Numerical Analysis

Aakash Jog

2015-16

Contents

---

1 Lecturer Information

Prof. Nir Sochen

Office: Schreiber 201
Telephone: +972 3-640-8044
E-mail: sochen@post.tau.ac.il
Office Hours: Sundays, 10:00–12:00

2 Required Reading

1. S. D. Conte and C. de Boor, Elementary Numerical Analysis, 1972

Part I
Representation of Numbers and Errors

1 Floating Point Representation

Exercise 1.
Represent 9.75 in base 2.

Solution 1.

$$\begin{aligned} 9.75 &= 8 + 1 + \frac{1}{2} + \frac{1}{4} \\ &= 2^3 + 2^0 + 2^{-1} + 2^{-2} \\ &= 2^3\left(2^0 + 2^{-3} + 2^{-4} + 2^{-5}\right) \\ &= \left(2^{11}\left(1 + 0.001 + 0.0001 + 0.00001\right)\right)_2 \\ &= \left(2^{11}\left(1.00111\right)\right)_2 \end{aligned}$$

Definition 1 (Double precision floating point representation)**.** A floating point representation which uses 64 bits for representation of a number is called a double precision floating point representation.
The standard form of double precision representation is

$$a = \underbrace{\pm}_{1\text{ bit}}\underbrace{1}_{1\text{ bit}}.\underbrace{\cdots}_{52\text{ bits}} \times w^{\underbrace{\pm}_{1\text{ bit}}\underbrace{\cdots}_{10\text{ bits}}}$$

Theorem 1 (Range of double precision floating point representation)**.** *The largest number which can be represented with double precision floating point representation is approximately* 10^{307} *and the smallest number which can be represented is approximately* 10^{-307}.

Proof. As the exponent has 10 bits for representation,

$$-\left(10^{10} - 1\right) \leq \text{exponent} \leq \left(10^{10} - 1\right)$$

Therefore,

$$-1023 \leq \text{exponent} \leq 1023$$

Therefore, the smallest number, in terms of absolute value, which can be represented, is

$$1.\underbrace{0\cdots0}_{52 \text{ bits}}\times 2^{-1024} \approx 10^{-307}$$

Therefore, the smallest number which can be represented is approximately 10^{-307}, and the largest number which can be represented is approximately 10^{307}. □

Definition 2 (Overflow)**.** If a result is larger than the largest number which can be represented, it is called overflow.

Definition 3 (Underflow)**.** If a result is smaller than the smallest number which can be represented, it is called underflow.

Definition 4 (Least significant digit)**.**

$$1 = 1.\underbrace{0\cdots0}_{52 \text{ zeros}}\times 2^{0}$$

Let 1_ε be the smallest number larger than 1, which can be represented in double precision floating point representation.
Therefore,

$$\begin{aligned}1 &= 1.\underbrace{0\cdots0}_{51 \text{ zeros}}1 \times 2^{0}\\ &= 1 + 2^{-52}\\ &\approx 1 + 2 \times 10^{-16}\end{aligned}$$

Therefore,

$$\begin{aligned}1 - 1_\varepsilon &= 2^{-52}\\ &\approx 2 \times 10^{-16}\end{aligned}$$

This number is called the least significant digit, or the machine precision. It is the maximum possible error in representation. It is represented by ε.

Definition 5 (Error)**.** Let the DPFP representation of a number x be $\widetilde{x}$. The absolute error in representation is defined as

$$\begin{aligned}\text{absolute error} &= |x - \widetilde{x}|\\ &= 0.0\cdots01 \times 2^{\text{exponent}}\end{aligned}$$

The relative error in representation is defined as

$$\begin{aligned}\delta &= \frac{|x-\widetilde{x}|}{x} \\ &= 0.0\cdots01 \\ &< \varepsilon\end{aligned}$$

The maximum error, $2^{-52} \approx 2 \times 10^{-16}$, is called the machine precision.
In general,

$$\widetilde{x}\,\widetilde{\star}\,\widetilde{y} = (x \star y)\,(1+\delta)$$

where δ is the relative error, ε is the machine precision, $\delta < \varepsilon$, and $\star$ is an operator.

1.1 Loss of Significant Digits in Addition and Subtraction

Exercise 2.
Represent $\pi + \frac{1}{30}$ in base 10 with 4 digits.

Solution 2.

$$\pi \approx 3.14159$$

Approximating by ignoring the last digits,

$$\widetilde{\pi} = 3.141$$

Similarly,

$$\widetilde{\frac{1}{30}} = 3.333 \times 10^{-2}$$

Therefore, adding,

$$\begin{aligned}\widetilde{\pi} + \widetilde{\frac{1}{30}} &= 3.141 + 0.03333 \\ &= 3.174\end{aligned}$$

Therefore,

$$\delta = \left| \frac{\left(\widetilde{\pi} + \widetilde{\frac{1}{30}}\right) - \left(\pi + \frac{1}{30}\right)}{\pi + \frac{1}{30}} \right|$$
$$= 0.0003$$

Therefore, $\delta < \varepsilon = 0.001$

Exercise 3.
Given

$$a = 1.435234$$
$$b = 1.429111$$

Find the relative error.

Solution 3.

$$a = 1.435234$$
$$b = 1.429111$$

Therefore,

$$a - b = 0.0061234$$

Approximating by ignoring the last digits,

$$\widetilde{a} = 1.435$$
$$\widetilde{b} = 1.429$$

Therefore,

$$\widetilde{a} - \widetilde{b} = 0.006$$

Therefore,

$$\delta = \left| \frac{(a-b) - \left(\widetilde{a} - \widetilde{b}\right)}{a - b} \right|$$

Therefore,

$$\delta > 10^{-3}$$
$$\therefore \delta > \varepsilon$$

Exercise 4.
Solve

$$x^2 + 10^8 x + 1 = 0$$

Solution 4.

$$x = \frac{-10^8 \pm \sqrt{10^{16} - 4}}{2}$$

Therefore,

$$x_- \approx -10^8$$

Therefore, by Vietta Rules,

$$x_1 x_2 = \frac{c}{a}$$
$$x_1 + x_2 = -\frac{b}{a}$$

Therefore,

$$x_+ x_- = 1$$
$$\therefore x_+ = \frac{1}{x_-}$$
$$\approx -10^{-8}$$

In MATLAB, this can be executed as x = **roots**([1,10^8,1])
This gives the result

$$x_+ = -7.45 \times 10^{-9}$$

Therefore, the absolute error is

$$|\tilde{x} - x| = \left| -7.45 \times 10^{-9} - \left(-10^{-8}\right) \right|$$
$$= 2.55 \times 10^{-9}$$

Therefore,

$$\delta = \left| \frac{\tilde{x} - x}{x} \right|$$
$$= \left| \frac{2.55 \times 10^{-9}}{10^{-8}} \right|$$
$$= 0.255$$
$$= 25\%$$

The algorithm used by MATLAB is

if $b \geq 0$ **then**

$x_1 = \frac{-b-\sqrt{b^2-4ac}}{2a}$

$x_2 = \frac{x}{ax_1}$

else

$x_2 = \frac{-b+\sqrt{b^2-4ac}}{2a}$

$x_1 = \frac{c}{ax_2}$

end if

This is done to avoid subtraction of numbers close to each other, and hence avoid the possible error.

