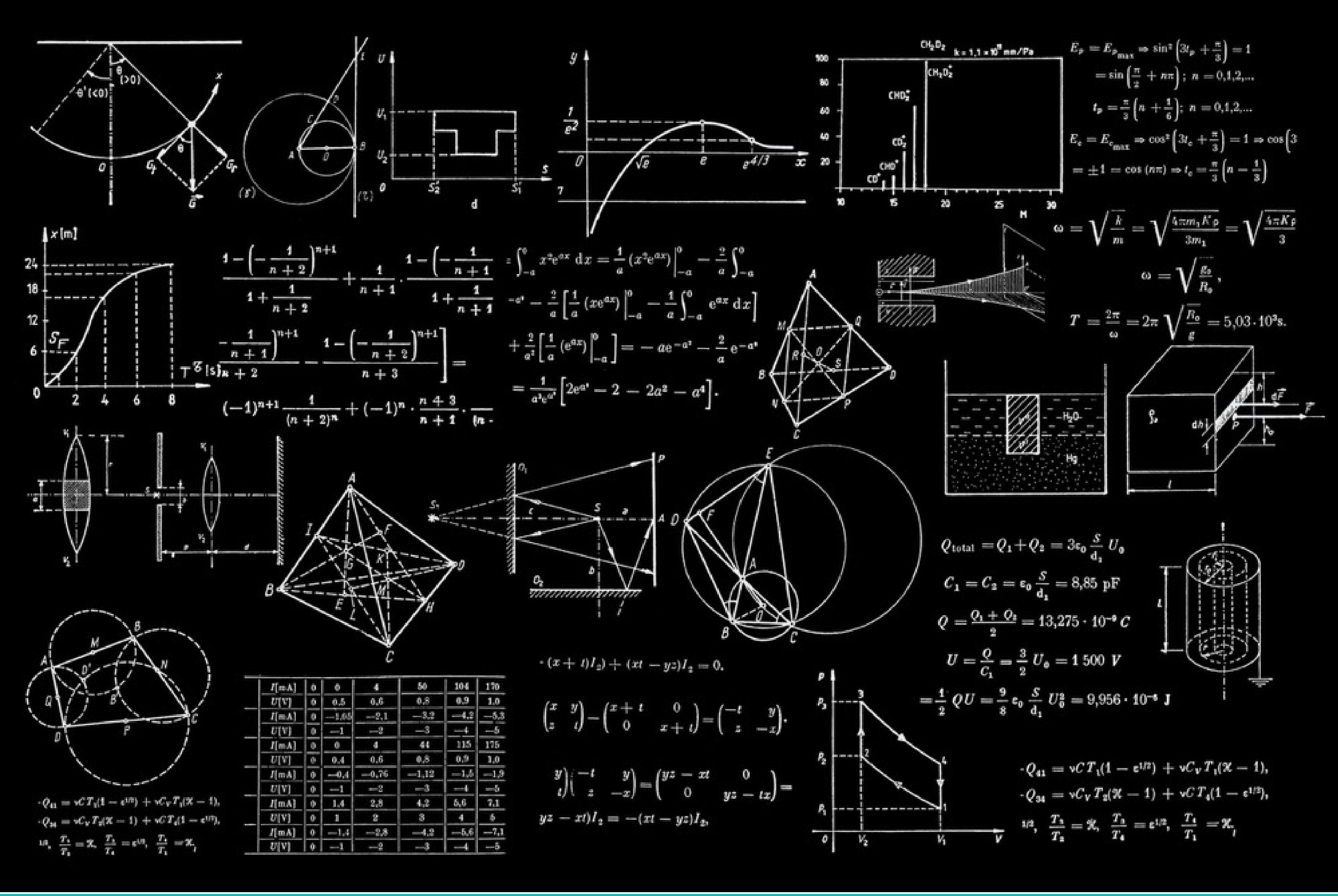

Numerical Analysis

Author: Zye Wang

Institute: YZU Mathematical

Date: January 3, 2024

Version: 0.3

Bio: Information

ElegantLaTeX Program

Stay hungry and stay foolish.

Contents

Introduction

The content of introduction.

Chapter 1 Mathematical foundations

1.1 Norm Space

1.1.1 Norm Space

Definition 1.1 (Norm space)

Let X be a complex (or real) linear space (vector space). A function $\|\cdot\|: X \to \mathbb{R}$ with the properties

1. $\|x\| \geq 0$, *(positivity)*
2. $\|x\| = 0$ *if and only if* $x = 0$,*(definiteness)*
3. $\|\alpha x\| = |\alpha|\|x\|$,*(homogeneity)*
4. $\|x + y\| \leqslant \|x\| + \|y\|$, *(triangle inequality)*

for all $x, y \in X$ and all $\alpha \in \mathbb{C}$ (or $\mathbb{R}$) is called a norm on X. A linear space X equipped with a norm is called a normed space. For $X = \mathbb{R}^n$ or $X = \mathbb{C}^n$ we will also call the norm a vector norm. ♣

Remark For each norm, the second triangle inequality

$$|\|x\| - \|y\|| \leqslant \|x - y\|$$

holds for all $x, y \in X$.

Proof From the triangle inequality we have

$$\|x\| = \|x - y + y\| \leqslant \|x - y\| + \|y\|,$$

whence $\|x\| - \|y\| \leqslant \|x - y\|$ follows. Analogously, by interchanging the roles of x and y we have $\|y\| - \|x\| \leqslant \|y - x\|$.

For two elements x, y in a normed space $\|x - y\|$ is called the distance between x and y.

1.1.2 Vector Norm

Definition 1.2 (Common vector norms)

1. *The 1-norm of a vector $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is defined as*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

2. *The 2-norm of a vector $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is defined as*

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

.

3. *The p-norm of a vector $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is defined as*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, where 1 \leqslant p < \infty$$

.

4. *∞-norm of a vector $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is defined as*

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$$

.

Example 1.1 assume $x = (1, -4, 0, 2)^{\mathrm{T}}$, calculate its vector norm $\|x\|_1, \|x\|_2, \|x\|_\infty$

Solution

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = 7$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{21}$$

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| = 4$$

Remark In $\mathbb{R}^n$ space, any two norms are equivalent.

Proposition 1.1

If a sequence of vectors converges in terms of one norm, then it converges in terms of any norm. $\lim_{k\to\infty} x^{(k)} = x^* \Longleftrightarrow \lim_{k\to\infty} \left\| x^{(k)} - x^* \right\| = 0$ *where* $\|\cdot\|$ *is any norm of a vector*

1.1.3 Matrix Norm

Definition 1.3 (Frobenius norm)

If we extend the concept of vector norms to matrices, then, based on the $2-$*norm of vectors in* $\mathbb{R}^n$*, we obtain a norm for matrices in* $\mathbb{R}^{n\times n}$ *defined as:*

$$F(A) = \|A\|_F = \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}$$

This is referred to as the Frobenius norm (or $\mathbf{F}$ *norm) of matrix* A*.*

Definition 1.4

For any real-valued function $\|\cdot\|$ *defined on the space* $\mathbb{R}^{n\times n}$ *with respect to any* $A, B \in \mathbb{R}^{n\times n}$*, satisfying the following conditions:*

1. *Positivity:* $\|A\| \geqslant 0; \|A\| = 0 \Leftrightarrow A = 0$
2. *Homogeneity:* $\|\alpha A\| = |\alpha| \|A\|, \forall \alpha \in \mathbb{R}$
3. *Triangle inequality:* $\|A + B\| \leqslant \|A\| + \|B\|$
4. *Compatibility:* $\|AB\| \leqslant \|A\| \cdot \|B\|$

then the real-valued function $\|\cdot\|$ *is termed a matrix norm on the space* $\mathbb{R}^{n\times n}$*.*

Remark Due to the frequent simultaneous consideration of matrices and vectors in most estimation-related problems, there is a desire to introduce a new matrix norm that is compatible with vector norms. Specifically, for any vector $x \in \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n\times n}$, it is required that the inequality $\|Ax\| \leqslant \|A\| \cdot \|x\|$ holds.

Definition 1.5 (Operator Norm)

Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Consider a vector norm $\|\|_\alpha$ where $\alpha = 1, 2, \infty$. Correspondingly, define a non-negative function for matrices as follows:

$$\|A\|_\alpha = \max_{x\neq\theta} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}$$

Here, $\|A\|_\alpha$ is a matrix norm on $\mathbb{R}^{n\times n}$ and is referred to as the operator norm of the matrix A. ♣

Definition 1.6 (Common Matrix Norm)

1. *1-Norm:*

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}| \quad \textit{(Column Sum Norm)}$$

2. *∞-Norm:*

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}| \quad \textit{(Row Sum Norm)}$$

3. *2-Norm:*

$$\|A\|_2 = \sqrt{\lambda_{max}\left(A^{\mathrm{T}}A\right)} \quad \textit{(Spectral Norm)}$$

♣

Example 1.2 Assume $A = \begin{pmatrix} 2 & -1 \\ -2 & 4 \end{pmatrix}$, calculate $\|A\|_1, \|A\|_2, \|A\|_\infty$.

Solution $\|A\|_1 = \max\{2 + |-2|, |-1| + 4\} = 5 \quad \|A\|_\infty = \max\{2 + |-1|, |-2| + 4\} = 6$

$$A^{\mathrm{T}}A = \begin{pmatrix} 2 & -2 \\ -1 & 4 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 8 & -10 \\ -10 & 17 \end{pmatrix}$$

$$\left|\lambda E - A^{\mathrm{T}}A\right| = 0 \Rightarrow \lambda_1 \approx 23.466, \lambda_2 \approx 1.534$$

$$\|A\|_2 = \sqrt{23.466} \approx 4.844$$

Definition 1.7 (Spectral Radius)

Let $\lambda_i (i = 1, 2, \ldots, n)$ be the eigenvalues of matrix A.

The quantity $\rho(A) = \max_{1\leqslant i\leqslant n}\{|\lambda_i|\}$ is referred to as the spectral radius of matrix A. ♣

Theorem 1.1

For any matrix norm, it holds that $\rho(A) \leqslant \|A\|$. ♡

Theorem 1.2

If matrix A is symmetric, then $\|A\|_2 = \rho(A)$. ♡

Remark In $R^{n\times n}$ space, any two matrix norms are equivalent

Definition 1.8 (Convergence of Matrix Sequences)

The convergence of matrix sequences is also defined as

$$\lim_{k\to\infty} A^{(k)} = A^* \Leftrightarrow \lim_{k\to\infty} \left\|A^{(k)} - A^*\right\| = 0 \left(\Leftrightarrow a_{ij}^{(k)} \to a_{ij}^*, i, j = 1, 2, \cdots, n\right).$$

♣

Theorem 1.3

Assume $A \in R^{n\times n}$, *then* $\lim\limits_{k\to\infty} A^k = O \iff \rho(A) < 1$.

