upper bound for j is $j \leq n - 1$. Thus, the range of j is determined as $-k \leq j \leq n - 1$, ensuring that the spline remains well-defined and accounts for boundary contributions at x_0 and x_n.

Hence we have $n + k$ (include $k = 0$) coefficients (unknowns) that satisfy the $n + 1$ interpolation condition,

$$p_k^{\mathbb{S}}(x_i) = f_i = \sum_{j=-k}^{n-1} c_{j,k} B_{j,k}(x_i) \quad i = 0, 1, \ldots, n.$$

The system becomes underdetermined for higher degrees k, meaning there are more unknowns than equations. As we observed in the cubic interpolation, we might need to impose more conditions to get a system of equations.

Chapter 2

System of Equations

In many applications of science and engineering, solving a system of equations is essential. One prominent example arises in Operations Research, where traffic flow modelling involves solving such systems. The foundational work of Ford and Fulkerson [4] introduced the maximum flow problem, which significantly advanced the theory and applications of system-solving techniques.

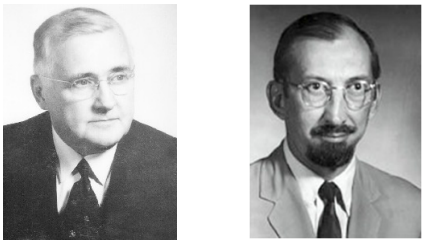

Figure 2.1: Lester Randolph Ford Jr. (23 September 1927–26 February 2017, left) and Delbert Ray Fulkerson (14 August 1924–10 January 1976, right).

Another important application arises in the discretisation of differential equations, a technique widely used in civil, mechanical, and electrical engineering. When differential equations are discretised, the resulting system of equations often takes the form of a *band matrix*[1]. Depending on the choice of polynomial approximation used in the discretisation, the resulting band matrix can be tridiagonal, pentadiagonal, or a more general band matrix. Efficiently solving these systems is crucial to obtaining solutions to the differential equations.

In this chapter, we first introduce *direct methods* for solving systems of equations, followed by iterative methods. Direct methods aim to find the exact solution theoretically in a finite number of steps, though practical computations are subject to round-off errors, which must be carefully managed to ensure accuracy.

[1]**Band Matrix:** A matrix $\{a_{ij}\}_{i,j=1}^{n}$ is called a band matrix if all elements outside a certain diagonal band are zero. The band is determined by:

$$a_{ij} = 0 \quad \text{if} \quad j < i - k_1 \quad \text{or} \quad j > i + k_2; \quad k_1, k_2 \geq 0,$$

where k_1 and k_2 are the lower and upper bandwidths, respectively. Special cases include diagonal matrices ($k_1 = k_2 = 0$) and tridiagonal matrices ($k_1 = k_2 = 1$).

2.1 Gaussian Elimination

The most fundamental method that one studies in linear algebra for solving the system of equation is the Gaussian elimination. Even though the method has been developed independently in ancient China and early modern Europe, it became popular due to Gauss and hence it was named Gaussian elimination by George Forsythe.

Figure 2.2: Carl Friedrich Gauss: 30 April 1777-23 February 1855.

Suppose we have a system of n equations for n variables of the form

$$\begin{aligned} \mathrm{R}_1 \quad &: a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \quad = b_1 \\ \mathrm{R}_2 \quad &: a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \quad = b_2 \\ &\qquad\qquad \vdots \qquad\qquad\qquad\quad \vdots \\ \mathrm{R}_n \quad &: a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \quad = b_n. \end{aligned} \tag{2.1}$$

In Eq. (2.1) the matrix $\{a_{ij}\}_{i,j=1}^{n}$ and the vector $\{b_i\}_{i=1}^{n}$ are given and the $\{x_i\}_{i=1}^{n}$ are the unknowns.

These system of equations follow certain rules due to which we can transform it into a "simpler" system of equations, i.e., easy to solve. We recall three properties that helps us to achieve this:

1. **Scalar Multiplication:** $\mathrm{R}_i \mapsto \lambda \mathrm{R}_i$ for $\lambda \in \mathbb{R}$.
2. **Scalar Multiplication and Adding:** $\mathrm{R}_i \mapsto \mathrm{R}_i + \lambda \mathrm{R}_j$ for some $j = 1, \ldots, n$ and $j \neq i$.
3. **Transposition:** $\mathrm{R}_i \leftrightarrow \mathrm{R}_j$ for $i \neq j$.

We can represent the system of equation presented in Eq. (2.1) as a $n \times (n+1)$ matrix $[\mathbf{A}, \mathbf{b}]$ and this is called as the *augmented matrix* and is given by

$$[\mathbf{A}, \mathbf{b}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \end{array}\right]. \tag{2.2}$$

The line is shown so as to represent the separation between $\mathbf{A}$ and $\mathbf{b}$. The idea of the Gaussian elimination with backward substitution is to reduce the system provided in Eq. (2.2) to an upper triangular matrix and then perform backward substitution.

Let us for the uniformity of the notation denote b_i by $a_{i,n+1}$ for $i = 1, 2, \ldots, n$ then

$$\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{b}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \ldots & a_{2n} & a_{2,n+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & a_{n,n+1} \end{array}\right]. \tag{2.3}$$

Provided $a_{11} \neq 0$ we perform the operations corresponding to

$$\mathrm{R}_j \mapsto \mathrm{R}_j - \frac{a_{j1}}{a_{11}}\mathrm{R}_1 \quad \text{for} \quad j = 2, 3, \ldots, n,$$

to eliminate the coefficient x_1 in each of the rows. Once the coefficients of x_1 are cancelled, we do the same for other rows and follow a sequential procedure for $i = 2, 3, \ldots, n-1$ and perform the operation

$$\mathrm{R}_j \mapsto \mathrm{R}_j - \frac{a_{ji}}{a_{ii}}\mathrm{R}_i \quad \text{for} \quad j = i+1, i+2, \ldots, n.$$

The resulting matrix has the form

$$\tilde{\tilde{\mathbf{A}}} = [\mathbf{A}, \mathbf{b}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & a_{1,n+1} \\ 0 & \tilde{a}_{22} & \ldots & \tilde{a}_{2n} & \tilde{a}_{2,n+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \tilde{a}_{nn} & \tilde{a}_{n,n+1} \end{array}\right].$$

This system of equation has the same solution set as Eq. (2.1). But the new system of equation has the form:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= a_{1,n+1} \\ \tilde{a}_{22}x_2 + \cdots + \tilde{a}_{2n}x_n &= \tilde{a}_{2,n+1} \\ \vdots \quad \vdots \quad \vdots & \\ \tilde{a}_{nn}x_n &= \tilde{a}_{n,n+1}. \end{aligned}$$

By backward substitution we get

$$x_n = \frac{\tilde{a}_{n,n+1}}{\tilde{a}_{nn}}.$$

Solving the $(n-1)^{\text{th}}$ equation for x_{n-1} and using the value of x_n we get

$$\begin{aligned} \tilde{a}_{n-1,n-1}x_{n-1} + \tilde{a}_{n-1,n}x_n &= \tilde{a}_{n-1,n+1} \\ x_{n-1} &= \frac{\tilde{a}_{n-1,n+1} - \tilde{a}_{n-1,n}x_n}{\tilde{a}_{n-1,n-1}}. \end{aligned}$$

Continuing this process we get

$$x_i = \frac{\tilde{a}_{i,n+1} - \sum_{j=i+1}^{n} \tilde{a}_{ij}x_j}{\tilde{a}_{ii}},$$

for $i = n-1, n-2, \ldots, 2, 1$ where for $i = 1$, $\tilde{a}_{1,n+1} = a_{1,n+1}$ and $\tilde{a}_{11} = a_{11}$.

Gaussian elimination can also be seen more precisely by forming a sequence of augmented matrices $\tilde{\mathbf{A}}^{(1)}, \tilde{\mathbf{A}}^{(2)}, \ldots, \tilde{\mathbf{A}}^{(n)}$ where $\tilde{\mathbf{A}}^{(1)}$ is the matrix given in Eq. (2.3) and a general $\tilde{\mathbf{A}}^{(k)}$ matrix for $k = 2, 3, \ldots, n$ is given by

$$\tilde{\mathbf{A}}^{(k)} = \left[\begin{array}{ccccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1,k-1}^{(1)} & a_{1,k}^{(1)} & \ldots & a_{1,n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2,k-1}^{(2)} & a_{2,k}^{(2)} & \ldots & a_{2,n}^{(2)} & a_{2,n+1}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \ldots & a_{k-1,n}^{(k-1)} & a_{k-1,n+1}^{(k-1)} \\ 0 & 0 & 0 & \ldots & 0 & a_{k,k}^{(k)} & \ldots & a_{k,n}^{(k)} & a_{k,n+1}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & a_{n,k}^{(k)} & \ldots & a_{n,n}^{(k)} & a_{n,n+1}^{(k)} \end{array}\right]. \tag{2.4}$$

where x_{k-1} has been eliminated from $\mathrm{R}_k, \ldots, \mathrm{R}_n$.

In general the matrix entries are given by

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{if } \; i = 1, 2, \ldots, k-1 \;\text{ and }\; j = 1, 2, \ldots, n+1, \\ 0 & \text{if } \; i = k, k+1, \ldots, n \;\text{ and }\; j = 1, 2, \ldots, k-1, \\ a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)} & \text{if } \; i = k, k+1, \ldots, n \;\text{ and }\; j = k, k+1, \ldots, n+1. \end{cases}$$

This procedure will fail if any of the elements $\{a_{ii}^{(i)}\}$ for $i = 1, 2, \ldots, n$ is zero as

$$\mathrm{R}_i \mapsto \mathrm{R}_i - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} \mathrm{R}_k$$

cannot be performed or the backward substitution fails.

The system may still have solution but the technique might be altered.

Example 2.1. Consider the augmented matrix

$$\tilde{\mathbf{A}} = \tilde{\mathbf{A}}^{(1)} = \left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right].$$

Performing the operations, $\mathrm{R}_3 \mapsto \mathrm{R}_3 - \mathrm{R}_1$, $\mathrm{R}_4 \mapsto \mathrm{R}_4 - \mathrm{R}_1$, and $\mathrm{R}_2 \mapsto \mathrm{R}_2 - 2\mathrm{R}_1$,

$$\tilde{\mathbf{A}} = \tilde{\mathbf{A}}^{(2)} = \left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

Here $a_{22}^{(2)}$ is zero and is called the *pivot* element. Hence the procedure cannot proceed. So we search the second column for first non-zero entry after 2^{nd} row. Since, $a_{32}^{(2)} \neq 0$, we perform $\mathrm{R}_2 \leftrightarrow \mathrm{R}_3$ and then proceed.

The above example shows what happens if $a_{kk}^{(k)} = 0$ for some $k = 1, 2, \ldots, n-1$. In this case we follow:

1. The k^{th} column of $\tilde{\mathbf{A}}^{(k-1)}$ is searched from the k^{th} row to the n^{th} row for first non zero entry, $a_{pk}^{(k)} \neq 0$ for $k+1 \leq p \leq n$.

2. Then $\mathrm{R}_p \leftrightarrow \mathrm{R}_k$ is performed to get a temporary matrix $\tilde{\mathbf{A}}^{(k-1)'}$ and then the usual elimination follows.

In the case $a_{pk}^{(k)} = 0$ for each p, then the system does not have a unique solution as two columns are the linearly dependent. Finally, if $a_{nn}^{(n)} = 0$ then the system does not have a unique solution.

The algorithm for the Gaussian elimination is provided in Algorithm 6. Although the algorithm looks like we are creating new matrices $\tilde{\mathbf{A}}^{(i)}$ for $i = 1, 2, \ldots, n$ but we can perform all the computation using only one $n \times (n+1)$ matrix for storage.

2.1.1 Computational Complexity

Now we look at the computational complexity of the Gaussian elimination. Generally time taken to perform a multiplication or division is generally more than addition or subtraction. Hence, we count these operations separately.

The arithmetic operations happens in Step 2.3:

1. **Computation of m_{ki}:** Requires division and $(n-i)$ operations.

2. **Multiplication of $m_{ki}\mathrm{R}_i$:** This multiplication happens with the non-zero entries of R_i which is $(n-i) \times (n-i+1)$ as non zero entries is given by $(n-i+1)$ in the R_i^{th} row.

3. **Subtraction for $\mathrm{R}_k - m_{ki}\mathrm{R}_i$:** These will also $(n-i+1) \times (n-i)$ as we subtract the non-zero entries.

The first two are multiplication and division and the last one is addition/subtraction.

Multiplication/Division Complexity

Now, $(n-i) + (n-i) \times (n-i+1) = (n-i) \times (n-i+2) = (n-i)^2 + 2(n-i)$. Summing i from 1 to $n-1$ we get

$$\begin{aligned}
\sum_{i=1}^{n-1}(n-i)(n-i+2) &= \sum_{i=1}^{n-1}(n-i)^2 + 2\sum_{i=1}^{n-1}(n-i) \\
&= \sum_{i=1}^{n-1} i^2 + 2\sum_{i=1}^{n-1} i \\
&= \frac{(n-1)n(2n-1)}{6} + \frac{2n(n-1)}{2} = \frac{2n^3+3n^2-5n}{6}.
\end{aligned}$$

In the above equation we have used the basic identities of summation, namely $\sum_{i=1}^{n} i^2$ and $\sum_{i=1}^{n} i$.

Algorithm 6 Gauss Elimination

Given: Matrix $\mathbf{A}$, right hand side $\mathbf{b}$ and dimension n.
Find: Solution $\mathbf{x}$.

Step 1:Create Augmented Matrix $\tilde{\mathbf{A}}$
Initialize $\tilde{\mathbf{A}}$ as a zero matrix of size $n \times (n+1)$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** $n+1$ **do**
    if $j \leq n$ **then**
      $\tilde{\mathbf{A}}_{ij} = \mathbf{A}_{ij}$
    else
      $\tilde{\mathbf{A}}_{ij} = \mathbf{b}_i$
    end if
  end for
end for

Step 2: Reduce the matrix to Row-Echelon form
for $i = 1$ **to** $n-1$ **do**

Step 2.1: Check Pivot
Initialize $p = -1$
for $q = i$ **to** n **do**
  if $\tilde{\mathbf{A}}_{qi} \neq 0$ **then**
    $p = q$
    break
  end if
end for
if $p = -1$ **then**
  Output("No Unique Solution")
  exit()
end if

Step 2.2: Exchange Rows $\mathrm{R}_i \leftrightarrow \mathrm{R}_p$
if $p \neq i$ **then**
  temp $= 0$
  for $j = 1$ **to** $n+1$ **do**
    temp $= \tilde{\mathbf{A}}_{ij}$
    $\tilde{\mathbf{A}}_{ij} = \tilde{\mathbf{A}}_{pj}$
    $\tilde{\mathbf{A}}_{pj} =$ temp
  end for
end if

Step 2.3: Matrix Reduction
$m_{ki} = 0$
for $k = i+1$ **to** n **do**
  $m_{ki} = \tilde{\mathbf{A}}_{ki}/\tilde{\mathbf{A}}_{ii}$
  for $j = i$ **to** $n+1$ **do**
    $\tilde{\mathbf{A}}_{kj} = \tilde{\mathbf{A}}_{kj} - m_{ki}\tilde{\mathbf{A}}_{ij}$
  end for
end for
end for

Step 3: Check for no Solution
if $\tilde{\mathbf{A}}_{nn} = 0$ **then**
  Output("No Unique Solution")
  exit()
end if

Step 4: Backward Substitution
Initialize x as a vector of size n
$x_n = \frac{\tilde{\mathbf{A}}_{n,n+1}}{\tilde{\mathbf{A}}_{nn}}$
for $i = n-1$ **to** 1 **do**
  sum $= 0$
  for $j = i+1$ **to** n **do**
    sum $=$ sum $+ \tilde{\mathbf{A}}_{ij}x_j$
  end for
  $x_i = \frac{\tilde{\mathbf{A}}_{i,n+1} - \text{sum}}{\tilde{\mathbf{A}}_{ii}}$
end for

---

return $\{x_i\}_{i=1}^{n}$

Addition/Subtraction Complexity

$$\begin{aligned}\sum_{i=1}^{n-1}(n-i)(n-i+1) &= \sum_{i=1}^{n-1}(n-i)^2+\sum_{i=1}^{n-1}(n-i)\\ &= \sum_{i=1}^{n-1}i^2+\sum_{i=1}^{n-1}i\\ &= \frac{n(n-1)(2n-1)}{6}+\frac{n(n-1)}{2}=\frac{n^3-n}{3}.\end{aligned}$$

Hence we notice that in Step 2.3 we require $\mathcal{O}(n^3)$ operations.

The next step that require arithmetic operations are the ones in backward substitution, i.e, Step 4. First is in the computation of x_n which requires one division. For the computation of rest of the $\{x_i\}$ we need $(n-i)$ multiplications and one division for each i and $(n-i-1)$ addition for each summation followed by one subtraction.

Multiplication/Division Complexity

$$\begin{aligned}1+\sum_{i=1}^{n-1}((n-i)+1) &= 1+\left(\sum_{i=1}^{n-1}(n-i)\right)+n-1\\ &= n+\sum_{i=1}^{n-1}(n-i)\\ &= n+\sum_{i=1}^{n-1}i\\ &= n+\frac{n(n-1)}{2}=\frac{n^2+n}{2}.\end{aligned}$$

Addition/Subtraction Complexity

$$\sum_{i=1}^{n-1}((n-i-1)+1) = \sum_{i=1}^{n-1}(n-i)=\frac{n^2-n}{2}.$$

Hence in total we require

$$\frac{2n^3+3n^2-5n}{6}+\frac{n^2+n}{2}=\frac{n^3}{3}+n^2-\frac{n}{3},$$

operations for multiplication and division; and

$$\frac{n^3-n}{3}+\frac{n^2-n}{2}=\frac{2n^3+3n^2-5n}{6},$$

for addition and subtraction. Hence we have $\mathcal{O}(n^3/3)$ computational complexity.

2.1.2 Gauss-Jordan Algorithm

Wilhelm Jordan was a geodesist who extended the basic Gaussian elimination to achieve a full row-reduced echelon form of a matrix. Do not confuse Wilhelm Jordan with Camille Jordan (who gave us Jordan Curve theorem and Jordan Canonical form).

Figure 2.3: Wilhelm Jordan: 1 March 1842-17 April 1899.

This method is a variation of Gaussian elimination where the variable x_i is not only removed from $\mathrm{R}_{i+1}, \mathrm{R}_{i+2}, \ldots, \mathrm{R}_n$ but also from $\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_{i-1}$. Upon this reduction the augmented matrix looks like

$$[\mathbf{A}, \mathbf{b}] = \left[\begin{array}{cccc|c} a_{11}^{(1)} & 0 & \ldots & 0 & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & 0 & a_{2,n+1}^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & a_{nn}^{(n)} & a_{n,n+1}^{(n)} \end{array}\right].$$

Then the solution can be easily obtained using

$$x_i = \frac{a_{i,n+1}^{(i)}}{a_{ii}^{(i)}}, \qquad \text{for } \; i = 1, 2, \ldots, n.$$

The Gauss-Jordan algorithm is presented in Algorithm 7.

2.2 Matrix Factorisation

Like polynomial interpolation, which was the basis for developing more efficient algorithms such as Lagrange and Newton divided differences, Gaussian elimination is the foundation for more advanced topics.

Gaussian elimination consists of two steps: the row-reduction step and backward substitution. The former has a computational complexity of $\mathcal{O}(n^3)$, while the latter requires only $\mathcal{O}(n^2)$. This means that if we have a triangular matrix, solving the system requires only $\mathcal{O}(n^2)$ operations.

2.2.1 LU Decomposition

Suppose that we have $\mathbf{A} = \mathbf{LU}$, meaning that $\mathbf{A}$ has been factored into a lower triangular matrix $(\mathbf{L})$[2] and an upper triangular matrix $(\mathbf{U})$[3]. Then, solving $\mathbf{Ax} = \mathbf{b}$ can be done in two

[2]**Lower Triangular Matrix:** A matrix $\mathbf{A} = \{a_{ij}\}_{i,j=1}^n$ is said to be lower triangular if $a_{ij} = 0$ for $i > j$.
[3]**Upper Triangular Matrix:** A matrix $\mathbf{A} = \{a_{ij}\}_{i,j=1}^n$ is said to be upper triangular if $a_{ij} = 0$ for $j > i$.

Algorithm 7 Gauss Jordan

Given: Matrix $\mathbf{A}$, right hand side $\mathbf{b}$ and dimension n.
Find: Solution $\mathbf{x}$.

Step 1:Create Augmented Matrix $\tilde{\mathbf{A}}$
Initialize $\tilde{\mathbf{A}}$ as a zero matrix of size $n \times (n+1)$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** $n+1$ **do**
    if $j \leq n$ **then**
      $\tilde{\mathbf{A}}_{ij} = \mathbf{A}_{ij}$
    else
      $\tilde{\mathbf{A}}_{ij} = \mathbf{b}_i$
    end if
  end for
end for

Step 2: Reduce the matrix to Row-Echelon form
for $i = 1$ **to** n **do**

Step 2.1: Check Pivot
Initialize $p = -1$
for $q = i$ **to** n **do**
  if $\tilde{\mathbf{A}}_{qi} \neq 0$ **then**
    $p = q$
    break
  end if
end for
if $p = -1$ **then**
  Output("No Unique Solution")
  exit()
end if

Step 2.2: Exchange Rows $\mathrm{R}_i \leftrightarrow \mathrm{R}_p$
if $p \neq i$ **then**
  temp $= 0$
  for $j = 1$ **to** $n+1$ **do**
    temp $= \tilde{\mathbf{A}}_{ij}$
    $\tilde{\mathbf{A}}_{ij} = \tilde{\mathbf{A}}_{pj}$
    $\tilde{\mathbf{A}}_{pj} =$ temp
  end for
end if

Step 2.3: Matrix Reduction
$m_{ki} = 0$
for $k = 1$ **to** n **do**
  if $k = i$ **then**
    continue
  else
    $m_{ki} = \tilde{\mathbf{A}}_{ki} / \tilde{\mathbf{A}}_{ii}$
    for $j = i$ **to** $n+1$ **do**
      $\tilde{\mathbf{A}}_{kj} = \tilde{\mathbf{A}}_{kj} - m_{ki}\tilde{\mathbf{A}}_{ij}$
    end for
  end if
end for
end for

Step 3: Check for no Solution
if $\tilde{\mathbf{A}}_{nn} = 0$ **then**
  Output("No Unique Solution")
  exit()
end if

Step 4: Backward Substitution
Initialize x as a zero vector of size n
for $i = 1$ **to** n **do**
  $x_i = \frac{\tilde{\mathbf{A}}_{i,n+1}}{\tilde{\mathbf{A}}_{ii}}$
end for

---

return $\{x_i\}_{i=1}^{n}$

steps:

- Solve $\mathbf{Ly} = \mathbf{b}$ for $\mathbf{y}$.
- Solve $\mathbf{Ux} = \mathbf{y}$ for $\mathbf{x}$.

Both steps require only $\mathcal{O}(n^2)$ operations.

Although several mathematicians introduced LU decomposition, the Polish mathematician Tadeusz Banachiewicz is credited with generalizing the method for arbitrary matrices.

Figure 2.4: Tadeusz Banachiewicz: 13 February 1882 - 17 November 1954.

LU decomposition reduces an $\mathcal{O}(n^3/3)$ problem to an $\mathcal{O}(2n^2)$ problem. This reduction is useful but comes at a cost: the factorization of $\mathbf{A}$ into $\mathbf{L}$ and $\mathbf{U}$ itself requires $\mathcal{O}(n^3/3)$ operations. However, once computed, the factorization can be stored and reused for multiple right-hand-side vectors $\mathbf{b}$.

To proceed with LU decomposition, we assume that $\mathbf{Ax} = \mathbf{b}$ can be solved using Gaussian elimination without row pivoting, i.e., $a_{ii}^{(i)} \neq 0$ for $i = 1, 2, \ldots, n$.

The first step in Gaussian elimination consists of performing, for each $j = 2, 3, \ldots, n$,

$$\mathrm{R}_j \mapsto \mathrm{R}_j - m_{j1}\mathrm{R}_1, \qquad \text{where} \quad m_{j1} = \frac{a_{j1}^{(1)}}{a_{11}^{(1)}}.$$

An equivalent way of viewing this is by multiplying $\mathbf{A}$ on the left by the matrix $\mathbf{M}^{(1)}$, where

$$\mathbf{M}^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -m_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -m_{n1} & 0 & \cdots & 1 \end{bmatrix}.$$

This is called the *first Gaussian transformation matrix*. The product of this matrix with $\mathbf{A}$ is denoted by $\mathbf{A}^{(2)}$, so that

$$\mathbf{A}^{(2)} = \mathbf{M}^{(1)}\mathbf{A}.$$

Similarly, the right-hand side vector is updated as

$$\mathbf{b}^{(2)} = \mathbf{M}^{(1)}\mathbf{b}.$$

Next, we construct $\mathbf{M}^{(2)}$ by replacing the subdiagonal entries in the second column of the identity matrix with the negative of the multipliers

$$m_{j2} = \frac{a_{j2}^{(2)}}{a_{22}^{(2)}}.$$

This process continues until we obtain an upper triangular matrix $\mathbf{A}^{(n)}$, given by

$$\mathbf{A}^{(n)} = \mathbf{M}^{(n-1)}\mathbf{M}^{(n-2)}\cdots\mathbf{M}^{(1)}\mathbf{A}.$$

At this point, we define $\mathbf{U} = \mathbf{A}^{(n)}$ as the upper triangular matrix in the LU factorization.

To compute the lower triangular matrix $\mathbf{L}$, we note that the inverse of each $\mathbf{M}^{(k)}$ matrix is given by

$$\mathbf{L}^{(k)} = \left[\mathbf{M}^{(k)}\right]^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & m_{k+1,k} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{n,k} & 0 & \cdots & 1 \end{bmatrix}.$$

The lower triangular matrix $\mathbf{L}$ is then obtained as

$$\mathbf{L} = \mathbf{L}^{(1)}\mathbf{L}^{(2)}\cdots\mathbf{L}^{(n-1)}.$$

Since each $\mathbf{L}^{(k)}$ is the inverse of $\mathbf{M}^{(k)}$, we confirm that

$$\mathbf{LU} = \mathbf{A}.$$

Theorem 2.2. (Doolittle LU Decomposition) *If Gaussian elimination can be performed on the system* $\mathbf{Ax} = \mathbf{b}$ *without row interchanges, then the matrix* $\mathbf{A}$ *can be factored as* $\mathbf{A} = \mathbf{LU}$*, where*

$$m_{ji} = \frac{a_{ji}^{(i)}}{a_{ii}^{(i)}},$$

and

$$\mathbf{U} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{(n)} \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & 1 \end{bmatrix}.$$

The above factorization is the Doolittle method, where $\mathbf{L}$ has ones on its diagonal. Alternatively, if the ones are placed on the diagonal of $\mathbf{U}$, the technique is called Crout's LU decomposition.

Once the LU factorization is obtained, the system $\mathbf{Ax} = \mathbf{LUx}$ is solved efficiently by first computing $\mathbf{y}$ from $\mathbf{Ly} = \mathbf{b}$ using forward substitution and then solving $\mathbf{Ux} = \mathbf{y}$ using backward substitution.

It is important to note that not all square matrices have an LU factorization. For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has no LU factorization. Suppose it did; then there would exist $\mathbf{L}$ and $\mathbf{U}$ such that $\mathbf{A} = \mathbf{LU}$. However, this would lead to a contradiction, as one of the factors would necessarily be singular while $\mathbf{A}$ is not.

Next, we note that the LU decomposition is not unique.

Theorem 2.3. *If a matrix has an LU decomposition, then it is not unique.*

Proof. Let $\mathbf{A}$ have an LU decomposition, i.e., $\mathbf{A} = \mathbf{LU}$. Then, we can write

$$\begin{aligned} \mathbf{A} &= \mathbf{LU} \\ &= \mathbf{LDD}^{-1}\mathbf{U} \\ &= (\mathbf{LD})\left(\mathbf{D}^{-1}\mathbf{U}\right), \end{aligned}$$

where $\mathbf{D}$ is any diagonal matrix. Since $\mathbf{LD}$ remains lower triangular and $\mathbf{D}^{-1}\mathbf{U}$ is still upper triangular, we obtain infinitely many LU decompositions of $\mathbf{A}$ by varying $\mathbf{D}$. □

PLU Decomposition

So far, we have assumed that LU decomposition is applicable to systems of equations that do not require pivoting. However, in general, pivoting is necessary. To introduce LU decomposition with pivoting, we first define the permutation matrix.

Definition 2.4. A *permutation matrix* $\mathbf{P} = \{p_{ij}\}_{i,j=1}^{n}$ is an $n \times n$ matrix obtained by rearranging the rows of the identity matrix.

For example, the matrix

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

is a 3×3 permutation matrix where the second and third rows are interchanged. For any 3×3 matrix $\mathbf{A}$, multiplying by $\mathbf{P}$ on the left swaps these two rows:

$$\mathbf{PA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

Let $k_1, k_2, \ldots, k_n$ be a permutation of $1, 2, \ldots, n$. The permutation matrix $\mathbf{P}$ is then defined as:

$$p_{ij} = \begin{cases} 1 & \text{if } \ j = k_i, \\ 0 & \text{otherwise.} \end{cases}$$

This satisfies the following properties:

1. **PA** permutes the rows of **A**:

$$\mathbf{PA} = \begin{bmatrix} a_{k_1 1} & a_{k_1 2} & \dots & a_{k_1 n} \\ a_{k_2 1} & a_{k_2 2} & \dots & a_{k_2 n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k_n 1} & a_{k_n 2} & \dots & a_{k_n n} \end{bmatrix}.$$

2. The inverse of a permutation matrix exists and is given by $\mathbf{P}^{-1} = \mathbf{P}^\top$.

In the previous section, we saw that for any nonsingular matrix $\mathbf{A}$, the linear system $\mathbf{Ax} = \mathbf{b}$ can be solved using Gaussian elimination with row interchanges. If the required row interchanges are known beforehand, we can apply them initially, allowing us to use LU decomposition without further row swaps. That is, for any nonsingular matrix $\mathbf{A}$, there exists a permutation matrix $\mathbf{P}$ such that the system

$$\mathbf{PAx} = \mathbf{Pb}$$

can be solved without row interchanges. Consequently, we can factorize $\mathbf{PA}$ as

$$\mathbf{PA} = \mathbf{LU}.$$

Since $\mathbf{P}$ is a permutation matrix, we have $\mathbf{P}^{-1} = \mathbf{P}^\top$, which implies

$$\mathbf{A} = \left(\mathbf{P}^\top \mathbf{L}\right) \mathbf{U}.$$

While $\mathbf{U}$ remains upper triangular, the matrix $\mathbf{P}^\top \mathbf{L}$ may not necessarily be lower triangular unless $\mathbf{P} = \mathbf{I}$.

Based on this, we establish the following lemma.

Lemma 2.5. *Let $\mathbf{A}$ be an $n \times n$ matrix. Then, there exists a permutation matrix $\mathbf{P}$ such that $\mathbf{PA}$ has an LU decomposition, i.e., $\mathbf{PA} = \mathbf{LU}$.*

The next theorem addresses the uniqueness of the LU decomposition.

Theorem 2.6. *Let $\mathbf{A}$ be an $n \times n$ matrix, and let $\mathbf{P}$ be an $n \times n$ permutation matrix such that $\mathbf{PA}$ has an LU decomposition. If $\mathbf{A}$ is invertible, then there exists a unique $n \times n$ lower triangular matrix $\mathbf{L}$ with all diagonal entries equal to 1, and a unique $n \times n$ upper triangular matrix $\mathbf{U}$ such that*

$$\mathbf{PA} = \mathbf{LU}.$$

Proof. The existence of the LU decomposition follows from Lemma 2.5. We now prove the uniqueness.