Part II
Approximation of Functions

1 Series of Approximations

1.1 Order of Convergence

Definition 6. Let $\{\alpha_n\}_{n=1}^{\infty}$ be a series. $\{\alpha_n\}$ is said to converge to α, denoted as $\alpha_n \to \alpha$, if $\forall \varepsilon > 0$, $\varepsilon \in \mathbb{R}$, $\exists n_0(\varepsilon) \in \mathbb{N}$, such that $\forall n \in \mathbb{N}$, $n > n_0(\varepsilon)$, $|\alpha_n - \alpha| < \varepsilon$.

Usually, the series $\{\alpha_n\}$ is compared to a simpler series such as $\frac{1}{n}, \frac{1}{n^{\beta}}, \ldots$.

Definition 7. α_n is said to be "big-O" of β_n, and is said to behave like β_n, if $\exists k \in \mathbb{R}$, $k > 0$, $\exists n_0 \in \mathbb{N}$, $n_0 > 0$, such that $\forall n > n_0$,

$$|\alpha_n| \leq k|\beta_n|$$

It is denoted as

$$\alpha_n = \mathrm{O}(\beta_n)$$

Definition 8. α_n is said to be "small-O" of β_n if

$$\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} = 0$$

It is denoted as

$$\alpha_n = \mathrm{o}(\beta_n)$$

Exercise 5.
Find the order of convergence of

$$\alpha_n = 2n^3 + 3n^2 + 4n + 5$$

Solution 5.

$$\begin{aligned}\alpha_n &= 2n^3 + 3n^2 + 4n + 5 \\ &\leq (2+3+4+5)n^3 \\ \therefore \alpha_n &\leq 14n^3\end{aligned}$$

Therefore, comparing to the standard form,

$$\begin{aligned}k &= 14 \\ \beta_n &= n^3\end{aligned}$$

Therefore, as $\forall n \geq 1$, $|a_n| \leq 14|\beta_n|$,

$$\alpha_n = \mathrm{O}(\beta_n)$$

Also,

$$\begin{aligned}\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} &= \lim_{n\to\infty} \frac{2n^3 + 2n^2 + 4n + 5}{n^3} \\ &= 2\end{aligned}$$

Therefore, as the limits is not zero,

$$\alpha_n \neq \mathrm{o}(\beta_n)$$

However, $\forall \delta > 0$,

$$\alpha_n = \mathrm{o}\left(n^{3+\delta}\right)$$

2 Representation of Polynomials

2.1 Power series

Definition 9 (Power series representation of polynomials)**.**

$$P_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

This representation may lead to loss of significant digits.

Exercise 6.
Let $P(x)$ represent a straight line.

$$P(6000) = \frac{1}{3}$$
$$P(6001) = -\frac{2}{3}$$

If only 5 decimal digits are used, show that there is a loss of significant digits, if the power series representation of the polynomial is used.

Solution 6.
$P(x)$ represents a straight line. Therefore,

$$P(x) = ax + b$$

Therefore,

$$6000a + b = \frac{1}{3}$$
$$6001a + b = -\frac{2}{3}$$

Therefore,

$$\begin{pmatrix} 6000 & 1 \\ 6001 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$$
$$\therefore \begin{pmatrix} a \\ b \end{pmatrix}$$
$$= \frac{1}{|A|} \begin{pmatrix} 1 & -1 \\ -6001 & 6000 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$$
$$= -\begin{pmatrix} 1 \\ -6000.3 \end{pmatrix}$$
$$= \begin{pmatrix} -1 \\ 6000.3 \end{pmatrix}$$

Therefore,

$$a = -1$$
$$b = 6000.3$$

Therefore,

$$P(x) = -x + 6000.3$$

Substituting 6000 and 6001 in this expression,

$$P(6000) = 0.3$$
$$P(6001) = 0.7$$

However, the most accurate values of $P(6000)$ and $P(6001)$, using 5 decimal digits only, should be

$$P(6000) = 0.33333$$
$$P(6001) = -0.66666$$

Therefore, there is a loss of significant digits.

2.2 Shifted Power Series

Definition 10 (Shifted power series representation of polynomials)**.**

$$P_n(x) = a_0 + a_1(x - c) + \cdots + a_n(x - c)^n$$

This representation is a power series shifted by c. Hence, this representation does not lead to loss of significant digits.

Exercise 7.
Let $P(x)$ be a straight line.

$$P(6000) = \frac{1}{3}$$
$$P(6001) = -\frac{2}{3}$$

If only 5 decimal digits are used, show that there is no loss of significant digits, if the shifted power series representation of the polynomial is used, with $c = 6000$.

Solution 7.
$P(x)$ represents a straight line. Therefore,

$$P(x) = a(x - 6000) + b$$

Therefore,

$$b = \frac{1}{3}$$
$$a + b = -0.66666$$
$$\therefore a = -0.99999$$

Therefore,

$$P(x) = -0.99999(x - 6000) + 0.33333$$

Substituting 6000 and 6001 in this expression,

$$P(6000) = 0.33333$$
$$P(6001) = -0.66666$$

Therefore, there is no loss of significant digits, as the values of $P(6000)$ and $P(6001)$ are the most accurate values possible, using 5 decimal digits.

2.3 Newton's Form

Definition 11 (Newton's form of representation of polynomials)**.**

$$P_n(x) = a_0 + a_1(x - c_1) + \cdots + a_n(x - c_1) \ldots (x - c_n)$$

The number of multiplications needed to calculate $P_n(x)$ is

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

The number of additions or subtractions needed to calculate $P_n(x)$ is

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

Therefore, the total number of operations needed to calculate $P_n(x)$ is $\mathrm{O}(n^2)$.

2.4 Nested Newton's Form

Definition 12 (Nested Newton's form of representation of polynomials)**.**

$$P_n(x) = a_0 + (x - c_1)\Big(a_1 + (x - c_2)\left(a_2 + (x - c_3)\left(\ldots\right)\right)\Big)$$

The number of multiplications needed to calculate $P_n(x)$ is

$$\sum_{i=1}^{n} 1 = n$$

The number of additions or subtractions needed to calculate $P_n(x)$ is

$$\sum_{i=1}^{n} 2 = 2n$$

Therefore, the total number of operations needed to calculate $P_n(x)$ is big-O of $\mathrm{O}(n)$.

2.5 Properties of Polynomials

Theorem 2. *For a polynomial in shifted power series form,*

$$P_n(x) = P_n(c) + (x - c)q_{n-1}(x)$$

Proof.

$$\begin{aligned} P_n(x) &= a_0 + a_1(x - c) + \cdots + a_n(x - c)^n \\ &= a_0 + (x - c)\left(a_1 + a_2(x - 2) + \cdots + a_n(x - c)^{n-1}\right) \\ &= a_0 + (x - c)q_{n-1}(x) \\ &= P_n(c) + (x - c)q_{n-1}(x) \end{aligned}$$

□

Theorem 3. *If c is a root of $P_n(x)$, i.e., if*

$$P_n(c) = 0$$

then

$$P_n(x) = (x - c)q_{n-1}(x)$$

If $c_1 \neq c_2$ are roots of $P_n(x)$, then

$$P_n(x) = (x - c_1)(x - c_2)r_{n-2}(x)$$

Similarly, if $P_n(x)$ has n different roots, then

$$P_n(x) = A(x - c_1)\dots(x - c_n)$$

where $A \in \mathbb{R}$.
If $P_n(x)$ has $n + 1$ different roots, then

$$P_n(x) = A(x - c_1)\dots(x - c_n)(x - c_{n+1})$$

where $A = 0$.