Example 1.3 prove $A = \begin{pmatrix} 1/2 & 0 \\ 1/4 & 1/2 \end{pmatrix}$ is a convergent matrix

Proof $A^2 = \begin{pmatrix} 1/4 & 0 \\ 1/4 & 1/4 \end{pmatrix}$ $A^3 = \begin{pmatrix} 1/8 & 0 \\ 3/16 & 1/8 \end{pmatrix}$ $A^4 = \begin{pmatrix} 1/16 & 0 \\ 1/8 & 1/16 \end{pmatrix}$ $\cdots$ $A^k = \begin{pmatrix} (1/2)^k & 0 \\ k/2^{k+1} & (1/2)^k \end{pmatrix}$

$\because \lim\limits_{k\to\infty} (1/2)^k = 0,\ \lim\limits_{k\to\infty} k/2^{k+1} = 0$ $\therefore A$ is convergent

Remark $\rho(A) = 1/2 < 1$

Proposition 1.2

if $\|A\| < 1$, *then* $I \pm A$ *is inverse, and*

$$\|(I \pm A)^{-1}\| \leqslant \frac{1}{1 - \|A\|}$$

where $\|\cdot\|$ *is the operator norm of a matrix.*

Proof if $I \pm A$ is not inverse, then $(I \pm A)x = 0$ has a nonzero solution, thus there exisits nonzero vector x_0 s.t.

$$\pm A x_0 = -x_0$$

now,

$$\frac{\|A x_0\|}{\|x_0\|} = 1, \|A\| \geqslant 1$$

what's more,

$$(I \pm A)^{-1} \pm A(I \pm A)^{-1} = (I \pm A)(I \pm A)^{-1} = I$$

$$\longrightarrow (I \pm A)^{-1} = I \mp A(I \pm A)^{-1}$$

$$\longrightarrow \|(I \pm A)^{-1}\| \leqslant 1 + \|A\| \cdot \|(I \pm A)^{-1}\|$$

$$\longrightarrow \|(I \pm A)^{-1}\| \leqslant \frac{1}{1 - \|A\|}$$

1.2

1.3

Chapter 2 Error Analysis

Introduction

- *Error analysis is the study of how well a model or an estimator fits the data.*
- *Error analysis is a fundamental part of the theory of statistics*

2.1 Basic Concepts of Error

Definition 2.1

Source and classification

1. *Model error*
2. *Measurement error*
3. *Truncation error*
4. *Round off error*

♣

Definition 2.2 (Absolute Error)

Suppose that x^* is an approximation to x, then

$$|e = x - x^*|$$

is called the absolute error of x^*

♣

Definition 2.3 (Relative Error)

If x is an approximation to x^*, then

$$e_r = \frac{x - x^*}{x}$$

is called the relative error of x^*

♣

Definition 2.4 (Relative Error Bound)

A positive number ε is called the relative error bound of x^* if

$$|e_r| = \left|\frac{x - x^*}{x}\right| \leqslant \varepsilon_r \quad or \quad |e_r| = \left|\frac{x - x^*}{x^*}\right| \leqslant \varepsilon_r$$

.

♣

2.2 Significant Digits

Definition 2.5

Suppose $x = \pm\left(a_1 \times 10^{-1} + a_2 \times 10^{-2} + \ldots a_n \times 10^{-n}\right) \times 10^m$

$where m \in \mathbb{Z}$, $a_i \in \{0, 1, 2 \cdots, 9\}$,$a_1 \neq 0$.

If x has n significant digits, the error can be represented as

$$|x^* - x| \leqslant \left(\frac{1}{2} \times 10^{-n}\right) \times 10^m$$

♣

Theorem 2.1

Suppose $x = \pm\left(a_1 \times 10^{-1} + a_2 \times 10^{-2} + \ldots a_n \times 10^{-n}\right) \times 10^m$ *is the approximation of* x^*

1. *If* x *has* l *significant digits, then the relative error bound* x *is*
$$\frac{1}{2a_1} \times 10^{-l+1}$$
2. *If the relative error bound of* x *is*
$$\frac{1}{2\left(a_1+1\right)} \times 10^{-l+1}$$
where $1 \leqslant l \leqslant n$, *then* x *has at least* l *significant digits.*

2.3 Machine Number System

Suppose the computer has an n-bit word length. using the β system. and the order code bit is p. Then the floating -point representation of numbers in a compute is

$$x = \pm\left(0.a_1a_2\cdots a_n\right)\beta^p$$

β is called the base of a floating -point number. $\alpha = \pm\left(0.a_1a_2\ldots a_n\right)$ is called the mantissa

The set composed by all floating-point number and zero is called the Machine Number System. denoted by

$$F(\beta, n, L, U) = \{0\} \cup \left\{x \mid x = \pm\left(0, a_1a_2\ldots a_n\right)\beta^P\right\}.$$

Proposition 2.1

1. $F(\beta, n, L, U)$ *is composed of limited number with the number of*
$$1 + 2(\beta-1)\beta^{n-1}(U-L+1)$$
2. *The number with the highest absolute value*
$$\pm\left(\frac{\beta-1}{\beta} + \frac{\beta-1}{\beta^2} + \ldots \frac{\beta-1}{\beta^n}\right)\beta^U = \pm\left(1-\beta^{-n}\right)\beta^U$$
3. *The None-zero number with the smallest absolute value*
$$\pm\left(\frac{1}{\beta} + \frac{0}{\beta^2} + \ldots \frac{0}{\beta^n}\right)\beta^L = \pm\beta^{-1+L}$$

Theorem 2.2

Suppose real number $x \neq 0$. *and floating -point number in* $F(\beta, n, L, U)$ *is* $fl(x)$. *then* e_r *is the relative error of* $fl(x)$ *satisfies*

$$|e_r| = \left|\frac{x - fl(x)}{x}\right| \leqslant \frac{1}{2}\beta^{1-n}$$

Let $\varepsilon = \frac{f(l)-x}{x}$ $\quad fl(x) = x(1+\varepsilon) \quad |\varepsilon| \leqslant \frac{1}{2}\beta^{1-n}$

Proposition 2.2

suppose x_1,x_2 *are flouting-point number, then*

1. $fl\left(x_1 + x_2\right) = \left(x_1 + x_2\right)\left(1+\varepsilon_1\right)$
2. $fl\left(x_1 - x_2\right) = \left(x_1 - x_2\right)\left(1+\varepsilon_2\right)$
3. $fl\left(x_1x_2\right) = \left(x_1x_2\right)\left(1+\varepsilon_3\right)$

4. $fl\left(x_1/x_2\right) = \left(x_1/x_2\right)\left(1+\varepsilon_4\right)$
   where $|\varepsilon_i| \leqslant \frac{1}{2}\beta^{1-n}$

Remark

1. When adding numbers of the same number, add the ones with smaller absolute value first.
2. In computer floating-point operations, the associative law addition may not necessarily satisfy

Example 2.1 Suppose $n=3, L=-5, U=5x=1.623, y=0.184, z=0.00362.$
find $u=(x+y)+z. \quad v=x+(y+z).$

Solution

$$fl(x) = 0.162\times 10^1$$

$$fl(y) = 0.184\times 10^0$$

$$fl(z) = 0.362\times 10^{-2}$$

$$fl(x)+fl(y) = 0.162\times 10^1 + 0.018\times 10^1 = 0.180\times 10^1$$

$$\begin{aligned} u &= (fl(x)+fl(y)+fl(z)) \\ &= 0.180\times 10^1 + 0.362\times 10^{-2} \\ &= 0.180\times 10^1 + 0.000\times 10^1 \\ &= 0.180\times 10^1 \end{aligned}$$

2.4 Numerical Stability

Example 2.2 Calculate the following integral

$$I_n = \int_0^1 \frac{x^n}{x+5}dx \quad ,n=0,1,2,\ldots,10.$$

Solution

$$\begin{aligned} I_n &= \int_0^1 \frac{x^n}{x+5}dx \\ &= \int_0^1 \frac{x^{n-1}(x+5)}{x+5} - 5\int_0^1 \frac{x^{n-1}}{x+5}dx \\ &= \frac{1}{n} - 5I_{n-1} \\ I_0 &= \int_0^1 \frac{1}{x+5}dx = \ln\left(\frac{6}{5}\right) \\ \tilde{I}_0 &\approx \ln 1.2. \quad \tilde{I}_1 = 1-5\tilde{I}_0\ldots \end{aligned}$$

$$suppose \quad e_n = I_n - \tilde{I}_n \to |e_n| = 5^n \mid e_0 \mid$$

Thus it's an unstable algorithm

Then use another method

$$I_{n-1} = \frac{1}{5}\left(\frac{1}{n} - I_n\right) \Rightarrow |e_{n-1}| = \frac{1}{5}|e_n|$$

$$|e_{10-k}| = \left(\frac{1}{5}\right)^k |e_{10}|$$

Let's calculate the approximate value of I_{10} below

By first Mean Value Theorem of Integrals

$$
\begin{aligned}
& I_n = \frac{1}{\xi_n + 5} \int_0^1 x^n dx = \frac{1}{\xi_n + 5} \cdot \frac{1}{n+1} \qquad (0 < \xi_n < 1) \\
& \frac{1}{6} \frac{1}{n+1} < I_n < \frac{1}{5} \frac{1}{n+1} \\
& let\ \tilde{I_n} = \frac{1}{2} \left(\frac{1}{6} \frac{1}{n+1} + \frac{1}{5} \frac{1}{n+1} \right) \Rightarrow \tilde{I_{10}} = \frac{1}{60} \\
& \quad \left| I_{10} - \tilde{I_{10}} \right| \leqslant \frac{1}{2} \left(\frac{1}{55} - \frac{1}{66} \right) = \frac{1}{660}
\end{aligned}
$$

Proposition 2.3

1. *Avoid the Loss of Accuracy.*
2. *Avoid the subtraction of Nearly Equal Numbers.*
3. *Avoid Big Numbers "swallowing" Small Numbers*
4. *Avoid Dividing by a Number with Small Absolute Value*

Example 2.3

$$
\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}
$$

Example 2.4 Using the quadratic formula and 4-digit rounding arithmetic: to find the roots of $x^2 + 62.10x + 1 = 0$

Solution

$$
\begin{aligned}
x_1^* &= \frac{-62.10 + \sqrt{(62.10)^2 - 4.000}}{2.000} \\
&= \frac{(-62.1)^2 - \left((62.10)^2 - 4.000\right)}{2.000 \times \left(-62.10 - \sqrt{(62.10)^2 - 4.000}\right.} 00 \\
&= \frac{2.000}{-62.10 - \sqrt{(62.10)^2 - 4.000}}
\end{aligned}
$$

Chapter 3 Solutions of Equations in One Variable

3.1 The Bisection Method

Lemma 3.1 (Intermediate value Theorem)

If $f \in C[a,b].k$ is a number between $f(a)$ and $f(b)$ then there exists at least one point x^*. s.t.

$$f(x^*) = k$$

♡

Corollary 3.1 (Zero-point Theorem)

If $f \in c[a,b], f(a)f(b) < 0$, then there exists at least one number c in (a,b) s.t.