Suppose $\mathbf{L}$ is not unit lower triangular. Then, we can express the decomposition as

$$\mathbf{PA} = \mathbf{LU}.$$

Rewriting,

$$\mathbf{PA} = \mathbf{LD}^{-1}\mathbf{DU},$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal entries match those of $\mathbf{L}$. Since $\mathbf{A}$ is invertible, $\mathbf{L}$ is also invertible, ensuring that $\mathbf{D}^{-1}$ exists. Defining

$$\mathbf{L}_1 = \mathbf{LD}^{-1}, \quad \mathbf{U}_1 = \mathbf{DU},$$

we obtain a new factorization with $\mathbf{L}_1$ as a unit lower triangular matrix and $\mathbf{U}_1$ as an upper triangular matrix:

$$\mathbf{PA} = \mathbf{L}_1\mathbf{U}_1.$$

Now, suppose there exists another decomposition:

$$\mathbf{PA} = \mathbf{L}_2\mathbf{U}_2,$$

where $\mathbf{L}_2$ is also unit lower triangular. Then, we equate the two decompositions:

$$\mathbf{L}_1\mathbf{U}_1 = \mathbf{L}_2\mathbf{U}_2.$$

Since $\mathbf{A}$ is invertible, both $\mathbf{L}_1$ and $\mathbf{P}$ are invertible, implying that $\mathbf{U}_1 = \mathbf{L}_1^{-1}\mathbf{PA}$ is also invertible.

Thus, we obtain

$$\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{U}_2\mathbf{U}_1^{-1}. \tag{2.5}$$

Since:

1. The inverse of a lower (upper) triangular matrix is lower (upper) triangular.
2. The product of lower (upper) triangular matrices remains lower (upper) triangular.

it follows that $\mathbf{L}_2^{-1}\mathbf{L}_1$ is lower triangular, and $\mathbf{U}_2\mathbf{U}_1^{-1}$ is upper triangular. Since $\mathbf{L}_2^{-1}\mathbf{L}_1$ is also unit diagonal, the only possibility is

$$\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{I} \Rightarrow \mathbf{L}_2 = \mathbf{L}_1.$$

Similarly, we obtain $\mathbf{U}_1 = \mathbf{U}_2$, proving uniqueness. □

2.2.2 LDL$^\top$ Decomposition

In linear algebra we have certain special matrices and they enjoy certain "good" properties. This is true with respect to their LU decomposition as well. We first mention certain matrices, followed by their properties, and then their special kind of factorisation.

Definition 2.7. A matrix $\mathbf{A}$ is said to be *diagonally dominant* when

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}| \qquad \forall \;\; i = 1, 2, \ldots, n.$$

If the inequality is strict then it is called as *strictly diagonally dominant.*

Algorithm 8 LU Decomposition with Partial Pivoting

Given: Matrix $\mathbf{A}$ of size $n \times n$.
Find: Matrices $\mathbf{L}$, $\mathbf{U}$, and $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{LU}$.

Step 1: Initialize Matrices
Initialize $\mathbf{L}$ as an $n \times n$ identity matrix.
Initialize $\mathbf{P}$ as an $n \times n$ identity matrix.
Initialize $\mathbf{U}$ as $\mathbf{A}$.

Step 2: Perform LU Decomposition
for $i = 1$ **to** $n - 1$ **do**

Step 2.1: Check Pivot
Initialize $p = -1$
for $q = i$ **to** n **do**
if $\mathbf{U}_{qi} \neq 0$ **then**
$p = q$
break
end if
end for
if $p = -1$ **then**
Output("Matrix is singular but the LU decomposition still exists!")
continue
end if

Step 2.2: Exchange Rows for $\mathbf{P}$ and $\mathbf{U}$, $\mathrm{R}_i \leftrightarrow \mathrm{R}_p$
if $p \neq i$ **then**
temp1 $= 0$; temp2 $= 0$.
for $j = 1$ **to** $n + 1$ **do**
temp1 $= \mathbf{P}_{ij}$; temp2 $= \mathbf{U}_{ij}$
$\mathbf{P}_{ij} = \mathbf{P}_{pj}$; $\mathbf{A}_{ij} = \mathbf{A}_{pj}$.
$\mathbf{P}_{pj} =$ temp1; $\mathbf{A}_{pj} =$ temp2.
end for
end if

Step 2.3: Matrix Reduction
for $k = i + 1$ **to** n **do**
$m_{ki} = \mathbf{U}_{ki}/\mathbf{U}_{ii}$
$\mathbf{L}_{ki} = m_{ki}$
for $j = i$ **to** n **do**
$\mathbf{U}_{kj} = \mathbf{U}_{kj} - m_{ki}\mathbf{U}_{ij}$
end for
end for
end for

---

return $\mathbf{L}, \mathbf{U}, \mathbf{P}$

Theorem 2.8. *A strictly diagonally dominant matrix* $\mathbf{A}$ *is non-singular. Moreover, in this case, Gaussian elimination can be performed on any linear system of the form* $\mathbf{Ax} = \mathbf{b}$ *to obtain it's unique solution without row or column interchanges, and the computations will be stable with respect to the growth of round-off errors.*

Proof. This theorem has three parts:

1. Non-Singularity of $\mathbf{A}$.
2. Unique solution using Gaussian Elimination and no row-interchange.
3. Stability of the solution.

We will prove the first two parts, as the proof of the third part is out of the scope of this lecture. For the first part we use the method of contradiction. Suppose $\mathbf{A}$ is singular. Then the system $\mathbf{Ax} = \mathbf{0}$ has non-trivial solution, say $\mathbf{x} = \{x_i\}$. Let k be an index for which

$$0 < |x_k| = \max_{1 \le j \le n} |x_j|.$$

As $\mathbf{Ax} = \mathbf{0}$, we get $\sum_{j=1}^{n} a_{ij}x_j = 0$ for $i = 1, 2, \ldots, n$. At $i = k$

$$\sum_{j=1}^{n} a_{kj}x_j = 0 \Rightarrow a_{kk}x_k = -\sum_{j=1, j\neq k}^{n} a_{kj}x_j.$$

From the triangular inequality we have

$$\begin{aligned} |a_{kk}||x_k| &= \left| \sum_{j=1, j\neq k}^{n} a_{kj}x_j \right| \\ &\le \sum_{j=1, j\neq k}^{n} |a_{kj}||x_j| \\ &< \sum_{j=1, j\neq k}^{n} |a_{kj}||x_k| \end{aligned}$$

Hence, $|a_{kk}| < \sum_{j=1, j\neq k}^{n} |a_{kj}|$ which is a contradiction as $\mathbf{A}$ is strictly diagonally dominant. Hence the matrix $\mathbf{A}$ is singular.

For the second part we show that the matrices $\mathbf{A}^{(k)}$ for $k = 2, 3, \ldots, n$ generated during the Gaussian elimination is strictly diagonally dominant. Hence it ensure that each pivot element is non-zero.

Since $\mathbf{A}$ is strictly diagonally dominant, $a_{11} \neq 0$ and $\mathbf{A}^{(2)}$ can be formed. Thus for each $i = 2, 3, \ldots, n$,

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad \text{for} \;\; 2 \le j \le n.$$

First, $a_{i1}^{(2)} = 0$. Now using the triangle inequality

$$\sum_{j=2,j\neq i}^{n} |a_{ij}^{(2)}| = \sum_{j=2,j\neq i}^{n} \left| a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}} \right| \leq \sum_{j=2,j\neq i}^{n} |a_{ij}^{(1)}| + \sum_{j=2,j\neq i}^{n} \left| \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}} \right|.$$

But since $\mathbf{A}$ is strictly diagonally dominant,

$$\begin{aligned} \sum_{j=1,j\neq i}^{n} |a_{ij}^{(1)}| &< |a_{ii}^{(1)}| \\ \sum_{j=2,j\neq i}^{n} |a_{ij}^{(1)}| &< |a_{ii}^{(1)}| - |a_{i1}^{(1)}|, \end{aligned}$$

and similarly

$$\begin{aligned} \sum_{j=1,j\neq i}^{n} |a_{1j}^{(1)}| &< |a_{11}^{(1)}| \\ \sum_{j=2,j\neq i}^{n} |a_{1j}^{(1)}| &< |a_{11}^{(1)}| - |a_{1i}^{(1)}|, \end{aligned}$$

so

$$\sum_{j=2,j\neq i}^{n} |a_{ij}^{(2)}| < |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|} \left(|a_{11}^{(1)}| - |a_{1i}^{(1)}| \right) = |a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|}.$$

The reverse triangle inequality implies

$$|a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \leq \left| a_{ii}^{(1)} - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \right| = |a_{ii}^{(2)}|,$$

which gives

$$\sum_{j=2,j\neq i} |a_{ij}^{(2)}| < |a_{ii}^{(2)}|.$$

This establish the strict diagonal dominance for rows $2, \ldots, n$. But the first row of $\mathbf{A}^{(2)}$ and $\mathbf{A}$ are the same, so $\mathbf{A}^{(2)}$ is strictly diagonally dominant.

We can continue this process inductively and see that the result holds. □

Definition 2.9. A matrix $\mathbf{A}$ is said to be *positive definite* if $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. If the matrix is symmetric then it is referred to as *symmetric positive definite*.

For the next few theorems and corollaries, we will not be presenting the proofs can be found in Linear Algebra books. If you are interested, you can refer to [6].

Theorem 2.10. (**Necessary Conditions for Symmetric Positive Definite**) *If* $\mathbf{A}$ *is a* $n \times n$ *symmetric positive definite matrix then*

1. $\mathbf{A}$ *has an inverse.*
2. $a_{ii} > 0$ *for each* $i = 1, 2, \ldots, n$.
3. $\max_{1 \le k,j \le n} |a_{kj}| \le \max_{1 \le i \le n} |a_{ii}|$
4. $(a_{ij})^2 < a_{ii}a_{jj}$ *for each* $i \neq j$.

These conditions are only necessary conditions. For sufficient and necessary condition we introduce the notion of leading principal sub-matrix.

Definition 2.11. A *leading principal sub-matrix* of a matrix $\mathbf{A}$ is a matrix of the form

$$\mathbf{A}_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix},$$

for some $k = 1, 2, \ldots, n$.

Theorem 2.12. **(Necessary and Sufficient Condition for Symmetric Positive Definite)** *A symmetric matrix* $\mathbf{A}$ *is symmetric positive definite if and only if it's leading principal sub-matrices have a positive determinant.*

Example 2.13. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}.$$

It has three principal sub-matrix, $\mathbf{A}_1, \mathbf{A}_2$, and $\mathbf{A}_3$ each having determinants as $2, 3$, and 4 respectively.

Theorem 2.14. *A symmetric matrix* $\mathbf{A}$ *is symmetric positive definite if and only if Gaussian elimination without row interchanges can be performed on the linear system* $\mathbf{Ax} = \mathbf{b}$ *with all positive pivot element. Moreover in this case, the computations are stable with respect to the growth of round-off errors.*

Again we are not interested in the proof of the above theorem but rather certain corollaries that come while proving this theorem.

Corollary 2.15. (**LDL$^\top$ Factorisation**) *The matrix* **A** *is symmetric positive definite if and only if* **A** *can be factored in the form* **LDL**$^\top$ *where* **L** *is a unit lower triangular matrix and* **D** *is a diagonal matrix with positive diagonal entries.*

Corollary 2.2.2 has a counterpart in case we have **A** as a symmetric matrix.

Corollary 2.16. *Let* **A** *be a symmetric matrix for which Gaussian elimination can be applied without row interchange. Then* **A** *can be factored into* **LDL**$^\top$*, where* **L** *is lower unit triangular matrix with ones on the diagonal and* **D** *is the diagonal matrix with* $a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{nn}^{(n)}$.

The algorithm to compute **LDL**$^\top$ is presented in Algorithm 9.

Algorithm 9 $\mathrm{LDL}^\top$ Decomposition

Given: Symmetric matrix $\mathbf{A}$ of size $n \times n$.
Find: Matrix $\mathbf{L}$ (with unit diagonal) and $\mathbf{D}$ such that $\mathbf{A} = \mathbf{LDL}^T$.

Step 1: Initialize Matrices
Initialize $\mathbf{L}$ as an identity matrix of size $n \times n$.
Initialize $\mathbf{D}$ as a zero matrix of size $n \times n$.

Step 2: Compute D and L
$\mathbf{D}_{11} = \mathbf{A}_{11}$
for $i = 1$ **to** n **do**

Step 2.1: Compute D
if $i \neq 1$ **then**
sum $= 0$
for $j = 1$ **to** $i - 1$ **do**
sum $=$ sum $+ \mathbf{D}_{jj}\mathbf{L}_{ij}^2$
end for
$\mathbf{D}_{ii} = \mathbf{A}_{ii} -$ sum
end if

Step 2.2: Compute L
for $j = i + 1$ **to** n **do**
sum $= 0$
if $i \neq 1$ **then**
for $k = 1$ **to** $i - 1$ **do**
sum $=$ sum $+ \mathbf{D}_{kk}\mathbf{L}_{ik}\mathbf{L}_{jk}$
end for
end if
$\mathbf{L}_{ji} = \frac{\mathbf{A}_{ji} - \text{sum}}{\mathbf{D}_{ii}}$
end for
end for

---

return $\mathbf{L}, \mathbf{D}$

Algorithm 9 is based on the computation of individual entries of $\mathbf{L}$ and $\mathbf{D}$. Let us take an example of how these entries actually look like or to be more precise how this algorithm is created.

Example 2.17. Let $\mathbf{A}$ be a 3×3 symmetric matrix having $\mathrm{LDL}^\top$ decomposition. Then

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & \ell_{21} & \ell_{31} \\ 0 & 1 & \ell_{32} \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} d_{11} & d_{11}\ell_{21} & d_{11}\ell_{31} \\ d_{11}\ell_{21} & d_{22} + d_{11}\ell_{21}^2 & d_{22}\ell_{32} + d_{11}\ell_{21}\ell_{31} \\ d_{11}\ell_{31} & d_{22}\ell_{32} + d_{11}\ell_{21}\ell_{31} & d_{11}\ell_{31}^2 + d_{22}\ell_{32}^2 + d_{33} \end{bmatrix}.\end{aligned}$$

We notice that $d_{11} = a_{11}$ and $\ell_{i1} = a_{i1}/d_{11}$ for $i = 2, 3$. After this we can compute d_{22} and then ℓ_{32}. Finally we compute d_{33}. Same process can be extended to a $n \times n$ matrix.

2.2.3 Cholesky Decomposition

From Theorem 2.14 we have another corollary related to symmetric positive definite matrix which gives another decomposition.

Corollary 2.18. (**Cholesky Decomposition**) *The matrix* $\mathbf{A}$ *is symmetric positive definite if and only if* $\mathbf{A}$ *can be factored in the form* $\mathbf{LL}^\top$ *where* $\mathbf{L}$ *is a lower triangular matrix with non-zero diagonal entries.*

The Cholesky decomposition was discovered by André-Louis Cholesky who was a French military officer (along with being a mathematician).

Figure 2.5: André-Louis Cholesky: 15 October 1875-31 August 1918.

The algorithm for the Cholesky decomposition can be found in Algorithm 10.

Algorithm 10 Cholesky Decomposition

Given: Symmetric positive definite matrix $\mathbf{A}$ of size $n \times n$.
Find: Matrices $\mathbf{L}$ such that $\mathbf{A} = \mathbf{L}\mathbf{L}^T$.

Step 1: Initialize Matrix
Initialize $\mathbf{L}$ as a zero matrix of size $n \times n$.

Step 2: Compute L
$\mathbf{L}_{11} = \sqrt{\mathbf{A}_{11}}$
for $j = 2$ **to** n **do**
  $\mathbf{L}_{j1} = \frac{\mathbf{A}_{j1}}{\mathbf{L}_{11}}$
end for
for $i = 2$ **to** n **do**

Step 2.1: Compute $\mathbf{L}_{ii}$
sum $= 0$
for $k = 1$ **to** $i - 1$ **do**
  sum $=$ sum $+ \mathbf{L}_{ik}^2$
end for
$\mathbf{L}_{ii} = \sqrt{\mathbf{A}_{ii} - \text{sum}}$
for $j = i + 1$ **to** n **do**

Step 2.2: Compute $\mathbf{L}_{ji}$
sum $= 0$
for $k = 1$ **to** $i - 1$ **do**
  sum $=$ sum $+ \mathbf{L}_{jk}\mathbf{L}_{ik}$
end for
$\mathbf{L}_{ji} = \frac{1}{\mathbf{L}_{ii}} (\mathbf{A}_{ji} - \text{sum})$
end for
end for

---

return $\mathbf{L}$

Algorithm 10 is based on the computation of individual entries of $\mathbf{L}$. Let us take an example of how these entries actually looks like.

Example 2.19. Let $\mathbf{A}$ be a 3×3 symmetric positive definite matrix having a $\mathrm{LL}^\top$ decomposition. Then

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ 0 & \ell_{22} & \ell_{32} \\ 0 & 0 & \ell_{33} \end{bmatrix} \\ &= \begin{bmatrix} \ell_{11}^2 & \ell_{11}\ell_{21} & \ell_{11}\ell_{31} \\ \ell_{11}\ell_{21} & \ell_{21}^2 + \ell_{22}^2 & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} \\ \ell_{11}\ell_{31} & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \end{bmatrix}.\end{aligned}$$

We notice that $\ell_{11} = \sqrt{a_{11}}$ and $\ell_{i1} = a_{i1}/\ell_{11}$ for $i = 2, 3$. After this we can compute ℓ_{22} and then ℓ_{32}. Finally we compute ℓ_{33}. Same process can be extended to a $n \times n$ matrix.

Until now we have not discussed about the computational complexity of any of the three factorisation methods. Table 2.1 gives the applicability, computational complexity, and advantages of the three factorisation methods. For brevity, we would not derive the computational complexity for the methods but interested students can try it for their own.

Method	**Applicability**	**Advantages**	**Computational Complexity**	
			M/D	**A/S**
LU	General square matrices	Works for any matrix but requires pivoting	$\frac{n^3}{3} - \frac{n}{3}$	$\frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6}$
$\mathrm{LDL}^\top$	Symmetric matrices	More stable than LU, reduces storage, avoids pivoting	$\frac{n^3}{6} + n^2 - \frac{7n}{6}$	$\frac{n^3}{6} - \frac{n}{6}$
Cholesky	Symmetric positive definite matrices	Fastest and most efficient, lowest computation cost	$\frac{n^3}{6} + \frac{n^2}{2} - \frac{2n}{3}$	$\frac{n^3}{6} - \frac{n}{6}$

Table 2.1: Applicability, Computational Complexity, and Advantages for LU, $\mathrm{LDL}^\top$, and Cholesky Decomposition. M/D: Multiplication and Division, A/S: Addition and Subtraction.

We notice that the Cholesky decomposition requires the least number of operations while factorisation but it can be a little misleading as it requires extracting n square roots. However the computation of square root is a linear factor of n and will decrease significantly as n increases.

2.3 Iterative Methods

Root-finding methods are a class of iterative methods that we are aware of. In this part of the chapter, we will translate these ideas into a system of equations. Before delving into iterative methods for a system of equations, we need to find a way to measure the distance between n-dimensional column vectors. This will help us determine the sequence of vectors that converge to the solution of the system.

Definition 2.20. A *vector norm* on $\mathbb{R}^n$ is a function, $\|\cdot\|$ from $\mathbb{R}^n$ into $\mathbb{R}$ with the following properties:

1. $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
3. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for all $\alpha \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$.
4. $\|\mathbf{x} + \mathbf{y}\| \leq \|x\| + \|y\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Definition 2.21. The ℓ_2 and the ℓ_∞ norm for the vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are defined by

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

If we define a unit ball in $\mathbb{R}^2$ using these norms, then they are given by $\|\mathbf{x}\|_2 \leq 1$ which is an unit disc centred at $(0, 0)$ and $\|\mathbf{x}\|_\infty \leq 1$ which is a square.

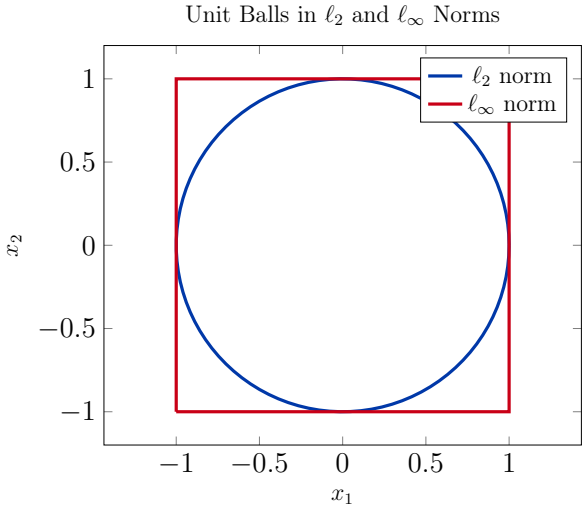

Figure 2.6: Unit balls in ℓ_2 and ℓ_∞ norm.

A fundamental property of these norms that is widely used is the Cauchy-Schwarz inequality.

Theorem 2.22. (**Cauchy-Schwarz Inequality**) *For each* $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ *and* $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ *in* $\mathbb{R}^n$

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i \leq \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2} \left\{ \sum_{i=1}^{n} y_i^2 \right\}^{1/2} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Proof. The result is immediate if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = 0\mathbf{0}$. Suppose $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$. Now, note that

for each $\lambda \in \mathbb{R}$ we have

$$0 \leq \|\mathbf{x} - \lambda \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (x_i - \lambda y_i)^2 = \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i y_i + \lambda^2 \sum_{i=1}^{n} y_i^2,$$

so that

$$2\lambda \sum_{i=1}^{n} x_i y_i \leq \|\mathbf{x}\|_2^2 + \lambda^2 \|\mathbf{y}\|_2^2.$$

As, $\|\mathbf{x}\|_2 > 0$ and $\|\mathbf{y}\|_2 > 0$ so we let $\lambda = \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2}$, which gives

$$\frac{2\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \left(\sum_{i=1}^{n} x_i y_i \right) \leq \|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{y}\|_2^2} \|\mathbf{y}\|_2^2 = 2\|\mathbf{x}\|_2^2,$$

which, after simplification, gives us the result. □

The norm of a vector measures the distance between an arbitrary vector and the zero vector. We define the distance between two vectors as

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left\{ \sum_{i=1}^{n} (x_i - y_i)^2 \right\}^{1/2} \quad \text{and} \quad \|\mathbf{x} - \mathbf{y}\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|.$$

Now, we define the convergence of a sequence of vectors in $\mathbb{R}^n$.

Definition 2.23. A sequence $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ of vectors in $\mathbb{R}^n$ is said to converge to $\mathbf{x}$ with respect to $\|\cdot\|$ if given for any $\varepsilon > 0$ there exist a $N(\varepsilon)$ such that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| < \varepsilon \quad \forall k \geq N(\varepsilon).$$

Next, we present the result regarding the equivalence of norms.

Theorem 2.24. *For each* $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n} \|\mathbf{x}\|_\infty.$$

Proof. Let x_j be the coordinate such that $\|\mathbf{x}\|_\infty = |x_j| = \max_{1 \leq i \leq n} |x_i|$.

Now,

$$\|\mathbf{x}\|_\infty^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 = \|\mathbf{x}\|_2^2.$$

Similarly,

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 \leq x_j^2 n \leq \|\mathbf{x}\|_\infty^2 n.$$

Hence, $\|\mathbf{x}\|_2 \leq \sqrt{n} \|\mathbf{x}\|_\infty$. □

Similar to vector norms, we also have matrix norms. The measure given to a matrix under a natural norm describes how the matrix stretches unit vectors relative to that norm. The maximum stretch is the norm of the matrix. The definition of the matrix norm is similar to that of the vector norm.

Theorem 2.25. *If $\|\cdot\|$ is a vector norm in $\mathbb{R}^n$, then*

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$$

is a matrix norm.

Note that we have an abuse of notation here; we denote $\|\cdot\|$ to show both the vector and the matrix norm.

Matrix norms defined by vector norms are called *natural* or *induced* matrix norm. We can also write the natural matrix norms as

$$\|\mathbf{A}\| = \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{y}\|}{\|\mathbf{y}\|},$$

as $\mathbf{y}/\|\mathbf{y}\|$ is a unit vector.

Corollary 2.26. *For any vector $\mathbf{y} \neq \mathbf{0}$, matrix $\mathbf{A}$ and any natural norm $\|\cdot\|$, we have*

$$\|\mathbf{A}\mathbf{y}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{y}\|.$$

The matrix norm that we consider are the ∞ norm, i.e.,$\|\mathbf{A}\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty$ and the ℓ_2 norm, i.e., $\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$.

Lastly, we define the $\|\cdot\|_\infty$ norm of a matrix.

Theorem 2.27. *If $\mathbf{A} = \{a_{ij}\}_{i,j=1}^n$ is a $n \times n$ matrix then*

$$\|\mathbf{A}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

A square matrix $\mathbf{A}$ takes the set of n-dimensional vectors into itself, which gives a linear function from $\mathbb{R}^n$ to $\mathbb{R}^n$. After this transformation, certain vectors might be parallel to the original vector, i.e., $\mathbf{x}$ is parallel to $\mathbf{A}\mathbf{x}$. It might be stretched, shrinked,or remains unchanged. The magnitude with which it stretches or shrinked is called the eigen or characteristic value. But why do we care about these eigenvalues? There is a close relation between these eigenvalues and the convergence of the iterative methods.

Definition 2.28. If $\mathbf{A}$ is a square matrix, then the *characteristic polynomial* of $\mathbf{A}$ is defined by

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}).$$

Definition 2.29. If $p(\lambda)$ is the characteristic polynomial of the matrix $\mathbf{A}$, the zeros of p are *eigenvalues* or *characteristic values* of the matrix $\mathbf{A}$. If λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x} \neq \mathbf{0}$ satisfies $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ then $\mathbf{x}$ is an *eigenvector* or characteristic vector of $\mathbf{A}$ corresponding the λ.

Note that if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ associated with λ and $\alpha \in \mathbb{R} \setminus \{0\}$ then $\alpha\mathbf{x}$ is an eigenvector since

$$\mathbf{A}(\alpha\mathbf{x}) = \alpha(\mathbf{A}\mathbf{x}) = \alpha(\lambda\mathbf{x}) = \lambda(\alpha\mathbf{x}).$$

As an immediate consequence of this is $\mathbf{x}$ is an eigenvector then we can choose $\alpha = \pm\|\mathbf{x}\|^{-1}$, which would make $\alpha\mathbf{x}$ an eigenvector with norm one. So far any eigenvalue and any vector norm we have eigenvectors with norm one.

Definition 2.30. The *spectral radius* $\rho(\mathbf{A})$ of a matrix is defined by

$$\rho(\mathbf{A}) = \max\{|\lambda|\},$$

where λ is an eigenvalue of $\mathbf{A}$. For $\lambda(:= \alpha + i\beta) \in \mathbb{C}$, $|\lambda| = (\alpha^2 + \beta^2)^{1/2}$.

Next, we have a relation between the spectral radius and the matrix norm.

Theorem 2.31. *If $\mathbf{A}$ is a $n \times n$ matrix then:*
1. $\|\mathbf{A}\|_2 = \left[\rho\left(\mathbf{A}^\top\mathbf{A}\right)\right]^{1/2}$.
2. $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$, *for any natural norm* $\|\cdot\|$.

If $\mathbf{A}$ is symmetric then $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$. Apart from spectral radius, another important property to study is how a matrix's power behaves.

Definition 2.32. We call $n \times n$ matrix $\mathbf{A}$ *convergent* if

$$\lim_{k\to\infty}\left(\mathbf{A}^k\right)_{ij} = 0 \quad \forall\ i = 1, 2, \ldots, n \quad \text{and} \quad j = 1, 2, \ldots, n.$$

Now, convergent matrices have a special connection with the spectral radius.

Theorem 2.33. *The following statements are equivalent:*
1. $\mathbf{A}$ *is a convergent matrix.*
2. $\lim_{n\to\infty}\|\mathbf{A}^n\| = 0$ *for some natural norm.*
3. $\lim_{n\to\infty}\|\mathbf{A}^n\| = 0$ *for all natural norms.*
4. $\rho(\mathbf{A}) < 1$.
5. $\lim_{n\to\infty}\mathbf{A}^n\mathbf{x} = \mathbf{0}$ *for all* $\mathbf{x}$.

2.3.1 Jacobi Method

After having a quick glance at the basics of linear algebra, we move back toward the domain of numerical analysis.

An iterative technique to solve the $n \times n$ linear system $\mathbf{Ax} = \mathbf{b}$ starts with an initial approximation $\mathbf{x}^{(0)}$ to the solution $\mathbf{x}$ and generates a sequence of vectors $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ such that $\mathbf{x}^{(k)} \to \mathbf{x}$ as $k \to \infty$.

Carl Gustav Jacob Jacobi was a German mathematician who proposed the Jacobi eigenvalue algorithm, an iterative method for calculating the eigenvalues and eigenvectors of a real symmetric matrix. The *Jacobi method* that we study is the stripped-down version of this algorithm only.

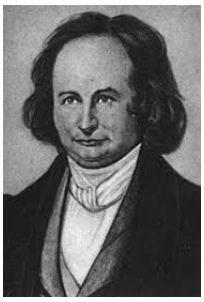

Figure 2.7: Carl Gustav Jacob Jacobi: 10 December 1804-18 February 1851

The Jacobi iterative method is obtained by solving the i^{th} equation in $\mathbf{Ax} = \mathbf{b}$ for x_i to obtain

$$x_i = \sum_{j=1, j\neq i}^{n} \left(-\frac{a_{ij}x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}} \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{2.6}$$

For each $k \geq 1$ we generate the components $x_i^{(k)}$ of $\mathbf{x}^{(k)}$ from the components of $\mathbf{x}^{(k-1)}$ by

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{j=1, j\neq i}^{n} \left(-a_{ij}x_j^{(k-1)} \right) + b_i \right] \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{2.7}$$

Example: Say we have the system of equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned}$$

Suppose $\mathbf{x}^{(0)}$ is the initial iterate, then the first iterative solution is given by

$$\begin{aligned} x_1^{(1)} &= \frac{1}{a_{11}}\left(b_1 - \left(a_{12}x_2^{(0)} + \cdots + a_{1n}x_n^{(0)}\right)\right) \\ x_2^{(1)} &= \frac{1}{a_{22}}\left(b_2 - \left(a_{21}x_1^{(0)} + \cdots + a_{2n}x_n^{(0)}\right)\right) \\ &\vdots \\ x_n^{(1)} &= \frac{1}{a_{nn}}\left(b_n - \left(a_{n1}x_1^{(0)} + \cdots + a_{n,n-1}x_{n-1}^{(0)}\right)\right), \end{aligned}$$

and similarly, we compute for $k \geq 1$.

In general, iterative techniques for solving linear systems of equations involve a process that converts $\mathbf{Ax} = \mathbf{b}$ into an equivalent system $\mathbf{x} = \mathbf{Tx}+\mathbf{c}$ for some fixed-matrix $\mathbf{T}$ and vector $\mathbf{c}$. Once the initial approximation is selected say $\mathbf{x}^{(0)}$ we get

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c} \quad \text{for each } k = 1, 2, \ldots.$$

We can have an equivalent formulation for the Jacobi method by splitting $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U},$$

where $\mathbf{L}$ is strict lower triangular part of $\mathbf{A}$, $\mathbf{U}$ is the strict upper triangular part of $\mathbf{A}$, and $\mathbf{D}$ is the diagonal. Say,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

Then,

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \cdots & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \cdots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Then we can re-write $\mathbf{Ax} = \mathbf{b}$ as

$$\mathbf{Dx} = (\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b},$$

and if $\mathbf{D}^{-1}$ exist, $\mathbf{x} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}$.

Then, the Jacobi iterative is given by

$$\mathbf{x}^{(k)} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k-1)} + \mathbf{D}^{-1}\mathbf{b} \quad \text{for} \quad k \geq 1. \tag{2.8}$$

Denoting $\mathbf{T}_{\mathrm{J}} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ and $\mathbf{c}_{\mathrm{J}} = \mathbf{D}^{-1}\mathbf{b}$ then we get the Jacobi iteration as

$$\mathbf{x}^{(k)} = \mathbf{T}_{\mathrm{J}}\mathbf{x}^{(k-1)} + \mathbf{c}_{\mathrm{J}} \quad \text{for} \quad k \geq 1. \tag{2.9}$$

We need $a_{ii} \neq 0$ for each $i = 1, 2, \ldots, n$. If one of the $a_{ii} = 0$ and the system is not singular, then the equations can be reordered so that no a_{ii} is zero.