Theorem 4. *If $p(x)$ and $q(x)$ are polynomials of degree at most n, that satisfy*

$$\begin{aligned} p(x_i) &= f(x_i) \\ q(x_i) &= f(x_i) \end{aligned}$$

for $i \in \{0, \dots, n\}$, then

$$p_n(x) \equiv q_n(x)$$

This means that there exists a unique polynomial with degree n which passes through $n + 1$ points, i.e. $n + 1$ points define a unique n degree polynomial.

Proof. Let

$$d_n(x) = p_n(x) - q_n(x)$$

Therefore, $d_n(x)$ is a polynomial of degree at most n, which has $n + 1$ roots. Therefore,

$$d_n(x) \equiv 0$$

Therefore,

$$p_n(x) \equiv q_n(x)$$

□

3 Interpolation

Theorem 5 (Weierstrass Approximation Theorem)**.** *Let $f(x) \in \mathrm{C}[a, b]$, i.e. it is continuous on $[a, b]$. Let $\varepsilon > 0$. Then there exists a polynomial $P(x)$ defined on $[a, b]$, such that $\forall x \in [a, b]$,*

$$|f(x) - P(x)| < \varepsilon$$

Definition 13 (Interpolating polynomial)**.** $p(x)$ is said to be the interpolating polynomial of $f(x)$, if for all sample points x_i,

$$f(x_i) = p(x_i)$$

Theorem 6. *Let $f(x)$ such that $\forall i \in \{0, \dots, n\}$,*

$$f(x_i) = y_i$$

Then, there exists a unique polynomial $p(x)$ of degree at most n, which interpolates $f(x)$ at all sample points x_i.

3.1 Direct Method

Definition 14 (Van der Monde matrix)**.** Let

$$p(x) = \sum_{i=0}^{n} a_i x^i$$

Let

$$f(x_i) = y_i$$

Therefore, as

$$p(x_i) = f(x_i)$$

the constraints are

$$\begin{aligned} a_0 + a_1 x_0 + \cdots + a_n {x_0}^n &= y_0 \\ a_1 + a_1 x_1 + \cdots + a_n {x_1}^n &= y_1 \\ &\vdots \\ a_n + a_1 x_n + \cdots + a_n {x_n}^n &= y_n \end{aligned}$$

Therefore,

$$\begin{pmatrix} 1 & x_0 & {x_0}^2 & \dots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & \dots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & \dots & {x_2}^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & {x_n}^2 & \dots & {x_n}^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

The matrix

$$V = \begin{pmatrix} 1 & x_0 & {x_0}^2 & \dots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & \dots & {x_1}^n \\ 1 & x_2 & {x_2}^2 & \dots & {x_2}^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & {x_n}^2 & \dots & {x_n}^n \end{pmatrix}$$

is called the Van der Monde matrix.

Theorem 7. *The Van der Monde matrix is invertible, and hence there exists a unique matrix of coefficients* $a_0, \dots, a_n$*, and hence the interpolating polynomial* $p(x)$ *is unique.*

3.2 Lagrange's Interpolation

Definition 15 (Lagrange polynomials)**.** Let

$$L_k(x) = \prod_{i=0; i \neq k}^{n} (x - x_i)$$

Therefore,

$$L_k(x_i) = \begin{cases} 0 & ; \quad i \neq k \\ 1 & ; \quad i = k \end{cases}$$

Let

$$l_k(x) = \frac{L_k(x)}{L_k(x_k)}$$

Therefore,

$$l_k(x_i) = \begin{cases} 0 & ; \quad i \neq k \\ 1 & ; \quad i = k \end{cases}$$

The polynomials $l_i(x)$ are called Lagrange polynomials.

Theorem 8. *Let*

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x)$$

where $l_i(x)$ are Lagrange polynomials.
Then, $p_n(x)$ is the interpolating polynomial of $f(x)$.

Exercise 8.
Which polynomial of degree 2 interpolates the below data?

x	$f(x)$
1	1
2	3
3	7

Solution 8.

$$L_k(x) = \prod_{i=0; i \neq k}^{n} (x - x_i)$$

Therefore,

$$L_1(x) = (x-2)(x-3)$$
$$L_2(x) = (x-1)(x-3)$$
$$L_3(x) = (x-1)(x-2)$$

Therefore,

$$\begin{aligned}
L_1(1) &= (1-2)(1-3) \\
&= 2 \\
L_2(2) &= (2-1)(2-3) \\
&= -1 \\
L_3(3) &= (3-1)(3-2) \\
&= 2
\end{aligned}$$

Therefore,

$$l_k(x) = \frac{L_k(x)}{L_k(x_k)}$$

Therefore,

$$\begin{aligned}
l_1(x) &= \frac{L_1(x)}{L_1(1)} \\
&= \frac{1}{2}(x-2)(x-3) \\
l_2(x) &= \frac{L_2(x)}{L_2(1)} \\
&= -(x-1)(x-3) \\
l_3(x) &= \frac{L_3(x)}{L_3(1)} \\
&= \frac{1}{2}(x-1)(x-2)
\end{aligned}$$

Therefore,

$$\begin{aligned}
p_2(x) &= \sum f(x_i) l_i(x) \\
&= \frac{1}{2}(x-2)(x-3) - 3(x-1)(x-3) + \frac{7}{2}(x-1)(x-2)
\end{aligned}$$

Exercise 9.
Given

$$k(z) = \int_0^{\frac{\pi}{2}} \frac{\mathrm{d}x}{\sqrt{1-(\sin z)^2(\sin x)^2}}$$

and

$$k(1) = 1.5709$$
$$k(4) = 1.5727$$
$$k(6) = 1.5751$$

approximate $k(3.5)$.

Solution 9.

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

Therefore,

$$l_1(x) = \frac{(x-4)(x-6)}{(1-4)(1-6)}$$
$$l_4(x) = \frac{(x-1)(x-6)}{(4-1)(4-6)}$$
$$l_6(x) = \frac{(x-1)(x-4)}{(6-1)(6-4)}$$

Therefore,

$$\begin{aligned} l_1(3.5) &= \frac{(3.5-4)(3.5-6)}{(1-4)(1-6)} \\ &= 0.08333 \\ l_4(3.5) &= \frac{(3.5-1)(3.5-6)}{(4-1)(4-6)} \\ &= 1.04167 \\ l_6(3.5) &= \frac{(3.5-1)(3.5-4)}{(6-1)(6-4)} \\ &= -0.125 \end{aligned}$$

Therefore,

$$\begin{aligned} p_2(x) &= \sum f(x_i) l_k(x) \\ \therefore p_2(3.5) &= \sum f(x_1) l_k(3.5) \\ &= (1.5709)(0.08333) + (1.5727)(1.04167) + (1.5751)(-0.125) \\ &= 1.57225 \end{aligned}$$

3.3 Hermite Polynomials

Definition 16. Let the given data be of the form $(x_i, f(x_i), f'(x_i))$, where $i = 0, \ldots, n$.
H_{2n+1} is called the Hermite polynomial of $f(x)$.
For H_{2n+1} to be the interpolation polynomial of $f(x)$, the constraints are

$$\begin{aligned} H_{2n+1}(x_i) &= f(x_i) \\ H'_{2n+1}(x_i) &= f'(x_i) \end{aligned}$$

Therefore, the number of constraints are $2n + 2$.
Hence, the polynomial is of degree at most $2n + 1$.