$$f(c) = 0$$

♡

The Algorithm

point.

set $a_0 = a, b_0 = b$

set $x_0 = a_0 + b_0/2 \leftarrow$

if $f(x_0) = 0$, then $x^* = x_0$

else $[a,b] = \begin{cases} [a_0, x_0], & \text{if } f(a_0) f(x_0) < 0 \\ [x_0, b_0], & \text{if } f(x_0) f(b_0) < 0 \end{cases}$

$$f(x_k) < \varepsilon / |b_k - a_k|$$

$$| \; yes$$

$$x^* = x_k$$

Remark

$$|x_k - x^*| \leqslant \frac{b_k - a_k}{2} = \frac{b-a}{2^{k+1}}$$

$$\text{if } |x_k - x^*| < \varepsilon, \quad \text{i.e.} \frac{b-a}{2^{k+1}} < \varepsilon \Rightarrow k > \log_2 \frac{b-a}{\varepsilon} - 1$$

Example 3.1 Use Bisection method to find solution of $f(x) = x^2 + 4x^2 - 10$ in $[1,2]$, the iteration is terminated when $|x_k - x^*| < \frac{1}{2} \times | \; 10^{-5}$

Solution $f(x) \in C[1,2] \quad f(1) = -5 < 0 \quad f(2) = 14 > 0$

By Intermediate Value Theorem, $\exists x^* \in (1,2) \quad s.t. f(x^*) = 0$, since $f'(x) = 3x^2 + 8x > 0$, for any $x \in (1,2)$, then x^* is unique,

n	a_n	b_n	x_n	$f(x_n)$
0	1^-	2^+	1.5^+	2.375
1	1^-	1.5^+	1.25^-	-1.79687
2	1.25^-	1.5^+	1.375^+	0.16211
3	1.25^-	1.315^+	1.3125^-	-0.84839

Remark

1. The Bisection Method is simple, effective, and easy to implement on a computer. However. if the equation has multiple roots on the root interval, this method only finds one of the roots.

2. If. there are even double routs. the method cannot be used.
3. The Bisection Method has a slow convergence speed and is often used to provide a good initial value for other iterative methods.

3.2 Fixed-Point Iteration

Theorem 3.1

Suppose $\varphi(x)$ satisfies

1. $\forall x \in [a,b], \varphi(x) \in [a,b]$
2. $\exists 0 \leqslant L < 1$, *s.t.* $\forall x \in (a,b) \quad |\varphi'(x)| \leqslant L < 1$

then. for any initial value $x_0 \in [a,b]$

the sequence $\{x_k = \varphi(x_{k-1})\}$ converges to the unique root x^*, s.t. $x^* = \varphi(x^*)$.

Proof

1. Existence

   consider $f(x) = \varphi(x) - x$

   since $f(a) = \varphi(a) - a \geqslant 0. \quad f(b) = \varphi(b) - b \leqslant 0$

   if $f(a)$ or $f(b) = 0$. then a or b is the fixed -point.

   if $f(a) > 0, f(b) < 0$,

   By Intermediate Value Theorem.

   $\exists$ fixed -point x^* s.t. $f(x^*) = 0$
2. Uniqueness

   Suppose $x^* = \varphi(x^*)$

$$
\begin{gathered}
y^* = \varphi(y^*) \\
|x^* - y^*| = |\varphi(x^*) - \varphi(y^*)| \leqslant L|x^* - y^*| \\
(1-L)|x^* - y^*| \leqslant 0.
\end{gathered}
$$

   since $1 - L > 0$, so $|x^* - y^*| = 0$

$$\text{ie. } x^* = y^*$$

   Astringency.

$$
\begin{aligned}
|x^* - x_k| = |\varphi(x^*) - \varphi(x_{k-1})| \leqslant L|x^* - x_{k-1}| \\
\cdots \leqslant L^k|x^* - x_0| \to 0
\end{aligned}
$$

Corollary 3.2

If φ satisfies the hypotheses of Theorem, then the following error bounds hold.

1. $$|x^* - x_k| \leqslant \frac{L}{1-L}|x_k - x_{k-1}|$$
2. $$|x^* - x_k| \leqslant \frac{L^k}{1-L}|x_1 - x_0|$$

Proof

1.

$$
\begin{aligned}
\left|x^{*}-x_{k}\right| &=\left|\varphi\left(x^{*}\right)-\varphi\left(x_{k-1}\right)\right| \\
& \leqslant L\left|x^{*}-x_{k-1}\right| \\
& \leqslant L\left(\left|x^{*}-x_{k}\right|+\left|x_{k}-x_{k-1}\right|\right) \\
\Rightarrow(1-L)\left|x^{*}-x_{k}\right| & \leqslant L\left|x_{k}-x_{k-1}\right|
\end{aligned}
$$

2.

$$
\begin{aligned}
\left|x_{k}-x_{k-1}\right| &=\left|\varphi\left(x_{k-1}\right)-\varphi\left(x_{k-2}\right)\right| \\
& \leqslant L\left|x_{k-1}-x_{k-2}\right| \\
& \leqslant L^{2}\left|x_{k-2}-x_{k-3}\right| \\
& \cdots \\
& \leqslant L^{k-1}\left|x_{1}-x_{0}\right| \\
\Rightarrow\left|x^{*}-x_{k}\right| & \leqslant \frac{L}{1-L}\left|x_{k}-x_{k-1}\right| \leqslant \frac{L^{k}}{1-L}\left|x_{1}-x_{0}\right|
\end{aligned}
$$

Example 3.2 Find the approximation of $\sqrt{2}\left(\varepsilon=10^{-5}\right)$

Solution *Suppose* $x=\sqrt{2}-1$, *then* $(x+2)x=1, f(x)=x^{2}+2x-1$

$$
\begin{aligned}
& x^{*} \in[0,0.5], \quad x=\frac{1}{x+2}=\varphi(x) \\
& \varphi(x) \in\left[\frac{2}{5}, \frac{1}{2}\right] \subset[0,0.5].
\end{aligned}
$$

thus. $\forall u, v \in[0,0.5]$,

$$
|\varphi(u)-\varphi(v)|=\left|\frac{1}{u+2}-\frac{1}{v+2}\right|=\left|\frac{u-v}{(u+2)(v+2)}\right| \leqslant \frac{1}{4}|u-v|
$$

thus. $\varphi(x)$ *is a compressed image on an Interval*

$$
x_{k+1}=\frac{1}{x_{k+2}} \quad \text{let } x_{0}=0
$$

$$
x^{*} \approx x_{8}=0.4142132 \quad \sqrt{2}=x^{*}+1 \approx 1.41421
$$

3.3 The convergence of iterative method

Theorem 3.2

Suppose the equation. $x=\varphi(x)$ *has a root* x^{*} *in* $[a, b]$, *if* $\left|\varphi^{\prime}(x)\right| \geqslant 1$ *for any* $x \in[a, b]$, *then for any* $x_{0} \in[a, b]\left(x_{0} \neq x^{*}\right)$, *the iterative equation* $x_{k+1}=\varphi\left(x_{k}\right)$ *must be divergence*

Theorem 3.3

Suppose the equation $x=\varphi(x)$ *has a root* x^{*} *in* $[a, b]$ *if* $\left|\varphi^{\prime}(x)\right| \leqslant L<1$ *for any* $x \in[a, b]$, *then for any* $x_{0} \in[a, b],\left(x_{0} \neq x^{*}\right)$, *the iterative equation* $x_{k+1}=\varphi\left(x_{k}\right)$ *must be convergence*

Definition 3.1 (Local Convergence)

The sequence $\{x_k\}_{k=0}^{\infty}$ *defined by* $x_k = \varphi(x_{k-1})$ *locally converges to* $x^* = \varphi(x^*)$ *if there exists* $\delta > 0$, *s.t.* $\{x_k\}_{k=0}^{\infty}$ *converges to* x^* *for any* $x_0 \in (x_0^* - \delta, x^* + \delta)$ ♣

Theorem 3.4

let x^* *be a root of* $x = \varphi(x)$, *If there exists a* $\delta > 0$. *s.t.* $\varphi'(x)$ *is continuous on* $(x^* - \delta, x^* + \delta)$ *and* $|\varphi'(x)| < 1$. *then the sequence locally converges to* x^* *for any* x *in* $\Omega = (x^* - \delta, x^* + \delta)$ ♡

Remark In most cases, if $|\varphi'(x)|$ is significantly smaller than 1 in the small areas near the root, then with the initial value x_0 in the area, $\{x_k\}$ always be convergence.

Definition 3.2 (Order of Convergence)

Suppose $\{x_k\}$ *is converges to* x^*, *denoted* $e_k = x^* - x_k$ *If positive constant c and p exist with*

$$\lim_{n\to\infty}\left|\frac{e_{n+1}}{e_n^p}\right| = C$$

Then $\{x_k\}$ *converges to* x^* *of order p with the asymptotic error constant c.*

1. *If* $p = 1$. *the sequence is linearly convergent*
2. *If* $p = 2$. *the sequence is quadratically convergent* ♣

Theorem 3.5

Consider iterative scheme $x_{n+1} = \varphi(x_n) \to x^*$
If $\exists\delta$, $\Omega = \{x \mid x \in (x^* - \delta, x^* + \delta)\}$ *s.t.* $\varphi'(x) \in C(\Omega)$ *and* $\varphi'(x) \neq 0$ *then the iterative scheme has lined convergence* ♡

Proof $e_{k+1} = x^* - x_{k+1} = \varphi(x^*) - \varphi(x_k) = \varphi'(\xi)(x^* - x_k) = \varphi'(\xi)e_k$

then

$$\lim_{n\to\infty}\frac{|e_{k+1}|}{|e_k|} = \lim_{n\to\infty}\varphi'(\xi) = \varphi'(x^*) = C \neq 0$$

Theorem 3.6

Consider iterative scheme $x_{n+1} = \varphi(x_n) \to x^*$
if there exists a $\delta > 0$, *s.t.* $\varphi(x)$ *is p times differentiable on* $(x^* - \delta, x^* + \delta)$, *and*

$$\varphi'(x^*) = \varphi''(x^*) = \ldots \varphi^{(p-1)}(x^*) = 0$$

but

$$\varphi^{(p)}(x^*) \neq 0$$

then $\{x_k\}$ *converges to* x^* *of under p, where* $p \geqslant 1$ *is an integer. and*

$$\lim_{k\to\infty}\left|\frac{e_{k+1}}{e_k^p}\right| = C = \frac{|\varphi^{(p)}(x^*)|}{p!}$$ ♡

Proof Taylor expansion of $\varphi(x)$ at x^*

$$\varphi(x)=\varphi\left(x^*\right)+\varphi'\left(x^*\right)\left(x-x^*\right)+\ldots\frac{\varphi^{(p-1)}\left(x^*\right)}{(p-1)!}\left(x-x^*\right)^{p-1}+\frac{\varphi^{(p)}\left(x^*\right)}{p!}\left(x-x^*\right)^{p}$$

$$\varphi(x)=\varphi\left(x^*\right)+\frac{\varphi^{(p)}(\xi)}{p!}\left(x-x^*\right)^{p}$$

$$\text{let}\quad x=x_k,\text{ then}$$

$$x_{k+1}-x^*=\frac{\varphi^{(p)}(\xi)}{p!}\left(x_k-x^*\right)^{p}$$

$$\text{Thus }\left|\frac{e_{k+1}}{e_k}\right|\rightarrow\frac{\left|\varphi^{(p)}\left(x^*\right)\right|}{p!}(k\rightarrow\infty)$$

Remark The above conclusion indicates that the convergence speed of the iterative format depends on the selection of the iterative function $\varphi(x)$.