The algorithm for the Jacobi method is provided in Algorithm 11.

Algorithm 11 Jacobi Iteration

Given: Matrix $\mathbf{A}$ with non-zero pivots, right hand side $\mathbf{b}$, dimension n, `max_iterations`, and `tolerance`.
Find: Solution $\mathbf{x}$.

Step 1: Jacobi Iterations
Initialize $\mathbf{x}^{\texttt{old}} = \mathbf{0}$
for $k = 1$ **to** `max_iterations` **do**
  for $i = 1$ **to** n **do**
    sum $= \mathbf{b}_i$
    for $j = 1$ **to** n **do**
      if $j \neq i$ **then**
        sum $=$ sum $- \mathbf{A}_{ij}\mathbf{x}_j^{\texttt{old}}$
      end if
    end for
    $\mathbf{x}_i = \frac{\text{sum}}{\mathbf{A}_{ii}}$
  end for
  `Error` $= \|\mathbf{x} - \mathbf{x}^{\texttt{old}}\|_\infty$
  if `Error` $<$ `tolerance` **then**
    Output("Convergence reached")
    break
  end if
  $\mathbf{x}^{\texttt{old}} = \mathbf{x}$
end for
if $k ==$ `max_iterations` **then**
  Output("Maximum Number of iterations reached")
end if

---

return $\mathbf{x}$

2.3.2 Gauss-Seidel Method

In the Jacobi method we require all the components of $\mathbf{x}^{(k-1)}$ are used to compute the components $x_i^{(k)}$ of $\mathbf{x}^{(k)}$. But, for $i > 1$, the component $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$ of $\mathbf{x}^{(k)}$ have already being computed. If we use these values, then it is expected to give better approximations to the actual solutions than $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$.

Then, it is reasonable to compute $x_i^{(k)}$ using the most recently calculated values.

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + b_i \right], \tag{2.10}$$

for $i = 1, 2, \ldots, n$. This is called the *Gauss Seidel method*. Gauss initially developed the

concept in the mid-1820s; it was only published and fully detailed by Seidel in 1874 through a private letter from Gauss to his student Gerling, making the method primarily attributed to both mathematicians.

Figure 2.8: Philipp Ludwig von Seidel: 24 October 1821-13 August 1896

Example: Say we have the system of equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned}$$

Suppose $\mathbf{x}^{(0)}$ is the initial iterate, then the first iterative solution is given by

$$\begin{aligned} x_1^{(1)} &= \frac{1}{a_{11}}\left(b_1 - \left(a_{12}x_2^{(0)} + a_{13}x_3^{(0)} + \cdots + a_{1n}x_n^{(0)}\right)\right) \\ x_2^{(1)} &= \frac{1}{a_{22}}\left(b_2 - \left(a_{21}x_1^{(1)} + a_{23}x_3^{(0)} + \cdots + a_{2n}x_n^{(0)}\right)\right) \\ &\vdots \\ x_n^{(1)} &= \frac{1}{a_{nn}}\left(b_n - \left(a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \cdots + a_{n,n-1}x_{n-1}^{(1)}\right)\right), \end{aligned}$$

and similarly, we compute for $k \geq 1$.

To write the Gauss-Seidel method in matrix form, we multiply Eq. (2.10) with a_{ii} and collect the k^{th} iterate term to get

$$a_{i1}x_1^{(k)} + a_{i2}x_2^{(k)} + \cdots + a_{ii}x_i^{(k)} = -a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i,$$

for $i = 1, 2, \ldots, n$. Then

$$\begin{aligned} a_{11}x_1^{(k)} &= -a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1, \\ a_{21}x_1^{(k)} + a_{22}x_2^{(k)} &= -a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2, \\ &\vdots \\ a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \cdots + a_{nn}x_n^{(k)} &= b_n. \end{aligned}$$

Then, we can write this system as

$$(\mathbf{D}-\mathbf{L})\mathbf{x}^{(k)} = \mathbf{U}\mathbf{x}^{(k-1)} + \mathbf{b},$$

and $\mathbf{x}^{(k)} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(k-1)} + (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}$ for $k \geq 1$, where $\mathbf{D}, \mathbf{L}$, and $\mathbf{U}$ are defined in the same way as Jacobi method. Then denoting $\mathbf{T}_{\text{GS}} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$ and $\mathbf{c}_{\text{GS}} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}$ we get the Gauss-Seidel method as

$$\mathbf{x}^{(k)} = \mathbf{T}_{\text{GS}}\mathbf{x}^{(k-1)} + \mathbf{c}_{\text{GS}}.$$

Now, $\mathbf{D}-\mathbf{L}$ is non singular if and only if $a_{ii} \neq 0$ for all $i = 1, 2, \ldots, n$.

It appears that the Gauss-Seidel method is always a better approximation to the Jacobi method, which is "mostly" true, but we have cases where this might not hold.

The algorithm for the Gauss-Seidel method is provided in Algorithm 12.

Algorithm 12 Gauss-Seidel Iteration

Given: Matrix $\mathbf{A}$ with non-zero pivots, right hand side $\mathbf{b}$, dimension n, `max_iterations`, and `tolerance`.
Find: Solution $\mathbf{x}$.

Step 1: Gauss-Seidel Iterations
Initialize $\mathbf{x}^{\texttt{old}} = \mathbf{0}$
for $k = 1$ **to** `max_iterations` **do**
  for $i = 1$ **to** n **do**
    sum $= \mathbf{b}_i$
    for $j = 1$ **to** n **do**
      if $j < i$ **then**
        sum $=$ sum $- \mathbf{A}_{ij}\mathbf{x}_j$
      else if $i < j$ **then**
        sum $=$ sum $- \mathbf{A}_{ij}\mathbf{x}_j^{\texttt{old}}$
      end if
    end for
    $\mathbf{x}_i = \frac{\text{sum}}{\mathbf{A}_{ii}}$
  end for
  `Error` $= \|\mathbf{x} - \mathbf{x}^{\texttt{old}}\|_\infty$
  if `Error` $<$ `tolerance` **then**
    Output("Convergence reached")
    break
  end if
  $\mathbf{x}^{\texttt{old}} = \mathbf{x}$
end for
if $k ==$ `max_iterations` **then**
  Output("Maximum Number of iterations reached")
end if

---

return $\mathbf{x}$

General Iteration Matrices

We need to analyze the formula

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c} \quad \text{for} \quad k \geq 1,$$

to study the convergence of general iteration techniques where $\mathbf{x}^{(0)}$ is arbitrary.

Lemma 2.34. *If the spectral radius $\rho(\mathbf{T}) < 1$ then $(\mathbf{I} - \mathbf{T})^{-1}$ exists, and*

$$(\mathbf{I} - \mathbf{T})^{-1} = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots = \sum_{j=0}^{\infty} \mathbf{T}^j.$$

Proof. Now, let λ be an eigenvalue of $\mathbf{T}$ with eigenvector $\mathbf{x}$ then

$$\mathbf{T}\mathbf{x} = \lambda\mathbf{x} \iff (\mathbf{I} - \mathbf{T})\mathbf{x} = (1 - \lambda)\mathbf{x}.$$

Hence, λ is an eigenvalue of $\mathbf{T}$ if and only if $1 - \lambda$ is a eigenvalue of $\mathbf{I} - \mathbf{T}$.

However, by the definition of spectral radius $|\lambda| \leq \rho(\mathbf{T}) < 1$, so $\lambda = 1$ is not an eigenvalue of $\mathbf{T}$ which implies 0 is not an eigenvalue of $\mathbf{I} - \mathbf{T}$.

Hence, $\mathbf{I} - \mathbf{T}$ is invertible. Let $\mathbf{S}_m = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots + \mathbf{T}^m$, then

$$(\mathbf{I} - \mathbf{T})\,\mathbf{S}_m = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots + \mathbf{T}^m - \mathbf{T} - \mathbf{T}^2 - \cdots - \mathbf{T}^{m+1} = \mathbf{I} - \mathbf{T}^{m+1}.$$

As $\rho(\mathbf{T}) < 1$ then by Theorem 2.33 we have $\mathbf{T}$ is convergent. Again using Theorem 2.33 we get

$$\lim_{m\to\infty} (\mathbf{I} - \mathbf{T})\mathbf{S}_m = \lim_{m\to\infty} (\mathbf{I} - \mathbf{T}^{m+1}) = \mathbf{I}.$$

Thus

$$(\mathbf{I} - \mathbf{T})^{-1} = \lim_{m\to\infty} \mathbf{S}_m = \sum_{j=0}^{\infty} \mathbf{T}^j.$$

□

Theorem 2.35. *For any $\mathbf{x}^{(0)} \in \mathbb{R}^n$ the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ defined by*

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c} \quad \text{for each } k \geq 1,$$

converges to the unique solution $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$ if and only if $\rho(\mathbf{T}) < 1$.

Proof. Let $\rho(\mathbf{T}) < 1$. Then

$$\begin{aligned}
\mathbf{x}^{(k)} &= \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c} \\
&= \mathbf{T}\left(\mathbf{T}\mathbf{x}^{(k-2)} + \mathbf{c}\right) + \mathbf{c} \\
&= \mathbf{T}^2\mathbf{x}^{(k-2)} + \mathbf{T}\mathbf{c} + \mathbf{c} \\
&\vdots \\
&= \mathbf{T}^k\mathbf{x}^{(0)} + \left(\mathbf{T}^{k-1} + \cdots + \mathbf{I}\right)\mathbf{c}. \qquad (2.11)
\end{aligned}$$

As $\rho(\mathbf{T}) < 1$ from Theorem 2.33 we get that $\mathbf{T}$ is convergent and $\lim_{k\to\infty} \mathbf{T}^k\mathbf{x}^{(0)} = \mathbf{0}$.

In Eq. (2.11) passing the limit of $k \to \infty$, and then using the previous lemma, we get,

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \lim_{k\to\infty} \mathbf{T}^{(k)}\mathbf{x}^{(0)} + \left(\sum_{j=0}^{\infty} \mathbf{T}^j\right)\mathbf{c} = \mathbf{0} + (\mathbf{I}-\mathbf{T})^{-1}\,\mathbf{c}.$$

Hence, $\{\mathbf{x}^{(k)}\} \to \mathbf{x}$ as $k \to \infty$ and $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$ is the unique limit.

Conversely we will show that for any $\mathbf{y} \in \mathbb{R}^n$ we have $\lim_{k\to\infty} \mathbf{T}^k\mathbf{y} = 0$ which is equivalent to $\rho(\mathbf{T}) < 1$.

Let $\mathbf{y} \in \mathbb{R}^n$ be arbitrary and $\mathbf{x}$ be the unique solution to $\mathbf{x} = \mathbf{T}\mathbf{x}+\mathbf{c}$. Define $\mathbf{x}^{(0)} = \mathbf{x}-\mathbf{y}$ and for $k \geq 1$

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}.$$

Now, by the hypothesis $\{\mathbf{x}^{(k)}\} \to \mathbf{x}$. Also,

$$\mathbf{x} - \mathbf{x}^{(k)} = (\mathbf{T}\mathbf{x} + \mathbf{c}) - \left(\mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}\right) = \mathbf{T}(\mathbf{x} - \mathbf{x}^{(k-1)}).$$

Inductively $\mathbf{x} - \mathbf{x}^{(k)} = \mathbf{T}\left(\mathbf{x} - \mathbf{x}^{(k-1)}\right) = \mathbf{T}^2\left(\mathbf{x} - \mathbf{x}^{(k-2)}\right) = \cdots = \mathbf{T}^k\left(\mathbf{x} - \mathbf{x}^{(0)}\right) = \mathbf{T}^k\mathbf{y}$. Hence, $\lim_{k\to\infty} \mathbf{T}^k\mathbf{y} = \lim_{k\to\infty} \mathbf{T}^k\left(\mathbf{x} - \mathbf{x}^{(0)}\right) = \lim_{k\to\infty}\left(\mathbf{x} - \mathbf{x}^{(k)}\right) = \mathbf{0}$.

As $\mathbf{y} \in \mathbb{R}^n$ was arbitrary. By Theorem 2.33 we get that $\rho(\mathbf{T}) < 1$. □

Based on this theorem, a nice corollary bounds the error.

Corollary 2.36. *If $\|\mathbf{T}\| < 1$ for any natural norm and $\mathbf{c}$ is a given vector, then the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ defined by $\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}$ converges, for any $\mathbf{x}^{(0)} \in \mathbb{R}^n$, to a vector $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$. Furthermore, the following error bounds hold:*

1. $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|\mathbf{T}\|^k\|\mathbf{x}^{(0)} - \mathbf{x}\|$.
2. $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|\mathbf{T}\|^k}{1-\|\mathbf{T}\|}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$.

Proof. Let $\|\mathbf{T}\| < 1$. By Theorem 2.31, we have $\rho(\mathbf{T}) < 1$. Therefore, the method converges to the solution $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$ by the previous theorem. We now prove the error bounds.

Starting from the iterative formula $\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}$, we subtract $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$ from both sides and taking the norm:

$$\begin{aligned}\|\mathbf{x}^{(k)} - \mathbf{x}\| &= \|\mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c} - \mathbf{T}\mathbf{x} - \mathbf{c}\| \\ &= \|\mathbf{T}(\mathbf{x}^{(k-1)} - \mathbf{x})\| \\ &\leq \|\mathbf{T}\| \cdot \|\mathbf{x}^{(k-1)} - \mathbf{x}\|.\end{aligned}$$

Applying this inequality recursively:

$$\begin{aligned}\|\mathbf{x}^{(k)} - \mathbf{x}\| &\leq \|\mathbf{T}\| \cdot \|\mathbf{x}^{(k-1)} - \mathbf{x}\| \\ &\leq \|\mathbf{T}\|^2 \cdot \|\mathbf{x}^{(k-2)} - \mathbf{x}\| \\ &\vdots \\ &\leq \|\mathbf{T}\|^k \cdot \|\mathbf{x}^{(0)} - \mathbf{x}\|.\end{aligned}$$

This proves the first error bound.

For the second error bound, consider the norm of the difference between successive iterates:

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| = \|\mathbf{T}(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)})\| \leq \|\mathbf{T}\| \cdot \|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}\|.$$

Applying this inequality recursively:

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \|\mathbf{T}\|^{k-1} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|.$$

For $m > k \geq 1$, we have:

$$\begin{aligned}
\|\mathbf{x}^{(m)} - \mathbf{x}^{(k)}\| &= \|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)} + \mathbf{x}^{(m-1)} - \cdots + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\
&\leq \|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}\| + \|\mathbf{x}^{(m-1)} - \mathbf{x}^{(m-2)}\| + \cdots + \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\
&\leq \|\mathbf{T}\|^{m-1} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| + \|\mathbf{T}\|^{m-2} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| + \cdots + \|\mathbf{T}\|^{k} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|.
\end{aligned}$$

Factoring out $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$:

$$\|\mathbf{x}^{(m)} - \mathbf{x}^{(k)}\| \leq \|\mathbf{T}\|^{k} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \left(1 + \|\mathbf{T}\| + \cdots + \|\mathbf{T}\|^{m-k-1}\right).$$

Taking the limit as $m \to \infty$:

$$\lim_{m\to\infty} \|\mathbf{x}^{(m)} - \mathbf{x}^{(k)}\| \leq \|\mathbf{T}\|^{k} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \sum_{i=0}^{\infty} \|\mathbf{T}\|^{i}.$$

Since $\|\mathbf{T}\| < 1$, we have:

$$\sum_{i=0}^{\infty} \|\mathbf{T}\|^{i} = \frac{1}{1 - \|\mathbf{T}\|}.$$

Thus:

$$\lim_{m\to\infty} \|\mathbf{x}^{(m)} - \mathbf{x}^{(k)}\| \leq \frac{\|\mathbf{T}\|^{k}}{1 - \|\mathbf{T}\|} \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|.$$

Since $\lim_{m\to\infty} \mathbf{x}^{(m)} = \mathbf{x}$, the second error bound follows. □

Hence, if $\rho(\mathbf{T}_{\mathrm{J}})$ and $\rho(\mathbf{T}_{\mathrm{GS}}) < 1$ then we get the convergence of these schemes. We also note from this corollary that the convergence of the method is based on $\|\mathbf{T}\|^{k}$. Now according to Theorem 2.31 we have $\rho(\mathbf{T}) \leq \|\mathbf{T}\|$. We have a more general result, i.e., for given $\varepsilon > 0$, $\rho(\mathbf{T}) \leq \|\mathbf{T}\| \leq \rho(\mathbf{T}) + \varepsilon$. Hence, we can say that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \approx \rho(\mathbf{T})^{k} \|\mathbf{x} - \mathbf{x}^{(0)}\|.$$

Thus, we would like to select methods with minimal $\rho(\mathbf{T})$.

We have a sufficient method to show the convergence of the Jacobi and the Gauss-Seidel method.

Theorem 2.37. *If* $\mathbf{A}$ *is strictly diagonally dominant then for any choice of* $\mathbf{x}^{(0)}$ *both the Jacobi and the Gauss-Seidel methods give sequences* $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ *that converge to the unique solution of* $\mathbf{Ax} = \mathbf{b}$.

There is no best method in general between the Gauss-Seidel and the Jacobi method. However, in exceptional cases, we have some results.

Theorem 2.38. *If* $a_{ij} \leq 0$ *for each* $i \neq j$ *and* $a_{ii} > 0$ *for each* $i = 1, 2, \ldots, n$, *then one and only one of the following statement holds:*

1. $0 \leq \rho(\mathbf{T}_{\text{GS}}) < \rho(\mathbf{T}_{\text{J}}) < 1$,
2. $1 < \rho(\mathbf{T}_{\text{J}}) < \rho(\mathbf{T}_{\text{GS}})$,
3. $\rho(\mathbf{T}_{\text{J}}) = \rho(\mathbf{T}_{\text{GS}}) = 0$,
4. $\rho(\mathbf{T}_{\text{J}}) = \rho(\mathbf{T}_{\text{GS}}) = 1$.

In the above theorem, we note that if 1. holds, then both methods converge together with $\mathbf{T}_{\text{GS}}$ being better, and if 2. holds, then both diverge and $\mathbf{T}_{\text{GS}}$ has "better" divergence.

2.3.3 Successive Over Relaxation

Definition 2.39. Suppose $\tilde{\mathbf{x}} \in \mathbb{R}^n$ is an approximation to the solution of the linear system defined by $\mathbf{Ax} = \mathbf{b}$. The *residual vector* for $\tilde{\mathbf{x}}$ with respect to the system is $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$.

In procedure such as Jacobi or Gauss-Seidel method, a residual vector is associated with each calculation of an approximate component to the solution. The true objective of an iterative method is to generate a sequence of approximation that allows the residual vector to converge rapidly to zero. Suppose

$$\mathbf{r}_i^{(k)} = \left(r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)}\right)^{\top},$$

denote the residual vector for the Gauss-Seidel method corresponding to the approximate solution vector $\mathbf{x}_i^{(k)}$ which is defined by

$$\mathbf{x}_i^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)}\right)^{\top}.$$

Now, the m^{th} component of $\mathbf{r}_i^{(k)}$ is

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)}, \tag{2.12}$$

or equivalently

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)},$$

for each $m = 1, 2, \ldots, n$.

In particular for $m = i$

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} - a_{ii}x_i^{(k-1)},$$

which is equivalent to

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}. \tag{2.13}$$

However, we know from the Gauss Seidel method that

$$x_i^{(k)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right]. \tag{2.14}$$

So we can write Eq. (2.13) as

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}.$$

Consequently the Gauss Seidel method can be characterised as choosing $\mathbf{x}_i^{(k)}$ to satisfy

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}. \tag{2.15}$$

Hence the update is determined by the residual at the current step.

Now let us look at the residual vector $\mathbf{r}_{i+1}^{(k)}$ associated with $\mathbf{x}_{(i+1)}^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)}\right)$ By Eq. (2.12) the i^{th} component of $\mathbf{r}_{i+1}^{(k)}$ is

$$\begin{aligned} r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^{i} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} - a_{ii}x_i^{(k)}. \end{aligned}$$

By Eq. (2.14) we notice that the RHS is zero. In a way $x_{i+1}^{(k)}$ is chosen such that the i^{th} component of $\mathbf{r}_{i+1}^{(k)}$ is zero. But here only one component is zero which may not be the most efficient way to reduce the norm of the vector $\mathbf{r}_{i+1}^{(k)}$. Hence if we modify Eq. (2.15) to

$$x_i^{(k)} = x_i^{(k-1)} + \omega\frac{r_{ii}^{(k)}}{a_{ii}}, \tag{2.16}$$

then for certain values of ω we can reduce the norm of the residual. Eq. (2.16) refers to *relaxation methods*. If $\omega \in (0, 1)$ we get *under relaxation method* and if $\omega > 1$ we get *over relaxation methods*. Generally we refer to them as *successive over relaxation (SOR) methods*.

The idea of the SOR methods were devised simultaneously by Stan Frankel and David M. Young Jr. in the 1950s but the idea of the relaxation methods can be traced back way


Figure 2.9: Stan Frankel (1919 – May 1978, left) and David M. Young Jr. (20 October 1923 – 21 December 2008, right).

earlier. Interestingly, Frankel was also a part of the Manhattan project and was a PostDoc under Oppenheimer.

We first reformulate the SOR method. By Eq. (2.16)

$$x_i^{(k)} = (1-\omega)\, x_i^{(k-1)} + \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right]$$

which is $a_{ii}x_i^{(k)} + \omega \sum_{j=1}^{i-1} a_{ij}x_j^k = (1-\omega)a_{ii}x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} + \omega b_i$. So in vectorise form this is

$$(\mathbf{D} - \omega\mathbf{L})\, \mathbf{x}^{(k)} = [(1-\omega)\mathbf{D} + \omega\mathbf{U}]\, \mathbf{x}^{(k-1)} + \omega\mathbf{b},$$

i.e.,

$$\mathbf{x}^{(k)} = (\mathbf{D} - \omega\mathbf{L})^{-1} [(1-\omega)\mathbf{D} + \omega\mathbf{U}]\, \mathbf{x}^{(k-1)} + \omega\, (\mathbf{D} - \omega\mathbf{L})^{-1}\, \mathbf{b}.$$

Letting $\mathbf{T}_\omega = (\mathbf{D} - \omega\mathbf{L})^{-1} [(1-\omega)\mathbf{D} + \omega\mathbf{U}]$ and $\mathbf{c}_\omega = \omega\, (\mathbf{D} - \omega\mathbf{L})^{-1}\, \mathbf{b}$ we get the SOR method as

$$\mathbf{x}^{(k)} = \mathbf{T}_\omega \mathbf{x}^{(k-1)} + \mathbf{c}_\omega. \tag{2.17}$$

Now, the next big question is what should be the appropriate value of ω. In general for $n \times n$ system we cannot say it but for particular cases we have the answer.

Theorem 2.40. *If $a_{ii} \neq 0$ for all $i = 1, 2, \ldots, n$ then $\rho(\mathbf{T}_\omega) \geq |\omega - 1|$. This means SOR can only converge if $0 < \omega < 2$.*

Proof. Let $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of $\mathbf{T}_\omega$. Then

$$\begin{aligned}
\rho(\mathbf{T}_\omega)^n \geq \prod_{i=1}^{n} \lambda_i &= \det(\mathbf{T}_\omega) \\
&= \det\left((\mathbf{D} - \omega\mathbf{L})^{-1} [(1-\omega)\mathbf{D} + \omega\mathbf{U}] \right) \\
&= \det (\mathbf{D} - \omega\mathbf{L})^{-1} \det\left((1-\omega)\mathbf{D} + \omega\mathbf{U} \right) \\
&= \det(\mathbf{D}^{-1}) \det\left((1-\omega)\mathbf{D} \right) \\
&= \frac{1}{\prod_{i=1}^{n} a_{ii}} (1-\omega)^n \prod_{i=1}^{n} a_{ii} = (1-\omega)^n.
\end{aligned}$$

Now, $\rho(\mathbf{T}_\omega) = \max_{1 \leq i \leq n} |\lambda_i| \geq |1 - \omega|$. Now the method will converge if $\rho(\mathbf{T}_\omega) < 1$. Hence $\omega \in (0, 2)$. □

Next, we present a theorem regarding the convergence of the SOR method for symmetric positive definite matrices.

Theorem 2.41. *If* $\mathbf{A}$ *is symmetric positive definite matrix and* $\omega \in (0, 2)$ *then the SOR method converges for any choice of initial approximation.*

Theorem 2.42. *If* $\mathbf{A}$ *is symmetric positive definite and tridiagonal, then* $\rho(\mathbf{T}_{\text{GS}}) = [\rho(\mathbf{T}_{\text{J}})]^2 < 1$ *and the optimal choice of* ω *for the SOR method is*

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(\mathbf{T}_{\text{J}})]^2}},$$

with this choice of ω *we have* $\rho(\mathbf{T}_\omega) = \omega - 1$.

The SOR algorithm is presented in Algorithm 13.

2.3.4 Condition Number

When solving systems of linear equations, either using *iterative solvers* or *direct solvers*, numerical errors are inevitable. In iterative solvers, errors can accumulate due to finite precision and approximation, while in direct solvers, round-off errors from finite precision arithmetic can also affect the solution.

Consider the system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

where $\mathbf{A}$ is an invertible matrix. Suppose we introduce a small perturbation $\delta\mathbf{b}$ to the right-hand side $\mathbf{b}$, resulting in a perturbed solution $\mathbf{x} + \delta\mathbf{x}$. The perturbed system is given by:

$$\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}.$$

Expanding this and subtracting the original system $\mathbf{A}\mathbf{x} = \mathbf{b}$ yields:

$$\mathbf{A}\delta\mathbf{x} = \delta\mathbf{b}.$$

Since $\mathbf{A}$ is invertible, we can solve for $\delta\mathbf{x}$ as:

$$\delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b}.$$

Taking the norm of both sides, we have:

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{b}\| \leq \|\mathbf{A}^{-1}\| \cdot \|\delta\mathbf{b}\|.$$

Dividing both sides by $\|\mathbf{x}\|$, we obtain the bound on the relative error in the solution:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|}{\|\mathbf{x}\|} \cdot \|\delta\mathbf{b}\|.$$

Algorithm 13 SOR Iteration

Given: Matrix $\mathbf{A}$ with non-zero pivots, right hand side $\mathbf{b}$, dimension n, ω, `max_iterations`, and `tolerance`.
Find: Solution $\mathbf{x}$.

Step 1: SOR Iterations
Initialize $\mathbf{x}^{\texttt{old}} = \mathbf{0}$
for $k = 1$ **to** `max_iterations` **do**
  for $i = 1$ **to** n **do**
    sum $= \mathbf{b}_i$
    for $j = 1$ **to** n **do**
      if $j < i$ **then**
        sum $=$ sum $- \mathbf{A}_{ij}\mathbf{x}_j$
      else if $i < j$ **then**
        sum $=$ sum $- \mathbf{A}_{ij}\mathbf{x}_j^{\texttt{old}}$
      end if
    end for
    $\mathbf{x}_i = (1-\omega)\mathbf{x}_i^{\texttt{old}} + \omega\frac{\text{sum}}{\mathbf{A}_{ii}}$
  end for
  `Error` $= \|\mathbf{x} - \mathbf{x}^{\texttt{old}}\|_\infty$
  if `Error` $<$ `tolerance` **then**
    Output("Convergence reached")
    break
  end if
  $\mathbf{x}^{\texttt{old}} = \mathbf{x}$
end for
if $k ==$ `max_iterations` **then**
  Output("Maximum Number of iterations reached")
end if

---

return $\mathbf{x}$

Since $\mathbf{Ax} = \mathbf{b}$, we know that:

$$\|\mathbf{b}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|.$$

Using this relationship, we can rewrite the relative error as:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

This inequality shows that the relative error in the solution is bounded by the relative error in the right-hand side, scaled by the factor $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$. We define the *condition number* of the matrix $\mathbf{A}$ as:

$$\kappa(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|.$$

The condition number $\kappa(\mathbf{A})$ provides a measure of how sensitive the solution $\mathbf{x}$ is to perturbations in $\mathbf{b}$. Specifically: - If $\kappa(\mathbf{A}) \approx 1$, the system is said to be *well-conditioned*, meaning small perturbations in $\mathbf{b}$ lead to small errors in $\mathbf{x}$. - If $\kappa(\mathbf{A})$ is large, the system is *ill-conditioned*, and even small perturbations in $\mathbf{b}$ may cause large errors in $\mathbf{x}$.

Preconditioners

To reduce the condition number and improve numerical stability, we can use a technique called *preconditioning* . Preconditioning involves transforming the system $\mathbf{Ax} = \mathbf{b}$ by multiplying both sides with a matrix $\mathbf{P}$ to obtain an equivalent system with a lower condition number. There are two common types of preconditioning:

1. **Left Preconditioning:** Multiply both sides of the system by $\mathbf{P}^{-1}$:
$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{P}^{-1}\mathbf{b}.$$

2. **Right Preconditioning:** Solve the system:
$$\mathbf{AP}^{-1}\mathbf{y} = \mathbf{b}, \quad \text{where} \quad \mathbf{x} = \mathbf{P}^{-1}\mathbf{y}.$$

A good preconditioner $\mathbf{P}$ should satisfy two key properties:

1. The convergence of the iterative method applied to the preconditioned system $\mathbf{P}^{-1}\mathbf{A}$ or $\mathbf{AP}^{-1}$ should be faster than for the original system.
2. Solving the system involving $\mathbf{P}$ should be computationally inexpensive.

In practice, a balance must be struck between these two requirements.

Some commonly used preconditioners are:

1. **Jacobi (Diagonal) Preconditioner:** $\mathbf{P} = \mathbf{D}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$.
2. **Forward Gauss-Seidel Preconditioner:** $\mathbf{P} = \mathbf{D} + \mathbf{L}$, where $\mathbf{L}$ is the strict lower triangular part of $\mathbf{A}$.
3. **Backward Gauss-Seidel Preconditioner:** $\mathbf{P} = \mathbf{D} + \mathbf{U}$, where $\mathbf{U}$ is the strict upper triangular part of $\mathbf{A}$.
4. **Symmetric Gauss-Seidel Preconditioner:** $\mathbf{P} = (\mathbf{D} + \mathbf{L})\mathbf{D}^{-1}(\mathbf{D} + \mathbf{U})$.

For convenience, we often denote the preconditioner by $\mathbf{P}^{-1}$ rather than $\mathbf{P}$. Preconditioning is a powerful tool in improving the stability and performance of numerical solvers, and iterative methods are often used in conjunction with preconditioners rather than as standalone solvers.

2.4 Least Square Methods

Least Square Problems has been quite helpful in different science areas, from physics to data science. In simple language, i.e., the mathematical language, we are trying to solve an overdetermined system, i.e., $\mathbf{Ax} = \mathbf{b}$, by minimizing the ℓ_2 norm of the residual.

Consider a linear system of equation $\mathbf{Ax} = \mathbf{b}$ with n unknowns and m equations with $m > n$, i.e.,

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

Hence we need to compute $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. A direct solution to such a problem does not exist as the invertibility of the matrix is in question. Hence, instead, we try to reduce the residual $\mathbf{r}$ given by

$$\mathbf{r} = \mathbf{b} - \mathbf{Ax} \in \mathbb{R}^m.$$

What do we mean by reduction? If we choose the ℓ_2 norm then the problem is: Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \geq n$ and $\mathbf{b} \in \mathbb{R}^m$, find $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{b} - \mathbf{Ax}\|_2$ is minimized.

The choice of ℓ_2 norm can be justified geometrically. We seek a vector $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{Ax} \in \mathbb{R}^m$ is closest to point $\mathbf{b}$ in range of $\mathbf{A}$ (see Fig. 2.4).