Theorem 9. *Let*

$$H_{2n+1}(x) = \sum_{i=0}^{n} f(x_i)\psi_{n,i}(x) + \sum_{i=0}^{n} f'(x_i)\varphi_{n,i}(x)$$

Let

$$\delta_{ij} = \begin{cases} 0 & ; \quad i \neq j \\ 1 & ; \quad i = j \end{cases}$$

If the polynomials ψ and φ satisfy

$$\begin{aligned} \psi_{n,i}(x_j) &= \delta_{ij} \\ {\psi_{n,i}}'(x_j) &= 0 \\ \varphi_{n,i}(x_j) &= 0 \\ {\varphi'_{n,i}}'(x_j) &= \delta_{ij} \end{aligned}$$

then the polynomial H_{2n+1} is the interpolation polynomial of $f(x)$.

3.4 Newton's Interpolation

Definition 17 (Newton's polynomial)**.** The polynomial

$$p_n(x) = \sum_{i=0}^{n} A_i \prod_{j=0}^{i-1} (x - x_j)$$

is called Newton's polynomial.

Theorem 10. *If $p_k(x)$, constructed based on $x_1, \dots, x_k$ is known, then $p_{k+1}(x)$, based on $x_1, \dots, x_{k+1}$ can be constructed as*

$$p_{k+1}(x) = p_k(x) + A_{k+1}(x - x_0) \dots (x - x_k)$$

Proof. For $i = 0, \dots, k$,

$$\begin{aligned} p_{k+1}(x_i) &= p_k(x_i) + A_{k+1} \prod_{j=0}^{k} (x_i - x_j) \\ &= p_k(x_i) + 0 \end{aligned}$$

$\forall i = 0, \dots, k$, $(x_i - x_i) = 0$. Therefore, if $i = j$, $(x_i - x_j) = 0$. Therefore, $\prod (x_i - x_j) = 0$

For $i = k + 1$,

$$\begin{aligned} p_{k+1}(x_{k+1}) &= p_k(x_{k+1}) + A_{k+1} \prod_{j=0}^{k} (x_{k+1} - x_j) \\ &= f(x_{k+1}) \end{aligned}$$

where A_{k+1} can be calculated using $p_k(x_{k+1})$ and $f(x_{k+1})$.
Therefore,
For $n = 1$,

$$\begin{aligned} p_0(x) &= A_0 \\ &= f(x_0) \end{aligned}$$

For $n = 2$,

$$\begin{aligned} p_1(x) &= p_0(x) + A_1(x - x_0) \\ &= f(x_0) - A_1(x - x_0) \\ &= f(x_1) \end{aligned}$$

Therefore,

$$\begin{aligned} A_1 &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\ &= f[x_0, x_1] \end{aligned}$$

For $n = 3$,

$$\begin{aligned} p_2(x) &= p_1(x) + A_2(x - x_0)(x - x_1) \\ &= f(x_0) + f[x_0, x_1](x - x_0) \\ &= f(x_0) + f[x_0, x_1](x - x_0) + A_2(x - x_0)(x - x_1) \\ &= f(x_2) \end{aligned}$$

Therefore,

$$
\begin{aligned}
A_2 &= \frac{1}{(x_2 - x_0)(x_2 - x_1)} \left(f(x_2) - f(x_0) - f[x_0, x_1](x_2 - x_0)\right) \\
&= f[x_0, x_1, x_2]
\end{aligned}
$$

and so on.
In general,

$$A_k = f[x_0, \dots, x_k]$$

□

Definition 18 (Divided difference)**.**

$$
\begin{aligned}
f[x_0, \dots, x_k] &= \frac{f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}]}{x_k - x_0} \\
f[x_0] &= f(x_0)
\end{aligned}
$$

is called the kth order divided difference of $f(x)$.

Exercise 10.
Given

$$k(z) = \int_0^{\frac{\pi}{2}} \frac{\mathrm{d}x}{\sqrt{1 - (\sin z)^2 (\sin x)^2}}$$

and

$$
\begin{aligned}
k(1) &= 1.5709 \\
k(4) &= 1.5727 \\
k(6) &= 1.5751
\end{aligned}
$$

approximate $k(3.5)$.

Solution 10.
For the first order divided differences,

$$k[x_i] = k(x_i)$$

Therefore,

$$\begin{aligned} k[1] &= k(1) \\ &= 1.5709 \\ k[4] &= k(4) \\ &= 1.5727 \\ k[6] &= k(6) \\ &= 1.5751 \end{aligned}$$

For the second order divided differences,

$$k[x_i, x_j] = \frac{k[i] - k[j]}{i - j}$$

Therefore,

$$\begin{aligned} k[1,4] &= \frac{k[1] - k[4]}{1 - 4} \\ &= \frac{1.5727 - 1.5709}{3} \\ k[4,6] &= \frac{k[4] - k[6]}{4 - 6} \\ &= \frac{1.5751 - 1.5727}{2} \end{aligned}$$

For the third order divided differences,

$$k[x_i, x_j, x_k] = \frac{k[i,j] - k[j,k]}{i - k}$$

Therefore,

$$k[1,4,6] = \frac{k[1,4] - k[4,6]}{1 - 6}$$

Hence,

$$\begin{aligned} A_0 &= k[1] \\ A_1 &= k[1,4] \\ A_2 &= k[1,4,6] \end{aligned}$$

4 Error in Interpolation

Definition 19 (Error in interpolation)**.** The error in interpolation is defined to be

$$e(x) = f(x) - p_k(x)$$

Theorem 11.

$$e(x) = f[x_0, \ldots, x_k, x] \prod_{i=0}^{k} (x - x_i)$$

Theorem 12 (Rolle's Theorem)**.** *Let f be continuous on $[a, b]$, with a continuous derivative on (a, b), and $f(a) = f(b) = 0$. Then, $\exists \varepsilon \in (a, b)$, such that*

$$f'(\varepsilon) = 0$$

Theorem 13 (Lagrange's Mean Value Theorem)**.** *Let f be continuous on $[a, b]$, with a continuous derivative on (a, b). Then, $\exists \varepsilon \in (a, b)$, such that*

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

This theorem is a general case of Lagrange's Mean Value Theorem.

Theorem 14. *Let f be continuous on $[a, b]$ with k continuous derivatives on (a, b). Then, $\exists \varepsilon \in (a, b)$, such that*

$$f[x_0, \ldots, x_k] = \frac{f^{(k)}(\varepsilon)}{k!}$$

Theorem 15. *Let f be continuous on $[a, b]$ with n continuous derivatives on (a, b), not necessarily distinct. Then, the interpolation polynomial is*

$$p_n(x) = \sum_{i=0}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j)$$

Theorem 16. *Let f be continuous on $[a, b]$ with k continuous derivatives on (a, b), not necessarily distinct.*
If

$$\left| \frac{f^{(k+1)}(\varepsilon)}{(k+1)!} \right| \leq M$$

then, for $\forall \varepsilon \in [x_0, x_k]$,

$$|e(x)| \leq \left| \frac{f^{(k+1)}(\varepsilon)}{(k+1)!} \prod_{i=0}^{k} (x - x_i) \right|$$

4.1 Minimizing the Maximum Error

Theorem 17. *The minimum error in interpolation is given by*

$$\min_{0 \le x_0 \le \cdots \le x_k} \left(\max \left| \prod_{i=0}^{k} (x - x_i) \right| \right) = \min_{0 \le x_0 \le \cdots \le x_k} \Big(\max |p_{k+1}(x)| \Big)$$

Definition 20 (Chebyshev polynomial)**.** The Chebyshev polynomial is defined as

$$T_n(x) = \cos(n \cos^{-1} x)$$

Theorem 18. *If $x = \cos\theta$,*

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ &\vdots \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) \end{aligned}$$

And hence,

$$T_n(x) = \prod_{i=0}^{n-1} (x - x_i)$$

where

$$x_i = \cos\left(\frac{(2i+1)\pi}{2n}\right)$$

$\forall i \in \{0, \ldots, n-1\}$.