If $\varphi'\left(x^*\right)\neq 0$, the scheme can only be linear convergence

Remark

1. If $\left|g'\left(x^*\right)\right|<1$, but $\left|g'\left(x^*\right)\right|\neq 0$, then the sequence $\{x_k\}_{k=0}^{\infty}$ is linearly convergent.
2. If $g'\left(x^*\right)=0$, put $g''\left(x^*\right)\neq 0$, then the sequence is quadratically convergent

3.4 Newton's Method and Secant Method.

3.4.1 Newton's Method

Taylor expansion

$$f(x)\approx f\left(x_0\right)+f'\left(x_0\right)\left(x-x_0\right)=0\Rightarrow x=x_0-\frac{f(x)}{f\left(x_0\right)}$$

$$f\in c^2[a,b]\quad f'\left(x_n\right)\neq 0$$

Theorem 3.7

let $f\in C^2[a,b]$, x^* is a simple root of $f(x)$ in $[a,b]$, and $f'\left(x^*\right)\neq 0$, then . the sequence generated by Newton's method converges to x^* for any initial value $x_0\in\Omega$.

Proof

$$\varphi(x)=x-\frac{f(x)}{f'(x)}$$

$$\varphi'\left(x^*\right)=\frac{f\left(x^*\right)f''\left(x^*\right)}{\left[f'\left(x^*\right)\right]^2}=0$$

$$\forall x\in\Omega.\quad\left|\varphi'(x)\right|<1.$$

Thus, Newton iteration is locally convergence

Example 3.3 Find the root of $x^3+4^2-10=0$ in $[1,2]$ by Newton iteration with 10^{-4} accuracy.

Solution *let* $f(x)=x^3+4x^2-10$

According to Newton Method

$$x_{n+1}=x_n-\frac{x_n^3+4x_n^2-10}{3x_n^2+8x_n}$$

By selecting $x_0 = 1.5$

n	x_n	$x_n - x_{n-1}$
0	1.5	0.126
1	1.37333	0.126
2	1.36526	0.007
3	1.36523	0.000

$$x^* \approx x_3 = 1.365303$$

3.4.2 Secant Method

Definition 3.3

The secant method is an iterative technique

$$x_k = x_{k-1} - \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})} f(x_{k-1}).$$

♣

Theorem 3.8

If x^* is a simple root of the equation $f(x) = 0$. $f \in c^2\Omega$, the sequence generated by secant method converges to x^* of order.

$$P = \frac{1+\sqrt{5}}{2} \approx 1.618$$

for any initial value $x_0, x_1 \in \Omega$, as δ sufficiently small.

♡

Theorem 3.9

$f \in c^2[a,b]$, if x^* is a simple root of $f(x) = 0$ in $[a,b]$, the Newton iteration with at least second-order convergence.
If $f'(x) \neq 0$

$$\lim_{k\to\infty} \left|\frac{x_{k+1} - x^*}{(x_k - x^*)^2}\right| = \left|\frac{f''(x^*)}{2f'(x^*)}\right|$$

♡

Proof

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \quad \varphi'(x) = \frac{f(x^*) f''(x^*)}{[f'(x^*)]^2} = 0$$

$$\varphi''(x) = \begin{cases} \dfrac{f''(x^*)}{f'(x^*)} & \text{if } f''(x^*) \neq 0 \\ 0 & \text{if } f''(x^*) = 0 \end{cases}$$

$$\left|\frac{e_{k+1}}{e_k p}\right| = \left|\frac{x^* - x_{k+1}}{(x^* - x_k)^p}\right| \Rightarrow \frac{|\varphi^{(p)}(x^*)|}{p!} = \left|\frac{f''(x^*)}{2f'(x^*)}\right|$$

3.4.3 Newton's Method for finding Multiple Roots

$$f(x) = (x - x^*)^m p(x)$$

1.

$$
\begin{aligned}
&[f(x)]^{\frac{1}{m}}=\left(x-x^{*}\right)[p(x)]^{\frac{1}{m}}=0 \\
&\text{let } g(x)=[f(x)]^{\frac{1}{m}} \Rightarrow g^{\prime}(x)=\frac{1}{m}[f(x)]^{\frac{1}{m}-1} f^{\prime}(x) \\
&x_{n+1}=x_{n}-\frac{[f(x)]^{\frac{1}{m}}}{\frac{1}{m}[f(x)]^{\frac{1}{m}-1} f^{\prime}(x)}=x_{n}-\frac{m f\left(x_{n}\right)}{f^{\prime}\left(x_{n}\right)}
\end{aligned}
$$

Downside: Need to know the multiplicity of roots beforehand.

2.

$$
g(x)=\frac{f(x)}{f^{\prime}(x)}=\frac{\left(x-x^{*}\right)^{m} p(x)}{m\left(x-x^{*}\right)^{m-1} p(x)+\left(x-x^{*}\right)^{m} p^{\prime}(x)}=\frac{\left(x-x^{*}\right) p(x)}{m p(x)+\left(x-x^{*}\right) p^{\prime}(x)}
$$

.

Apparently $g^{\prime}\left(x^{*}\right)=\frac{1}{m} \neq 0$. x^{*} is a simple root of $g(x)=0$

$$
x_{k+1}=x_{k}-\frac{g\left(x_{k}\right)}{g^{\prime}\left(x_{k}\right)}=x_{k}-\frac{f\left(x_{k}\right) f^{\prime}\left(x_{k}\right)}{\left[f^{\prime}\left(x_{k}\right)\right]^{2}-f\left(x_{k}\right) f^{\prime \prime}\left(x_{k}\right)}
$$

Advantage. Not necessary to know the zero root multiplicity of $f(x)=0$. in advance and its also applicate to the case of a single root.

Chapter 4 Direct method for Linear System

4.1 Gaussian Elimination

$$\Rightarrow \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{pmatrix} \to \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ & & \ddots & \vdots & \vdots \\ & & & a_{nn}^{(n)} & b_n^{(n)} \end{pmatrix} \text{(Forward Elimination)}$$

$$\Rightarrow \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{pmatrix} \quad \text{(Backward substitution)}$$

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}, x_i = \frac{b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j}{a_{ii}^{(i)}}, i = n-1, \ldots, 2, 1$$

Theorem 4.1

If all the leading principal minors of the coefficient matrix A are non-zero, then Gaussian elimination can proceed sequentially, resulting in a unique solution.

Remark In fact, as long as A is non-singular, meaning A is invertible, the system of equations can be transformed into a triangular system through stepwise elimination and row exchanges, allowing the unique solution to be determined.

Remark The computational complexity of Gaussian elimination

$$\frac{1}{3}\left(n^3 + 3n^2 - n\right)$$

Proposition 4.1 (Column Pivoting Elimination)

Choosing a column pivot before each round of elimination

1. *Note* $|a_{i_1,1}| = \max\limits_{1 \leqslant i \leqslant n} |a_{i,1}|$, *To perform transformations on the augmented matrix.* $r_{i_1} \leftrightarrow r_1$
2. *Note* $\left|a_{i_2,2}^{(2)}\right| = \max\limits_{2 \leqslant i \leqslant n} \left|a_{i2}^{(2)}\right|$, *To perform transformations on the augmented matrix.* $r_{i_2} \leftrightarrow r_2$
3. *Note* $\left|a_{i_k,k}^{(k)}\right| = \max\limits_{k \leqslant i \leqslant n} \left|a_{ik}^{(k)}\right|$, *To perform transformations on the augmented matrix.* $r_{i_k} \leftrightarrow r_k$

Remark *It does not alter the solutions of the system of equations, while effectively overcoming the shortcomings of the Gaussian elimination method.*

4.2 Doolittle Decomposition

Definition 4.1

To decompose a non-singular matrix A into the product of a lower triangular matrix L and an upper triangular matrix U, i.e., $A = LU$, is known as the triangular decomposition or LU decomposition of matrix A

Remark *L is a unit lower triangular matrix, and U is a general upper triangular matrix in the triangular decomposition, known as the Doolittle decomposition.*

♣

Theorem 4.2

Let A be an n-order square matrix. If all the leading principal minors of A are non-zero, then A can be uniquely decomposed into the product of a unit lower triangular matrix L and an upper triangular matrix U.

♡

Proposition 4.2

1. *Calculate the elements in the first row of U: $u_{1j} = a_{1j}, \quad j = 1, 2, \cdots, n$.*
2. *Calculate the elements in the first column of L: $l_{i1} = a_{i1}/u_{11}, \quad i = 2, 3, \cdots, n$.*
3. *Calculate the elements in the i-th row of U for $i = 2, 3, \cdots, n$:*

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \quad j = i, i+1, \cdots, n$$

.

4. *Calculate the elements in the i-th column of L for $i = 2, 3, \cdots, n$:*

$$l_{ji} = \frac{\left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki}\right)}{u_{ii}}, \quad j = i+1, i+2, \cdots, n, \quad i \neq n$$

.