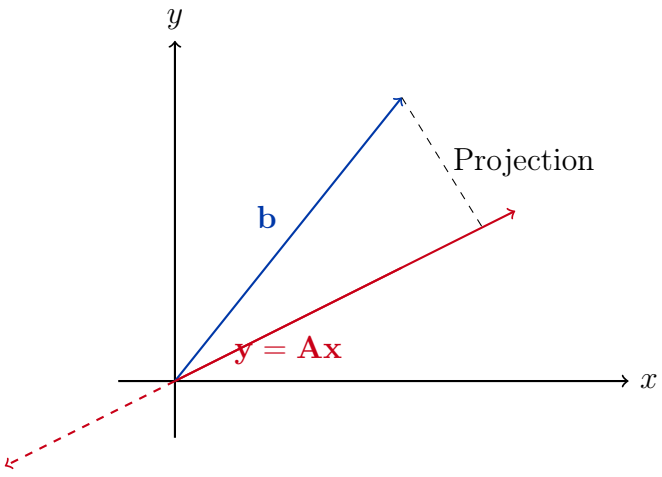

Figure 2.10: Orthogonal projection.

We want to find $\mathbf{Ax}$ in the range of $\mathbf{A}$ such that $\mathbf{r} = \mathbf{Ax} - \mathbf{b}$ is minimum. It is clear from the geometry that $\mathbf{Ax} = \mathbf{Pb}$ is the solution where $\mathbf{P} \in \mathbb{R}^{m \times m}$ is the orthogonal projection operator that maps $\mathbb{R}^m$ to range of $\mathbf{A}$. In other words, the residual must be orthogonal to range(A).

Theorem 2.43. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ *and* $\mathbf{b} \in \mathbb{R}^m$ *be given. A vector* $\mathbf{x} \in \mathbb{R}^n$ *minimizes the residual norm* $\|\mathbf{r}\|_2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2$ *thereby solving the least square problem if and only if* $\mathbf{r} \perp \text{range}(\mathbf{A})$*, i.e.,*

$$\mathbf{A}^\top \mathbf{r} = \mathbf{0}, \tag{2.18}$$

or

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{b}, \tag{2.19}$$

or

$$\mathbf{P}\mathbf{b} = \mathbf{A}\mathbf{x}, \tag{2.20}$$

where $\mathbf{P} \in \mathbb{R}^{m\times m}$ *is the orthogonal projection onto* range$(\mathbf{A})$*. The* $n \times n$ *system Eq.* (2.19) *known as* normal equation *is non-singular if and only if* $\mathbf{A}$ *has full rank.*

We have not talked about the orthogonal projection $\mathbf{P}$, but we will use its certain formulations, the major one being $\mathbf{P} = \mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$. Interested readers can read any standard Linear Algebra book to read more about it; see [5, Section 6.6].

Now how do we actually solve Eq. (2.18), Eq. (2.19) or Eq. (2.20)? If $\mathbf{A}$ has full rank, then the solution to the least square problem is unique and given by

$$\mathbf{x} = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{b}.$$

The matrix $\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$ is called the *pseudoinverse* of $\mathbf{A}$ and is denoted by $\mathbf{A}^+$. This is a matrix of size $n \times m$. The problem is to compute one or both vectors

$$\mathbf{x} = \mathbf{A}^+\mathbf{b} \quad \mathbf{y} = \mathbf{P}\mathbf{b},$$

where $\mathbf{A}^+$ is the pseudoinverse of $\mathbf{A}$.

First, look at Eq. (2.19) and try to solve it. Now, we have that $\mathbf{A}^\top\mathbf{A}$ is a symmetric and positive definite matrix. Hence, we can apply Cholesky Decomposition (see 10) to write

$$\mathbf{A}^\top\mathbf{A} = \mathbf{L}\mathbf{L}^\top$$

and then solve $(\mathbf{L}\mathbf{L}^\top)\mathbf{x} = \mathbf{A}^\top\mathbf{b}$ to get $\mathbf{x}$. Here, it is important to note that we need to solve two systems of equations.

2.4.1 QR Decomposition

We have seen that matrix factorization has certain advantages. There is one factorization that is useful for least square methods.

We recall from Linear Algebra that give linearly independent vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ in $\mathbb{R}^n$ we can compute an orthogonal linearly independent set of vectors $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ using Gram Schmidt orthogonalization [5, Theorem 6.4]. Further,we can also compute an orthonormal set of linearly independent vectors $\{\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots, \hat{\mathbf{q}}_n\}$. In matrix notation it means if we denote $\mathbf{A}|_{m\times n}$ by columns of $\{\mathbf{A}_i\}_{i=1}^n$ and $\mathbf{Q}|_{m\times n}$ by $\{\hat{\mathbf{q}}_i\}_{i=1}^n$ then

$$\mathbf{A} = \mathbf{Q}\mathbf{R},$$

where $\mathbf{R}$ is a $n \times n$ matrix. To compute $\mathbf{R}$, we re-write the Gram-Schmidt orthogonalization as

$$\begin{aligned}
\mathbf{A}_1 &= \hat{\mathbf{q}}_1 \|\mathbf{q}_1\| \\
\mathbf{A}_2 &= \hat{\mathbf{q}}_2 \|\mathbf{q}_2\| + \langle \mathbf{A}_2, \hat{\mathbf{q}}_1 \rangle \hat{\mathbf{q}}_1 \\
\mathbf{A}_3 &= \hat{\mathbf{q}}_3 \|\mathbf{q}_3\| + \langle \mathbf{A}_3, \hat{\mathbf{q}}_1 \rangle \hat{\mathbf{q}}_1 + \langle \mathbf{A}_3, \hat{\mathbf{q}}_2 \rangle \hat{\mathbf{q}}_2,
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the ℓ_2 inner product. Looking at the pattern, we notice that $\mathbf{R}$ is an upper triangular matrix such that

$$r_{ij} = \begin{cases} \langle \mathbf{A}_j, \mathbf{q}_i \rangle & \text{if } i < j, \\ \|\mathbf{q}_i\| & \text{if } i = j, \\ 0 & \text{else.} \end{cases}$$

The existence of the QR factorization comes from Gram-Schmidt orthogonalization, and the uniqueness follows the same logic as in Theorem 2.6.

Now, to solve Eq. (2.20) using $\mathbf{A} = \mathbf{QR}$ and the definition of $\mathbf{P}$ we note that

$$\begin{aligned}
\mathbf{P} &= \mathbf{A} \left(\mathbf{A}^\top \mathbf{A} \right)^{-1} \mathbf{A}^\top \\
&= (\mathbf{QR}) \left[(\mathbf{QR})^\top \mathbf{QR} \right]^{-1} (\mathbf{QR})^\top \\
&= \mathbf{QR} \left[\mathbf{R}^\top \mathbf{Q}^\top \mathbf{QR} \right]^{-1} \mathbf{R}^\top \mathbf{Q}^\top \\
&= \mathbf{QR} \left[\mathbf{R}^\top \mathbf{R} \right]^{-1} \mathbf{R}^\top \mathbf{Q}^\top \\
&= \mathbf{Q} \left(\mathbf{RR}^{-1} \right) \left(\mathbf{R}^\top \right)^{-1} \mathbf{R}^\top \mathbf{Q}^\top \\
&= \mathbf{QQ}^\top.
\end{aligned}$$

The above result holds as $\mathbf{Q}^\top$ is the left-inverse of $\mathbf{Q}$. Using this, we re-write

$$\begin{aligned}
\mathbf{Pb} &= \mathbf{Ax} \\
\mathbf{QQ}^\top \mathbf{b} &= \mathbf{QRx} \\
\mathbf{Q}^\top \mathbf{b} &= \mathbf{Rx}.
\end{aligned}$$

Hence, we get $\mathbf{x} = \mathbf{R}^{-1} \mathbf{Q}^\top \mathbf{b}$. Here, we note that we only need to solve one system of equations.

The algorithm for QR decomposition using Gram-Schmidt orthogonalization is present in Algorithm 14.

Algorithm 14 QR Decomposition

Given: Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$.
Find: Orthonormal matrix $\mathbf{Q}$ and upper triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$
Step 1: Initialize Matrices
Initialize $\mathbf{Q}$ as a zero matrix of size $m \times n$.
Initialize $\mathbf{R}$ as a zero matrix of size $n \times n$.
for $j = 1$ **to** n **do**
Set $\mathbf{q}_j = \mathbf{A}_j$
for $i = 1$ **to** $j - 1$ **do**
$r_{ij} = \langle \mathbf{q}_i, \mathbf{A}_j \rangle$
$\mathbf{q}_j = \mathbf{q}_j - r_{i,j}\mathbf{q}_i$
end for
$r_{jj} = \|\mathbf{q}_j\|_2$.
$\hat{\mathbf{q}}_j = \mathbf{q}_j / r_{jj}$.
end for

---

return $\mathbf{Q} = [\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots, \hat{\mathbf{q}}_n]$, $\mathbf{R}$

Chapter 3

Differential Equations

In 1998, with the sponsorship of Boston businessman Landon T. Clay, the Clay Mathematics Institute (CMI) was established. CMI is a private, non-profit foundation dedicated to advancing and disseminating mathematical knowledge. On May 24, 2000, CMI announced the *Millennium Prize Problems*, a list of seven fundamental unsolved problems in mathematics.


Figure 3.1: Claude-Louis Navier (February 10, 1785 – August 21, 1836, left) and George Gabriel Stokes (August 13, 1819 – February 1, 1903, right).

One of these problems concerns the existence and smoothness of solutions to the *Navier-Stokes equations*. These equations form a fundamental system in fluid dynamics and are named after the French mathematician Claude-Louis Navier and the Irish mathematician George Gabriel Stokes.

Let $\mathbf{v}(\mathbf{x}, t)$ be a three-dimensional velocity field, and let $p(\mathbf{x}, t)$ denote the pressure. The Navier-Stokes equations are given by

$$\frac{\partial \mathbf{v}}{\partial t} - \nu \Delta \mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla p + \mathbf{f}(\mathbf{x}, t), \tag{3.1}$$

where $\nu > 0$ is the viscosity and $\mathbf{f}(\mathbf{x}, t)$ represents an external force field. Since there are four unknowns ($\mathbf{v} = (v_1, v_2, v_3)^\top$ and p) but only three equations, an additional equation is required. This comes from the incompressibility condition:

$$\nabla \cdot \mathbf{v} = 0. \tag{3.2}$$

CMI offers a prize of US $1,000,000 for resolving the following problem:

"Prove or give a counterexample of the following statement: In three space dimensions and time, given an initial velocity field, there exists a vector velocity and a scalar pressure field, which are both smooth and globally defined, that solve the Navier–Stokes equations."

Since the Navier-Stokes equations are highly complex partial differential equations, obtaining their solutions often requires numerical methods. Before diving into numerical PDE methods, we begin with simpler cases by exploring numerical approaches for ordinary differential equations (ODEs).

For many ODEs, exact analytical solutions are either unavailable or highly complex. Consider the following ODEs:

$$\frac{dy}{dx} = y - x^2, \qquad \frac{dy}{dx} = x - y^2. \tag{3.3}$$

The first equation is easily solvable, but the second one does not admit a simple closed-form solution. In fact, using Wolfram Alpha, we obtain the solution to the second ODE as

$$y(x) = -\frac{\iota x^{3/2}\left\{-c_1 J_{-4/3}\left(\frac{2}{3}\iota x^{3/2}\right) + c_1 J_{2/3}\left(\frac{2}{3}\iota x^{3/2}\right) - 2J_{-2/3}\left(\frac{2}{3}\iota x^{3/2}\right)\right\} - c_1 J_{-1/3}\left(\frac{2}{3}\iota x^{3/2}\right)}{2x\left(J_{-1/3}\left(\frac{2}{3}\iota x^{3/2}\right) + J_{1/3}\left(\frac{2}{3}\iota x^{3/2}\right)\right)}, \tag{3.4}$$

where $J_n(\cdot)$ denotes the Bessel function of the first kind, and c_1 is a constant.

In the first part of the course we consider initial value problems (IVP) of the form:

$$\frac{dy}{dt} = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha. \tag{3.5}$$

Here, α is referred to as the initial condition. The reason they are called IVP is because the solution depend on the initial value at $t = \alpha$.

Before moving to the numerics we recall certain results regarding the existence and uniqueness of ODEs. We first recall what Lipschitz condition is.

Definition 3.1. A function $f(t, y)$ is said to satisfy the *Lipschitz condition* in the variable y on a set $D \subset \mathbb{R}^2$ if there exists a constant $L > 0$ such that

$$|f(t, y_1) - f(t, y_2)| \le L|y_1 - y_2|, \tag{3.6}$$

for all $(t, y_1), (t, y_2) \in D$. The constant L is called a *Lipschitz constant* for $f(t, y)$.

Example 3.2. Let $D = \{(t, y) : 0 \le t \le 1, -2\pi \le y \le 2\pi\}$ then $f = y\cos(t)$ is a Lipschitz function with Lipschitz constant 1. Let $(t, y_1), (t, y_2) \in D$

$$|f(t, y_1) - f(t, y_2)| = |y_1\cos(t) - y_2\cos(t)| \le |\cos(t)||y_1 - y_2| \le |y_1 - y_2|,$$

as $|\cos(t)| \le 1$ on $[-2\pi, 2\pi]$.

We need Lipschitz condition for showing the existence and uniqueness of the solution of Eq. (3.5). Next we present a result for the existence and uniqueness of solution of first order ODEs.

Theorem 3.3. *Suppose $D = \{(t, y) : a \leq t \leq b,\ -\infty < y < \infty\}$ and $f(t, y)$ is continuous on D. If f satisfies a Lipschitz condition in y on D, then the initial value problem (3.5) has a unique solution $y(t)$ for $a \leq t \leq b$.*

After defining the existence and uniqueness of the ODE we need to talk about how well the solution behaves if the solution has some error or it is perturbed. This is referred to as well-posedness.

Definition 3.4. The initial value problem (3.5) is said to be a *well-posed* problem (or Lyapunov Stable) if:

1. A unique solution $y(t)$ exists.
2. There exist constants $\varepsilon_0 > 0$ and $k > 0$ such that for every $\varepsilon_0 > \varepsilon > 0$ and for all continuous perturbations $\delta(t)$ satisfying $|\delta(t)| < \varepsilon$, and for all initial perturbations $|\delta_0| < \varepsilon$, the perturbed problem

$$\frac{dz}{dt} = f(t, z) + \delta(t), \quad a \leq t \leq b, \quad z(a) = \alpha + \delta_0, \tag{3.7}$$

has a unique solution satisfying $|z(t) - y(t)| < k\varepsilon$ for all $t \in [a, b]$.

Eq. (3.7) is referred as a *perturbed problem* associated with Eq. (3.5). We notice that for well posedness we require the existence and uniqueness of the solution. Then the definition states that if we have some error in $f(t, y)$ or in the initial condition α, then the error introduced in the solution is still bounded. The constant ε_0 shows that the error in $f(t, y)$ and α are bounded and the constant k helps in bounding the error in the solution.

3.1 One-Step Methods

Numerical methods for IVP can be broadly divided into two classes. We discuss the first class of these methods referred to as one-step methods. The reason they are called one-step is because the solution at time t_{i+1} depends only on the solution at time t_i.

3.1.1 Euler Method

To compute a solution to the differential equation numerically, we do not compute a continuous approximation to the well-posed problem. Instead, we compute the solution at some pre-defined points called *mesh points*.

Definition 3.5. Given an interval $[a, b]$, let $\{t_i\}_{i=0}^n$ be points defined in $[a, b]$ such that $t_0 = a$, $t_n = b$ and $\{t_i\}_{i=0}^n \in [a, b]$. The set $\{t_i\}_{i=0}^n$ is referred to as *mesh points*.

We assume that the mesh points are equally distributed in $[a, b]$, which can be ensured by choosing a positive integer n and selecting the mesh points

$$t_i = a + ih$$

for each $i = 0, 1, 2, \ldots, n$. Here $h = \frac{b-a}{n} = t_{i+1} - t_i$ is referred as the *step size*. We can also find non-uniform meshing in the literature, such as geometrical mesh.

The early method for computing a numerical solution to the ODE can be traced back to Leonhard Euler in 1760 when he approximated the solution to a given ODE using tangent lines. This method becomes the basis for future numerical methods and is referred to as the *Euler's method*.

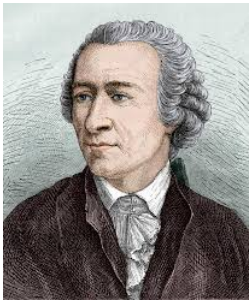

Figure 3.2: Leonhard Euler: 15 April 1707 – 18 September 1783.

Euler's method can be derived in several ways. Geometrically, we can see that the tangent can be used to approximate the solution at t_1 using (t_0, α), see Fig. .3.3.

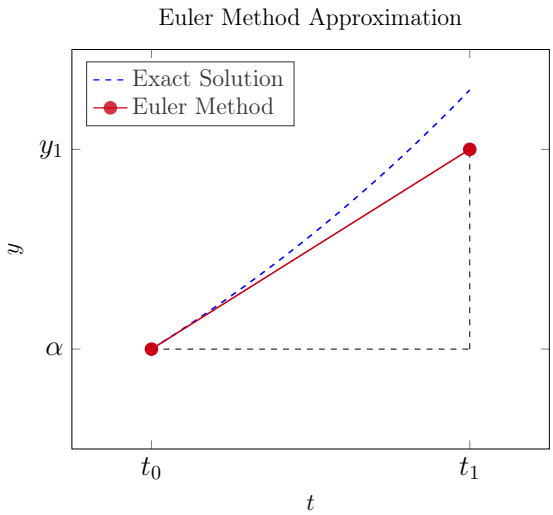

Figure 3.3: Geometrical view of the Euler Method

Another way is to use the Taylor's theorem. Let $y(t)$ be the unique solution to the initial value problem and suppose it has two continuous derivatives on $[a, b]$ so that for each $i = 0, 1, 2, \ldots, n-1$,

$$y_{i+1} = y_i + (t_{i+1} - t_i)y'(t_i) + \frac{(t_{i+1} - t_i)^2}{2}y''(\xi_i),$$

for some $\xi \in (t_i, t_{i+1})$. Unless mentioned otherwise we will denote $y(t_i)$ by y_i. As $h = t_{i+1} - t_i$ and $y'(t) = f(t, y)$ we have

$$y_{i+1} = y_i + hf(t_i, y_i) + \frac{h^2}{2}y''(\xi_i),$$

for $i = 0, 1, \ldots, n-1$. Ignoring the remainder term, we have the classic Euler method as

$$\begin{aligned} y_0^h &= \alpha, \\ y_{i+1}^h &= y_i^h + hf(t_i, y_i^h), \quad i \in \{0, 1, \ldots, n-1\}, \end{aligned} \tag{3.8}$$

where $\{y_i^h\}_{i=0}^n$ is the numerical approximation of y_i at $t = t_i$. Eq. (3.8) is called the *difference equation* associated with the Euler method.

The algorithm for the Euler method is given in Algorithm 15.

Algorithm 15 Euler Method

Given: Endpoints a, b, Number of mesh points: n, Initial condition: α, right-hand side: $f(t, y)$.
Find: Solution $\{y_i^h\}_{i=0}^n$.

Step 1: Meshing
Initialize $h = \frac{b-a}{n}$
Initialize $\mathbf{y}^h$ as zero vector of size $(n+1)$
Set $t = a$, and $y_0^h = \alpha$

Step 2: Euler Method
for $i = 0$ **to** $n-1$ **do**
$y_{i+1}^h = y_i^h + hf(t, y_i^h)$

Step 2.1: Update t
$t = t + h$
end for

---

return $\mathbf{y}^h$

Example 3.6. Consider $y' = te^{3t} - 2y$, $0 \le t \le 1$ and $y(0) = 0$ with $h = 0.5$. Now, $n = \frac{1}{0.5} = 2$. Hence, using the Euler method, we get

$$\begin{aligned} y_0^h &= 0 \\ y_1^h &= y_0^h + hf(t_0, y_0^h) \\ &= 0 + 0.5 \times 0 = 0 \\ y_2^h &= y_1^h + hf(t_1, y_1^h) \\ &= 0.5(0.5e^{1.5}) = 0.41218. \end{aligned}$$

Now, we try to bound the error of the obtained numerical solution. To derive the error bound, we need two lemmas.

Lemma 3.7. *For all* $x \ge -1$ *and* $m \ge 0$ *we have* $0 \le (1+x)^m \le e^{mx}$.

Proof. Let $f(x) = e^x$ and $x_0 = 0$. Using the Taylor's theorem for $f(x)$ at $x_0 = 0$ and $n = 1$ we get

$$e^x = 1 + x + \frac{x^2}{2!}e^\xi,$$

for some ξ in between x and zero. Now,

$$0 \leq 1 + x \leq 1 + x + \frac{x^2}{2}e^\xi = e^x.$$

As $(1 + x) \geq 0$ we have $(1 + x)^m \leq e^{mx}$. □

Lemma 3.8. *If $s, t \geq 0$ and $\{a_i\}_{i=0}^k$ is a sequence satisfying $a_0 \geq -\frac{t}{s}$ and*

$$a_{i+1} \leq (1 + s)a_i + t, \tag{3.9}$$

for each $i = 0, 1, 2, \ldots, k - 1$. Then

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

Proof. Let i be fixed. Then from Eq. (3.9) we have

$$\begin{aligned} a_{i+1} &\leq (1 + s)a_i + t \\ &\leq (1 + s)\left((1 + s)a_{i-1} + t\right) + t \\ &= (1 + s)^2 a_{i-1} + \left[(1 + s) + 1\right] t \\ &\vdots \\ &\leq (1 + s)^{i+1} a_0 + \left[\sum_{j=0}^{i} (1 + s)^j\right] t. \end{aligned}$$

We note that $\sum_{j=0}^{i}(1 + s)^j$ is a geometric series with $(1 + s)$ as ratio. Hence,

$$\sum_{j=0}^{i}(1 + s)^j = \frac{1 - (1 + s)^{i+1}}{1 - (1 + s)} = \frac{1}{s}\left((1 + s)^{i+1} - 1\right).$$

Thus $a_{i+1} \leq (1 + s)^{i+1}a_0 + \frac{t}{s}\left[(1 + s)^{i+1} - 1\right] = (1 + s)^{i+1}\left[a_0 + \frac{t}{s}\right] - \frac{t}{s}$.

Using Lemma 3.7, we have the result. □

Theorem 3.9. *Suppose f is continuous and satisfies a Lipschitz condition with constant L on $D = \{(t, y) : a \leq t \leq b \quad \text{and} \quad -\infty < y < \infty\}$ and a constant M exist with $|y''(t)| < M$ for all $t \in [a, b]$, where $y(t)$ is the unique solution to the IVP*

$$y'(t) = f(t, y) \quad a \leq t \leq b, \quad y(a) = \alpha.$$

Let $\{y_i^h\}_{i=0}^n$ be the approximations generated by the Euler method for $n \geq 0$. Then for each $i = 0, 1, 2, \ldots, n$

$$|y_i - y_i^h| \leq \frac{hM}{2L}\left(e^{L(t_i - a)} - 1\right).$$

Proof. For $i = 0$, $y_i^h = \alpha = y_i$. Hence, the results are trivial. Now, using Taylor's approximation, we have

$$y_{i+1} = y_i + hf(t_i, y_i) + \frac{h^2}{2}y''(\xi),$$

for $i = 0, 1, \ldots, n-1$ and from the Euler method

$$y_{i+1}^h = y_i^h + hf(t_i, y_i^h).$$

Subtracting these equations, we have

$$y_{i+1} - y_{i+1}^h = y_i - y_i^h + h\left[f(t_i, y_i) - f(t_i, y_i^h)\right] + \frac{h^2}{2}y''(\xi_i).$$

Hence,

$$|y_{i+1} - y_{i+1}^h| \leq |y_i - y_i^h| + h\left|\left[f(t_i, y_i) - f(t_i, y_i^h)\right]\right| + \frac{h^2}{2}|y''(\xi_i)|.$$

Now, f satisfies a Lipschitz condition in y and hence if L denotes the Lipschitz constant we have

$$|f(t_i, y_i) - f(t_i, y_i^h)| \leq L|y_i - y_i^h|,$$

and as $|y''(x)| \leq M$ we get

$$|y_{i+1} - y_{i+1}^h| \leq |y_i - y_i^h| + hL|y_i - y_i^h| + \frac{h^2M}{2} = (1 + hL)|y_i - y_i^h| + \frac{h^2M}{2}.$$

Now, from Lemma 3.8 for $s = hL$, $t = h^2M/2$ and $a_j = |y_j - y_j^h|$, for each $j = 0, 1, 2, \ldots, n$ we have

$$\begin{aligned} |y_{i+1} - y_{i+1}^h| &\leq e^{(i+1)hL}\left(0 + \frac{h^2M}{2hL}\right) - \frac{h^2M}{2hL} \\ &= e^{(i+1)hL}\frac{hM}{2L} - \frac{hM}{2L} \\ &= \frac{hM}{2L}\left(e^{(t_{i+1} - a)L} - 1\right), \end{aligned}$$

where we have used that $t_{i+1} = a + (i + 1)h$ for $i = 0, 1, 2, \ldots, n$. □

We notice that in the Theorem 3.9, we require M, which may not be possible as the solution is unknown. But as we know $f(t,y)$ we can compute $y''(t)$ by

$$y''(t) = \frac{d}{dt}\left(y'(t)\right) = \frac{d}{dt}\left(f(t,y)\right) = \frac{\partial f}{\partial t} f\left(t, y(t)\right) + \frac{\partial f}{\partial y}\frac{dy}{dt} = \frac{\partial f}{\partial t} f\left(t, y(t)\right) + \frac{\partial f}{\partial y} f\left(t, y(t)\right).$$

We also notice that the error is of $\mathcal{O}(h)$, i.e., linear.

In the above theorem, we are missing the influence of the round-off errors. In reality, we solve the system

$$\begin{aligned} u_0^h &= \alpha + \delta_0 \\ u_{i+1}^h &= u_i^h + hf(t_i, u_i^h) + \delta_{i+1}, \quad \text{for } \; i = 0, 1, \ldots, n-1, \end{aligned} \tag{3.10}$$

where δ_i is the round-off error associated with u_i^h, i.e., $y_i - u_i^h = \delta_i$. In this case, we still have an error bound to the form.

Theorem 3.10. *Let $y(t)$ denote the unique solution to the IVP*

$$y'(t) = f\left(t, y(t)\right), \quad a \le t \le b, \quad y(a) = \alpha,$$

and $\{u_i^h\}_{i=0}^n$ be the approximations obtained from Eq. (3.10). If $|\delta_i| < \delta$ for all $i = 0, 1, \ldots, n$ and the hypothesis of Theorem 3.9 hold for this IVP, then

$$|y_i - u_i^h| \le \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)\left[e^{L(t_i - a)} - 1\right] + |\delta_0| e^{L(t_i - a)},$$

for each $i = 0, 1, 2, \ldots, n$.

The above theorem states that if the round-off errors are bounded for each i, then the numerical solution will also be bounded, but we notice that the error is not linear as

$$\lim_{h \to 0}\left(\frac{hM}{2} + \frac{\delta}{h}\right) = \infty.$$

Hence, the error increases for small h. However, we can find an optimal value of h. Let $E(h) = \frac{hM}{2} + \frac{\delta}{h} \Rightarrow E'(h) = \frac{M}{2} - \frac{\delta}{h^2}$. If $h < \sqrt{2\delta/M}$, then $E'(h) < 0$ which implies that $E(h)$ is decreasing. Similarly if $h > \sqrt{2\delta/M}$ then $E'(h) > 0$ which implies that $E(h)$ is increasing.

Hence, $h = \sqrt{2\delta/M}$ is the minimal value of h.

So, the error will increase if h is below this value. However, in reality, δ is very small, so it does not affect Euler's method.

3.1.2 Higher Order Taylor Method

Once we have discretized our differential equation into the difference equation, we want to check how well our analytical solution (y) satisfies the difference equation. This seems counter-intuitive as we want to see if our numerical solution (y_h) approximates the analytical solution

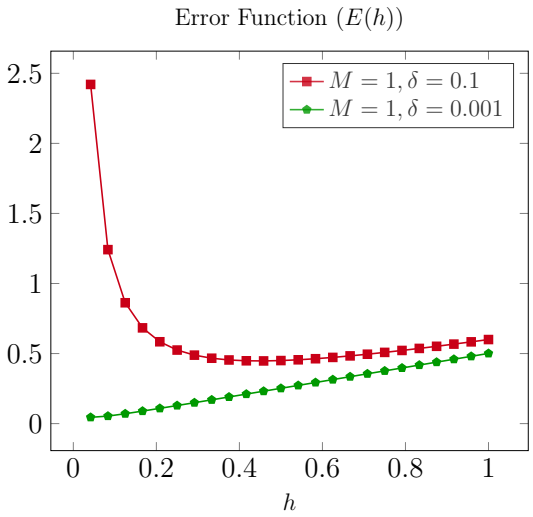

Figure 3.4: Plot of $E(h)$ for $M = 1$ and varying δ.

well. However, as the analytical solution is unknown, we cannot guarantee this, so we check for local truncation errors.

Definition 3.11. The difference method

$$\begin{aligned} y_0^h &= \alpha \\ y_{i+1}^h &= y_i^h + h\Phi(t_i, y_i^h; h) \qquad \text{for} \quad i = 0, 1, \ldots, n-1, \end{aligned}$$

has *local truncation error*

$$\tau_{i+1}(h) = \frac{y_{i+1} - (y_i + h\Phi(t_i, y_i; h))}{h} = \frac{y_{i+1} - y_i}{h} - \Phi(t_i, y_i; h),$$

for $i = 0, 1, 2, \ldots, n-1$.

We call it a local error because it measures the accuracy of the method at a specific step, assuming that the method was exact at the previous step. Now, the local truncation error for the Euler method is

$$\tau_{i+1}(h) = \frac{y_{i+1} - y_i}{h} - f(t_i, y_i) \qquad \text{for} \quad i = 0, 1, 2, \ldots, n-1.$$

For deriving the Euler method, we assume that y has two continuous derivatives, and hence, by Taylor approximation, we have

$$\tau_{i+1}(h) = \frac{h}{2} y''(\xi) \qquad \text{for} \quad \xi \in (t_i, t_{i+1}).$$

If $y''(t)$ is known and bounded by M on $[a, b]$, this implies

$$|\tau_{i+1}(h)| \leq \frac{Mh}{2},$$

so the local truncation error is $\mathcal{O}(h)$.

One way to select a difference method for solving ODEs is in such a manner that their local truncation error are of $\mathcal{O}(h^p)$, for large values of p while keeping the number and complexity of calculations of the methods within a reasonable bound. The reason for this is to say, for example, $h = 0.01$, then for $p = 1$, we have an error of order 0.01, but for $p = 2$, we get 10^{-4}, which is a drastic reduction.

Now, the question is: How do we develop methods with higher-order truncation errors? Notice that the derivation of the Euler method is based on using Taylor's theorem for $n = 1$, so one idea is to use Taylor's theorem for large n.

Suppose the solution $y(t)$ of the IVP $y'(t) = f(t, y)$, $a \leq t \leq b$, and $y(a) = \alpha$ has $(n + 1)$ continuous derivatives. We expand $y(t)$ in terms of its n^{th} Taylor polynomial about t_i and evaluate it at t_{i+1} to obtain

$$y_{i+1} = y_i + hy_i' + \frac{h^2}{2}y_i'' + \cdots + \frac{h^n}{n!}y_i^{(n)} + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_i), \tag{3.11}$$

for $\xi_i \in (t_i, t_{i+1})$.

Using successive differentiation of the solution $y(t)$ gives $y'(t) = f(t, y(t))$, $y''(t) = f'(t, y(t))$, and generally $y^{(k)}(t) = f^{(k-1)}(t, y(t))$. Substituting this in Eq. (3.11) we get

$$y_{i+1} = y_i + hf(t_i, y_i) + \frac{h^2}{2}f'(t_i, y_i) + \cdots + \frac{h^n}{n!}f^{(n-1)}(t_i, y_i) + \frac{h^{n+1}}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)). \tag{3.12}$$

We recover the difference equation if we ignore the last term in Eq. (3.12). Based on this, we get the Taylor method of order n as

$$\begin{aligned} y_0^h &= \alpha \\ y_{i+1}^h &= y_i^h + hT^{(n)}(t_i, y_i), \qquad \text{for } \; i = 0, 1, 2, \ldots, n-1, \end{aligned}$$

where

$$T^{(n)}(t_i, y_i^h) = f(t_i, y_i^h) + \frac{h}{2}f'(t_i, y_i^h) + \cdots + \frac{h^{n-1}}{n!}f^{(n-1)}(t_i, y_i^h).$$

Note that for $n = 1$, we get the Taylor method of order one, which is the Euler method. Now we consider the same example as earlier and compute the Taylor method of order 2 and 3.