Part III

Solutions of Equations

1 Solving Non-linear Equations

1.1 Bisection Method

Algorithm 1 Bisection Method

Let f be continuous on $[a, b]$, such that $f(a)f(b) < 0$.
$m \leftarrow \frac{a_n+b_n}{2}$
if $f(a_n)f(m) < 0$ **then**
    $a_{n+1} \leftarrow a_n$
    $b_{n+1} \leftarrow m$
    $r_n \leftarrow b_{n+1}$
else
    $a_{n+1} \leftarrow m$
    $b_{n+1} \leftarrow a_n$
    $r_n \leftarrow a_{n+1}$
end if
$r \leftarrow \lim_{n\to\infty} r_n$
r is a root of the equation $f(x) = 0$

Theorem 19. *Let f be continuous on $[a, b]$, such that $f(a)f(b) < 0$, where $\{r_n\}$ are generated by the bisection algorithm. Then*

$$\lim_{n\to\infty} r_n = r$$

such that $f(r) = 0$, and

$$|r_n - r| < \frac{b-a}{2^n}$$

where $n \in \mathbb{N}$.

1.2 Regula Falsi

Algorithm 2 Regula Falsi Method

Let f be continuous on $[a,b]$, such that $f(a)f(b) < 0$.
if $f(a_n)f(x_n) < 0$ **then**
    $b_{n+1} \leftarrow x_n$
else
    $a_{n+1} \leftarrow x_n$
end if
Solve $p_1(x) = f(a_n) + f[a_n, b_n](x - a_n)$ for x_n
$x_n \leftarrow \frac{f(b_n)a_n - f(a_n)b_n}{f(b_n) - f(a_n)}$
$r \leftarrow \lim\limits_{n\to\infty} r_n$

2 Newton-Raphson Method

Algorithm 3 Newton-Raphson Method

Choose $x_0 \in \mathbb{R}$ to be the first approximation of $f(x)$.
$x_{n+1} \leftarrow x_n - \frac{f(x_n)}{f'(x_n)}$

Exercise 11.
Solve

$$x = a^{\frac{1}{m}}$$

using Newton-Raphson method, and hence find $\sqrt{2}$.

Solution 11.

$$\begin{aligned} x &= a^{\frac{1}{m}} \\ \therefore x^m &= a \end{aligned}$$

Therefore, let

$$f(x) = x^m - a$$

Therefore, the solution to the equation is the solution to

$$f(x) = 0$$

Therefore,

$$f(x) = x^m - a$$
$$\therefore f'(x) = mx^{m-1}$$

Therefore,

$$\begin{aligned}
x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\
&= \frac{{x_n}^m - a}{m{x_n}^{m-1}} \\
&= \frac{m{x_n}^m - {x_n}^m + a}{m{x_n}^{m-1}} \\
&= \frac{1}{m}\left(\frac{a}{{x_n}^{m-1}} + (m-1)x_n\right)
\end{aligned}$$

Therefore, if $m = 2$,

$$x_{n+1} = \frac{1}{2}\left(\frac{a}{x_n} + x_n\right)$$

Therefore, if $a = 2$,

$$x_{n+1} = \frac{1}{2}\left(\frac{2}{x_n} + x_n\right)$$

Therefore, let

$$x_0 = 2$$

Therefore,

$$x_1 = 1.5$$
$$x_2 = 1.41666$$
$$x_3 = 1.414215685$$

2.1 Fixed Point Iterations

Definition 21 (Fixed point)**.** A fixed point of a function $g(x)$ is a point which satisfies

$$x = g(x)$$

Theorem 20 (Fixed point theorem)**.** *Let g be a continuous function in $[a, b]$ such that*

1. $\forall x \in [a, b]$*,* $g(x) \in [a, b]$*.*
2. $g'(x)$ *exists and* $\forall x \in [a, b]$*,* $|g'(x)| < 1$*, or* $g(x)$ *is Lipschitz, i.e.* $|g(x) - g(y)| \leq k|x - y|$*.*

then,

1. $\exists! \xi$*, such that* $\xi \in [a, b]$ *is a fixed point of* $g(x)$*.*
2. $\forall x \in [a, b]$*, the series* $x_{n+1} = g(x_n)$ *converges to* ξ*.*

2.2 Secant Method

Algorithm 4 Secant Method

1: Choose $x_0 \in \mathbb{R}$ to be the first approximation of $f(x)$.
2: $x_{n+1} \leftarrow x_n - \frac{f(x_n)}{f[x_{n-1}, x_n]}$

3 Rate of Convergence

Definition 22 (Rate of convergence)**.** Let the series x_n converge to ξ. If

$$\lim_{n \to \infty} \frac{|e_{n+1}|^2}{|e_n|^p} = c$$

where $c \neq 0 \in \mathbb{R}$. Then, p is the rate of convergence. The rate of convergence is said to be linear if $p = 1$, and quadratic if $p = 2$.

3.1 Newton's Method

Theorem 21. *The rate of convergence of Newton's method is 2.*

Proof. Let ξ be the root of $f(\xi)$.
Using the Taylor Series,

$$\begin{aligned} 0 &= f(\xi) \\ &= f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2} f''(\eta)(\xi - x_n)^2 + \ldots \end{aligned}$$

where $\eta \in [x_n, \xi]$.
Let $f(x)$ be continuous with a continuous derivative, such that $f'(\xi) \neq 0$.
Therefore $f'(x_n) \neq 0$, for $x_n \approx \xi$.
Therefore,

$$\begin{aligned}
0 &= f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2}f''(\eta)(\xi - x_n)^2 \\
\therefore -f(x_n) &= f'(x_n)(\xi - x_n) + \frac{1}{2}f''(\eta)(\xi - x_n)^2 + \dots \\
\therefore -\frac{f(x_n)}{f'(x_n)} &= (\xi - x_n) + \frac{1}{2}\frac{f''(\eta)}{f'(x_n)}(\xi - x_n) \\
\therefore \xi - \left(x_n - \frac{f(x_n)}{f'(x_n)}\right) &= -\frac{1}{2}\frac{f''(\eta)}{f'(x_n)}(\xi - x_n)^2 \\
\therefore \xi - x_{n+1} &= -\frac{1}{2}\frac{f''(\eta)}{f'(x_n)}(\xi - x_n)^2 \\
\therefore e_{n+1} &= -\frac{1}{2}\frac{f''(\eta)}{f'(x_n)}{e_n}^2 \\
\therefore \frac{e_{n+1}}{{e_n}^2} &= \frac{1}{2}\frac{f''(\eta)}{f'(x_n)}
\end{aligned}$$

Therefore, assuming $f''(\xi) \neq 0$,

$$\begin{aligned}
\therefore \lim_{n\to\infty}\left|\frac{e_{n+1}}{{e_n}^2}\right| &= \lim_{n\to\infty}\left|\frac{f''(\eta)}{2f'(x_n)}\right| \\
&= \frac{f''(\xi)}{2f'(\xi)} \\
&= c \\
&\neq 0
\end{aligned}$$

Therefore the rate of convergence of Newton's Method is 2. □

3.2 Fixed Point Iterations

Theorem 22. *The rate of convergence of fixed point iterations is 1.*

Proof.