♠

Example 4.1 Solve the system of equations using the Doolittle decomposition method:

$$\begin{cases} 2x_1 + x_2 + 2x_3 = 6 \\ 4x_1 + 5x_2 + 4x_3 = 18 \\ 6x_1 - 3x_2 + 5x_3 = 5 \end{cases}$$

Solution

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -2 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 2 & 1 & 2 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$\Rightarrow$ *Solve the equation $Ly = b$, resulting in $y = (6, 6, -1)^{\mathrm{T}}$.*

$\Rightarrow$ *Solve the equation $Ux = y$, resulting in $x = (1, 2, 1)^{\mathrm{T}}$.*

4.3 Square Root Method

Definition 4.2

Given an n-order real symmetric matrix A, for any non-zero vector x of length n, the condition $x^{\mathrm{T}}Ax > 0$ always holds, then matrix A is called a symmetric positive definite matrix. ♣

Proposition 4.3

Testing method:

1. *If A is symmetric and all leading principal minors are greater than 0, then A is a symmetric positive definite matrix.*
2. *If A is symmetric and all eigenvalues are greater than 0, then A is a symmetric positive definite matrix.* ♠

Theorem 4.3

If A is a symmetric positive definite matrix, then there exists a non-singular lower triangular matrix G such that $A = GG^{\mathrm{T}}$. ♡

Proof

$$A = LU \quad L = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix}, U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix} \text{ and } u_{ii} > 0$$

$$A = LD\overline{U} \quad D = \begin{pmatrix} u_{11} & & & \\ & u_{22} & & \\ & & \ddots & \\ & & & u_{nn} \end{pmatrix}, \overline{U} = \begin{pmatrix} 1 & u_{12}/u_{11} & \cdots & u_{1n}/u_{11} \\ & 1 & \cdots & u_{2n}/u_{22} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}$$

$$A = A^{\mathrm{T}} \Rightarrow LD\overline{U} = \overline{U}^{\mathrm{T}} DL^{\mathrm{T}}$$

$$\text{Decomposition Uniqueness } \Rightarrow \mathrm{L}^{\mathrm{T}} = \overline{U}$$

$$\Rightarrow A = LDL^{\mathrm{T}}$$

$$\Rightarrow D = D^{\frac{1}{2}}D^{\frac{1}{2}} \quad \text{where,} D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{u_{11}} & & & \\ & \sqrt{u_{22}} & & \\ & & \ddots & \\ & & & \sqrt{u_{nn}} \end{pmatrix}$$

$$A = LD^{\frac{1}{2}}D^{\frac{1}{2}}L^{\mathrm{T}} = LD^{\frac{1}{2}}\left(D^{\frac{1}{2}}\right)^{\mathrm{T}}L^{\mathrm{T}} = LD^{\frac{1}{2}}\left(LD^{\frac{1}{2}}\right)^{\mathrm{T}}$$

denote $G = LD^{\frac{1}{2}}$, G is a non-singular lower triangular matrix , $A = GG^{\mathrm{T}}$

Theorem 4.4

If A is an n-order symmetric positive definite matrix, then there exists a real non-singular lower triangular matrix G such that $A = GG^{\mathrm{T}}$. When the diagonal elements of G are constrained to be positive, this *decomposition is unique, and it is referred to as the Cholesky decomposition of A.*

Proposition 4.4

Direct Triangular Decomposition Method

$$G=\begin{pmatrix} g_{11} & & & \\ g_{21} & g_{22} & & \\ \vdots & \vdots & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{pmatrix} \quad A=\begin{pmatrix} g_{11} & & & \\ g_{21} & g_{22} & & \\ \vdots & \vdots & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{pmatrix}\begin{pmatrix} g_{11} & g_{21} & \cdots & g_{n1} \\ & g_{22} & \cdots & g_{n2} \\ & & \ddots & \vdots \\ & & & g_{nn} \end{pmatrix}$$

$$\text{matrix multiply} \Longrightarrow a_{ij}=\sum_{k=1}^{j-1} g_{ik}g_{jk}+g_{jj}g_{ij}$$

$$\Longrightarrow \begin{cases} g_{jj}=\sqrt{a_{jj}-\sum\limits_{k=1}^{j-1} g_{jk}^2}, j=1,2,\ldots,n \\ g_{ij}=\left(a_{ij}-\sum\limits_{k=1}^{j-1} g_{ik}g_{jk}\right)/g_{ij}, i=j+1,\ldots,n \end{cases}$$

Example 4.2 Solve the system of equations using the Square Root Method:

$$\begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix} x=\begin{pmatrix} 4 \\ 6 \\ 7.25 \end{pmatrix}$$

Solution *A is symmetric positive defined* $\quad A=GG^{\mathrm{T}}=\begin{pmatrix} 2 & & \\ -0.5 & 2 & \\ 0.5 & 1.5 & 1 \end{pmatrix}\begin{pmatrix} 2 & -0.5 & 0.5 \\ & 2 & 1.5 \\ & & 1 \end{pmatrix}$

Solve $Gy=\mathrm{b} \Longrightarrow y=(2,3.5,1)^{\mathrm{T}}$
Solve $G^{\mathrm{T}}x=y \Longrightarrow x=(1,1,1)^{\mathrm{T}}$

4.4 Tridiagonal matrix algorithm

Definition 4.3 (Diagonally dominant Matrices)

$A=(a_{ij})_{n\times n}$

1. *If the elements of A satisfy $|a_{ii}|>\sum\limits_{\substack{j=1\\j\neq i}}^{n}|a_{ij}|$ for each $i=1,2,\ldots,n$, then A is called a strictly diagonally dominant matrix.*
2. *If the elements of A satisfy $|a_{ii}|\geqslant\sum\limits_{\substack{j=1\\j\neq i}}^{n}|a_{ij}|$ and at least one of these inequalities holds strictly for each $i=1,2,\ldots,n$, then A is called a weakly diagonally dominant matrix.*

Definition 4.4 (diagonally dominant tridiagonal matrix)

The system of equations $Ax = d$ for a diagonally dominant tridiagonal matrix is given by:

$$\begin{bmatrix} a_1 & b_1 & & & & \\ c_2 & a_2 & b_2 & & & \\ & c_3 & a_3 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & & c_n & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix}$$

Conditions:

1. $|a_1| > |b_1|$
2. $|a_i| \geqslant |b_i| + |c_i|, \quad b_i \cdot c_i \neq 0, \quad i = 2, \dots, n-1$
3. $|a_n| > |c_n|$

♣

Proposition 4.5

The matrix A can be decomposed into Doolittle form: $A = LU$,

$$L = \begin{bmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & l_3 & 1 & & \\ & & \ddots & \ddots & \\ & & & l_n & 1 \end{bmatrix}, \quad U = \begin{bmatrix} u_1 & b_1 & & & \\ & u_2 & b_2 & & \\ & & u_3 & b_3 & \\ & & & \ddots & \ddots \\ & & & & u_n \end{bmatrix}$$

where $u_1 = a_1$, $l_i = \frac{c_i}{u_{i-1}}$, and $u_i = a_i - l_i b_{i-1}$, for $i = 2, 3, \dots, n$.

♠

Example 4.3 Solve the tridiagonal system of equations using the Thomas algorithm:

$$\begin{bmatrix} 3 & 1 & & \\ 2 & 3 & 1 & \\ & 2 & 3 & 1 \\ & & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Solution *Perform the LU decomposition $A = LU$*

$$L = \begin{bmatrix} 1 & & & \\ 2/3 & 1 & & \\ & 6/7 & 1 & \\ & & 7/15 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 3 & 1 & & \\ & 7/3 & 1 & \\ & & 15/7 & 1 \\ & & & 38/15 \end{bmatrix}$$

Solving $Ly = d$, where $\begin{bmatrix} 1 & & & \\ 2/3 & 1 & & \\ & 6/7 & 1 & \\ & & 7/15 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$,

we obtain $y_1 = 1, y_2 = -2/3, y_3 = 11/7, y_4 = -11/15$.

Solving $Ux = y$, where $\begin{bmatrix} 3 & 1 & & \\ & 7/3 & 1 & \\ & & 15/7 & 1 \\ & & & 38/15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ -2/3 \\ 11/7 \\ -11/15 \end{bmatrix}$,
we obtain $x_4 = -11/38, x_3 = 33/38, x_2 = -25/38, x_1 = 21/38$.

Chapter 5 Iterative method for Linear System

To be continued

Chapter 6 Interpolation

6.1 Lagrange Interpolation

Theorem 6.1 (Lagrange Interpolation)

Given $n+1$ distinct points $x_0, \ldots, x_n \in [a, b]$ and $n+1$ values $y_0, \ldots, y_n \in \mathbb{R}$, there exists a unique polynomial $p_n \in P_n$ with the property

$$p_n(x_j) = y_j, \quad j = 0, \ldots, n.$$

In the Lagrange representation, this interpolation polynomial is given by

$$p_n = \sum_{k=0}^{n} y_k \ell_k$$

with the Lagrange factors

$$\ell_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}, \quad k = 0, \ldots, n$$

♡

Remark

$$\ell_i(x_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad i, j = 0, 1, \cdots, n$$

Theorem 6.2 (Remainder of Lagrange interpolation)

If $x_0, x_1, x_2, \ldots x_n$ are $(n+1)$ distinct points in $[a, b]$ and $f \in C^{n+1}[a, b]$, then for any $x \in [a, b]$, there exists $\xi(x) \in (a, b)$, s.t.

$$R_n(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{k=0}^{n} (x - x_k).$$

♡

Proof since $P_n(x_i) = f(x_i), i = 0, 1, 2, \ldots n$.

i.e.

$$R_n(x_i) = f(x_i) - P_n(x_i) = 0$$

Suppose

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

let

$$\varphi(t) = f(t) - P_n(t) - k(x)(t - x_0)(t - x_1) \ldots (t - x_n)$$

$t = x_i (i = 1, 2, \ldots n)$ are zero points of $\varphi(t)$.
By generalized Rolle's theorem, there exists $\xi(x) \in (a, b)$.
s.t.

$$\varphi^{(n+1)}(\xi(x)) = 0$$

where

$$\varphi^{(n+1)}(t) = f^{(n+1)}(t) - k(x)(n+1)!$$

let $t = \xi(x)$

$$k(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}$$

i.e.

$$R_n(x) = k(x) \prod_{k=0}^{n} (x - x_k)$$

6.2 Newton Interpolation

Suppose $x_0, x_1 \ldots x_n$ are $(n+1)$ distinct points,
Construct $P_n(x)$ satisfy

$$P_n(x_i) = f(x_i)$$

$$P_n(x) = \sum_{i=0}^{n} \left(a_i \prod_{k=0}^{i-1} (x - x_k) \right)$$

$$\begin{cases} p_n(x_0) = a_0 & = y_0 \\ p_n(x_1) = a_0 + a_1(x_1 - x_0) & = y_1 \\ p_n(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) & = y_2 \\ \cdots & \cdots \\ p_n(x_n) = a_0 + a_1(x_n - x_0) + \cdots + a_n(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}) & = y_n \end{cases}$$

$$p_n(x) = f(x_0) + \sum_{i=1}^{n} \left(f[x_0, x_1, \cdots, x_i] \prod_{k=0}^{i-1} (x - x_k) \right)$$

where

$$f[x_0, x_1, x_2, \cdots, x_i] = \frac{f[x_1, x_2, \cdots, x_i] - f[x_0, x_1, \cdots x_{i-1}]}{x_i - x_0}$$

x_k	$f(x_k)$	1st	2nd	3rd	4th
x_0	$f(x_0)$				
x_1	$f(x_1)$	$f[x_0, x_1]$			
x_2	$f(x_2)$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
x_3	$f(x_3)$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$	
x_4	$f(x_4)$	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$f[x_0, x_1, x_2, x_3, x_4]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Example 6.1 Find $P_4(x)$ which passes through $(1, 0), (2, 2), (4, 12)(5, 20).\ (6, 70)$

Solution *make divided-difference form*

x_i	$f(x_i)$	1st	2rd	3rd	4th
1	0				
2	2	2			
4	12	5	1		
5	20	8	1	0	
6	70	50	21	5	1

$$\begin{aligned} P_4(x) &= 0+2(x-1)+1(x-1)(x-2)+0+1(x-1)(x-2)(x-4)(x-5) \\ &= x^4+2x^3+50x^2-79x+40 \end{aligned}$$

Theorem 6.3 (Remainder of Newton Interpolation)

$$R_n(x)=f(x)-p_n(x)=f\left[x_0,x_1\ldots x_n,x\right)]\prod_{k=0}^{n}(x-x_k).$$

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

6.3 Piecewise Polynomial Interpolation

Definition 6.1 (Piecewise Polynomial Interpolation)

$$\varphi(x)=\begin{cases} \varphi_0(x), & x\in[x_0,x_1] \\ \varphi_1(x), & x\in[x_1,x_2] \\ \vdots & \\ \varphi_{n-1}(x), & x\in[x_{n-1},x_n] \end{cases}$$

if $\varphi(x)$ satisfies following conditions:

1. $\varphi(x)\in C[a,b]$
2. $\varphi(x_j)=f(x_j), j=0,1,\cdots,n$
3. *in each interval* $[x_k,x_{k+1}]$ $(k=0,1,\cdots,n-1)$, $\varphi_k(x)$ *is linear polynomial*

we call $\varphi(x)$ Piecewise linear interpolation function

Remark Disadvantages: $varphi(x)$ only continuous and not smooth, derivative does not exist at nodes.