Example 3.12. Consider the IVP

$$y'(t) = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

First we need to compute $f'(t, y(t))$ which is equal to $y'(t) - 2t = y(t) - t^2 + 1 - 2t$ and then similarly $f''(t, y(t)) = y(t) - t^2 - 2t - 1$.Now,

$$\begin{aligned} T^{(2)}(t_i, y_i^h) &= f(t_i, y_i^h) + \frac{h}{2} f'(t_i, y_i^h) \\ &= y_i^h - t_i^2 + 1 + \frac{h}{2}\left(y_i^h - t_i^2 + 1 - 2t_i\right) \\ &= y_i^h \left(1 + \frac{h}{2}\right) - t_i^2 \left(1 + \frac{h}{2}\right) + \left(1 + \frac{h}{2}\right) - t_i h. \end{aligned}$$

Similarly,

$$T^{(3)}(t_i, y_i^h) = \left(y_i^h - t_i^2 + 1\right)\left(1 + \frac{h}{2} + \frac{h^2}{6}\right) - t_i h - \frac{h^2}{3} t_i - \frac{h^2}{3}.$$

We plot the solution obtained with the order two and order three methods in Fig. 3.5.

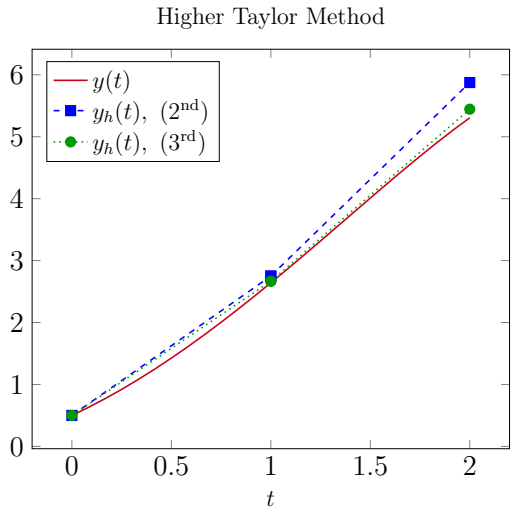

Figure 3.5: Taylor Methods of order 2 and 3.

Theorem 3.13. *If Taylor's method of order n is used to approximate the solution to*

$$y'(t) = f(t, y(t)), \quad a \le t \le b, \quad y(a) = \alpha,$$

with step size h and if $y \in \mathcal{C}^{n+1}[a, b]$, then the local truncation error is of $\mathcal{O}(h^n)$.

Proof. The local truncation error of the Taylor method is

$$\tau_{i+1}(h) = \frac{y_{i+1} - y_i}{h} - T^{(n)}(t_i, y_i) = \frac{h^n}{(n+1)!} f^{(n)}\left(\xi_i, y(\xi_i)\right),$$

for each $i = 0, 1, 2, \ldots, n-1$. Since $y \in \mathcal{C}^n[a, b]$ we have $y^{(n+1)}(t) = f^{(n)}(t, y(t))$ is bounded on $[a, b]$ and $\tau_i(h) = \mathcal{O}(h^n)$ for $i = 0, 1, 2, \ldots, n$. □

In Algorithm 16 we present the Higher-Order Taylor method for order n.

Algorithm 16 Higher-Order Taylor Method

Given: Endpoints a, b, Number of mesh points: n, Initial condition: α, right-hand side: $f(t, y)$, k derivatives of $f(t, y)$, Taylor method order: υ.
Find: Solution $\{y_i^h\}_{i=0}^n$.

Step 1: Meshing
Initialize $h = \frac{b-a}{n}$
Initialize y^h as zero vector of size $(n+1)$
Set $t = a$, and $y_0^h = \alpha$

Step 2: Higher-Order Taylor Method
for $i = 0$ **to** n **do**
  sum $= 0$
  for $k = 0$ **to** υ **do**
    sum $=$ sum $+ \frac{h^k}{k!} f^{(k)}(t, y_i^h)$
  end for
  $y_{i+1}^h = y_i^h + h \times$ sum

  Step 2.1: Update t
  $t = t + h$
end for

---

return $\mathbf{y}^h$

3.1.3 Runge Kutta Methods

The Taylor methods described in the previous subsection have the desirability of higher-order truncation error but require a lot of computations for the derivative of $f(t, y)$. In the late 1800s, Carl Runge derived various numerical formulas for approximating the solution to initial value problems, and later, in 1901, Martin Wilhelm Kutta generalized these methods to incorporate systems of first-order differential equations, hence giving rise to Runge Kutta methods (RK methods).

The RK methods have higher-order local truncation errors, but they avoid the computation of higher-order derivatives of $f(t, y)$. As seen throughout this course, we start with a basic method and refine it. Similarly, one can consider the RK methods to be a refinement of the Euler method. Before we derive these methods, we recall or present Taylor's theorem in two variables.

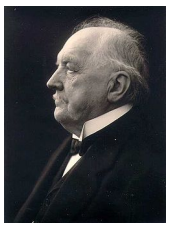

Figure 3.6: Martin Wilhelm Kutta: November 3, 1867 – December 25, 1944.

Theorem 3.14. *Suppose that $f(t,y)$ and all its partial derivatives of the order less than or equal to $n+1$ are continuous on $D = \{(t,y) : a \le t \le b, c \le y \le d\}$ and let $(t_0, y_0) \in D$. For every $(t,y) \in D$ there exist a ξ between t and t_0 and μ between y and y_0 with*

$$f(t,y) = P_n(t,y) + R_n(t,y),$$

where

$$\begin{aligned} P_n(t,y) &= f(t_0,y_0) + \left[(t-t_0)\frac{\partial f}{\partial t}(t_0,y_0) + (y-y_0)\frac{\partial f}{\partial y}(t_0,y_0)\right] \\ &\quad + \left[\frac{(t-t_0)^2}{2}\frac{\partial^2 f}{\partial t^2}(t_0,y_0) + (t-t_0)(y-y_0)\frac{\partial^2 f}{\partial t \partial y}(t_0,y_0) + \frac{(y-y_0)^2}{2}\frac{\partial^2 f}{\partial y^2}(t_0,y_0)\right] \\ &\quad + \cdots + \left[\frac{1}{n!}\sum_{j=0}^{n} {}^n\mathrm{C}_j (t-t_0)^{n-j}(y-y_0)^j \frac{\partial^n f}{\partial t^{n-j}\partial y^j}(t_0,y_0)\right], \end{aligned}$$

and

$$R_n(t,y) = \frac{1}{(n+1)!}\sum_{j=0}^{n+1} {}^{n+1}\mathrm{C}_j (t-t_0)^{n+1-j}(y-y_0)^j \frac{\partial^{n+1} f}{\partial t^{n+1-j}\partial y^j}(\xi,\mu).$$

The function $P_n(t,y)$ is called as the n^{th} Taylor polynomial in two-variable for the function f about (t_0,y_0), and $R_n(t,y)$ is the remainder term associated with $P_n(t,y)$.

RK methods, in general, can be found in any order. Hence, we first start by deriving RK methods of order two. We notice that for the Taylor method of order two, we have

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2}f'(t,y).$$

Now, the idea is to approximate the derivative $f'(t,y)$ (and $f(t,y)$) by a function of type $a_1 f(t+\alpha_1, y+\beta_1)$ such that the error is not greater than of $\mathcal{O}(h^2)$. Since

$$f'(t,y) = \frac{d}{dt}f(t,y) = \frac{\partial f}{\partial y}(t,y)\frac{dy}{dt} + \frac{\partial f}{\partial t} = f(t,y)\frac{\partial f}{\partial y}(t,y) + \frac{\partial f}{\partial t}(t,y).$$

Substituting this in $T^{(2)}(t,y)$ we get

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2}\left[f(t,y)\frac{\partial f}{\partial y}(t,y) + \frac{\partial f}{\partial t}(t,y)\right]. \tag{3.13}$$

Now, using the Taylor expansion of $f(t+\alpha_1, y+\beta_1)$ of degree one about (t,y) we get

$$a_1 f(t+\alpha_1, y+\beta_1) = a_1 \left[f(t,y) + \alpha_1 \frac{\partial f}{\partial t}(t,y) + \beta_1 \frac{\partial f}{\partial y}(t,y) + R_1(t+\alpha_1, y+\beta_1) \right], \tag{3.14}$$

where

$$R_1(t+\alpha_1, y+\beta_1) = \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \alpha_1\beta_1 \frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial y^2}(\xi,\mu), \tag{3.15}$$

for some ξ between t and $t+\alpha_1$ and μ between y and $y+\beta_1$.

Matching the coefficients in Eq. (3.13) and Eq. (3.14) we get

$$a_1 = 1, \quad \beta_1 = \frac{h}{2} f(t,y), \quad \alpha_1 = \frac{h}{2},$$

so

$$T^{(2)}(t,y) = f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t,y)\right) - R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t,y)\right),$$

and from Eq. (3.15)

$$R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t,y)\right) = \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \frac{h^2}{4} f(t,y)\frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{h^2}{8}\left(f(t,y)\right)^2 \frac{\partial^2 f}{\partial y^2}(\xi,\mu).$$

If all the second order partial derivatives of f are bounded we get $R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t,y)\right)$ is of $\mathcal{O}(h^2)$. Hence, our new method has the same order as the Taylor method of order two but does not require the derivatives of $f(t,y)$.

This method is also referred to as the *midpoint method* and is given by

$$\begin{aligned} y_0^h &= \alpha \\ y_{i+1}^h &= y_i^h + hf\left(t_i + \frac{h}{2}, y_i^h + \frac{h}{2} f(t_i, y_i^h)\right), \quad \text{for } \; i = 0, 1, 2, \ldots, n-1. \end{aligned}$$

This is called the midpoint method because we are evaluating the value of y at t_{i+1} using the mid-point between t_i and t_{i+1}.

Midpoint method is not the only $\mathcal{O}(h^2)$ method that is used in practice. We also have a refined form of the Euler method called the *modified Euler's method* and is given by

$$\begin{aligned} y_0^h &= \alpha \\ y_{i+1}^h &= y_i^h + \frac{h}{2}\left[f(t_i, y_i^h) + f\left(t_{i+1}, y_i^h + hf(t_i, y_i^h)\right)\right], \quad \text{for } \; i = 0, 1, 2, \ldots, n-1. \end{aligned}$$

Here first, we compute the solution at t_{i+1} and then take the average of the tangent at (t_i, y_i^h) and (t_{i+1}, y_{i+1}^h) and then compute the new solution. Fig. 3.7 shows the graphical representation of the midpoint and the modified euler methods.

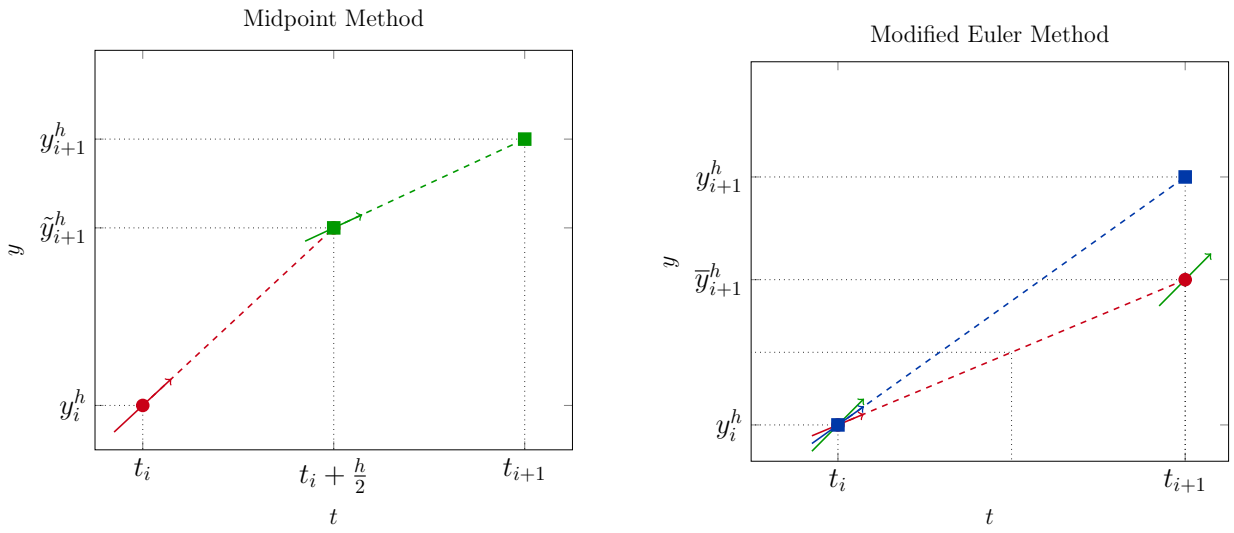

Figure 3.7: Geometrical view of the Midpoint and the Modified Euler Method. $\tilde{y}_{i+1}^h$ refers to $y_i^h + \frac{h}{2} f(t_i, y_i^h)$, $\overline{y}_{i+1}^h$ refers to $y_i^h + h f(t_i, y_i^h)$, and the arrows represent the direction of the vector field.

Example 3.15. Consider the IVP

$$y'(t) = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

We apply the midpoint method to compute the solution for $h = 0.5$. Let $\tilde{y}_{i+1}^h = y_i^h + \frac{h}{2} f(t_i, y_i^h)$. Then we can write $y_{i+1}^h = y_i^h + hf\left(t_i + \frac{h}{2}, \tilde{y}_{i+1}^h\right)$.
Now, for the midpoint method we have

$$y_0^h = 0.5,$$

For y_1^h we first compute

$$\tilde{y}_1^h = y_0^h + \frac{h}{2} f(t_0, y_0^h) = 0.5 + \frac{0.5}{2} f(0, 0.5) = 0.875,$$

and finally

$$y_1^h = y_0^h + hf\left(t_0 + \frac{h}{2}, \tilde{y}_1^h\right) = 0.5 + 0.5 \times f\left(0 + \frac{0.5}{2}, 0.875\right) = 1.40625.$$

For y_2^h we first compute

$$\tilde{y}_2^h = y_1^h + \frac{h}{2} f(t_1, y_1^h) = 1.40625 + \frac{0.5}{2} f(0.5, 1.40625) = 1.9453125,$$

and finally

$$y_2^h = y_1^h + hf\left(t_1 + \frac{h}{2}, \tilde{y}_2^h\right) = 1.40625 + 0.5 \times f\left(0.5 + \frac{0.5}{2}, 1.9453125\right) = 2.59765625.$$

Based on the ideas of RK-2, we can generalize the idea to get RK-ν methods for $\nu \geq 1$. The difference equation for a RK-ν method is given by

$$\begin{aligned} y_0^h &= \alpha \\ k_1 &= y_i^h \\ k_2 &= y_i^h + ha_{2,1}f(t_i, k_1) \\ k_3 &= y_i^h + ha_{3,1}f(t_i, k_1) + ha_{3,2}f(t_i + c_2h, k_2) \\ &\vdots \\ k_\nu &= y_i^h + h\sum_{j=1}^{\nu-1} a_{\nu,j}f(t_i + c_jh, k_j), \end{aligned}$$

and

$$y_{i+1}^h = y_i^h + h\sum_{j=1}^{\nu} b_j f(t_i + c_jh, k_j),$$

The matrix $\{a_{i,j}\}_{i,j=1}^{\nu}$ is called the *RK Matrix*, $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots b_\nu \end{bmatrix}$ is called the *RK weights*, and $\mathbf{c} = \begin{bmatrix} c_1 & c_2 & \cdots & c_\nu \end{bmatrix}^\top$ is called the *RK nodes*. Generally we write the RK-method as a table called the *Butcher tableau*,

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\top \end{array},$$

named after a New Zealand mathematician, John Charles Butcher.

Figure 3.8: John Charles Butcher: March 31, 1933-.

Let us look at some of the Butcher tableau for $\nu = 1, 2, 3$, and 4.

- **RK-1**: RK-1 method is the Euler method which is given by

  $$y_{i+1}^h = y_i^h + hf(t_i, y_i^h),$$

  hence $\nu = 1$, $b_1 = 1$, and $c = 0$. There is no $a_{1,1}$. Therefore, the butcher table is

  $$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}.$$

- **RK-2**: RK-2 method is the midpoint method which is given by

  $$\begin{aligned} k_1 &= y_i^h \\ k_2 &= y_i^h + \frac{h}{2}f(t_i, k_1) \\ y_{i+1}^h &= y_i^h + hf\left(t_i + \frac{h}{2}, k_2\right). \end{aligned}$$

Here $\nu = 2$, $b_1 = 0$, $b_2 = 1$, $c_1 = 0$, $c_2 = 1/2$, $a_{2,1} = 1/2$. Therefore, the Butcher table is

$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}.$$

- **RK-3**: Now we look at a method of order $\mathcal{O}(h^3)$. This is referred to as the *Heunn method* named after the German mathematician Karl Heunn. This method is given by

$$\begin{aligned} k_1 &= y_i^h \\ k_2 &= y_i^h + \frac{h}{3} f(t_i, k_1) \\ k_3 &= y_i^h + \frac{2h}{3} f\left(t_i + \frac{h}{3}, k_2\right) \\ y_{i+1}^h &= y_i^h + \frac{h}{4} f(t_i, k_1) + \frac{3h}{4} f\left(t_i + \frac{2h}{3}, k_3\right), \end{aligned}$$

and the Butcher tableau is given by

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array}.$$

Fig. 3.9 shows the geometrical view of the RK-3 method.

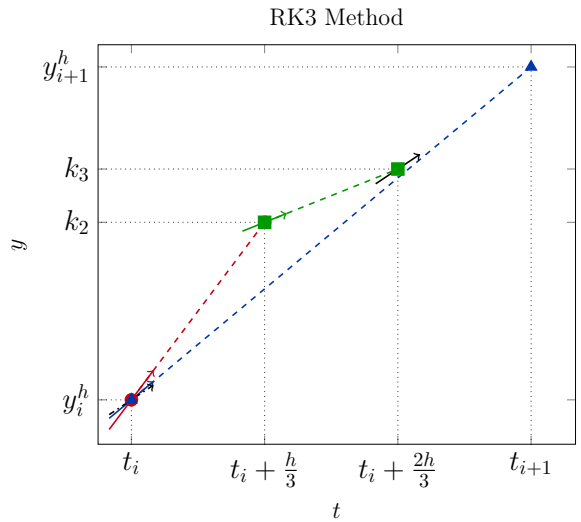

Figure 3.9: Geometrical view of the RK-3 Method.

- **RK-4**: The RK-4 method is also referred to as the *classic RK method* and is the most widely used RK-scheme. We will not give the method but rather present its Butcher tableau, which is given by

Evaluation per step	**Error Order**
2	$\mathcal{O}(h^2)$
3	$\mathcal{O}(h^3)$
4	$\mathcal{O}(h^4)$
$5 \leq n \leq 7$	$\mathcal{O}(h^{n-1})$
$8 \leq n \leq 9$	$\mathcal{O}(h^{n-2})$
$10 \leq n$	$\mathcal{O}(h^{n-3})$

Table 3.1: Different RK schemes with error order and number of evaluations.

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}.$$

In practice, one usually uses the RK-4 method only. But why? There are two axes on which one needs to compare the methods: error and evaluation. In Table 3.1, we provide evaluation per step and error order.

Now, for RK-5 and higher order methods, we notice that the number of evaluations increases, but there is no very drastic difference in error order. For RK-5, we have the same error as RK-4, but an extra evaluation is required. To get $\mathcal{O}(h^5)$ convergence, we need the RK-6 method. And therefore, RK-4 hits the sweet spot.

A keen-eye viewer will notice that the Butcher tableau has certain properties. The derivation of these properties comes from Taylor expansion, but we would just use them.

1. $\sum_{i=1}^{\nu} b_i = 1$: Geometrically, this means that on the interval $[t_i, t_{i+1}]$ we are taking an average weight of the intermediate steps, and hence they should be equal to one.

2. $c_i = \sum_{j=1}^{i-1} a_{i,j}, \quad \text{for } i = 2, 3, \ldots, \nu$. This means that the time fraction c_i at which the function is evaluated is determined by the sum of the weights $a_{i,j}$ of the previous slopes. Since explicit methods do not have implicit dependencies, each stage i only depends on previously computed values.

3. $c_i \leq c_{i+1}$: This holds as we march forward in time or improve the slopes at an intermediate time using previous and intermediate ones (see RK-4).

Algorithm 17 give the algorithm for a RK-ν method whose Butcher table is given by $\mathbf{A}, \mathbf{b}$, and $\mathbf{c}$.

Algorithm 17 Runge-Kutta Method

Given: Endpoints a, b, Number of mesh points: n, Initial condition: α, Right-hand side: $f(t, y)$, RK Matrices and Vectors: $\mathbf{A}, \mathbf{b}, \mathbf{c}$.
Find: Solution $\{y_i^h\}_{i=0}^n$.

Step 1: Meshing
Compute step size: $h = \frac{b-a}{n}$
Initialize y^h as a zero vector of size $(n+1)$
Set $t = a$, and $y_0^h = \alpha$

Step 2: Runge-Kutta Method
for $i = 1$ **to** n **do**

Step 2.1: Compute RK Steps
Initialize $\mathbf{k}$ as a zero vector of size ν.
Initialize $k_1 = y_{i-1}^h$
for $\ell = 2$ **to** ν **do**
$\text{sum}_k = 0$
for $j = 1$ **to** $\ell - 1$ **do**
$\text{sum}_k = \text{sum}_k + \mathbf{A}_{\ell,j} \cdot f(t + \mathbf{c}_j \cdot h, k_j)$
end for
$k_\ell = k_1 + h \cdot \text{sum}_k$
end for

Step 2.2: Compute Weighted Sum
$\text{sum} = 0$
for $j = 1$ **to** ν **do**
$\text{sum} = \text{sum} + \mathbf{b}_j \cdot f(t + \mathbf{c}_j \cdot h, k_j)$
end for
$y_i^h = y_{i-1}^h + h \cdot \text{sum}$

Step 2.3: Update t
$t = t + h$
end for

---

return $\mathbf{y}^h$

Example 3.16. Consider the IVP

$$y'(t) = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

In Table 3.2 we present the errors computed using RK-ν for $\nu = 1, 2, 3, 4$. Here we consider $h = 0.2$. We notice that as our ν increases, the error corresponding to t_i decreases.

t_i	$\nu = 1$			$\nu = 2$			$\nu = 3$			$\nu = 4$		
	y_i^h	y_i	$\lvert y_i - y_i^h\rvert$	y_i^h	y_i	$\lvert y_i - y_i^h\rvert$	y_i^h	y_i	$\lvert y_i - y_i^h\rvert$	y_i^h	y_i	$\lvert y_i - y_i^h\rvert$
0.00	0.50000	0.50000	0.00000	0.50000	0.50000	0.00000	0.50000	0.50000	0.00000	0.50000	0.50000	0.00000
0.20	0.80000	0.82930	0.02930	0.82800	0.82930	0.00130	0.82924	0.82930	0.00005	0.82929	0.82930	0.00001
0.40	1.15200	1.21409	0.06209	1.21136	1.21409	0.00273	1.21398	1.21409	0.00011	1.21408	1.21409	0.00001
0.60	1.55040	1.64894	0.09854	1.64466	1.64894	0.00428	1.64877	1.64894	0.00018	1.64892	1.64894	0.00002
0.80	1.98848	2.12723	0.13875	2.12128	2.12723	0.00595	2.12699	2.12723	0.00024	2.12720	2.12723	0.00003
1.00	2.45818	2.64086	0.18268	2.63317	2.64086	0.00769	2.64056	2.64086	0.00030	2.64082	2.64086	0.00004
1.20	2.94981	3.17994	0.23013	3.17046	3.17994	0.00948	3.17958	3.17994	0.00036	3.17989	3.17994	0.00005
1.40	3.45177	3.73240	0.28063	3.72117	3.73240	0.01123	3.73198	3.73240	0.00042	3.73234	3.73240	0.00006
1.60	3.95013	4.28348	0.33336	4.27062	4.28348	0.01286	4.28302	4.28348	0.00046	4.28341	4.28348	0.00007
1.80	4.42815	4.81518	0.38702	4.80096	4.81518	0.01422	4.81470	4.81518	0.00048	4.81509	4.81518	0.00009
2.00	4.86579	5.30547	0.43969	5.29037	5.30547	0.01510	5.30501	5.30547	0.00047	5.30536	5.30547	0.00011

Table 3.2: Comparison of computed values y_i^h and exact values y_i for different ν.

3.1.4 Stability

In the previous sections, we examined different one-step methods. In this section, we discuss their convergence and stability. Recalling from early, we can write a one-step difference equation method used to solve the IVP as

$$y_{i+1}^h = y_i^h + h\Phi(t_i, y_i^h; h), \quad 0 \le i \le n-1, \quad y_0^h = \alpha, \tag{3.16}$$

where $\Phi(\cdot, \cdot; \cdot)$ is called the increment function.

We first state the notion of consistency.

Definition 3.17. A one-step difference equation method with local truncation error $\tau_i(h)$ at the i^{th} step is said to be *consistent* with the differential equation it approximates if

$$\lim_{h \to 0} \max_{1 \le i \le n} |\tau_i(h)| = 0.$$

Let $\tau(h) = \max_{1 \le i \le n} |\tau_i(h)|$, then we refer to $\tau(h)$ as the global truncation error. Hence, a method is said to be consistent if $\lim_{h \to 0} \tau(h) = 0$. Moreover, we say that the scheme has order p if $\tau(h) = \mathcal{O}(h^p)$ for $h \to 0$.

By consistency, we mean that the difference equation for the method approaches the differential equation as h tends to zero. Hence, consistency is a property of the discretization method, not of the solution.

Apart from consistency, another important concept is the stability of the numerical method. Similar to the well-posedness of ODEs, we have the concept of zero stability of the one-step methods.

Definition 3.18. The numerical method Eq. (3.16) for the approximation of the IVP is *zero stable* if for every $\varepsilon > 0$, there exists an $h_0 > 0$ and $C > 0$ such that for all $h \in (0, h_0]$

$$|\tilde{y}_i^h - y_i^h| \leq C\varepsilon \quad \text{for} \quad 0 \leq i \leq n, \tag{3.17}$$

where $\tilde{y}_i^h$ is the solution to the perturbed problem

$$\begin{aligned} \tilde{y}_0^h &= \alpha + \delta_0, \\ \tilde{y}_{i+1}^h &= \tilde{y}_i^h + h\left[\Phi(t_i, \tilde{y}_i^h; h) + \delta_{i+1}\right], \end{aligned}$$

and y_i^h is the solution to the problem

$$\begin{aligned} y_0^h &= \alpha, \\ y_{i+1}^h &= y_i^h + h\Phi(t_i, y_i^h; h), \end{aligned}$$

for $i = 0, 1, 2, \ldots, n-1$ under the assumption $|\delta_k| \leq \varepsilon$ for $0 \leq k \leq n$.

This definition means that if we have controlled perturbations (bounded by ε) in our data, then the solution remains similarly controlled. The zero-stability requires Eq. (3.17) holds for any value $h \leq h_0$.

Next, we give a theorem without proof that provides a sufficient condition for a method to be stable.

Theorem 3.19. *Suppose the IVP is approximated by a one-step difference method in the form*

$$\begin{aligned} y_0^h &= \alpha, \\ y_{i+1}^h &= y_i^h + h\Phi(t_i, y_i^h, h). \end{aligned}$$

Suppose also that a number $h_0 > 0$ exists and that $\Phi(t, w, h)$ is continuous and satisfies a Lipschitz condition in the second variable w with Lipschitz constant L on the set

$$D = \{(t, w, h) \mid a \leq t \leq b,\ -\infty < w < \infty,\ 0 \leq h \leq h_0\}.$$

Then the one-step method is zero stable.

After consistency and stability of a one-step method, we deal with its convergence.

Definition 3.20. A one-step numerical method is said to be *convergent* with respect to the differential equation it approximates if

$$\lim_{h \to 0} \max_{1 \leq i \leq n} |y_i^h - y_i| = 0.$$

Let $C(h) = \max_{1 \leq i \leq n} |y_i^h - y_i|$. Then we say the method is convergent with order p if there exists a constant $\mathcal{C} > 0$ is independent of h such that $C(h) = \mathcal{C}\mathcal{O}(h^p)$.

Theorem 3.21. (**Convergence of One-Step Method**) *Under the same assumptions as in Theorem 3.19, we have*

$$|y_i - y_i^h| \leq \frac{\tau(h)}{L}\left(e^{iLh} - 1\right), \quad 1 \leq i \leq n.$$

Therefore, if the method is consistent and zero stable, then the method is convergent as well and it has the same order of convergence as that of consistency.

The proof of this theorem is similar to the one presented for the Euler method and is left as an exercise.

The above theorem states that if a method is consistent and zero-stable, then it is convergent. This property is called the *Lax Equivalence Theorem*, named after the Hungarian mathematician Peter Lax.

Fun fact: Peter Lax is the only numerical analyst (in the true sense) to have won the Abel Prize.

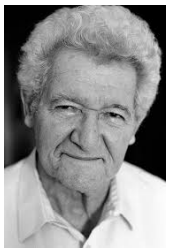

Figure 3.10: Lax Péter Dávid: May 11, 1926-.

Absolute Stability

Until now, we have dealt with the situation of stability in the case of a bounded time interval, but what if we do not? For example, in the case of population modeling, $t \to \infty$. Then, our notion of stability should also change.

Absolute stability deals with the asymptotic behavior of y_i^h as opposed to zero-stability, which dealt with stability on a fixed time interval. Let us consider the linear IVP:

$$\begin{aligned} y'(t) &= \lambda y(t), \quad t > 0, \\ y(0) &= 1, \end{aligned} \tag{3.18}$$

with $\lambda \in \mathbb{C}$. We know that $e^{\lambda t}$ is the analytical solution to it, and $\lim_{t\to\infty} |y(t)| = 0$ if $\mathrm{Re}(\lambda) < 0$.

Definition 3.22. A numerical method for approximating Eq. (3.18) is *absolutely stable* if

$$|y_i^h| \to 0 \quad \text{as} \quad t_i \to \infty. \tag{3.19}$$

Let h be the discretization step size. The numerical solution y_i^h of Eq. (3.18) depends on h and λ. The *region of absolute stability* of the numerical method is the subset of the complex plane

$$\mathcal{A} = \{z = h\lambda \in \mathbb{C} \mid \text{Eq. (3.19) holds}\}.$$

Thus, $\mathcal{A}$ is the set of values of $h\lambda$ for which the numerical method finds solutions that decay to zero as $t_i \to \infty$.

Let us look at the region of stability for a few methods.

- **Euler Method**: The Euler method (3.8) applied to Eq. (3.18) gives

$$\begin{aligned} y_{i+1}^h &= y_i^h + h\lambda y_i^h, \\ &= y_i^h(1 + h\lambda), \qquad \text{for } i \geq 0. \end{aligned}$$

Applying this recursively, we get

$$y_{i+1}^h = (1 + h\lambda)^{i+1},$$

as $y_0^h = 1$. Let $\varphi(z) = 1+z$, we call this function as the *stability function*. Now, Eq. (3.19) is satisfied if $\varphi(z) < 1$. From Complex Analysis, we know that $|1 + z| < 1$ if z lies within the unit circle centered at $(-1, 0)$. This means $h\lambda \in \mathbb{C}^-$.

Let $\lambda = a + \iota b$, then we have

$$\begin{aligned} |1 + h\lambda|^2 &= (1 + h(a + \iota b))\,\overline{(1 + h(a + \iota b))} \\ &= (1 + h(a + \iota b))\,(1 + h(a - \iota b)) \\ &= (1 + ah)^2 + (bh)^2. \end{aligned}$$

If $|1 + h\lambda| < 1$, then we have $1 + 2ah + a^2h^2 + b^2h^2 < 1$, which implies

$$h < -\frac{2a}{a^2 + b^2} = -\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}.$$

Hence, the region of stability for the Euler method is all those $h\lambda$ such that

$$h\lambda \in \mathbb{C}^- \quad \text{and} \quad 0 < h < -\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}.$$

Fig. 3.11 shows the region of stability for the Euler method.