$$\begin{aligned}
\xi &= g(\xi) \\
&= g(x_n) + g'(\eta)(\xi - x_n) \\
\therefore \xi - g(x_n) &= g'(\eta)(\xi - x_n) \\
\therefore \xi - x_{n+1} &= g'(\eta)(\xi - x_n) \\
\therefore e_{n+1} &= g'(\eta)e_n
\end{aligned}$$

If $g'(\xi) \neq 0$, then

$$\begin{aligned}
\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} &= \lim_{n\to\infty} \left|g'(\eta)\right| \\
&= g'(\xi) \\
&= c \\
&\neq 0
\end{aligned}$$

Therefore the rate of convergence if 1. □

3.3 Secant Method

Let

$$\begin{aligned}
f(x) &= p_1(x) + \text{error} \\
&= f(x_n) + f[x_n, x_{n-1}](x - x_n) + f[x_n, x_{n-1}, x](x - x_n)(x - x_{n-1})
\end{aligned}$$

Therefore,

$$\begin{aligned}
0 &= f(\xi) \\
&= f(x_n) + f[x_n, x_{n-1}](\xi - x_n) + f[x_n, x_{n-1}, x](\xi - x_n)(\xi - x_{n-1})
\end{aligned}$$

Therefore,

$$\begin{aligned}
-\frac{f(x_n)}{f[x_n, x_{n-1}]} &= \xi - x_n + \frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n-1}]}(\xi - x_n)(\xi - x_{n-1}) \\
\therefore \xi - x_n + \frac{f(x_n)}{f[x_n, x_{n-1}]} &= -\frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n-1}]}(\xi - x_n)(\xi - x_{n-1}) \\
\therefore \xi - x_{n+1} &= -\frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n-1}]}(\xi - x_n)(\xi - x_{n-1}) \\
\therefore e_{n+1} &= -\frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n+1}]} e_n e_{n-1}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n||e_{n-1}|} &= \left|\frac{f[\xi, \xi, \xi]}{f[\xi, \xi]}\right| \\
&= \left|\frac{f''(\xi)}{2\varphi'(\xi)}\right| \\
&= c
\end{aligned}$$

Let c be non zero.
Therefore,

$$|e_{n+1}| = c|e_n||e_{n-1}|$$

Let the rate of convergence be p.
Therefore,

$$|e_n| = b|e_{n-1}|^p$$

For a large n,

$$|e_{n+1}| = b|e_n|^p$$

Therefore,

$$\begin{aligned} e_{n+1} &= c\,|b|e_{n-1}|^p|\,|e_{n-1}| \\ &= bc|e_{n-1}|^{p+1} \end{aligned}$$

Therefore,

$$\begin{aligned} |e_{n+1}| &= b\,|b|e_{n-1}|^p|^p \\ &= bb^p|e_{n-1}|^{p^2} \end{aligned}$$

Therefore,

$$c = b^p$$

Therefore,

$$p^2 = p + 1$$

Therefore, the rate of convergence is

$$\rho = \frac{1+\sqrt{5}}{2}$$

Part IV

Linear Systems and Matrices

Theorem 23. *Let A be a $n \times n$ matrix. Then, the following statements are equivalent.*

1. *For any vector b there is a unique solution for $Ax = b$.*
2. *The homogeneous system $Ax = 0$ has only the trivial solution $x = 0$.*
3. *A is invertible.*
4. $\det A \neq 0$.

1 Direct Methods

1.1 Back Substitution

Algorithm 5 Back Substitution

Input: $b_{n\times 1}$, upper triangular $A_{n\times n}$

Output: $Ax = b$

1: $x_n \leftarrow \frac{b_n}{a_{nn}}$

2: **for all** $0 < k \leq n$ **do**

3: $\quad x_k \leftarrow \frac{b_k - \sum_{j=k+1}^{n} a_{kj}x_j}{a_{kk}}$

4: **end for**

1.2 LU Decomposition/Gaussian Elimination

Algorithm 6 LU Decomposition/Gaussian Elimination

Input: invertible $A_{n\times n}$

Output: lower triangular $L_{n\times n}$, and upper triangular $U_{n\times n}$, such that $LU = A$

procedure ROWOPERATION((P, i, j))
    $R_i \leftarrow R_i - m_{ij} R_j$ ▷ R_i and R_j are the ith and jth rows of P
end procedure

$A^{(1)} \leftarrow A$
$b^{(1)} \leftarrow b$
for $k = 1, \ldots, n-1$ **do**
    for $i = k+1, \ldots, n$ **do**
        $m_{ik} \leftarrow \frac{{a_{ik}}^{(k)}}{{a_{kk}}^{(k)}}$
        $A^{(k+1)} \leftarrow$ ROWOPERATION($A^{(k)}, i, k$).
    end for
end for

if $i > j$ **then**
    $L_{ij} \leftarrow m_{ij}$
else if $i = j$ **then**
    $L_{ij} \leftarrow 1$
else
    $L_{ij} \leftarrow 0$
end if
$U \leftarrow A^{(n)}$

Theorem 24. *Let the LU Decomposition/Gaussian Elimination of A be*

$$A = LU$$

Then the solution to the matrix equation

$$Ax = bj$$

is given by

$$Ly = b$$

where

$$Ux = y$$

Theorem 25. *The number of operations required for solving the matrix equation* $A_{n\times n}x_{n\times 1} = b_{n\times 1}$ *using LU Decomposition/Gaussian Elimination is* $\mathrm{O}\left(\frac{2}{3}n^3\right)$.

2 Error Analysis

Definition 23. The norm of the vector is defined to be a function from $\mathbb{R}^n$ to $\mathbb{R}$ which satisfies all of the following.

1. $\forall x \in \mathbb{R}^n$, $\|x\| \geq 0$.

2. $\|x\| = 0 \iff x = 0$.

3. $\forall x \in \mathbb{R}$, $\forall \alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha|\|x\|$.

4. $\forall x, y \in \mathbb{R}$, $\|x + y\| \leq \|x\| + \|y\|$.

Definition 24 (Infinity norm)**.** The function $\max\limits_{1\leq i\leq n} |y_i|$ is defined to be the infinity norm of the vector y.

Definition 25 (L_1 norm)**.** The function $\sum\limits_{i=1}^{n} |y_i|$ is defined to be the L_1 norm of the vector y.

Definition 26 (L_2 norm)**.** The function $\sqrt{\sum\limits_{i=1}^{n} {y_i}^2}$ is defined to be the L_2 norm of the vector y.

Definition 27 (Matrix norm)**.** A function from $\mathbb{R}^{n^2}$ to $\mathbb{R}$, which for every $A, B \in \mathbb{R}^{n^2}$ and for any $\alpha \in \mathbb{R}$, satisfies the following conditions is called the matrix norm of a matrix A.

1. $\|A\| \geq 0$.

2. $\|A\| = 0 \iff A = 0$.

Theorem 26. *If* $\|\cdot\|$ *is a vector norm on* $\mathbb{R}^n$*, then the function*

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

is a matrix norm.

Definition 28 (Induced norm)**.** Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$. The function

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

is called the induced norm.

Definition 29 (Induced infinity norm)**.** The function

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty$$

is called the induced infinity norm.

Theorem 27.

$$\sup_{\|x\|=1} \|Ax\| = \sup_{\|x\|\leq 1} \|Ax\| = \sup_{\|x\|\neq 0} \frac{\|Ax\|}{\|x\|}$$

Theorem 28.

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

where $A = (a_{ij})$*.*

Theorem 29.