Theorem 6.4 (Remainder of Piecewise Linear Polynomial Interpolation)

Suppose $f\in C^2[a,b]$, *let* $M_2=\max\limits_{a\leqslant x\leqslant b}\left|f''(x)\right|$ *for any* $x\in[a,b]$

According Lagrange interpolation in each $[x_k,x_{k+1}]$.

$$\begin{aligned} |R_1(x)|=|f(x)-\varphi_k(x)| &= \left|\frac{1}{2}f''(\xi)(x-x_k)(x-x_{k+1})\right| \\ &\leqslant \frac{1}{8}\left|f''(\xi)\right|h_k^2 \hookrightarrow x_{k+1}-x_k \end{aligned}$$

let $h=\max h_k$, *then in* $[a,b]$

$$\max_{a\leqslant x\leqslant b}|f(x)-\varphi(x)|\leqslant\frac{M_2}{8}h^2$$

6.4 Hermite Interpolation

3.3. suppose $(n+1)$ distinct point $x_0, x_1, x_2, \ldots . x_n$.
interpolating condition $f\left(x_i\right)=y_i \quad f'\left(x_i\right)=m_i$
in each $[x_{i-1}, X_i]$, there hold 4 conditions. which can determine a 3-rod-degree poly nominal.

$$H_i(x)=\varphi_{i-1}(x) y_{i-1}+\varphi_i(x) y_i+\psi_{i-1}(x) m_{i-1}+\psi_i(x) m_i .$$

$$\varphi_{i-1}\left(x_{i-1}\right)=1 \quad \varphi_{i-1}\left(x_i\right)=0. \quad \varphi_{i-1}'\left(x_{i-1}\right)=0 \quad \varphi_{i-1}'\left(x_i\right)=0.$$

$$\varphi_i\left(x_{i-1}\right)=0 \quad \psi_i\left(x_i\right)=1 \quad \varphi_i'\left(x_{i-1}\right)=0 \quad \varphi_i'\left(x_i\right)=0.$$

$$\psi_{i-1}\left(x_{i-1}\right)=0 \quad \varphi_{i-1}\left(x_i\right)=0. \quad \psi_{i-1}'\left(x_{i-1}\right)=1 \quad \psi_{i-1}'\left(x_i\right)=0.$$

$$\psi_i\left(x_{i-1}\right)=0 \quad \psi_i\left(x_i\right)=u \quad \psi_i'\left(x_{i-1}\right)=0, \quad \psi_i'\left(x_i\right)=1$$

$$\varphi_{i-1}(x)=(kx+b)\left(x x_{i-1}\right)^2 .$$

Chapter 7 Curve Fitting

7.1 Least-square Method

For $(x_k, y_k) \quad k=1,2,\dots m$ to construct

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \dots a_n x^n (m \gg n)$$

by satisfying

$$\min Q = \sum_{k=1}^{m} \left| p_n(x_k) - y_k \right|^2$$

$$Q(a_0, a_1, a_2, \dots a_n) = \sum_{k=1}^{m} \left(a_0 + a_1 x_k + a_2 x_k^2 + \dots a_n x_k^n - y_k \right)^2$$

$$\min Q \Leftrightarrow \begin{cases} 0 = \frac{\partial Q}{\partial a_0} = 2 \sum\limits_{k=1}^{m} \left(a_0 + a_1 x_k + \dots a_1 x_k^n - y_k \right) \\ 0 = \frac{\partial Q}{\partial a_1} = 2 \sum\limits_{k=1}^{m} \left(a_0 + a_1 x_k + a_2 x_k^2 + \dots a_n x_k^n - y_k \right) \cdot x_k \\ \vdots \\ 0 = \frac{\partial Q}{\partial a_n} = 2 \sum\limits_{k=1}^{m} \left(a_0 + a_1 x_k + a_2 x_k^2 + \dots a_n x_k^n - y_k \right) \cdot x_k^n. \end{cases}$$

Normal equation is:

$$\begin{cases} \left(\sum\limits_{k=1}^{m} 1 \right) a_0 + \left(\sum\limits_{k=1}^{m} x_k \right) a_1 + \left(\sum\limits_{k=1}^{m} x_k^2 \right) a_2 + \dots \left(\sum\limits_{k=1}^{m} x_k^n \right) a_n = \sum\limits_{k=1}^{m} y_k \\ \left(\sum\limits_{k=1}^{m} x_k \right) a_0 + \left(\sum\limits_{k=1}^{m} x_k^2 \right) a_1 + \left(\sum\limits_{k=1}^{m} x_k^3 \right) a_2 + \dots \left(\sum\limits_{k=1}^{m} x_k^{n+1} \right) a_n = \sum\limits_{k=1}^{m} x_k y_k \\ \left(\sum\limits_{k=1}^{m} x_k^2 \right) a_0 + \left(\sum\limits_{k=1}^{m} x_k^3 \right) a_1 + \left(\sum\limits_{k=1}^{m} x_k^4 \right) a_2 + \dots \left(\sum\limits_{k=1}^{m} x_k^{n+2} \right) a_n = \sum\limits_{k=1}^{m} x_k^2 y_k \\ \vdots \\ \left(\sum\limits_{k=1}^{m} x_k^n \right) a_0 + \left(\sum\limits_{k=1}^{m} x_k^{n+1} \right) a_1 + \left(\sum\limits_{k=1}^{m} x_k^{n+2} \right) a_2 + \dots \left(\sum\limits_{k=1}^{m} x_k^{2n} \right) a_n = \sum\limits_{k=1}^{m} x_k^n y_k \end{cases}$$

The Matrix Form is:

$$\begin{pmatrix} \sum\limits_{k=1}^{m} x_k^0 & \sum\limits_{k=1}^{m} x_k^1 & \sum\limits_{k=1}^{m} x_k^2 & \dots & \sum\limits_{k=1}^{m} x_k^n \\ \sum\limits_{k=1}^{m} x_k^1 & \sum\limits_{k=1}^{m} x_k^2 & \sum\limits_{k=1}^{m} x_k^3 & \dots & \sum\limits_{k=1}^{m} x_k^{n+1} \\ \vdots & & & & \\ \sum\limits_{k=1}^{m} x_k^n & \sum\limits_{k=1}^{m} x_k^{n+1} & \sum\limits_{k=1}^{m} x_k^{n+2} & \dots & \sum\limits_{k=1}^{m} x_k^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^{m} y_k \\ \sum\limits_{i=1}^{m} x_k y_k \\ \vdots \\ \sum\limits_{i=1}^{m} x_k^n y_k \end{pmatrix}$$

Alternatively, it can also be represented as the following matrix form:

$$X^\top X a = X^\top y$$

where

$$X=\begin{pmatrix}1 & x_0 & x_0^2 & \dots & x_0^n\\ 1 & x_1 & x_1^2 & \dots & x_1^n\\ 1 & x_2 & x_2^2 & \dots & x_2^n\\ \vdots & & & & \\ 1 & x_m & x_m^2 & \dots & x_m^n\end{pmatrix}, c=\begin{pmatrix}c_0\\ c_1\\ c_2\\ \vdots\\ c_n\end{pmatrix}, y=\begin{pmatrix}y_0\\ y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}$$

Thus by solving the normal equation, we can get the coefficients $a_0, a_1, \dots a_n$ of the polynomial $P_n(x)$.

Theorem 7.1

The least-square method is the only method that can be used to solve the linear regression problem. ♡

Example 7.1 Use $P_1(x) = a_0 + a_1 x$ to fit

x_k	1	2	3	4
y_k	4	10	18	26

Solution

$$\textit{minimize}\quad Q\left(a_0, a_1\right)=\sum_{k=1}^{4}\left|p_1\left(x_k\right)-y_k\right|^2=\sum_{k=1}^{4}\left|a_0+a_1 x-y_k\right|^2$$

$\Longleftrightarrow$ *to solve*

$$\begin{pmatrix}\sum\limits_{k=1}^{4} 1 & \sum\limits_{k=1}^{4} x_k\\ \sum\limits_{k=1}^{4} x_k & \sum\limits_{k=1}^{4} x_k^2\end{pmatrix}\begin{pmatrix}a_0\\ a_1\end{pmatrix}=\begin{pmatrix}\sum\limits_{k=1}^{\infty} y_k\\ \sum\limits_{k=1}^{\infty} x_k y_k\end{pmatrix}$$

$$\begin{pmatrix}4 & 10\\ 10 & 30\end{pmatrix}\begin{pmatrix}a_0\\ a_1\end{pmatrix}=\begin{pmatrix}58\\ 182\end{pmatrix}$$

$$\begin{cases}a_0=-4\\ a_1=7.4\end{cases}$$

Definition 7.1 (contradicting equations)

give such a linear system

$$\begin{cases}a_0+a_1 x_1+\dots a_n x_1^n=y_1\\ a_0+a_1 x_2+\dots a_n x_2^n=y_2\\ \vdots\\ a_0+a_1 x_m+\dots a_n x_m^n=y_m\end{cases}$$

if $n \leqslant m. R(A) \neq R(\bar{A})$. *then the equation is called contradicting equation* ♣

Proposition 7.1

$A_{m\times n}x = b$ *is contradictory equation with* $R(A) = n \ll m$

(1) $A^\top A$ *is symmetric positive definite.*

(2) $A^\top Ax = A^\top b$ *has unique solution*

(3) $Q=\sum\limits_{i=1}^{m}\left(\sum\limits_{j=0}^{n} a_j x_i^j-y_j\right)^2$ *has minimal value at sol of*

$$A^\top Ax = A^\top b.$$

♠

Example 7.2 Find least-squares solution of $\begin{cases} 2x_1 + 4x_2 = 11 \\ 3x_1 - 5x_2 = 3 \\ x_1 + 2x_2 = 6 \\ 2x_1 + x_2 = 7 \end{cases}$