- **Implicit Euler Method**: Until now, we have discussed only explicit methods. Explicit methods are called explicit because the solution y_{i+1}^h depends only on the previous solution at t_i. Now, we can also have methods that are implicit in nature, i.e., the solution at y_{i+1}^h depends on y_{i+1}^h itself. We will not cover implicit methods in detail, but we will

discuss one of the simplest implicit methods, also known as the *Implicit Euler Method*. The solution given by the implicit Euler method applied to Eq. (3.5) is given by

$$\begin{aligned} y_0^h &= \alpha, \\ y_{i+1}^h &= y_i^h + hf(t_{i+1}, y_{i+1}^h), \quad \text{for } i = 0, 1, 2, \ldots, n-1. \end{aligned} \tag{3.20}$$

Now applying Eq. (3.20) to Eq. (3.18), we get

$$y_{i+1}^h = \frac{y_i^h}{1 - h\lambda}.$$

Applying this recursively and using $y_0^h = 1$, we get

$$y_{i+1}^h = \frac{1}{(1 - h\lambda)^{i+1}}.$$

Here the stability function is given by $\varphi(z) = \frac{1}{|1-z|}$. Now, Eq. (3.19) is satisfied if $1 < |1 - z|$, which holds true for every $z \in \mathbb{C} \setminus B_{(1,0)}(1)$ (here $B_{(1,0)}(1)$ denotes the unit circle centered at $(1, 0)$ in the complex plane). Fig. 3.11 shows the region of stability for the implicit Euler method.

- **Runge-Kutta Methods**: We will not derive the stability function for the RK method but it is given by

$$\varphi(z) = \frac{\det(\mathbf{I} - z\mathbf{A} + z\mathbf{e}\mathbf{b}^\top)}{\det(\mathbf{I} - z\mathbf{A})},$$

where $\mathbf{e}$ denotes the vector with all ones. In case of RK methods that were discussed in these notes, $\mathbf{A}$ is strictly lower triangular and hence, $\varphi(z)$ is a polynomial function.

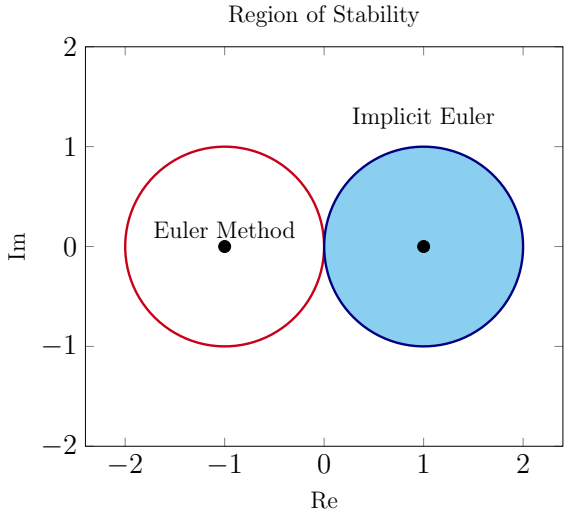

Figure 3.11: Region of Stability for the Euler and Implicit Euler Method

The last concept we want to handle is the concept of A-stability.

Definition 3.23. We say that a method has *A-Stability* if $\mathcal{A} \cap \mathbb{C}^- = \mathbb{C}^-$, i.e., $\mathrm{Re}(\lambda) < 0$.

From the above definition we note that the Euler method is not A-stable whereas implicit Euler method is A-stable. The reason it is called A-stable because it requires that the absolute stability holds for *all* values of $h\lambda \in \mathbb{C}^-$. Hence A-stability implies absolute stability but not vice-versa.

Example 3.24. Let us consider the IVP

$$y'(t) = -5y(t), \quad \text{for } t \geq 0, \quad y(0) = 1.$$

We know that the explicit Euler method has the region of stability if

$$h < -\frac{2\text{Re}(\lambda)}{|\lambda^2|} = -\frac{(-2 \times 5)}{25} = \frac{2}{5}.$$

Therefore, if $h > 0.4$ we should get unstable solutions. Fig. 3.12(left) shows the solution obtained with $h = 0.2, 0.4$, and 0.5. We notice that the solution is not stable for $h = 0.5$ and 0.4 but for $h = 0.4$ the solution is not growing.
In Fig. 3.12(right) we use the implicit Euler method. We notice that the solution is stable. We also show how the method is actually implemented. Now, the implicit Euler method is given by

$$y_{i+1}^h = y_i^h + hf(t_{i+1}, y_{i+1}^h), \quad i \geq 0,$$

and $y_0^h = 1$. For $f(t, y) = -5y(t)$, we get

$$y_{i+1}^h = \frac{y_i^h}{1 + 5h}.$$

Let $h = 0.5$, then

$$y_1 = \frac{y_0}{1 + 5 \times 0.5} = 1/3.5 = 0.2857,$$

and

$$y_2 = \frac{y_1}{1 + 5 \times 0.5} = 0.2857/3.5 = 0.08162.$$

3.2 Multistep Methods

One-step methods are called *one-step* because the solution at the point t_{i+1} depends only on the solution at t_i. Even though we use slope values at intermediate time points, we do not store these values for, say, computations at t_{i+2}. We also observe that as i increases, the error grows. Hence, it is reasonable to develop methods that consider not only the values at t_i but also at $t_0, t_1, \ldots, t_{i-1}$, thereby giving rise to multistep methods. The solution obtained by a multistep method, rather than relying solely on t_{i-1}, depends on the previous m time steps.

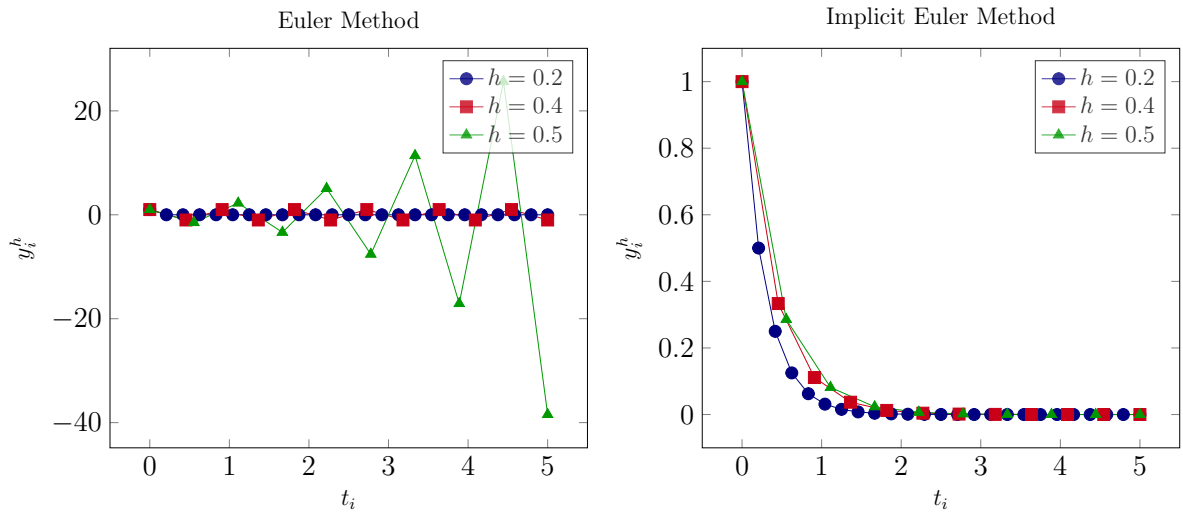

Figure 3.12: Numerical solution obtained with the Euler method (left) and the implicit Euler method (right).

Definition 3.25. An *m-step multistep method* for solving the IVP

$$y'(t) = f(t, y(t)), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{3.21}$$

is represented by the following equation:

$$y_{i+1}^h = \sum_{j=0}^{m-1} a_j y_{i+1-m+j}^h + h \sum_{j=0}^{m} b_j f(t_{i+1-m+j}, y_{i+1-m+j}^h), \tag{3.22}$$

for $i = m-1, m, \ldots, n-1$, where m is an integer greater than one, y_{i+1}^h is the approximation at the mesh point t_{i+1}, $h = (b-a)/n$, and $\{a_j\}_{j=0}^{m-1}$, $\{b_j\}_{j=0}^{m}$ are constants. The starting values are denoted by

$$y_0^h = \alpha, \quad y_1^h = \alpha_1, \ldots, \quad y_{m-1}^h = \alpha_{m-1}, \tag{3.23}$$

which are specified.

Let us look at an example. For $m = 3$, we have

$$y_{i+1}^h = a_2 y_i^h + a_1 y_{i-1}^h + a_0 y_{i-2}^h + h \sum_{j=0}^{3} b_j f(t_{i-2+j}, y_{i-2+j}^h). \tag{3.24}$$

If $b_m = 0$, then we get an explicit m-step multistep method as the solution at t_{i+1} completely depends on previous values. If $b_m \neq 0$, then we get an implicit method as the solution depends implicitly. Some of the most commonly used multistep methods include:

1. **Adams-Bashforth Method** This is a four-step explicit method given by

$$\begin{aligned} y_0^h &= \alpha, \quad y_1^h = \alpha_1, \quad y_2^h = \alpha_2, \quad y_3^h = \alpha_3, \\ y_{i+1}^h &= y_i^h + \frac{h}{24}\left[55f(t_i, y_i^h) - 59f(t_{i-1}, y_{i-1}^h) + 37f(t_{i-2}, y_{i-2}^h) - 9f(t_{i-3}, y_{i-3}^h)\right] \end{aligned}$$

   for $i = 3, 4, \ldots, n-1$.

2. **Adams-Moulton Method** This is an implicit three-step method given by

$$\begin{aligned} y_0^h &= \alpha, \quad y_1^h = \alpha_1, \quad y_2^h = \alpha_2, \\ y_{i+1}^h &= y_i^h + \frac{h}{24}\left[9f(t_{i+1}, y_{i+1}^h) + 19f(t_i, y_i^h) - 5f(t_{i-1}, y_{i-1}^h) + f(t_{i-2}, y_{i-2}^h)\right] \end{aligned}$$

   for $i = 2, 3, 4, \ldots, n-1$.

The above methods are named after John Couch Adams, who also discovered Neptune, Francis Bashforth, who was a priest (as well as a mathematician), and Forest Ray Moulton, who has a crater named after him on the moon.

Figure 3.13: John Couch Adams (June 5, 1819 – January 21, 1892, left), Francis Bashforth (January 8, 1819 – February 12, 1912, center), and Forest Ray Moulton (April 29, 1872 – December 7, 1942, right).

Both schemes are of fourth order. In reality, the Adams-Bashforth method and the Adams-Moulton method are a class of multistep methods, i.e., there exist second-, third-, and higher-order Adams-Bashforth and Adams-Moulton methods. We will not present the detailed derivation of these methods, but the main idea is to write the IVP as

$$y_{i+1} = y_i + \int_{t_i}^{t_{i+1}} f(t, y(t))dt, \tag{3.25}$$

and then interpolate $(t, f\,(t, y(t)))$ using the previous time step values and then integrate them. Depending on the number of integration points, we get different order Adams-Bashforth and Adams-Moulton methods.

Until now, we have not discussed the values of α_i. Due to the initial condition, we know y_0^h, but for the remaining α_i, we need to use either a Taylor method or a Runge-Kutta (RK) method before starting with the multistep method.

The Adams-Moulton method is an implicit method, i.e., y_{i+1}^h also appears on the RHS. Now, if $f(t, y)$ is nonlinear in $y(t)$, then we cannot solve this equation easily, say if $f(t, y) = e^y$. In this case, we use a predictor-corrector strategy. We first deploy an Adams-Bashforth method to get a predictor of $y_{i+1}^{h,\mathrm{P}}$ and then use this $y_{i+1}^{h,\mathrm{P}}$ to compute y_{i+1}^h in the Adams-Moulton method. The reason we do this is that implicit methods inherently provide better accuracy, and we want to leverage that.

3.2.1 Stability

For multistep methods, the concept of stability, consistency, and convergence are compounded as the solution at $(i+1)^{\text{th}}$ step not only depends on the i^{th} step but other previous steps. A general m-multistep method for approximating the solution to the IVP

$$y'(t) = f(t,y), \quad a \le t \le b, \quad y(a) = \alpha$$

has the form

$$\begin{aligned} y_0^h &= \alpha, \quad y_1^h = \alpha_1, \cdots, y_{m-1}^h = \alpha_{m-1}, \\ y_{i+1}^h &= a_{m-1}y_i^h + a_{m-2}y_{i-1}^h + \cdots + a_0 y_{i+1-m}^h + hF(t_i, h, y_{i+1}^h, y_i^h, \ldots, y_{i+1-m}^h), \end{aligned} \quad (3.26)$$

for $i = m-1, m, \ldots, n-1$ where $\{a_j\}_{j=0}^{m-1}$ are constants, $h = (b-a)/n$, and $t_i = a + ih$.

We first extend the notion of local truncation error for a multistep method expressed as above is given by

$$\tau_{i+1}(h) = \frac{y_{i+1} - a_{m-1}y_i - \cdots - a_0 y_{i+1-m}}{h} - F(t_i, h, y_{i+1}, y_i, \ldots, y_{i+1-m}),$$

for $i = m-1, m, \ldots, n-1$.

We are not going to derive the truncation error for the Adam-Bashforth and Adam-Moulton methods, but using the error estimates from interpolation theory, we get for the Adam-Bashforth method the truncation error as:

$$\tau_{i+1}(h) = \frac{251}{720} y^{(5)}(\mu_i) h^4, \quad \text{for some } \mu_i \in (t_{i-3}, t_{i+1}),$$

whereas the Adam-Moulton method

$$\tau_{i+1}(h) = -\frac{19}{720} y^{(5)}(\mu_i) h^4, \quad \text{for some } \mu_i \in (t_{i-2}, t_{i+1}).$$

Hence, if $y^{(5)}$ is bounded, we get the consistency error of the order four.

For the convergence analysis, we make two assumptions:

1. If $f = 0$ (i.e., the differential equation is homogeneous), then $F = 0$ also.
2. F satisfies Lipschitz condition with respect to $\{y_j^h\}$, in the sense that a constant L exists and for every pair of sequence $\{v_j\}_{j=0}^n$ and $\{\overline{v}_j\}_{j=0}^n$ and for $i = m-1, m, \ldots, n-1$, we have

$$|F(t_i, h, v_{i+1}, \ldots, v_{i+1-m} - F(t_i, h, \overline{v}_{i+1}, \ldots, \overline{v}_{i+1-m}| \le \sum_{j=0}^{m} |v_{i+1-j} - \overline{v}_{i+1-j}|.$$

The Adam-Bashforth and the Adam-Moulton satisfy these conditions provided f satisfies a Lipschitz condition.

After introducing the basic definition we first present the notion of convergence. Now this notion is the same for the multistep method as the one-step method, i.e., if $\{y_i^h\}_{i=0}^n$ is the numerical solution, then the multistep method is said to be convergent if

$$\lim_{h \to 0} \max_{0 \le i \le n} |y_i^h - y_i| = 0.$$

The notion of consistency differs slightly for the multistep method. Here, we want the difference method to approach the differential equation as $h \to 0$, but we also want the starting steps to satisfy the same condition. Since only y_0^h is known and is exact, we need to require that the errors in the starting values $\{\alpha_i\}$ also approach zero as step size decreases, i.e.,

$$\lim_{h\to 0} |\tau_i(h)| = 0 \quad \text{for} \quad i = m, m+1, \ldots, n, \quad \text{and}$$
$$\lim_{h\to 0} |\alpha_i - y_i^h| = 0 \quad \text{for} \quad i = 1, 2, \ldots, m-1.$$

Note that the consistency of multistep methods is based on the consistency of its one-step method.

Before we start with the discussion of the stability of multistep methods, we want to introduce certain notions. Associated with Eq. (3.26) is a polynomial called as the *characteristic polynomial* of the method given by

$$p(\lambda) = \lambda^m - a_{m-1}\lambda^{m-1} - a_{m-2}\lambda^{m-2} - \cdots - a_1\lambda - a_0.$$

Definition 3.26. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ denote the (not necessarily distinct) roots of the characteristic polynomial.

$$p(\lambda) = \lambda^m - a_{m-1}\lambda^{m-1} - a_{m-2}\lambda^{m-2} - \cdots - a_1\lambda - a_0,$$

associated with the multistep method (3.26). If $|\lambda_i| \leq 1$ for each $i = 1, 2, \ldots, m$ and all roots with absolute value 1 are simple roots, then the difference method is said to satisfy the *root condition*.

Definition 3.27.
1. Methods that satisfy the root condition and have $\lambda = 1$ as the only root of the characteristic equation with magnitude one are called *strongly stable*.
2. Methods that satisfy the root condition and have more than one distinct root with magnitude one are called *weakly stable*.
3. Methods not satisfying the root condition are called *unstable*.

The idea behind the use of the characteristic polynomial comes from recurrence relations. Now, if $\{y_i\}$ is the analytical solution and $\{y_i^h\}$ is the numerical solution. Then the error $e_{i+1} = y_{i+1}^h - y_{i+1}$ satisfy a homogenous recurrence relation of the form

$$e_{i+1} - \sum_{j=0}^{m-1} a_{m-1-j} e_{i-j} = 0.$$

Here, we used our second assumption on F. Using the theory of recurrence relation, we get the idea of the root condition.

Theorem 3.28. *A multi-step method of the form*

$$\begin{aligned} y_0^h &= \alpha, \quad y_1^h = \alpha_1, \cdots, y_{m-1}^h = \alpha_{m-1}, \\ y_{i+1}^h &= a_{m-1}y_i^h + a_{m-2}y_{i-1}^h + \cdots + a_0 y_{i+1-m}^h + hF(t_i, h, y_{i+1}^h, y_i^h, \ldots, y_{i+1-m}^h), \end{aligned}$$

is stable if and only if it satisfies the root condition. Moreover, if the difference method is consistent with the differential equation, then the method is stable if and only if it is convergent.

The above theorem is called the Dahlquest equivalence theorem, named after Swedish mathematician Germund Dahlquest, and is in the same spirit as the Lax-Equivalence theorem.

Example 3.29.

1. The fourth-order Adam-Bashforth method has the characteristic polynomial
$$p(\lambda) = \lambda^4 - \lambda^3 = \lambda^3(\lambda - 1),$$
and has roots $\{0, 0, 0, 1\}$ and is strongly stable.
2. The fourth-order Adam-Milne method given by
$$y_{i+1}^h = y_{i-3}^h + \frac{4h}{3}\left[2f(t_i, y_i^h) - f(t_{i-1}, y_{i-1}^h) + 2f(t_{i-2}, y_{i-2}^h)\right],$$
has the characteristic polynomial $p(\lambda) = \lambda^4 - 1$ has four roots $\pm 1, \pm\iota$. As all the roots have magnitude one, the method satisfies the root condition, but as we have multiple roots with magnitude one, the method is only weakly stable.

3.3 Finite Difference Methods

Until now, we have focussed intensively on differential equations with one derivative and one initial condition, but we know that some PDEs have higher-order terms. The conditions are defined on the boundary points rather than at the initial point. This section will concentrate on differential equations over the entire boundary, hence the name Boundary Value Problems (BVP). In general, physical problems that are space-dependent rather than time are modeled using BVPs, such as beam bending.

The two-point BVP that we will consider in this course is given by

$$y''(x) = f(x, y(x), y'(x)), \quad \text{for} \quad a \leq x \leq b, \tag{3.27}$$

with $y(a) = \alpha$ and $y(b) = \beta$.

We first start with certain existence and uniqueness results for the BVP.

Theorem 3.30. *Suppose that the function f in the BVP*

$$y''(x) = f(x, y(x), y'(x)), \quad \text{for} \quad a \le x \le b,$$

with $y(a) = \alpha$ and $y(b) = \beta$ is continuous on the set

$$D = \{(x, y, y') | \text{ for } \ a \le x \le b, \ \text{ with } \ -\infty < y < \infty \ \text{ and } \ -\infty < y' < \infty\},$$

and the partial derivatives f_y and $f_{y'}$ are also continuous on D. If

1. *$f_y(x, y, y') > 0$ for all $(x, y, y') \in D$, and*
2. *a constant M exists, with*

$$|f_{y'}(x, y, y')| \le M \ \ \forall \ (x, y, y') \in D,$$

then the BVP has a unique solution.

In the case $f(x, y, y')$ is linear in y and y', i.e., there exists $p(x), q(x)$, and $r(x)$ exist with

$$f(x, y, y') = p(x)y' + q(x)y + r(x),$$

then Theorem 3.30 can be simplified.

Corollary 3.31. *Suppose the linear BVP*

$$y''(x) = p(x)y' + q(x)y + r(x), \quad \text{for} \quad a \le x \le b,$$

with $y(a) = \alpha$ and $y(b) = \beta$ satisfies

1. *$p(x), q(x)$, and $r(x)$ are continuous on $[a, b]$.*
2. *$q(x) > 0$ on $[a, b]$.*

Then, the BVP has a unique solution.

3.3.1 Finite Difference Method for Linear Problems

The finite difference method (FDM) is the most popular method for solving a BVP numerically. Methods involving finite differences for solving BVP replace each derivative in the differential equation with an appropriate difference-quotient approximation. The FDM for the linear second-order BVP

$$y''(x) = p(x)y'(x) + q(x)y(x) + r(x), \quad \text{for} \quad a \le x \le b,$$

with $y(a) = \alpha$ and $y(b) = \beta$ requires the difference-quotient approximation be used to approximate both y' and y''.

We start with meshing over domain $[a, b]$ into equidistant sub-intervals. Hence. we divide $[a, b]$ into sub-intervals using a step size $h = (b - a)/(n + 1)$ such that

$$x_i = a + ih, \quad i = 0, 1, \ldots, n + 1.$$

Here $x_0 =$ and $x_{n+1} = a + (n+1)h = b$. At the interior mesh points, x_i for $i = 1, 2, \ldots, n$ the differential equation to be approximated is

$$y_i'' = p_i y_i' + q_i y_i + r_i.$$

We have used y_i'' to denote $y''(x_i)$ and similarly for p_i, q_i, r_i, y_i, and y_i'. We start by using a Taylor polynomial approximation of $y(x)$ at x_i and evaluate it at x_{i+1} to get

$$y_{i+1} = y(x_i + h) = y_i + hy_i' + \frac{h^2}{2}y_i'' + \frac{h^3}{6}y_i^{(3)} + \frac{h^4}{24}y^{(4)}(\xi_i^+), \tag{3.28}$$

for some $\xi_i^+ \in (x_i, x_{i+1})$. Similarly we use the Taylor approximation to evaluate at x_{i-1}, i.e,

$$y_{i-1} = y(x_i - h) = y_i - hy_i' + \frac{h^2}{2}y_i'' - \frac{h^3}{6}y_i^{(3)} + \frac{h^4}{24}y^{(4)}(\xi_i^-), \tag{3.29}$$

for some $\xi_i^- \in (x_{i-1}, x_i)$ For both the cases we assumed that $y \in \mathcal{C}^4[x_{i-1}, x_{i+1}]$.

Adding Eq. (3.28) and Eq. (3.29) we get

$$y_i'' = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{h^2}{24}\left[y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-)\right].$$

Using the intermediate value theorem on $y^{(4)}(x)$ we have there exists a $\xi_i \in (x_{i-1}, x_{i+1})$ such that

$$y_i'' = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{h^2}{24}y^{(4)}(\xi_i).$$

This is called the *central-difference formula* for y_i''.

If we subtract Eq. (3.29) from Eq. (3.28) we get

$$y_i' = \frac{y_{i+1} - y_{i-1}}{2h} - \frac{h^2}{6}y^{(3)}(\eta_i)$$

for some $\eta_i \in (x_{i-1}, x_{i+1})$. This is the central difference approximation of $y'(x)$. We also have forward and backward difference approximations of $y'(x)$ in the literature. However, the central difference is used as it has a higher order.

Using the central difference approximation, we get

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = p_i\left(\frac{y_{i+1} - y_{i-1}}{2h}\right) + q_i y_i + r_i - \frac{h^2}{12}\left(2p_i y^{(3)}(\eta_i) - y^{(4)}(\xi_i)\right).$$

Ignoring the remainder term, we get an FDM with a truncation error of $\mathcal{O}(h^2)$ along with the boundary conditions $y(a) = \alpha$ and $y(b) = \beta$ to define the system of the linear equation as

$$y_0^h = \alpha, \quad y_{n+1}^h = \beta,$$

and

$$-\frac{y_{i+1}^h - 2y_i^h + y_{i-1}^h}{h^2} + p_i\left(\frac{y_{i+1}^h - y_{i-1}^h}{2h}\right) + q_i y_i^h = -r_i,$$

for $i = 1, 2, \ldots, n$.

Rearranging the terms and using the boundary conditions, we get

$$-\left(1+\frac{h}{2}p_i\right)y_{i-1}^h+\left(2+h^2q_i\right)y_i^h-\left(1-\frac{h}{2}p_i\right)y_{i+1}^h=-h^2r_i,$$

and the resulting system of equations is expressed in the form of a $n\times n$ tridiagonal matrix

$$\mathbf{Ay}=\mathbf{b},$$

where

$$\mathbf{A}=\begin{bmatrix} 2+h^2q_1 & -1+\frac{h}{2}p_1 & 0 & \cdots & 0 & 0 & 0\\ -1-\frac{h}{2}p_2 & 2+h^2q_2 & -1+\frac{h}{2}p_2 & \cdots & 0 & 0 & 0\\ 0 & -1-\frac{h}{2}p_3 & 2+h^2q_3 & \cdots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1-\frac{h}{2}p_{n-1} & 2+h^2q_{n-1} & -1+\frac{h}{2}p_{n-1}\\ 0 & 0 & 0 & \cdots & 0 & -1-\frac{h}{2}p_n & 2+h^2q_n \end{bmatrix},$$

$$\mathbf{y}=\begin{bmatrix} y_1^h\\ y_2^h\\ \vdots\\ y_{n-1}^h\\ y_n^h \end{bmatrix},\quad \text{and}\quad \mathbf{b}=\begin{bmatrix} -h^2r_1+\left(1+\frac{h}{2}p_1\right)y_0^h\\ -h^2r_2\\ \vdots\\ -h^2r_{n-1}\\ -h^2r_n+\left(1-\frac{h}{2}p_n\right)y_{n+1}^h \end{bmatrix}.$$

The following results mention the existence and uniqueness of the solution of the above method.

Theorem 3.32. *Suppose that $p(x), q(x)$ and $r(x)$ are continuous on $[a,b]$. If $q(x)\geq 0$ on $[a,b]$, then the system of equation $\mathbf{Ay}=\mathbf{b}$ has a unique solution provided that $h<2/L$ where $L=\max_{a\leq x\leq b}|p(x)|$.*

Example 3.33. Consider the BVP

$$y''(x) = 4(y(x) - x), \quad 0 \le x \le 1, \quad y(0) = 0, \quad y(1) = 2.$$

We solve the above problem using a mesh with 5 points, i.e., $h = 0.25$. The finite difference approximation gives

$$\frac{y_{i+1}^h - 2y_i^h + y_{i-1}^h}{h^2} = 4(y_i^h - x_i), \quad i = 1, 2, 3,$$

with $y_0^h = 0$ and $y_4^h = 2$. For $i = 1$, we have

$$\frac{y_2^h}{h^2} - \left(\frac{2}{h^2} + 4\right) y_1^h = -1,$$

and similarly for $i = 2, 3$ we get

$$\frac{y_3^h}{h^2} - \left(\frac{2}{h^2} - 4\right) y_2^h + \frac{y_1^h}{h^2} = -2,$$

and

$$\frac{2}{h^2} + \frac{y_2^h}{h^2} - \left(\frac{2}{h^2} - 4\right) y_3^h = -3.$$

Hence, the system of equations is

$$\begin{bmatrix} -\left(\frac{2}{h^2} + 4\right) & \frac{1}{h^2} & 0 \\ \frac{1}{h^2} & -\left(\frac{2}{h^2} - 4\right) & \frac{1}{h^2} \\ 0 & \frac{1}{h^2} & -\left(\frac{2}{h^2} - 4\right) \end{bmatrix} \begin{bmatrix} y_1^h \\ y_2^h \\ y_3^h \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \\ -3 - \frac{2}{h^2} \end{bmatrix}.$$

Algorithm 18 presents the algorithm for the FDM for linear BVPs.

Algorithm 18 Finite Difference Method for Linear Problem

Given: Endpoints a, b, Number of internal mesh points: n, Boundary conditions: $\alpha,\ \beta$, Functions: $p(x), q(x)$, and $r(x)$.
Find: Solution $\{y_i^h\}_{i=0}^{n+1}$.

Step 1: Meshing and Initialization
Initialize $h = \frac{b-a}{n+1}$
Initialize $\overline{\mathbf{y}}^h$ as zero vector of size n
Initialize $\mathbf{b}$ as zero vector of size n
Initialize $\mathbf{A}$ as a $n \times n$ zero matrix

Step 2: Finite Difference Method

Step 2.1: Create A and b
for $i = 1$ **to** n **do**
  $x_i = a + ih$
  $\mathbf{A}_{ii} = 1 + h^2 q(x_i)$
  if $i > 1$ **then**
    $\mathbf{A}_{i,i-1} = -1 - \frac{h}{2} p(x_i)$
  end if
  if $i < n$ **then**
    $\mathbf{A}_{i,i+1} = -1 + \frac{h}{2} p(x_i)$
  end if
  $\mathbf{b}_i = -h^2 r(x_i)$
end for

Step 2.2: Adjust Boundary Conditions
$\mathbf{b}_1 = \mathbf{b}_1 + \left(1 + \frac{h}{2} p(x_1)\right) \alpha$
$\mathbf{b}_n = \mathbf{b}_n + \left(1 - \frac{h}{2} p(x_n)\right) \beta$

Step 2.3: Solve $\mathbf{A}\overline{\mathbf{y}}_h = \mathbf{b}$
Solve the system $\mathbf{A}\overline{\mathbf{y}}_h = \mathbf{b}$ to get coefficient vector $\overline{\mathbf{y}}_h$.

Step 2.4: Concatenate
Concatenate $\mathbf{y}_h = \begin{bmatrix} \alpha & \overline{\mathbf{y}}_h & \beta \end{bmatrix}$.

---

return $\mathbf{y}^h$

3.3.2 Finite Difference Method for Non-Linear Problems

In the last section, we dealt with linear BVPs. In this, we will consider the general nonlinear BVP

$$y'' = f(x, y, y') \quad \text{for} \quad a \leq x \leq b,$$

with $y(a) = \alpha$ and $y(b) = \beta$. For this the difference method remains the same. However, the system of equations is nonlinear, so we need to solve it iteratively.

We make certain assumptions on f for the further development of the methods, namely:

1. f and the partial derivatives f_y and $f_{y'}$ are all continuous on

$$D = \{(x, y, y') | \text{ for } \ a \leq x \leq b, \ \text{ with } \ -\infty < y < \infty \ \text{ and } \ -\infty < y' < \infty\}.$$

2. $f_y(x, y, y') \geq \delta$ on D for some $\delta > 0$, i.e., the partial derivative is bounded away from zero.

3. Constants K and L exist, with

$$K = \max_{(x,y,y') \in D} |f_y(x, y, y')| \quad \text{and} \quad L = \max_{(x,y,y') \in D} |f_{y'}(x, y, y')|,$$

i.e., the partial derivatives are bounded.

Hence, the hypothesis of Theorem 3.30 is satisfied, and we have the existence of a unique solution.

For the FDM, we start similarly, i.e., by meshing our domain $[a, b]$ into $(n+1)$ equidistant sub-intervals. Assuming that $y \in \mathcal{C}^4[a, b]$ we can replace $y''(x)$ and $y'(x)$ by their central difference approximation to get

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h} - \frac{h^2}{6} y^{(3)}(\eta_i)\right) + \frac{h^2}{12} y^{(4)}(\xi_i),$$

for some $\xi_i, \eta_i \in (x_{i-1}, x_{i+1})$.