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

Theorem 30. $\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij})^2}$ *is not an induced norm, for any vector norm.*

Definition 30 (Frobinus norm)**.**

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij})^2}$$

is called the Frobinus norm of A.

Theorem 31. *The Frobinus norm is a matrix norm.*

Definition 31. The spectral radius of a matrix A is defined as

$$\rho(A) = \max_{1\leq i\leq n} |\lambda_i|$$

where λ_i are the eigenvalues of A.

Theorem 32.

$$\|A\|_2 = \sqrt{\rho\left(A^\mathsf{T} A\right)}$$

Theorem 33. *For any matrix induced norm*

$$\rho(A) \leq \|A\|$$

Theorem 34. *For any $\varepsilon > 0$, there exists a norm for which*

$$\|A\| \leq \rho(A) + \varepsilon$$

2.1 Error in b

Let x be the ideal solution, and let $\widetilde{x}$ be the calculated solution.

$$e = x - \widetilde{x}$$

Therefore, the ideal system is

$$Ax = b$$

and the calculated system is

$$A\widetilde{x} = \widetilde{b}$$

Therefore,

$$e = x - \widetilde{x}$$

Let

$$\begin{aligned} r &= b - \widetilde{b} \\ &= b - A\widetilde{x} \end{aligned}$$

be the residue.
Therefore,

$$\begin{aligned} Ae &= A(x - \widetilde{x}) \\ &= Ax - A\widetilde{x} \\ &= b - A\widetilde{x} \\ &= r \end{aligned}$$

Therefore,

$$e = A^{-1} r$$

Therefore,

$$\begin{aligned}\|e\| &= \left\|A^{-1} r\right\| \\ &\leq \left\|A^{-1}\right\| \|r\|\end{aligned}$$

Therefore,

$$\frac{\|e\|}{\|x\|} = \frac{\|x - \widetilde{x}\|}{\|x\|}$$

Therefore,

$$\begin{aligned}\|b\| &= \|Ax\| \\ &\leq \|A\|\|x\| \\ \therefore \frac{1}{\|x\|} &\leq \|A\| \frac{1}{\|b\|} \\ \therefore \frac{\|e\|}{\|x\|} &\leq \|e\|\|A\| \frac{1}{\|b\|} \\ &\leq \|A\| \frac{1}{\|b\|} \left\|A^{-1}\right\| \|r\| \\ &\leq \|A\| \left\|A^{-1}\right\| \frac{\|r\|}{\|b\|}\end{aligned}$$

Definition 32 (Condition number)**.**

$$\operatorname{cond}(A) = \|A\| \left\|A^{-1}\right\|$$

is called the condition number of A.

Theorem 35. *For any matrix A,*

$$\operatorname{cond}(A) \geq 1$$

2.2 Estimation of $\operatorname{cond}(A)$

Theorem 36. *The eigenvalues of A^{-1} are $\frac{1}{\lambda_i}$, where λ_i are the eigenvalues of A.*

Proof. Let u_i be the eigenvectors of A, corresponding to λ_i.
Therefore,

$$Au_i = \lambda_i u_i$$

Therefore

$$\begin{aligned}
A^{-1}Au_i &= A^{-1}\lambda_i u_i \\
\therefore u_i &= A^{-1}\lambda_1 u_i \\
\therefore \frac{1}{\lambda_i} u_i &= A^{-1}u_i
\end{aligned}$$

Therefore, the eigenvalues of A^{-1}, corresponding to u_i, are $\frac{1}{\lambda_i}$. □

Theorem 37.

$$\text{cond}(A) \geq \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}$$

where λ_i are the eigenvalues of A.

Proof.

$$\begin{aligned}
\rho(A) &= \max_i |\lambda_i| \\
\therefore \rho\left(A^{-1}\right) &= \max_i \frac{1}{|\lambda_i|} \\
&= \frac{1}{\min\limits_i |\lambda_i|}
\end{aligned}$$

Therefore,

$$\rho(A)\rho\left(A^{-1}\right) = \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}$$

Therefore, as $\rho(A) \geq \|A\|$, and $\rho\left(A^{-1}\right) \geq \left\|A^{-1}\right\|$,

$$\begin{aligned}
\text{cond}(A) &\geq \rho(A)\rho\left(A^{-1}\right) \\
\therefore \text{cond}(A) &\geq \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}
\end{aligned}$$

□

Theorem 38. *For any non-invertible matrix B,*

$$\text{cond}(A) \geq \frac{\|A\|}{\|A - B\|}$$

Proof. If B is non-invertible, then $\exists x \neq 0$, such that

$$Bx = 0$$

Therefore,

$$\begin{aligned}\|A - B\|\|x\| &\geq \|(A - B)x\| \\ &\geq \|Ax\| \\ &\geq \frac{\|x\|}{\|A^{-1}\|}\end{aligned}$$

Therefore, as $x \neq 0$,

$$\|x\| \neq 0$$

Therefore,

$$\begin{aligned}\|A - B\| &\geq \frac{1}{\|A^{-1}\|} \\ \therefore \|A\| \left\|A^{-1}\right\| &\geq \|A\| \frac{1}{\|A - B\|} \\ \therefore \text{cond}(A) &\geq \|A\| \frac{1}{\|A - B\|}\end{aligned}$$

□

2.3 Error in A

Let x be the ideal solution, and let $\widetilde{x}$ be the calculated solution.
Let

$$\varepsilon = (\varepsilon_{ij})$$

be the error in A.
Let

$$e = x - \widetilde{x}$$

Therefore, the ideal system is

$$Ax = b$$

and the calculated system is

$$(A + \varepsilon)\widetilde{x} = b$$

Therefore,

$$\begin{aligned}
(A + \varepsilon)\widetilde{x} - Ax &= 0 \\
\therefore A\widetilde{x} - Ax + \varepsilon\widetilde{x} &= 0 \\
\therefore \varepsilon\widetilde{x} &= A\left(x - \widetilde{x}\right) \\
&= Ae
\end{aligned}$$

Therefore,

$$e = A^{-1}\varepsilon\widetilde{x}$$

Therefore

$$\begin{aligned}
\|e\| &= \left\|A^{-1}\right\| \|\varepsilon\| \|\widetilde{x}\| \\
\therefore \frac{\|e\|}{\|\widetilde{x}\|} &\leq \|A\| \left\|A^{-1}\right\| \frac{\|\varepsilon\|}{\|A\|} \\
\therefore \frac{\|e\|}{\|\widetilde{x}\|} &\leq \operatorname{cond}(A) \frac{\|\varepsilon\|}{\|A\|}
\end{aligned}$$

2.4 Iterative Improvement

Algorithm 7 Iterative Improvement

function LUSOLUTION($Ax = b$)
    $L, U \leftarrow$ LU DECOMPOSITION/GAUSSIAN ELIMINATION(A)
    Solve $Ly = b$
    Solve $Ux = y$ **return** x
end function
Solve $Ax = b$
$\widetilde{x}^{(1)} \leftarrow x$
for $i = 1, 2, \ldots$ **do**
    $r^{(n)} \leftarrow b - A\widetilde{x}^{(n)}$
    LUSOLUTION($Ae^{(n)} = r^{(n)}$)
    LUSOLUTION($Ae^{(n)} = r^{(n)}$)
end for
$\widetilde{x}^{(n+1)} \leftarrow \widetilde{x}^{(n)} + e^{(n)}$