Solution

$$A = \begin{pmatrix} 2 & 4 \\ 3 & -5 \\ 1 & 2 \\ 2 & 1 \end{pmatrix} \quad b = \begin{pmatrix} 11 \\ 3 \\ 6 \\ 7 \end{pmatrix}$$

$$A^\top A = \begin{pmatrix} 2 & 3 & 1 & 2 \\ 4 & -5 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ 3 & -5 \\ 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 18 & -3 \\ -3 & 46 \end{pmatrix}$$

$$A^\top b = \begin{pmatrix} 2 & 3 & 1 & 2 \\ 4 & -5 & 2 & 1 \end{pmatrix} \begin{pmatrix} 11 \\ 3 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} 51 \\ 48 \end{pmatrix}$$

$$A^\top A x = A^\top b.$$

Thus

$$\begin{pmatrix} 18 & -3 \\ -3 & 46 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 51 \\ 48 \end{pmatrix}$$

$$\begin{cases} x_1 = \frac{830}{273} \\ x_2 = \frac{113}{91} \end{cases}$$

Example 7.3 use $\varphi(x) = ae^{bx}$ to fit the following points

k	1	2	3	4	5	6
x_k	0.0	0.5	1.0	1.5	2.0	2.5
φ_k	2.0	1.2	0.9	0.6	0.4	0.3

Solution

$$\ln(\varphi(x)) = \ln a + bx$$

then is obviously

Chapter 8 Numerical Differentiation and Integration

Introduction

- ***Numerical differentiation*** *is a method to approximate the derivative of a function $f(x)$ at a point x_0 by a finite difference formula.*

8.1 Numerical Integration

Definition 8.1 (Numerical Integration)

Suppose $a = x_0 < x_1 < \ldots < x_n = b$, and f is integral in $[a, b]$

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i)$$

is called a numerical integration

$$E[f] = \int_a^b f(x)dx - \sum_{i=0}^{n} A_i f(x_i)$$

is called truncation error (remainder) (A_i is called quadrature coefficients / weight) ♣

Definition 8.2 (Degree of Precision)

The degree of precision of $\int_a^b f(x)dx \approx \sum\limits_{i=0}^{n} A_i f(x_i)$ is m

1. *When $f(x) = x^k, k = 0, 1, 2, \ldots m$, $\int_a^b x^k dx = \sum\limits_{i=0}^{m} A_i {x_i}^k$*
2. *When $f(x) = x^{m+1}$, $\int_a^b x^{m+1} dx \neq \sum\limits_{i=0}^{n} A_i x_i^{m+1}$* ♣

Example 8.1 $\int_0^h f(x)dx \approx A_0 f(0) + A_1 f\left(\frac{h}{2}\right) + A_2 f(h)$

Solution

$$f(x) = 1 \Longrightarrow h = \int_0^h 1dx = A_0 + A_1 + A_2$$

$$f(x) = x \Longrightarrow \frac{1}{2}h^2 = \int_0^h xdx = \frac{h}{2}A_1 + hA_2$$

$$f(x) = x^2 \Longrightarrow \frac{1}{3}h^3 = \int_0^h x^2 dx = \frac{1}{4}h^2 A_1 + h^2 A_2$$

$$\begin{cases} A_0 = \frac{h}{6} \\ A_1 = \frac{4}{6}h \\ A_2 = \frac{h}{6} \end{cases}$$

$$\int_a^b f(x)dx \approx \frac{h}{6}\left(f(0) + 4f\left(\frac{h}{2}\right) + f(h)\right)$$

when $f(x) = x^3$ $\quad$ *left* $= \int_0^h x^3 dx = \frac{h^4}{4}$ $\quad$ *right* $= \frac{h}{6}\left(\frac{h^3}{2} + h^3\right) = \frac{h^4}{4}$

when $f(x) = x^4$ $\quad$ *left* $= \int_0^h x^4 dx = \frac{h^5}{5}$ $\quad$ *right* $= \frac{h}{6}\left(\frac{h^4}{4} + h^4\right) = \frac{5}{24}h^5$

Approximate $I = \int_a^b f(x)dx$

first use Lagrange Interpolatory Polynomial $L_n(x)$ to approximate $f(x)$

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

denote. $w_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$

$$l_i(x) = \frac{w_{n+1}(x)}{(x - x_i)\, w'_{n+1}(x_i)}$$

$$\begin{aligned}
\Longrightarrow \int_a^b f(x)dx &\approx \int_a^b L_n(x)dx \\
&= \int_a^b f(x_i)\, l_i(x)dx \\
&= \sum_{i=0}^{n} f(x_i) \int_u^b l_i(x)dx \\
&= \sum_{i=0}^{n} f(x_i)\, A_i
\end{aligned}$$

let

$$\begin{aligned}
A_i &= \int_a^b l_i(x)dx \\
&= \int_a^b \frac{w_{n+1}(x)}{(x - x_i)\, w'_{n+1}(x_i)} dx \\
E[f] &= \int_a^b f(x)dx - \int_a^b L_n(x)dx \\
&= \int_a^b \frac{f^{(n+1)}(\xi_x)}{(n+1)!} w_{n+1}(x)dx
\end{aligned}$$

Definition 8.3

Suppose $a = x_0 < x_1 < \ldots x_n = b$, f *is integral in* $[a, b]$ *and* $f \in C^{n+1}[a, b]$

then

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i)$$

with

$$A_i = \int_a^b l_i(x)dx = \int_a^b \frac{w_{n+1}(x)}{(x - x_i)\, w'_{n+1}(x_i)} dx$$

is called Interpolatory numerical quadrature.

Its truncation error

$$E[f] = \int_a^b \frac{f^{(n+1)}(\xi_x)}{(n+1)!} w_{n+1}(x)dx$$

♣

Theorem 8.1

The Interpolatory numerical quadrature has at least n degrees of precision

Proof The remainder of interpolating numerical quadrature

$$E[f]=\int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}w_{n+1}(x)\,dx$$

when $f(x)=x^k, k=0,1,2,\dots n$

$$f^{(n+1)}(x)=\left[x^k\right]^{(n+1)}=0\Longrightarrow E[f]=0$$

$$\Longrightarrow \int_a^b x^k\,dx=\sum_{i=0}^{n}A_i x_i^k$$

when $f(x)=x^{n+1}$ $\quad f^{(n+1)}(x)=(n+1)\,!$

$$\begin{aligned}E[f]&=\int_a^b \frac{(n+1)!}{(n+1)!}w_{n+1}(x)\,dx\\&=\int_a^b w_{n+1}(x)\,dx\neq 0\end{aligned}$$

8.2 Newton-Cotes formula

8.2.1 Trapezoidal rule

Definition 8.4

$n=1$

$$\begin{aligned}\int_a^b f(x)dx&\approx\int_{x_0}^{x_1}\left[f\left(x_0\right)l_0\left(x\right)+f\left(x_1\right)l_1\left(x\right)\right]dx\\&=f\left(x_0\right)\int_{x_0}^{x_1}\frac{x-x_1}{x_0-x_1}dx+f\left(x_1\right)\int_{x_0}^{x_1}\frac{x-x_0}{x_1-x_0}dx\\&=(b-a)\left(\frac{f(a)+f(b)}{2}\right)\end{aligned}$$

Proposition 8.1

Remainder of Trapezoidal rule is

$$E[f]=\int_a^b\left[f(x)-L_1(x)\right]dx=-\frac{(b-a)^3}{12}f''(\eta)$$

8.2.2 Simpson's rule

Definition 8.5

$n=2$

$$\begin{aligned}
\int_a^b f(x)dx &\approx \int_a^b \left[f\left(x_0\right)l_0\left(x\right)+f\left(x_1\right)l_1\left(x\right)+f\left(x_2\right)l_2(x)dx\right] \\
&= f\left(x_0\right)\int_a^b l_0(x)dx + f\left(x_1\right)\int_a^b l_1(x)dx + f\left(x_2\right)\int_a^b l_2(x)dx \\
&= \frac{b-a}{6}f\left(x_0\right)+\frac{4(b-a)}{6}f\left(x_1\right)+\frac{b-a}{6}f\left(x_2\right)
\end{aligned}$$

Thus

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right)$$

Proposition 8.2

Degree of precision of Simpson's rule is 3

Proposition 8.3

Remainder of Simpson's rule

$$E[f]=\int_a^b [f(x)-L_2(x)]dx = -\frac{1}{90}h^5 f^{(4)}(\eta)$$

8.2.3 Newton-Cotes formula in general

let

$$h=\frac{b-a}{n}, x_i=a+ih$$

It can he written

$$\int_0^2 \frac{2(t-1)h(t-2)h}{(-h)(-2h)}hdt \int_0^2 \frac{th(t-2)h}{h(-h)}hdt \int_0^2 \frac{th(t-1)h}{2hh}hdt$$

Definition 8.6

Suppose $a=x_0<x_1<\ldots<x_n=b$ *are* $(n+1)$ *distinct points with equal division i.e* $x_i=a+ih, x=a+th.$ *then*

$$\begin{aligned}
\int_a^b f(x)dx &\approx \sum_{i=0}^{n} A_i f\left(x_i\right) \\
&= (b-a)\sum_{i=0}^{n} C_i^{(n)} f\left(x_i\right)
\end{aligned}$$

with

$$C_i^{(n)} = \frac{(-1)^{n-i}}{i!(n-i)!n}\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n}(t-j)dt$$

called cotes coefficient

Proof

$$\begin{aligned} A_i &= \int_a^b l_i(x)dx \\ &= \int_a^b \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}dx \\ &= \int_0^n \frac{th(t-1)h\dots(t-(i-1))h(t-(i+1))h\dots(t-n)h}{ih(i-1)h\dots h(-h)(-2)h\dots(-(n-i))h}hdt \\ &= h\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(t-j)}{i!(-1)^{n-i}(n-i)!}dt \\ &= \frac{(-1)^{n-i}(b-a)}{i!(n-i)!n}\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n}(t-j)dt \end{aligned}$$

apparently

$$C_i^{(n)} = \frac{(-1)^{n-i}}{i!(n-i)!n}\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n}(t-j)dt$$

Proposition 8.4

$C_i^{(n)}$: *Cotes coefficient* $\begin{cases} C_i^{(n)} = C_{n-i}^{(n)} \\ \sum\limits_{i=0}^{n} C_i^{(n)} = 1 \end{cases}$

Example 8.2

$n=1 \quad C_0^{(1)} = \frac{1}{2} \quad C_1^{(1)} = \frac{1}{2}$

$n=2 \quad C_0^{(2)} = \frac{1}{6} \quad C_1^{(2)} = \frac{4}{6} \quad C_2^{(2)} = \frac{1}{6}$

$n=3 \quad C_0^{(3)} = \frac{1}{8} \quad C_1^{(3)} = \frac{3}{8} \quad C_2^{(3)} = \frac{3}{8} \quad C_3^{(3)} = \frac{1}{8}$ (Simpson $\frac{3}{8}$ rule)