If we omit the error term, we get a difference method of the form

$$y_0^h = \alpha, \quad y_{n+1}^h = \beta,$$

and

$$-\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + f\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right), \ \text{ for } \ i = 1, 2, \ldots, n.$$

In a compact form, we get a system of equations of the form

$$\mathbf{F}(\mathbf{y}^h) = \mathbf{0}, \tag{3.30}$$

where $\mathbf{F}(\mathbf{y}^h) = \begin{bmatrix} f_1(\mathbf{y}^h) & f_2(\mathbf{y}^h) & \cdots f_n(\mathbf{y}^h) \end{bmatrix}^\top$ and

$$\begin{aligned}
f_1(\mathbf{y}^h) &= 2y_1^h - y_2^h + h^2 f\left(x_1, y_1^h, \frac{y_2^h - \alpha}{2h}\right) - \alpha \\
f_2^h(\mathbf{y}^h) &= -y_1^h + 2y_2^h - y_3^h + h^2 f\left(x_2, y_2^h, \frac{y_3^h - y_1^h}{2h}\right) \\
&\vdots \\
f_{n-1}^h(\mathbf{y}^h) &= -y_{n-2}^h + 2y_{n-1}^h - y_n^h + h^2 f\left(x_{n-1}, y_{n-1}^h, \frac{y_n^h - y_{n-2}^h}{2h}\right) \\
f_n(\mathbf{y}^h) &= -y_{n-1}^h + 2y_n^h + h^2 f\left(x_n, y_n^h, \frac{\beta - y_{n-1}^h}{2h}\right) - \beta.
\end{aligned}$$

Solving a nonlinear system of equations is not straightforward. One of the methods, i.e., widely used, is the Newton method. In Newton method, we generate a sequence of approximations $\{\mathbf{y}_{(j)}^h\}_{j=0}^k$ which approximate Eq. (3.30) provided the initial approximation $\{\mathbf{y}_{(0)}^h\}$ is sufficiently close to the solution.

Before going into the formula for the Newton method, let us revisit root-finding methods. Given a continuous function f, a root-finding problem $f(x) = 0$ can be modified into a fixed-point finding problem, i.e., find a p such that $g(p) = p$ for some function $g(x)$. The equivalence of the two can be seen by choosing $g(x)$ in such a way that

$$g(x) = x - \gamma f(x),$$

for some $\gamma \neq 0$. If p is a zero of f, then p is a fixed point of g and vice versa. We can replace γ with a function $\varphi(x)$ such that $\varphi(p) \neq 0$. The Newton method is based on this ideology and replaces $\varphi(x)$ by $[f'(x)]^{-1}$. Hence, you get your standard Newton-Raphson method as

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}, \quad \text{for } n \geq 1.$$

In general, to compute the fixed point p, we write the function iteration as

$$x_{n+1} = g(x_n), \quad \text{for } n \geq 0,$$

and as g is continuous, it will converge to p.

The same ideas extend to systems of equations as well. Here, the derivative of a function is replaced by the Jacobian . Let

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix}$$

then the Jacobian is given by

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \partial_{x_1} f_1(\mathbf{x}) & \partial_{x_2} f_1(\mathbf{x}) & \cdots & \partial_{x_n} f_1(\mathbf{x}) \\ \partial_{x_1} f_2(\mathbf{x}) & \partial_{x_2} f_2(\mathbf{x}) & \cdots & \partial_{x_n} f_2(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{x_1} f_n(\mathbf{x}) & \partial_{x_2} f_n(\mathbf{x}) & \cdots & \partial_{x_n} f_n(\mathbf{x}) \end{bmatrix},$$

and the Newton iterative process is

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \left[\mathbf{J}(\mathbf{x}^{(k-1)})\right]^{-1} \mathbf{F}(\mathbf{x}^{(k-1)}).$$

Now, the entire problem boils down to the computation of the Jacobian. Each f_i has contributions from three points y_{i-1}^h, y_i^h, and y_{i+1}^h. Hence, the Jacobian will be a tridiagonal matrix with entries

$$\mathbf{J}_{ij} = \begin{cases} -1 + \frac{h}{2} f_{y'}\left(x_i, y_i^h, \frac{y_{i+1}^h - y_{i-1}^h}{2h}\right), & \text{for } \; i = j-1, \; \text{ and } \; j = 2, \ldots, n \\ 2 + h^2 f_y\left(x_i, y_i^h, \frac{y_{i+1}^h - y_{i-1}^h}{2h}\right), & \text{for } \; i = j, \; \text{ and } \; j = 1, 2, \ldots, n \\ -1 - \frac{h}{2} f_{y'}\left(x_i, y_i^h, \frac{y_{i+1}^h - y_{i-1}^h}{2h}\right), & \text{for } \; i = j+1, \; \text{ and } \; j = 1, \ldots, n-1 \end{cases}$$

with $y_0^h = \alpha$ and $y_{n+1}^h = \beta$. Now we compute the solution to the system

$$\mathbf{J}(\mathbf{y}_{(k-1)}^h)\mathbf{v} = \mathbf{F}(\mathbf{y}_{(k)}^h),$$

and then we get the new iterate as

$$\mathbf{y}_{(k)}^h = \mathbf{y}_{(k-1)}^h - \mathbf{v}.$$

To stop the nonlinear loop, we use the condition if

$$\|\mathbf{y}_{(k)}^h - \mathbf{y}_{(k-1)}^h\|_2 < \texttt{tol},$$

for some $\texttt{tol} \ll 1$.

Example 3.34. Consider the nonlinear BVP

$$y''(x) = 2y^3(x), \quad -1 \le x \le 0, \quad y(-1) = \frac{1}{2} \text{ and } y(0) = \frac{1}{3}.$$

We apply the FDM for nonlinear BVP with $h = 0.25$ to get the system of equation as

$$\frac{y_{i-1}^h - 2y_i^h + y_{i+1}^h}{h^2} - 2(y_i^h)^3 = 0, \quad \text{for } i = 1, 2, 3.$$

Then

$$\begin{aligned} i &= 1, \quad f_1(\mathbf{y}^h) &&= \frac{1}{2} - 2y_1^h - 2h^2(y_1^h)^3 + y_2^h, \\ i &= 2, \quad f_2(\mathbf{y}^h) &&= y_1^h - 2y_2^h + y_3^h - 2h^2(y_2^h)^3, \quad \text{and} \\ i &= 3 \quad f_3(\mathbf{y}^h) &&= y_2^h - 2y_3^h - 2h^2(y_3^h)^3 + \frac{1}{3}. \end{aligned}$$

The Jacobian elements can be computed, and the Jacobian is given by

$$\mathbf{J} = \begin{bmatrix} -2 - 6h^2(y_1^h)^2 & 1 & 0 \\ 1 & -2 - 6h^2(y_2^h)^2 & 1 \\ 0 & 1 & -2 - 6h^2(y_3^h)^2 \end{bmatrix}.$$

Taking an initial value $\mathbf{y}_{(0)}^h$ we solve the system

$$\mathbf{J}(\mathbf{y}_{(0)}^h)\mathbf{v} = \mathbf{F}(\mathbf{y}_{(0)}^h),$$

to get $\mathbf{v}$ and then compute $\mathbf{y}_{(1)}^h = \mathbf{y}_{(0)}^h - \mathbf{v}$. We continue this process unless we reach convergence or maximum iterations.

The algorithm for the Newton method is similar to Algorithm 18 where the matrix $\mathbf{A}$ is replaced by $\mathbf{J}$.

3.4 Rayleigh-Ritz Method

The idea of the Rayleigh-Ritz method can be traced back to engineering methods. John William Strut Rayleigh was a noble-prize-winning physicist who discovered Argon, and Walther Ritz was also a physicist who made significant contributions to BVP numerics in his short life span.

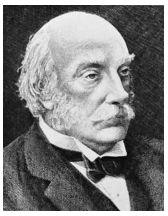
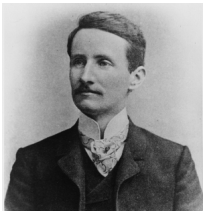

Figure 3.14: John William Strutt Rayleigh (November 12, 1842 – June 30, 1919, left) and Walther Ritz (February 22, 1878 – July 7, 1909, right).

The basic idea of the Rayleigh-Ritz method is to reformulate the problem as a minimization problem where we minimize certain integrals on a set of sufficiently differentiable functions satisfying the boundary conditions.

In this section, we will consider the second-order BVP of the form

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y(x) = f(x) \quad \text{for} \quad 0 \le x \le 1, \tag{3.31}$$

with homogeneous boundary conditions $y(0) = y(1) = 0$.

Theorem 3.35. *Let $p \in \mathcal{C}^1[0,1]$, $q(x), f(x) \in \mathcal{C}[0,1]$, and*

$$p(x) \ge \delta > 0, \quad q(x) \ge 0 \quad \text{for} \quad 0 \le x \le 1.$$

The function $y \in \mathcal{C}_0^2[0,1]$ is the unique solution to the differential equation (3.31) if and only if $y(x)$ is the unique function in $\mathcal{C}_0^2[0,1]$ that minimizes the integral

$$y(x) = \min_{u \in \mathcal{C}_0^2[0,1]} I[u] = \int_0^1 \left(p(x)\left[u'(x)\right]^2 + q(x)\left[u(x)\right]^2 - 2f(x)u(x)\right) dx. \tag{3.32}$$

The idea of taking this $I[u]$ comes from the Calculus of Variations. The proof of this theorem is done by first showing that the solution to Eq. (3.31) is a solution to the problem

$$\int_0^1 f(x)u(x)dx = \int_0^1 \left(p(x)\frac{dy}{dx}\frac{du}{dx} + q(x)y(x)u(x)\right) dx \quad \forall u(x) \in \mathcal{C}_0^2[0,1]. \tag{3.33}$$

Eq. (3.33) is called as the *variational problem* of Eq. (3.31). Next, we show that the solution $y \in \mathcal{C}_0^2[0,1]$ is a solution to Eq. (3.32) if and only if it is a solution to Eq. (3.33). After that, we show the uniqueness of the solution of Eq. (3.33) and hence finally, because of equivalence of solutions $y(x)$ is a unique solution to Eq. (3.31).

The Rayleigh-Ritz method computes the solution $y(x)$ by minimizing the integral, not over all the functions in $\mathcal{C}_0^2[0,1]$ but over a finite dimension subset $V_h \subset \mathcal{C}_0^2[0,1]$ consisting of linear combinations of certain basis functions $\{\varphi_i\}_{i=0}^n$. As these functions form a basis, they are linearly independent, and as they are a subset of $\mathcal{C}_0^2[0,1]$, they satisfy

$$\varphi_i(0) = \varphi_i(1) = 0, \quad \text{for} \quad i = 1, 2, \ldots, n.$$

The idea of the Rayleigh-Ritz method is to approximate $y(x)$ by $y^h(x)$ which is given by

$$y^h(x) = \sum_{i=1}^{n} c_i \varphi_i(x),$$

where $\{c_i\}_{i=1}^n$ are unknowns and then using $y^h(x)$ we minimize the integral.

From Eq. (3.32)

$$\begin{aligned} I[y^h] &= I\left[\sum_{i=1}^{n} c_i\varphi_i(x)\right] \\ &= \int_0^1 \left\{ p(x)\left[\sum_{i=1}^{n} c_i\varphi_i'(x)\right]^2 + q(x)\left[\sum_{i=1}^{n} c_i\varphi_i(x)\right]^2 - 2f(x)\left[\sum_{i=1}^{n} c_i\varphi_i(x)\right] \right\} dx. \end{aligned} \tag{3.34}$$

We know from Calculus that for a minimum to occur, we need that

$$\frac{\partial I}{\partial c_j} = 0 \quad \text{for} \quad j = 1, 2, \ldots, n.$$

Differentiating Eq. (3.34) we get

$$\frac{\partial I}{\partial c_j} = \int_0^1 \left\{ p(x)2\left[\sum_{i=1}^{n} c_i\varphi_i'(x)\right]\varphi_j'(x) + 2q(x)\left[\sum_{i=1}^{n} c_i\varphi_i(x)\right]\varphi_j(x) - 2f(x)\varphi_j(x) \right\} dx.$$

As $^{\partial I}/_{\partial c_j} = 0$ we get

$$0 = \sum_{i=1}^{n} \left[\int_0^1 2p(x)\varphi_i'(x)\varphi_j'(x) + 2q(x)\varphi_i(x)\varphi_j(x) \right] c_i dx - 2\int_0^1 f(x)\varphi_j(x)dx, \tag{3.35}$$

for each $j = 1, 2, \ldots, n$.

The equations defined in Eq. (3.35) are called as *normal equations* and produce a $n \times n$ linear system of the form

$$\mathbf{Ac} = \mathbf{b},$$

where

$$\mathbf{A}_{ij} = \int_0^1 \left[p(x)\varphi_i'(x)\varphi_j'(x) + q(x)\varphi_i(x)\varphi_j(x) \right] dx$$

and

$$b_i = \int_0^1 f(x)\varphi_i(x)dx.$$

We note that $\mathbf{A}$ is symmetric. Now the question remains: how to choose $\{\varphi_i(x)\}$?

3.4.1 Piecewise Linear Basis

The simplest choice of basis functions involves piecewise linear polynomials. We start by meshing the domain $[0, 1]$ into a non-equidistant mesh, i.e.,

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1.$$

Let $h_i = x_{i+1} - x_i$ for $i = 0, 1, \ldots, n$. We define the basis functions $\{\varphi_i\}_{i=1}^n$ as

$$\varphi_i(x) = \begin{cases} 0 & , \text{ if } \ 0 \le x \le x_{i-1} \\ \frac{1}{h_{i-1}}(x - x_{i-1}) & , \text{ if } \ x_{i-1} < x \le x_i \\ \frac{1}{h_i}(x_{i+1} - x) & , \text{ if } \ x_i < x \le x_{i+1} \\ 0 & , \text{ else.} \end{cases} \tag{3.36}$$

Functions $\varphi_i(x)$ are piecewise-linear (see Fig. 3.15), so the derivatives, while not continuous but are constant on (x_j, x_{j+1}) for each $j = 0, 1, \ldots, n$ and

$$\varphi_i'(x) = \begin{cases} 0 & , \text{ if } \ 0 \le x \le x_{i-1} \\ \frac{1}{h_{i-1}} & , \text{ if } \ x_{i-1} < x \le x_i \\ -\frac{1}{h_i} & , \text{ if } \ x_i < x \le x_{i+1} \\ 0 & , \text{ else,} \end{cases}$$

for $i = 1, 2, \ldots, n$.

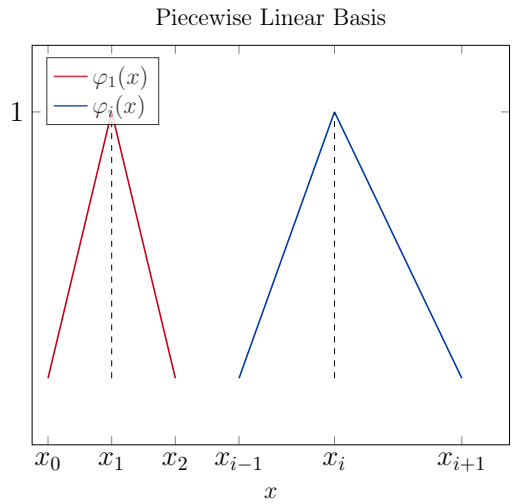

Figure 3.15: Piecewise linear basis functions

We note that $\text{supp}(\varphi_i) = \text{supp}(\varphi_i') = [x_{i-1}, x_{i+1}]$. Hence

$$\varphi_i(x)\varphi_j(x) = 0 \quad \text{and} \quad \varphi_i'(x)\varphi_j'(x) = 0,$$

for $j \neq i-1, i$, or $i+1$. Consequently, the linear system $\mathbf{Ac} = \mathbf{b}$ reduces to a tridiagonal linear

equation system. The non-zero entries are given by

$$\begin{aligned}
a_{ii} &= \int_0^1 \left\{ p(x)\left[\varphi_i'(x)\right]^2 + q(x)\left[\varphi_i(x)\right]^2 \right\} dx \\
&= \int_{x_{i-1}}^{x_i} \left\{ \frac{p(x)}{h_{i-1}^2} + \frac{q(x)}{h_{i-1}^2}(x - x_{i-1})^2 \right\} dx + \int_{x_i}^{x_{i+1}} \left\{ \frac{p(x)}{h_i^2} + \frac{q(x)}{h_i^2}(x_{i+1} - x)^2 \right\} dx \\
&= Q_{4,i} + Q_{2,i} + Q_{4,i+1} + Q_{3,i}, \qquad \text{for} \quad i = 1, 2, \ldots, n. \\
a_{i,i+1} &= \int_0^1 \left\{ p(x)\varphi_i'(x)\varphi_{i+1}'(x) + q(x)\varphi_i(x)\varphi_{i+1}(x) \right\} dx \\
&= \int_{x_i}^{x_{i+1}} -\frac{p(x)}{h_i^2} dx + \int_{x_i}^{x_{i+1}} \frac{q(x)}{h_i^2}(x_{i+1} - x)(x - x_i) dx \\
&= -Q_{4,i+1} + Q_{1,i}, \qquad \text{for} \quad i = 1, 2, \ldots, n-1. \\
a_{i,i-1} &= \int_0^1 \left\{ p(x)\varphi_i'(x)\varphi_{i-1}'(x) + q(x)\varphi_i(x)\varphi_{i-1}(x) \right\} dx \\
&= \int_{x_{i-1}}^{x_i} -\frac{p(x)}{h_{i-1}^2} dx + \int_{x_{i-1}}^{x_i} \frac{q(x)}{h_{i-1}^2}(x - x_{i-1})(x_i - x) dx \\
&= -Q_{4,i} + Q_{1,i-1}, \qquad \text{for} \quad i = 2, \ldots, n.
\end{aligned}$$

Finally, the entries of $\mathbf{b}$ are given by

$$\begin{aligned}
b_i &= \int_0^1 f(x)\varphi_i(x) dx \\
&= \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} \frac{f(x)}{h_{i-1}}(x - x_{i-1}) dx + \frac{1}{h_i} \int_{x_i}^{x_{i+1}} \frac{f(x)}{h_i}(x_{i+1} - x) dx \\
&= Q_{5,i} + Q_{6,i},
\end{aligned}$$

for $i = 1, 2, \ldots, n$.

We have six integrals to compute. Hence, if we have n internal points, we need to compute $6n$ integrals. We can use numerical quadrature to compute these solutions.

However, there is an alternative way in which we can approximate $p(x), q(x)$, and $r(x)$ by piecewise linear interpolation. Let $\mathbb{P}_q(x)$ be the approximation of $q(x)$, i.e.,

$$\mathbb{P}_q(x) = \sum_{i=0}^{n+1} q(x_i)\varphi_i(x),$$

where $\{\varphi_i(x)\}_{i=1}^n$ are defined as in Eq. (3.36) and

$$\varphi_0(x) = \begin{cases} \frac{x_1 - x}{x_1} & , \text{ if } \ 0 \le x \le x_1, \\ 0 & , \text{ else} \end{cases} \qquad \text{and} \quad \varphi_{n+1}(x) = \begin{cases} \frac{x - x_n}{1 - x_n} & , \text{ if } \ x_n \le x \le 1, \\ 0 & , \text{ else} \end{cases}$$

As the interval of integration is $[x_i, x_{i+1}]$, so the piecewise polynomial $\mathbb{P}_q(x)$ reduces to

$$\mathbb{P}_q(x) = q_i\varphi_i(x) + q_{i+1}\varphi_{i+1}(x).$$

Now, say we need to compute $Q_{i,1}$, then

$$\begin{aligned}
Q_{1,i} &= \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1}-x)(x-x_i)q(x)dx \\
&\approx \frac{1}{h_i^2}\int_{x_i}^{x_{i+1}} (x_{i+1}-x)(x-x_i)\mathbb{P}_q(x)dx \\
&= \frac{1}{h_i^2}\int_{x_i}^{x_{i+1}} (x_{i+1}-x)(x-x_i)\left(q_i\varphi_i(x)+q_{i+1}\varphi_{i+1}(x)\right)dx \\
&= \frac{q_i}{h_i^2}\int_{x_i}^{x_{i+1}} (x_{i+1}-x)(x-x_i)\frac{(x_{i+1}-x)}{h_i}dx + \frac{q_{i+1}}{h_i^2}\int_{x_i}^{x_{i+1}} (x_{i+1}-x)\frac{(x-x_i)^2}{h_i}dx \\
&= \frac{q_i}{h_i^3}\frac{h_i^4}{12} + \frac{q_{i+1}}{h_i^3}\frac{h_i^4}{12} \\
&= \frac{h_i}{12}\left(q_i+q_{i+1}\right).
\end{aligned}$$

Similarly, other integrals can be computed as

$$\begin{aligned}
Q_{2,i} &\approx \frac{h_{i-1}}{12}\left(3q_i+q_{i-1}\right), \\
Q_{3,i} &\approx \frac{h_i}{12}\left(q_i+q_{i+1}\right), \\
Q_{4,i} &\approx \frac{h_{i-1}}{2}\left(p_i+p_{i-1}\right), \\
Q_{5,i} &\approx \frac{h_{i-1}}{6}\left(2f_i+f_{i-1}\right), \\
Q_{6,i} &\approx \frac{h_i}{6}\left(2f_i+f_{i+1}\right).
\end{aligned}$$

After the computation of $Q_{k,i}$ for $k = 1, 2, \ldots, 6$, we can compute the matrix entries and the right-hand side to finally found the solution vector $\mathbf{c}$ and then

$$y^h = \sum_{i=1}^{n} c_i\varphi_i(x).$$

The algorithm for this method is similar to Algorithm 18 where we replace the matrix entries and right-hand side by matrix $\mathbf{A}$ and $\mathbf{b}$.

We can choose higher-order polynomials instead of piecewise Lagrange; this will result in better convergence, but the computation will increase. For a quadratic Lagrange polynomial we have $\text{supp}(\varphi_i) \subset [x_{i-2}, x_{i+2}]$. Hence, if instead of a tridiagonal system, we will have a penta-diagonal system as the node x_i will have contribution from nodes $x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}$.

The theory of solving BVP is numerically vast, and it includes Finite Element Methods, which are generally considered more general and sophisticated.

Chapter 4

System of Nonlinear Equations

We saw in Chapter 3 that we get a system of nonlinear equations while dealing with nonlinear differential equations. We looked at the Newton method in that chapter. Now, solving a system of nonlinear equations arises in other areas of mathematics and, more recently, in data science, where we need to minimize a system of nonlinear equations. In this chapter, we will delve deep into this area and discuss methods to solve nonlinear systems of equations.

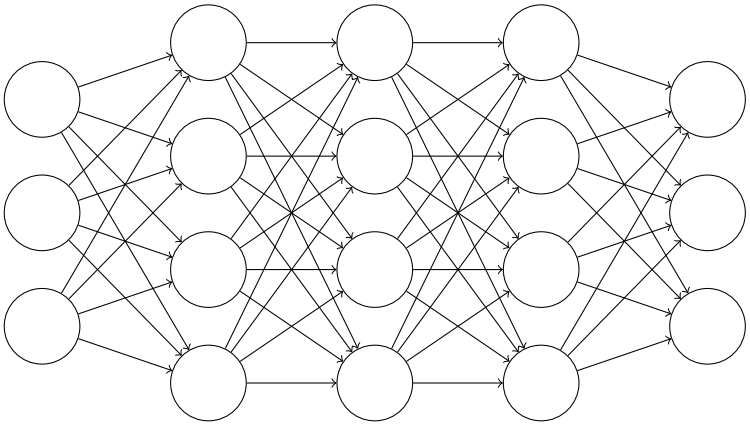

Figure 4.1: Neural Network Architecture

4.1 Fixed Point Iteration

A system of nonlinear equations has the form

$$\mathbf{F}(\mathbf{x}) = \mathbf{0},$$

where $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_n(\mathbf{x}) \end{bmatrix}^\top$ where $\mathbf{F}$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ and $f_i : \mathbb{R}^n \to \mathbb{R}$. The function $f_1, f_2, \ldots, f_n$ are called the *coordinate functions* of $\mathbf{F}$.

Before delving into the details of numerical methods, we start with some general results of multi-variable calculus.

Definition 4.1. Let f be a function defined on a set $\Omega \subset \mathbb{R}^n$ and maps into $\mathbb{R}$, i.e., $f : \Omega \to \mathbb{R}$. The function f is said to have the limit L at $\mathbf{x}_0$, written

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = L,$$

if given $\varepsilon > 0$ there exists a $\delta > 0$ such that $|f(\mathbf{x}) - L| < \varepsilon$ whenever $\mathbf{x} \in \Omega$ and $0 \leq \|\mathbf{x} - \mathbf{x}_0\|_{\ell^2} < \delta$.

Definition 4.2. Let f be a function from $\Omega \subset \mathbb{R}^n$ into $\mathbb{R}$. The function f is *continuous* at $\mathbf{x}_0 \in \Omega$ provided $\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x})$ exists and $\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$. Moreover, f is continuous on a set Ω if f is continuous at every point of Ω. This can be written as $f \in \mathcal{C}(\Omega)$.

We now define limit and continuity for a multivariable function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$.

Definition 4.3. Let $\mathbf{F}$ be a function from $\Omega \subset \mathbb{R}^n$ into $\mathbb{R}^n$ of the form

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_n(\mathbf{x}) \end{bmatrix}^\top,$$

where f_i is a mapping from $\mathbb{R}^n$ into $\mathbb{R}$ for each i. We say that $\mathbf{L}$ is the limit of $\mathbf{F}$, i.e.,

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{L} = \begin{bmatrix} L_1 & L_2 & \cdots & L_n \end{bmatrix}^\top,$$

if and only if $\lim_{\mathbf{x} \to \mathbf{x}_0} f_i(\mathbf{x}) = L_i$ for each $i = 1, 2, \ldots, n$.

Similarly, we can extend the definition of continuity. The function $\mathbf{F}$ is continuous at $\mathbf{x}_0 \in \Omega$ provided $\lim_{\mathbf{x} \to \mathbf{x}_0} \mathbf{F}(\mathbf{x})$ exists and $\lim_{\mathbf{x} \to \mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0)$. Similarly, $\mathbf{F}$ is continuous on Ω if $\mathbf{F}$ is continuous at each point $\mathbf{x}$ in Ω.

Solving a system of nonlinear equations extends the ideas of root-finding methods. Hence, we first talk about fixed points.

Definition 4.4. A function $\mathbf{G}$ from $\Omega \subset \mathbb{R}^n$ into $\mathbb{R}^n$ has a fixed point at $\mathbf{p} \in \Omega$ if $\mathbf{G}(\mathbf{p}) = \mathbf{p}$.

Next, we present a theorem about the existence of a fixed point.

Theorem 4.5. *Let* $\Omega = \left\{ \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top \mid a_i \le x_i \le b_i, \text{ for } i = 1, 2, \ldots, n \right\}$ *for some collection of constants* $\{a_i\}_{i=1}^n$ *and* $\{b_i\}_{i=1}^n$. *Suppose* $\mathbf{G}$ *is a continuous function from* $\Omega \subset \mathbb{R}^n$ *into* $\mathbb{R}^n$ *with the property that* $\mathbf{G}(\mathbf{x}) \in \Omega$ *for* $\mathbf{x} \in \Omega$. *Then,* $\mathbf{G}$ *has a fixed point in* Ω. *Moreover, suppose that all the coordinate functions of* $\mathbf{G}$ *has continuous partial derivatives and a constant* $\alpha < 1$ *exists with*

$$\left| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \right| \le \frac{\alpha}{n}, \quad \text{whenever } \mathbf{x} \in \Omega,$$

for each $i, j = 1, 2, \ldots, n$ *and each component function* g_i. *Then the sequence* $\{\mathbf{x}^{(k)}\}_{k=0}^\infty$ *defined by an arbitrary selected* $\mathbf{x}^{(0)} \in \Omega$ *and generated by*

$$\mathbf{x}^{(k)} = \mathbf{G}\left(\mathbf{x}^{(k-1)}\right), \quad \text{for } k \ge 1$$

converges to the unique fixed point $\mathbf{p} \in \Omega$ *and*

$$\|\mathbf{x}^{(k)} - \mathbf{p}\|_\infty \le \frac{\alpha^k}{1-\alpha} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty.$$

The above theorem is a generalization of the existence and uniqueness theorem for fixed points in one dimension. It says that if $\mathbf{G}$ is continuous on a bounded domain, then we have the existence of a fixed point. If, in addition, the partial derivative of $\{g_i\}_{i=1}^n$ is bounded above by one, then the fixed point is unique, and the error is bounded and converges to zero as $k \to \infty$.

Example 4.6. Consider the nonlinear system of equations

$$\begin{aligned} x_1^2 - 10x_1 + x_2^2 + 8 &= 0 \\ x_1x_2 + x_1 - 10x_2 + 8 &= 0. \end{aligned}$$

We write this in the form $\mathbf{G}(\mathbf{x})$ with

$$g_1(\mathbf{x}) = \frac{x_1^2 + x_2^2 + 8}{10} \quad \text{and} \quad g_2(\mathbf{x}) = \frac{x_1x_2^2 + x_1 + 8}{10}.$$

Now, we compute the partial derivatives

$$\frac{\partial g_1}{\partial x_1} = \frac{x_1}{5}, \quad \frac{\partial g_1}{\partial x_2} = \frac{x_2}{5}, \quad \frac{\partial g_2}{\partial x_1} = \frac{x_2^2 + 1}{10}, \quad \frac{\partial g_2}{\partial x_2} = \frac{x_1x_2}{5}.$$

In the domain $\Omega = \left\{ \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \mid \ 0 \le x_1, x_2 \le 1.5 \right\}$ we notice that

$$\max_{\mathbf{x}\in\Omega} \left| \frac{\partial g}{\partial x_1} \right| = 0.3, \quad \max_{\mathbf{x}\in\Omega} \left| \frac{\partial g_1}{\partial x_2} \right| = 0.3, \quad \max_{\mathbf{x}\in\Omega} \left| \frac{\partial g_2}{\partial x_1} \right| = 0.325, \quad \max_{\mathbf{x}\in\Omega} \left| \frac{\partial g_2}{\partial x_2} \right| = 0.45.$$

Hence for $\alpha = 0.95$ we get $\left| \frac{\partial g_i}{\partial x_j} \right| \le \frac{\alpha}{n}$, for $i, j = 1, 2$. Hence, $\mathbf{G}(\mathbf{x})$ has a unique fixed point.

The reason we talk about fixed points is that we can translate a root-finding problem into a fixed-point problem, i.e., if $\mathbf{p}$ is a zero of $\mathbf{F}(\mathbf{x})$ then $\mathbf{p}$ is a fixed-point of $\mathbf{G}(\mathbf{x}) = \mathbf{x} - \mathbf{F}(\mathbf{x})$ and vice-versa.

To use a fixed-point method we start with an initial iterate $\begin{bmatrix} x_1^{(0)} & x_2^{(0)} & \cdots & x_n^{(0)} \end{bmatrix}^\top$ and compute the new iterate $\mathbf{x}^{(1)} = \mathbf{G}(\mathbf{x}^{(0)})$. We continue this process until

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < \texttt{tol},$$

for some $\texttt{tol} \ll 1$.

Example 4.7. We consider the previous example and start with $\mathbf{x}^{(0)} = \mathbf{0}$, then

$$\mathbf{x}_1^{(1)} = g_1\left(\mathbf{x}^{(0)}\right) = \frac{\left[x_1^{(0)}\right]^2 + \left[x_2^{(0)}\right]^2 + 8}{10} = 0.8$$

$$\mathbf{x}_2^{(1)} = g_2\left(\mathbf{x}^{(0)}\right) = \frac{\left[x_1^{(0)}\right]\left[x_2^{(0)}\right]^2 + \left[x_1^{(0)}\right] + 8}{10} = 0.8.$$

We continue in this process to compute $\mathbf{x}^{(k)}$.