Theorem 39. *Consider a fixed point method*

$$f(x) = Ax - b$$

where A is a matrix, and x and b are vectors.
If g maps a closed set $S \subset \mathbb{R}^n$ to itself, and g is contracting, i.e. for $k < 1$,

$$\|g(x) - g(y)\| \leq k\|x - y\|$$

then,

1. *There exists a fixed point ξ in S.*
2. *The fixed point ξ is unique.*
3. *All series of the form $x^{(0)}, x^{(1)}, \ldots$, such that $x^{(n+1)} = g\left(x^{(n)}\right)$ converge to the fixed point ξ, i.e.,*

   $$\lim_{n\to\infty} \left\|\xi - x^{(n)}\right\| = 0$$

   i.e.,

   $$\begin{aligned}\left\|\xi - x^{(n)}\right\| &\leq \frac{k}{1-k}\left\|x^{(n)} - x^{(n-1)}\right\| \\ &\leq \frac{k^n}{1-k}\left\|x^{(1)} - x^{(0)}\right\|\end{aligned}$$

Theorem 40. *As the LU decomposition of A needs to be calculated only once, the algorithm is $\mathrm{O}\left(n^2\right)$.*

3 Gauss-Jacobi Method

Definition 33. A matrix C is called an approximate inverse to the matrix A if in some norm,

$$\|I - CA\| = k$$

such that

$$k < 1$$

Theorem 41. *If C is an approximate inverse to A, then A and C are invertible matrices.*

Theorem 42. *Let D be the matrix containing only the diagonal elements of A. Then, D^{-1} is an approximate inverse to A.*

Definition 34 (Gauss-Jacobi Method)**.** The iterative method

$$x^{(n+1)} = x^{(n)} + D^{-1}\left(b - Ax^{(n)}\right)$$

is called the Gauss-Jacobi Method.

Theorem 43. *The number of operations in the Gauss-Jacobi Method is* $\mathrm{O}\left(n^2\right)$*.*

Theorem 44. *Let D be the matrix containing only the diagonal elements of A. Then*

$$D^{-1}_{ij} = \frac{1}{a_{ii}}\delta_{ij}$$

where δ_{ij} is the Kronecker delta function.

Algorithm 8 Gauss-Jacobi Method

1: Find lower triangular L, diagonal D, and upper triangular U, such that $A = L + D + U$
2: $C \leftarrow D^{-1}$
3: $B_J \leftarrow (I - CA) = -C(L + U)$ ▷ $\|B_J\|$ is called the contraction coefficient.
4: $d_J \leftarrow Cb$
5: $x^{(n+1)} \leftarrow Bx^{(n)} + d$

Algorithm 9 Gauss-Seidel Method

1: Find lower triangular L, diagonal D, and upper triangular U, such that $A = L + D + U$
2: $C \leftarrow (L + D)^{-1}$
3: $B_{GS} \leftarrow (I - CA) = -CU$
4: $d_{GS} \leftarrow Cb$
5: $x^{(n+1)} \leftarrow Bx^{(n)} + d$

Part V

Numerical Differentiation and Integration

1 Rule, Nodes, and Weights

Consider a linear operator L, i.e.,

$$L(af + bg) = aL(f) + bL(g)$$

where f and g are two functions.
Let p_k be the interpolation polynomial of $f(x)$.
Therefore,

$$e(x) = f(x) - p_k(x)$$
$$\therefore L(e) = L(f) - L(p_k)$$

For example, for Lagrange interpolation,

$$p_k(x) = \sum_{i=0}^{k} f(x_i) l_i(x)$$

where all l_i are Lagrange polynomials with respect to the corresponding x_i.
Therefore,

$$L(p_k) = \sum_{i=0}^{k} f(x_i) L(l_i)$$

Therefore,

$$L(f) \approx \sum_{i=0}^{k} w_i f(x_i)$$

where $f(x_i)$ are called the nodes, w_i are called the weights, and the entire expression is called the rule.

2 Numerical Differentiation

2.1 $k = 1$

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0)$$

Therefore,

$$\begin{aligned} D_a(f) &\approx D_a(p_1) \\ \therefore f'(x) &\approx f[x_0, x_1] \end{aligned}$$

Let

$$\begin{aligned} a &= x_0 \\ h &= x_1 - x_0 \end{aligned}$$

Therefore,

$$\begin{aligned} f'(a) &\approx f[a, a+h] \\ &\approx \frac{f(a+h) - f(a)}{h} \end{aligned}$$

Therefore,

$$|E(f)| = \left| \frac{1}{2} h f''(\eta) \right|$$

where $\eta \in [a, a+h]$.
This is called the forward difference scheme.

Let

$$\begin{aligned} a &= x_0 \\ h &= x_0 - x_1 \end{aligned}$$

Therefore,

$$\begin{aligned} f'(a) &\approx f[a, a-h] \\ &\approx \frac{f(a) - f(a-h)}{h} \end{aligned}$$

Therefore,

$$|E(f)| = \left| \frac{1}{2} h f''(\eta) \right|$$

where $\eta \in [a, a+h]$.
This is called the backward difference scheme.

Let $a = \frac{x_0 - x_1}{2}$, and $h = \frac{x_1 - x_0}{2}$.

$$\begin{aligned} a &= \frac{x_0 - x_1}{2} \\ h &= \frac{x_1 - x_0}{2} \end{aligned}$$

Therefore,

$$\begin{aligned} f'(a) &\approx f[a-h, a+h] \\ &\approx \frac{f(a-h) - f(a+h)}{2h} \end{aligned}$$

Therefore,

$$|E(f)| = \left| \frac{h^2}{6} f'''(\eta) \right|$$

where $\eta \in [a, a+h]$.
This is called the central difference scheme.

2.2 $k = 2$

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$

Therefore,

$$\begin{aligned} D_a(f) &\approx D_a(p_2) \\ \therefore f'(x) &\approx f[x_0, x_1] + f[x_0, x_1, x_2](x - x_1 + x - x_0) \end{aligned}$$

Let

$$a = x_0$$

Therefore,

$$f'(a) \approx f[a, x_1] + f[a, x_1, x_2](a - x_1)$$

2.3 Error Analysis

Let

$$
\begin{aligned}
f(x) &= p_k(x) + e(x) \\
&= p_k(x) + f[x_0, \dots, x_k, x] \prod_{i=0}^{k} (x - x_i)
\end{aligned}
$$

Let

$$\psi_k(x) = \prod_{i=0}^{k} (x - x_i)$$

Therefore,

$$f(x) = p_k(x) + f[x_0, \dots, x_k, x]\psi_k(x)$$

Therefore,

$$f'(x) = {p_k}'(x) + \frac{\mathrm{d}}{\mathrm{d}x} \left(f[x_0, \dots, x_k, x]\psi_k(x) \right)$$

By definition,

$$\frac{\mathrm{d}}{\mathrm{d}x} f[x_0, \dots, x_k, x] = f[x_0, \dots, x_k, x, x]$$

Therefore,

$$f'(x) = {p_k}'(x) + f[x_0, \dots, x_k, x, x]\psi(x) + f[x_0, \dots, x_k, x]{\psi_k}'(x)$$

Therefore,

$$
\begin{aligned}
e(x) &= f'(x) - {p_k}'(x) \\
&= f[x_0, \dots, x_k, x, x]\psi_k(x) + f[x_0, \dots, x_k, x]{\psi_k}'(x)
\end{aligned}
$$

Therefore,

$$e(x) = \frac{f^{(k+2)}(\xi)}{(k+2)!}\psi_k(x) + \frac{f^{(k+1)}(\eta)}{(k+1)!}{\psi_k}'(x)$$

where $\xi, \eta \in [x_0, x_k]$.