We can also accelerate this method if we use a Gauss-Seidel-like approach, i.e.,

$$\begin{aligned}
\mathbf{x}_1^{(k)} &= g_1\left(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}\right) \\
\mathbf{x}_2^{(k)} &= g_2\left(x_1^{(k)}, x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}\right) \\
\mathbf{x}_3^{(k)} &= g_3\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}\right) \\
&\vdots \\
\mathbf{x}_n^{(k)} &= g_n\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)} \ldots, x_{n-1}^{(k)}, x_n^{(k-1)}\right)
\end{aligned}$$

This modification usually accelerates the method, but it is not always true, as seen in Theorem 2.38. In Theorem 2.38, we noted that the convergence depends on the spectral radius of the iteration matrix. The nonlinear system will depend on the spectral radius of the Jacobian of $\mathbf{G}$.

The algorithm for the fixed-point method using Gauss-Seidel approximation is given in Algorithm 19

Algorithm 19 Fixed Point Method with Gauss-Seidel Acceleration

Given: System of equations $\mathbf{G}(\mathbf{x})$; number of equations n; initial approximation $\mathbf{x}^{(0)}$; maximum number of iterations $N_{\max}$; tolerance `tol` $\ll 1$.
Find: Solution $\mathbf{x}$ or Error Message.

Step 1: Initialization
Initialize $\mathbf{x}$ and $\mathbf{x}^{\text{old}}$ as a zero vector of size n

---

Step 2: Fixed Point Method
for $k = 1$ **to** $N_{\max}$ **do**
  $\mathbf{x}^{\text{old}} = \mathbf{x}^{(0)}$

  Step 2.1: Gauss-Seidel Update
  for $i = 1$ **to** n **do**
    $\mathbf{x}_i = g_i(\mathbf{x}^{(0)})$
    $\mathbf{x}_i^{(0)} = \mathbf{x}_i$
  end for

  Step 2.2: Check for Convergence
  if $\|\mathbf{x} - \mathbf{x}^{\text{old}}\|_\infty <$ `tol` **then**
    `break`
  else
    $\mathbf{x}^{(0)} = \mathbf{x}$
  end if
end for
if $k == N_{\max} + 1$ **then**
  Output("Maximum Number of Iterations Reached")
  exit()
end if

---

return $\mathbf{x}$

4.1.1 Convergence Analysis

Lastly, we want to recall the notion of convergence.

Definition 4.8. Suppose $\{\mathbf{x}^{(n)}\}_{n=0}^{\infty}$ is a sequence that converges to $\mathbf{x}$ with $\mathbf{x}^{(n)} \neq \mathbf{x}$ for all n. If $\lambda > 0$ and $\alpha > 0$ exist with

$$\lim_{n\to\infty} \frac{\|\mathbf{x}^{(n+1)} - \mathbf{x}\|}{\|\mathbf{x}^{(n)} - \mathbf{x}\|^{\alpha}} = \lambda,$$

then $\{\mathbf{x}^{(n)}\}$ converges to $\mathbf{x}$ of order α, with asymptotic error constant λ.

Theorem 4.9. *Let the conditions of Theorem 4.5 be satisfied. Then the sequence of approximations* $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ *obtained using the fixed-point iteration converges linearly to the solution* $\mathbf{x}$.

Proof. Let $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ be the sequence obtained using the fixed-point iterations. Then using the mean value theorem for several variables, we have

$$\begin{aligned}
\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty &= \|\mathbf{G}\left(\mathbf{x}^{(k-1)}\right) - \mathbf{G}(\mathbf{x})\|_\infty \\
&= \max_{1\le i\le n} |g_i\left(\mathbf{x}^{(k-1)}\right) - g_i(\mathbf{x})| \\
&= \max_{1\le i\le n} |\nabla g_i\left(\xi\right)\cdot(\mathbf{x}^{(k-1)} - \mathbf{x})| \\
&= \max_{1\le i\le n} \left|\sum_{j=1}^{n} \frac{\partial g_i(\xi)}{\partial x_j}\cdot(x_j^{(k-1)} - x_j)\right| \\
&\le \max_{1\le i\le n} \sum_{j=1}^{n} \left|\frac{\partial g_i(\xi)}{\partial x_j}\right| \cdot |x_j^{(k-1)} - x_j| \\
&\le \frac{\alpha}{n} \max_{1\le i\le n} \sum_{j=1}^{n} |x_j^{(k-1)} - x_j| \\
&\le \frac{\alpha}{n}\cdot n \cdot \max_{1\le j\le n} |x_j^{(k-1)} - x_j| \\
&= \alpha\|\mathbf{x}^{(k-1)} - \mathbf{x}\|_\infty,
\end{aligned}$$

where ξ lies between $\mathbf{x}^{(k-1)}$ and $\mathbf{x}$. Hence

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty}{\|\mathbf{x}^{(k-1)} - \mathbf{x}\|_\infty} \le \alpha.$$

Therefore the method is linearly convergent. □

4.2 Newton Method

For a fixed-point iteration to work on a nonlinear equation system, we need to transform it into an algebraic system like $\mathbf{x} = \mathbf{G}(\mathbf{x})$. This is not always feasible for say $\mathbf{F}(\mathbf{x}) = \begin{bmatrix}\sin(x_1) & \cos(x_2)\end{bmatrix}^\top$. Hence, we study more general methods like the Newton method.

Now, to develop the Newton method for $\mathbf{F}(\mathbf{x})$, we follow the same strategy as that for nonlinear equations, i.e., write

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - \mathbf{A}(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x}),$$

where

$$\mathbf{A}(x) = \begin{bmatrix} a_{11}(\mathbf{x}) & a_{12}(\mathbf{x}) & \cdots & a_{1n}(\mathbf{x}) \\ a_{21}(\mathbf{x}) & a_{22}(\mathbf{x}) & \cdots & a_{2n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(\mathbf{x}) & a_{n2}(\mathbf{x}) & \cdots & a_{nn}(\mathbf{x}) \end{bmatrix},$$

is non-singular at fixed-point $\mathbf{p}$.

For the derivation of the Newton method, we want the same property as that of the Newton-Rapshon method, i.e., quadratic convergence for finding zeros to $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

Theorem 4.10. *Let $\mathbf{p}$ be a solution of $\mathbf{G}(\mathbf{x}) = \mathbf{x}$. Suppose a number $\delta > 0$ exits with*

1. *$\frac{\partial g_i}{\partial x_j}$ is continuous on $N_\delta = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{p}\|_{\ell^2} < \delta\}$, for each $i, j = 1, 2, \ldots, n$.*
2. *$\frac{\partial^2 g_i}{\partial x_j \partial x_k}(\mathbf{x})$ is continuous and $\left|\frac{\partial^2 g_i}{\partial x_j \partial x_k}(\mathbf{x})\right| \leq M$ for some constant M, whenever $\mathbf{x} \in N_\delta$ for each $i, j, k = 1, 2, \ldots, n$.*
3. *$\frac{\partial g_i}{\partial x_k}(\mathbf{p}) = 0$ for $i, k = 1, 2, \ldots, n$.*

Then a number $\hat{\delta} \leq \delta$ exists such that the sequence generated by $\mathbf{x}^{(k)} = \mathbf{G}\left(\mathbf{x}^{(k-1)}\right)$ converges quadratically to $\mathbf{p}$ for any choice $\mathbf{x}^{(0)}$, provided that $\|\mathbf{x}^{(0)} - \mathbf{p}\|_{\ell^2} < \hat{\delta}$. Moreover,

$$\|\mathbf{x}^{(k)} - \mathbf{p}\|_\infty \leq \frac{n^2 M}{2} \|\mathbf{x}^{(k-1)} - \mathbf{p}\|_\infty^2,$$

for each $k \geq 1$.

The above theorem states that if for a fixed-point method, the first and second partial derivatives of $\mathbf{G}$ are continuous and also if the second partial derivative is bounded for every $i \in \{1, 2, \ldots, n\}$, and the first partial derivative is zero at the fixed point, then the fixed point method converges quadratically.

Now, the question boils down to the entries of $\mathbf{A}^{-1}(\mathbf{x})$. Let the entries of $\mathbf{A}^{-1}(\mathbf{x})$ be denoted by $b_{ij}(\mathbf{x})$. We rewrite the equation

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - \mathbf{A}(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x}),$$

as $g_i(\mathbf{x}) = x_i - \sum_{j=1}^n b_{ij}(\mathbf{x}) f_j(\mathbf{x})$. First, we compute the partial derivatives of $g_i(\mathbf{x})$ with respect to x_k.

$$\frac{\partial g_i(\mathbf{x})}{\partial x_k} = \begin{cases} -\sum_{j=1}^n \left[b_{ij}(\mathbf{x}) \frac{\partial f_j(\mathbf{x})}{\partial x_k} + f_j(\mathbf{x}) \frac{\partial b_{ij}(\mathbf{x})}{\partial x_k} \right] & \text{if } \; i \neq k \\ 1 - \sum_{j=1}^n \left[b_{ij}(\mathbf{x}) \frac{\partial f_j(\mathbf{x})}{\partial x_k} + f_j(\mathbf{x}) \frac{\partial b_{ij}(\mathbf{x})}{\partial x_k} \right] & \text{if } \; i = k \end{cases}$$

According to Theorem 4.10 we need $\frac{\partial g_i}{\partial x_k}(\mathbf{p}) = 0$ for $i, k = 1, 2, \ldots, n$. Hence

$$\begin{aligned} \sum_{j=1}^n b_{kj}(\mathbf{p}) \frac{\partial f_j(\mathbf{p})}{\partial x_k} &= 1 \quad \text{if} \quad i = k \\ \sum_{j=1}^n b_{kj}(\mathbf{p}) \frac{\partial f_j(\mathbf{p})}{\partial x_k} &= 0 \quad \text{if} \quad i \neq k \end{aligned}$$

The above two conditions imply that

$$\mathbf{A}(\mathbf{p})^{-1}\mathbf{J}(\mathbf{p}) = \mathbb{I}_{n \times n},$$

where $\mathbf{J}(\mathbf{p})$ is the Jacobian matrix of $\mathbf{F}(\mathbf{x})$ and $\mathbb{I}_{n\times n}$ is the identity matrix of size $n \times n$. So $\mathbf{A}(\mathbf{p}) = \mathbf{J}(\mathbf{p})$.

Hence, the Newton method is defined as

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - \mathbf{J}(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x}),$$

and the new iterate is computed using

$$\mathbf{x}^{(k)} = \mathbf{G}\left(\mathbf{x}^{(k-1)}\right) = \mathbf{x}^{(k-1)} - \mathbf{J}\left(\mathbf{x}^{(k-1)}\right)^{-1}\mathbf{F}\left(\mathbf{x}^{(k-1)}\right),$$

for $k \geq 1$ and some initial iterate $\mathbf{x}^{(0)}$.

Example 4.11. Consider the system of equation

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1^2 - 10x_1 + x_2^2 + 8 \\ x_1x_2^2 - 10x_2 + 8 \end{bmatrix}.$$

Now,

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \partial_1 f_1 & \partial_2 f_1 \\ \partial_1 f_2 & \partial_2 f_2 \end{bmatrix} = \begin{bmatrix} 2x_1 - 10 & 2x_2 \\ x_2^2 & 2x_1x_2 - 10 \end{bmatrix}.$$

Say $\mathbf{x}^{(0)} = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ then,

$$\mathbf{F}\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} 8 \\ 8 \end{bmatrix}, \quad \text{and} \quad \mathbf{J}\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} -10 & 0 \\ 0 & -10 \end{bmatrix}.$$

Now, $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \mathbf{J}\left(\mathbf{x}^{(0)}\right)^{-1}\mathbf{F}\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} 0.8 & 0.88 \end{bmatrix}^\top$.

We notice that in Theorem 4.10, the result holds if the initial approximation $\mathbf{x}^{(0)}$ is close to the fixed-point $\mathbf{p}$ and hence the Newton method depends heavily on the initial choice.

4.3 Steepest Descent Method

One of the disadvantages of Newton's method is that a good initial approximation needs to be known to ensure convergence. Steepest Descent methods are numerical methods that converge to the solution for any initial approximation. The method only converges linearly, whereas the Newton method converges quadratically. Hence, the solution obtained with the steepest descent can be used as an initial approximation for the Newton method.

The basic idea of the steepest descent is to find a local minimum for a multivariable function of the form $g : \mathbb{R}^n \to \mathbb{R}$. However, what is the relation between a minimization algorithm and root-finding methods? The problem $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ has a solution $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$ precisely when the function $g(x)$ defined by

$$g(\mathbf{x}) = \sum_{i=1}^{n} \left[f_i\left(\mathbf{x}\right)\right]^2$$

has the minimal value 0.

The intitution behind the steepest descent method can be described as

1. Evaluate g at an initial approximation $\mathbf{x}^{(0)} = \begin{bmatrix} x_1^{(0)} & x_2^{(0)} & \cdots & x_n^{(0)} \end{bmatrix}^\top$.
2. Determine a direction from $\mathbf{x}^{(0)}$ that results in a decrease in the value of g.
3. Move in that direction appropriately and call the new value $\mathbf{x}^{(1)}$.
4. Repeat the above three steps with $\mathbf{x}^{(0)}$ replaced by $\mathbf{x}^{(1)}$.

Now, the question remains in which direction we should move. For this, we first recall the idea of the gradient.

Definition 4.12. For $g : \mathbb{R}^n \to \mathbb{R}$ the *gradient* of g at $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ denoted by $\nabla(g(\mathbf{x}))$ is defined by

$$\nabla g(\mathbf{x}) = \begin{bmatrix} \frac{\partial g}{\partial x_1}(\mathbf{x}) & \frac{\partial g}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial g}{\partial x_n}(\mathbf{x}) \end{bmatrix}^\top.$$

The notion of gradient in multivariable calculus is the same as that of derivative in a single variable, i.e., it defines the direction of the steepest ascent.

We first define the directional derivative of a scalar field $g(\mathbf{x})$ in the direction of a vector $\mathbf{v} \in \mathbb{R}^n$:

$$\nabla_{\mathbf{v}} g(\mathbf{x}) = \lim_{h \to 0} \frac{g(\mathbf{x} + h\mathbf{v}) - g(\mathbf{x})}{h} = \mathbf{v}^\top \nabla g(\mathbf{x})$$

where $\mathbf{v}$ is a unit vector in $\mathbb{R}^n$, i.e., $\|\mathbf{v}\|_{\ell^2} = 1$. Now, $\nabla g(\mathbf{x}) \cdot \mathbf{v} = \|\nabla g(\mathbf{x})\| \|\mathbf{v}\| \cos(\theta) = \|\nabla g(\mathbf{x})\| \cos(\theta)$. Hence the value of $\nabla_{\mathbf{v}} g(\mathbf{x})$ will be maximized when $\cos(\theta) = 1$, i.e., when $\mathbf{v}$ is pointing in the same direction as that of $\nabla g(\mathbf{x})$, and minimized when $\mathbf{v} = -\nabla g(\mathbf{x})$, i.e., in the direction of maximum descent. The object in the modulus of $\nabla g(\mathbf{x})$ gives the rate of maximum change, while $\nabla g(\mathbf{x})$ gives the direction of maximum change.

To find the minimum value of $g(\mathbf{x})$, we go in the direction opposite to that of $\nabla g(\mathbf{x})$, i.e., update as:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha \nabla g\left(\mathbf{x}^{(0)}\right)$$

for some $\alpha > 0$. The choice of α plays an important role. If α is small, we reach the minimum, but the process is very slow. The process will be faster for a large α but can overshoot, leading to change in local minima. The optimal value of α is a very open area of research.

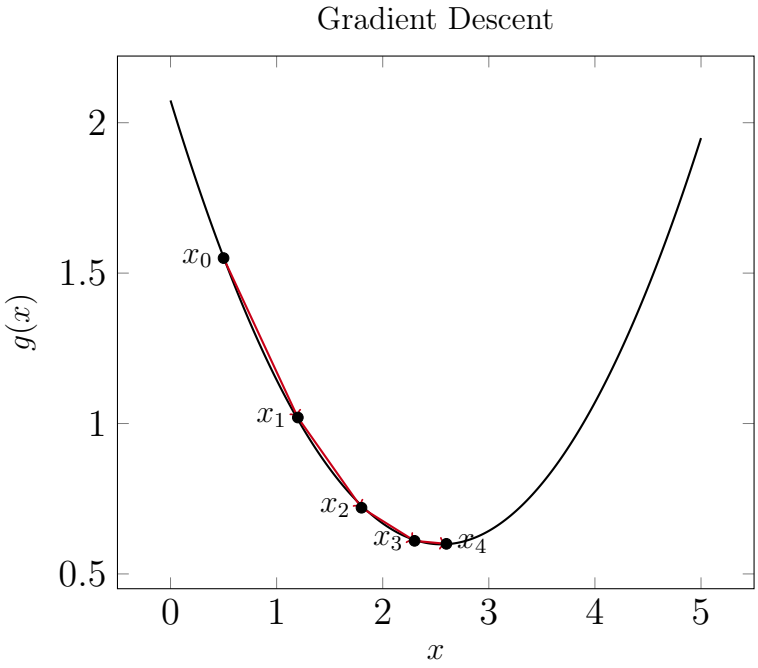

Figure 4.2: Minimization with Gradient Descent.

Chapter 5

Computing

The word **computing** has different meanings based on context and definition. According to Wikipedia:

> *Computing is any goal-oriented activity requiring, benefitting from, or creating computing machinery.*

This definition creates a recursive loop, as it uses "computing" to define itself. To break this loop, let us explore what a **computer** is. Wikipedia defines it as:

> *A computer is a machine that can be programmed to automatically carry out sequences of arithmetic or logical operations (computation).*

Here, two terms stand out: *arithmetic* and *logical*. These are fundamental concepts that mathematicians are familiar with. Thus, we have some basic understanding of computing.

In this course, we will not delve deeply into the workings of a computer. It is assumed that students are familiar with components like the keyboard, mouse (or trackpad), CPU, and monitor. For a refresher, see Figure 5.1.

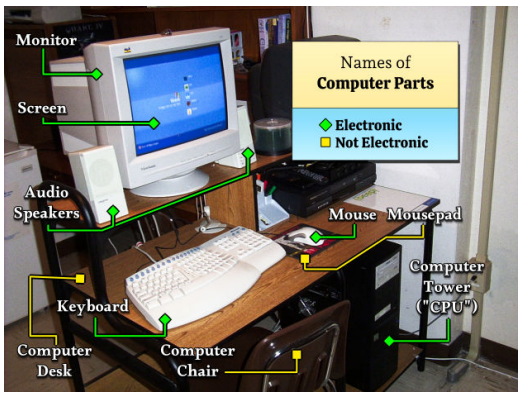

Figure 5.1: Parts of a computer from the early 2000s.

The primary aim of this course is to teach the fundamentals of programming using the Python language. This is not an *Introduction to Python* course. Instead, the focus is on how

to think about coding problems, understand common paradigms across languages, adopt good coding practices, maintain code, and debug effectively.

5.1 Good Practices in Coding

Coding is an art form, and like any art, its true audience is those who interact with it (in this case, the users of the code). A good codebase should be **well-documented**. Comments should explain the purpose of each function or variable. The beginning of the code should clearly state its objective.

5.1.1 Variable Initialization and Naming

Variables and functions should have **descriptive names**. For example, if a variable represents the number of oranges, naming it `n_oranges` is much clearer than simply using `n`. Additionally, variables should be initialized to prevent the use of garbage values.

Indentation of the conditional and iterative statements are important as it helps to differentiate different loops (or if-else statements).

In Python, indentation is mandatory, making this practice less error-prone. However, for languages like C++, proper formatting and indentation are crucial. Below we give an example of bad coding vs good coding in C++.

Bad Example:

```
int n_oranges;
for (int i = 0; i < 10; i++)
{
std::cout << i;
for(int j=0;j<2;j++
{
std::cout<<i+j;
}
std::cout<<n_oranges;
}
```

Good Example:

```
int n_oranges = 0;
for (int i = 0; i < 10; i++)
{
    std::cout << i;
    for (int j = 0; j < 2; j++)
    {
        std::cout << i + j;
    }
    std::cout<<n_oranges;
}
```

5.1.2 Reusability and Modularity

Code should be **reusable** and **modular** . For instance, consider a program that computes the Taylor series of a function. Instead of hardcoding the factorial computation in the main function, create a separate function for it. This is called *modularity*. This approach makes the code reusable. If another project requires the computation of ${}^{2}\mathrm{C}_{k}$, the factorial function can be reused without rewriting it.

5.2 Testing and Continuous Integration

Testing is a critical aspect of programming to ensure correctness and reliability. Continuous integration ensures that code changes do not break existing functionality.

After writing code, how do we know if it is correct? One effective approach is to verify the solution produced by the code against a pre-existing known solution. For example, if we write code to find the roots of a function, we can test its accuracy by using values with known solutions, such as $x^2 = 2$.

It is always advisable to run the code on multiple test cases to validate its correctness. Once the code is verified, we can create specific *test routines* to ensure its reliability in various scenarios.

5.3 Introduction to Computing Using Python

This course covers various aspects of computing, but we begin with the basics to build a strong foundation.

5.3.1 Variables

In Python, there are three commonly used variable types:

- `int`: Represents integers.
- `str`: Represents strings (text).
- `float`: Represents floating-point numbers (decimals).

Python does not require explicit variable declaration; you can assign values directly. For example:

```
n_oranges = 10  # An integer
price_oranges = 10.4  # A floating-point number
```

To define strings, use double quotes (`"`):

```
fruit = "oranges"  # A string
```

To check the type of a variable, use the `type()` function:

```
print(type(n_oranges))  # Output: <class 'int'>
```

Another important variable type is the **list**, which can contain multiple values of different types:

```
list_fruits = [n_oranges, price_oranges, fruit]  # A list with mixed
    types
```

While there are more variable types in Python, these four are essential for now.

Note: The `print()` function is used to display information. We will explore more advanced printing techniques later.

5.3.2 Arithmetic Operations

Arithmetic operations are fundamental in any programming language. Python provides the following operations:

Operation	Description	Example
$+$	Addition	$2 + 2 = 4$
$-$	Subtraction	$6 - 2 = 4$
$*$	Multiplication	$2 * 2 = 4$
$/$	Division	$2/2 = 1$
$**$	Exponentiation	$2 * *2 = 4$
$==$	Equality comparison	$2 == 2$
$\%$	Modulus (remainder)	$3\%2 = 1$

Table 5.1: Arithmetic Operations in Python

Additionally, the $!=$ operator means "not equal to," as in $3 \neq 2$.

These operations allow us to build more advanced functions and logic.

5.3.3 Compound Assignment

Python supports shorthand operations for self-assignment:

```
A = 10
A = A + 10  # Equivalent to:
A += 10
```

This shorthand applies to all arithmetic operations (`+=`, `-=`, `*=`, `/=`, etc.).

Note: When performing operations on variables of different types, such as `int` and `float`, Python automatically converts the result to `float`. For example:

```
result = 10 + 10.5  # result is a float (20.5)
```

However, operations combining `str` with `int` or `float` will result in errors:

```
"10" + 10  # This will raise a TypeError
```

Experiment with different cases to understand how Python handles these scenarios. Also, remember that Python follows the BODMAS convention.

5.3.4 Logical Operations

Another important class of operations is **logical operations**. These operations are used when you need to run specific parts of the code based on multiple conditions. For example:

- To check if a number is greater than 5 **and** divisible by 3.
- To check if a number is greater than 5 **or** divisible by 3.

In both cases, you have a logical expression to evaluate. Python provides three logical operators to handle such cases:

(*i*) **and**: Evaluates to `True` if both conditions A and B are true, otherwise `False`.

(*ii*) **or**: Evaluates to `True` if at least one of the conditions A or B is true, otherwise `False`.

(*iii*) **not**: Returns the negation of condition A.

The truth table for these logical operators is shown below:

A	**B**	**A and B**	**A or B**	**not A**
T	T	T	T	F
T	F	F	T	F
F	T	F	T	T
F	F	F	F	T

Table 5.2: Truth table for logical operators.

Logical operations can also involve more than two conditions. For example, suppose you have three conditions: **A**, **B**, and **C**. In such cases, you can group conditions using parentheses to control the order of evaluation. For instance:

- Check **(A and B)** first, and then combine the result with **C**.
- Evaluate **A or (B and C)** to prioritize **B and C**.

This flexibility allows for constructing complex logical expressions tailored to your requirements.

5.4 Conditional Statements

Conditional statements allow executing specific code blocks depending on conditions. In Python, the syntax is as follows:

```
if condition_1:
    execute_1
else:
    execute_2
```

Example: Checking if a number is even or odd:

```
eval_point = 5
if eval_point % 2 == 0:
    print(f"The number {eval_point} is even.")
else:
    print(f"The number {eval_point} is odd.")
```

For multiple conditions, we use `if-elif-else`:

```
eval_point = 5
if eval_point % 2 == 0:
    print(f"The number {eval_point} is divisible by 2.")
elif eval_point % 3 == 0:
    print(f"The number {eval_point} is divisible by 3.")
else:
    print(f"The number {eval_point} is not divisible by 2 or 3.")
```

Note: An `else` statement is usually not necessary for an `if` statement. Suppose we want to check if a number is even we can use

```
eval_point = 5
if eval_point % 2 == 0:
    print(f"The number {eval_point} is divisible by 2.")
```

Here we want to check if the number is even without checking if it odd or not.

Note: Here we have introduced a new way to print. The `print(f"...")` command is printing a formatted string. It prints the characters as well as the variable values defined in the curly braces $\{\cdot\}$.

5.5 Recursive Statements

Recursive Statements allow repetitive execution of code blocks.

5.5.1 For Loop

The syntax for a `for` loop in Python is:

```
for i in range(a, b):
    execute_1
```

Here:

- `i`: The loop iterator.
- `range(a, b)`: Specifies the range of values, starting at `a` and stopping before `b`, i.e., it goes over $\mathtt{a}, \mathtt{a}+1, \ldots, \mathtt{b}-1$.

Example: Summing numbers from 1 to 9:

```
total_sum = 0
for i in range(1, 10):
    print(i)
    total_sum += i
print(f"The summation of 9 points: {total_sum}")
```

5.5.2 Custom Step Size

The default step size of a `for` loop is one. If we want to use a custom step-size then we can use the following modification:

```
for i in range(a, b, step):
    execute_1
```

Example: Summing odd numbers from 1 to 9:

```
total_sum = 0
for i in range(1, 10, 2):
    print(i)
    total_sum += i
print(f"The summation of 10 points with step 2: {total_sum}")
```

5.5.3 Break and Continue

While using loops there can be cases when we want to exit the loop due to some condition. Also we can have cases when we want to skip some iteration. In this case we use `break` and `continue`, respectively.

- `break` : Exits the loop entirely.
- `continue` : Skips the current iteration and moves to the next.

Example: Adding even numbers up to 10 but stopping at 7:

```
total_sum = 0
for i in range(10):
    if i == 7:
        break
    if i % 2 == 1:
        continue
    print(i)
    total_sum += i
print(f"The summation of even numbers: {total_sum}")
```

5.5.4 Nested Loops

A `for` loop can be nested within another `for` loop. For example, to generate multiplication tables:

```
for i in range(1, 5):
    print(f"The table of {i}")
    for j in range(1, 11):
        print(f"{i} x {j} = {i * j}")
```

5.6 Functions

Until now, we have focused on sequential coding. However, to enhance reusability and maintainability, modular coding is essential. Functions enable modular programming by allowing code reuse. The syntax for creating a function is:

```
def function_name(input_1, input_2):
    # Function body
    result = some_operation(input_1, input_2)
    return result
```

Example: Function to check if a number is even:

```
def is_even(number):
        if number % 2 == 0:
                return True
        else:
                return False

value = 20
result = is_even(value)
print(f"The number {value} is even: {result}")
```

Note:

- A function can accept multiple inputs, a single input, or no input at all.
- A function may include multiple `return` statements or omit a `return` entirely, in which case it returns `None` by default.

Example: Function with multiple `return` statements:

```
def analyze_number(number):
    if number > 0:
        return "positive", number
    elif number < 0:
        return "negative", number
    else:
        return "zero", number

result_type, result_value = analyze_number(-5)
print(f"The number {result_value} is {result_type}.")
```

5.7 NumPy Library

In this section we provide an introduction to the NumPy library, a fundamental Python library for mathematical computations.

Importing the Library

To use NumPy in Python, the library must be imported. The standard convention is to import it with the alias `np`:

Import the NumPy library as follows: `import numpy as np`.

5.7.1 Arrays and Matrices

NumPy arrays are versatile tools used as vectors (one-dimensional) or matrices (two-dimensional). For instance:

- A one-dimensional array can be thought of as a row vector, e.g., $[1, 2, 3, 4]$. This kind of array can be created using `temp_array` = `np.array`$([1, 2, 3, 4])$.
- A two-dimensional matrix is defined by nesting arrays, e.g., $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$. This kind of array can be created using `temp_array` = `np.array`$([[1, 2, 3, 4], [5, 6, 7, 8]])$.

Key commands include:

- `np.size(array)`: Returns the total number of elements in an array.
- `np.shape(array)`: Provides the dimensions of an array.

Special Arrays: Zeros and Ones

It is common to initialize arrays with default values such as zeros or ones. This helps avoid uninitialized or garbage values in computations. For example:

- A zero matrix of size 10×10 can be created using `np.zeros((10, 10))`.
- Similarly, a ones matrix of the same size is created with `np.ones((10, 10))`.

Indexing in Arrays and Matrices

In NumPy:

- Indexing starts at 0.
- Negative indexing allows access to elements from the end, e.g., -1 refers to the last element.

For matrices, indexing uses row and column coordinates. For example, the element at row 0, column 0 in a matrix is accessed as [0][0]. Suppose we have the following $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} & \dots & a_{0\ n-1} \\ a_{10} & a_{11} & \dots & a_{1\ n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1\ 0} & a_{n-1\ 1} & \dots & a_{n-1\ n-1} \end{bmatrix},$$

then the $a_{n-1\ n-1}$ entry can be accessed using `A`$[-1][-1]$ as well as `A`$[$`n` $- 1][$`n` $- 1]$. Similarly, $a_{0\ \mathtt{n}-1}$ can be accessed using `A`$[0][-1]$ and `A`$[0][$`n` $- 1]$.

5.7.2 Linspace

To generate arrays with evenly spaced points, the `np.linspace` function is used:

> *For an array between a and b with n elements, use the syntax:*
> `np.linspace(a, b, n)`.

This is particularly useful for numerical methods. It it important to note that the end points a, b are included and the spacing between the points is $(b - a)/(n - 1)$.

5.7.3 Mathematical Functions

NumPy provides a wide range of mathematical functions, including:

- Trigonometric functions: `np.sin`, `np.cos`, etc.
- Hyperbolic functions: `np.sinh`, `np.cosh`, etc.
- Absolute value: `np.abs`.

For example, the sine and absolute value of $-\pi$ can be computed using `np.sin(-np.pi)` and `np.abs(-np.pi)`. For a comprehensive list of available mathematical routines, refer to the official documentation: `https://numpy.org/doc/stable/reference/routines.math.html`.

List of Algorithms

Index

Bibliography

[1] Robert G. Bartle and Donald R. Sherbert. *Introduction to real analysis*. Second. John Wiley & Sons, Inc., New York, 1992, pp. xii+404. ISBN: 0-471-51000-9.

[2] Jean-Paul Berrut and Lloyd N. Trefethen. "Barycentric Lagrange interpolation". In: *SIAM Rev.* 46.3 (2004), pp. 501–517. ISSN: 0036-1445,1095-7200. DOI: 10.1137/S0036144502417715. URL: https://doi.org/10.1137/S0036144502417715.

[3] J. Douglas Faires and Richard Burden. *Numerical methods*. Second. With 1 IBM-PC floppy disk (3.5 inch; HD). Brooks/Cole Publishing Co., Pacific Grove, CA, 1998, pp. xii+594. ISBN: 0-534-35187-5.

[4] L. R. Ford Jr. and D. R. Fulkerson. "Maximal flow through a network". In: *Canadian J. Math.* 8 (1956), pp. 399–404. ISSN: 0008-414X,1496-4279. DOI: 10.4153/CJM-1956-045-5. URL: https://doi.org/10.4153/CJM-1956-045-5.

[5] S.H. Friedberg, A.J. Insel, and L.E. Spence. *Linear Algebra*. Pearson Education, 2014. ISBN: 9780321998897. URL: https://books.google.co.in/books?id=KyB0DAAAQBAJ.

[6] Gilbert Strang. *Linear algebra and its applications*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1976, pp. xi+374.