$n=4 \quad C_0^{(4)} = \frac{7}{90} \quad C_1^{(4)} = \frac{32}{90} \quad C_2^{(4)} = \frac{12}{90} \quad C_3^{(4)} = \frac{32}{90} \quad C_4^{(4)} = \frac{7}{90}$

Cotes rule

Theorem 8.2

The degree of precision of $\int_a^b f(x)dx \approx (b-a)\sum\limits_{i=0}^{n} C_i^{(n)} f(x_i) = \begin{cases} n & n \text{ is odd} \\ n+1 & n \text{ is even} \end{cases}$

Remark $n=3 \quad E[f] = -\frac{8}{495}h^7 f^{(6)}\eta$

8.3 Composite Numerical Integration

8.3.1 Composite Trapezoidal rule

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx (x_{k+1}-x_k)\left[\frac{1}{2}f(x_k)+\frac{1}{2}f(x_{k+1})\right] = \frac{h}{2}(f_k+f_{k+1})$$

$$
\begin{aligned}
\int_a^b f(x)dx &= \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)\,dx \\
&\approx \sum_{k=0}^{n-1} \frac{h}{2}\left(f_k + f_{k+1}\right) \\
&= \frac{h}{2}\left(f_0 + 2\sum_{k=1}^{n-1} f_k + f_n\right)
\end{aligned}
$$

Example 8.3 Consider $f(x) = 2 + \sin(2\sqrt{x})$, Use composite trapezoidal rule with 11 nodes to compute an approximation of $\int_1^6 f(x)dx$

Solution $a = 1, b = 6, n = 10, \quad h = \frac{b-a}{n} = \frac{1}{2}$

$$
\int_1^6 f(x)dx \approx T_n = \frac{h}{2}\left[f_0 + 2\sum_{i=1}^{9} f_i + f_{10}\right]
$$

8.3.2 Composite Simpson's rule

$$
\begin{aligned}
\int_{x_k}^{x_{k+1}} f(x)dx &= (x_{k+1} - x_k)\left[\frac{1}{6}f(x_k) + \frac{4}{6}f\left(x_{k+\frac{1}{2}}\right) + \frac{1}{6}f(x_{k+1})\right] \\
&= \frac{h}{6}\left[f_k + 4f_{k+\frac{1}{2}} + f_{k+1}\right] \\
\int_a^b f(x)dx &= \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)dx \\
&= \frac{h}{6}\sum_{k=0}^{n-1}\left[f_k + 4f_{k+\frac{1}{2}} + f_{k+1}\right] \\
&= \frac{h}{6}\left[f_0 + 4\sum_{i=1}^{n} f_{\frac{i}{2}} + 2\sum_{i=1}^{n-1} f_i + f_n\right]
\end{aligned}
$$

Use Another representation, subdivide molecular nodes again

$n = 2m \quad$ subinterval, $h = \frac{b-a}{n} = \frac{b-a}{2m}$

$$
\begin{aligned}
\int_a^b f(x)dx &= \sum_{k=0}^{m-1} \int_{x_{2k}}^{x_{2k+2}} f(x)dx \\
&\approx \sum_{k=0}^{m-1} (x_{2k+2} - x_{2k})\left[\frac{1}{6}f_{2k} + \frac{4}{6}f_{2k+1} + \frac{1}{6}f_{2k+2}\right] \\
&= \frac{h}{3}\left[f_0 + 4\sum_{k=0}^{m-1} f_{2k+1} + 2\sum_{k=1}^{m-1} f_{2k} + f_{2m}\right]
\end{aligned}
$$

8.3.3 Remainder Estimation

Definition 8.7 (Convergence order)

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i)$$

If remainder $R[f] = \int_a^b f(x)dx - \sum_{i=0}^{n} A_i f(x_i)$ *satisfies*

$$\lim_{h \to 0} \frac{R[f]}{h^p} = C, \quad C \neq 0$$

say numerical quadrature is p *th convergent.* ♣

Proposition 8.5

The convergence order of composite Trapezoidal rule is $-\frac{b-a}{12} h^2 f''(\xi)$ ♠

$$\begin{aligned}
R[f] &= -\sum_{k=0}^{n-1} \frac{h^3}{12} f''(\eta_k) \\
&= -\sum_{k=0}^{n-1} \frac{h^2}{12} \frac{b-a}{n} f''(\eta_k) \\
&= -\frac{b-a}{12} h^2 \left(\frac{1}{n} \sum_{k=0}^{n-1} f''(\eta_k) \right) \\
&= -\frac{b-a}{12} h^2 f''(\xi) \\
&\sim O\left(h^2\right)
\end{aligned}$$

Proposition 8.6

The convergence order of composite Simpson 's rule is $-\frac{b-a}{180}\left(\frac{h}{2}\right)^4 \cdot f^{(4)}(\xi)$ ♠

$$\begin{aligned}
R[f] &= -\sum_{k=0}^{n-1} \frac{\left(\frac{h}{2}\right)^5}{90} f^{(4)}(\eta_k) \\
&= -\sum_{k=0}^{n-1} \frac{\left(\frac{h}{2}\right)^4}{90} \frac{b-a}{2n} f^{(4)}(\eta_k) \\
&= -\frac{\left(\frac{h}{2}\right)^4}{90} \frac{b-a}{2} \cdot \frac{1}{n} \sum_{k=0}^{n} f^{(4)}(\eta_k) \\
&= -\frac{b-a}{180} \left(\frac{h}{2}\right)^4 \cdot f^{(4)}(\xi) \\
&\sim O\left(h^4\right)
\end{aligned}$$

8.4 Romberg integration

8.4.1 Recursive trapezoidal rule

$h = \frac{b-a}{n}$

$$T_n = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)dx \approx \sum_{k=0}^{n-1} \frac{h}{2} [f_k + f_{k+1}]$$

$$\begin{aligned} T_{2n} &= \sum_{k=0}^{n-1} \left(\int_{x_k}^{x_{k+\frac{1}{2}}} f(x)dx + \int_{x_{k+\frac{1}{2}}}^{x_{k+1}} f(x)dx \right) \\ &\approx \sum_{k=0}^{n-1} \left(\frac{\frac{h}{2}}{2} \left(f_k + f_{k+\frac{1}{2}} \right) + \frac{\frac{h}{2}}{2} \left(f_{k+\frac{1}{2}} + f_{k+1} \right) \right) \\ &= \sum_{k=0}^{n-1} \frac{h}{4} \left[f_k + 2f_{k+\frac{1}{2}} + f_{k+1} \right] \end{aligned}$$

$$T_{2n} = \frac{1}{2} T_n + \sum_{k=0}^{n-1} \frac{h}{2} f_{k+\frac{1}{2}}$$

$$\frac{R_{T_n}(f)}{R_{T_{2n}}(f)} \approx \frac{4}{1}$$

$$\frac{I - T_n(f)}{I - T_{2n}(f)} \approx \frac{4}{1} \Longrightarrow I \approx \frac{4}{3} T_{2n}(f) - \frac{1}{3} T_n(f)$$

Chapter 9 Numerical method for Ordinary differential equation

Introduction

- *Euler's method*

9.1 The Existence of Solutions to Initial Value Problems

$$\begin{cases} \frac{dy}{dx} = f(x,y), x \in [a,b] \\ y(a) = y_0 \end{cases} \quad \begin{cases} y' = f(x,y) \quad x \in [a,b] \\ y(a) = y_0 \end{cases} \quad \to \text{ numerical solution}$$

Definition 9.1 (Lipschitz condition)

f satisfies a Lipschitz condition in variable y,

$$\textit{if there exists } L > 0, \textit{ st. } \left| f\left(x, y_1\right), f\left(x, y_2\right) \right| \leqslant L \left| y_1 - y_2 \right|$$

Theorem 9.1 (The Existence Theorem of Solutions to Initial Value Problems)

Suppose $D = \{(x,y) \mid a \leqslant x \leqslant b, -\infty < y < +\infty\}$ and f is continuous on D. If f satisfies Lipschitz condition on D in variable y
i.e. $\exists L > 0$, s.t. $\left| f\left(x, y_1\right) - f\left(x, y_2\right) \right| \leqslant L \left| y_1 - y_2 \right|, \forall \left(x, y_1\right), \left(x, y_2\right) \in D$
the IVP has a unique solution $y(x)$

9.2 Euler Method

9.2.1 Euler's method

Proposition 9.1

Euler's method for $\begin{cases} y'(x) = f(x,y) \\ y(a) = y_0 \end{cases}, x \in [a,b]$
step size $h = \frac{b-a}{n}$ node $x_i = a + ih \quad i = 0, 1, 2 \ldots n$.
Thus the iterative scheme is

$$y_{i+1} = y_i + hf(x_i, y_i)$$

Proof

1.

$$\int_{x_0}^{x_1} y'(x) dx = \int_{x_0}^{x_1} f\left(x, y(x)\right) dx$$

$$\Rightarrow y(x_1) - y(x_0) = \int_{x_0}^{x_1} f(x, y(x))\, dx \approx hf(x_0, y(x_0))$$
$$\Rightarrow y(x_1) \approx y(x_0) + hf(x_0, y(x_0))$$

$$y(x_{i+1}) \approx y(x_i) + hf(x_i, y(x_i))$$

$$y_{i+1} = y_i + h(f(x_i, y_i))$$

2. $y'(x) = f(x, y)$

$$y'(x_0) \approx \frac{y(x_1) - y(x_0)}{h} \quad \text{forward divided-difference}$$

$$\Rightarrow hy'(x_0) \approx y(x_1) - y(x_0)$$
$$y(x_1) \approx y(x_0) + hy'(x_0)$$
$$y(x_1) \approx y(x_0) + hf(x_0, y(x_0))$$

Example 9.1 Use Euler's method to approximate $y' = y - x^2 + 1, 0 \leqslant x \leqslant 2, y(0) = 0.5$ with $n = 10$

Proposition 9.2

The local truncation error of Euler method is $O(h^2)$

Proof Suppose $y_i = y(x_i)$

$$\begin{aligned} R_{i+1} &= y(x_{i+1}) - y_{i+1} \\ &= y(x_{i+1}) - [y_i + hf(x_i, y_i)] \\ &= y(x_{i+1}) - [y(x_i) + hf(x_i, y(x_i)] \\ &= y(x_i) + hy'(x_i) + \frac{h^2}{2} y''(\xi_i) - [y(x_i) + hf(x_i, y(x_i))] \\ &= \frac{h^2}{2} y''(\xi_i). \end{aligned}$$

Definition 9.2 (the accuracy of numerical method)

If the local truncation error of one numerical method is $O\left(h^{p+1}\right)$
we call this numerical method has p order accuracy.

Proposition 9.3

1. *The local truncation error of Euler's Method is* $O(h^2)$
2. *The global truncation error of Euler's Method is* $O(h)$
3. *The accuracy of Euler's Method is* 1 *order.*

9.2.2 Implicit Euler's Method

Proposition 9.4

The iterative scheme of Implicit Euler's Method is

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$$

Proof

1.

$$\begin{aligned} y'(x) &= f(x,y) \\ \int_{x_i}^{x_{i+1}} y'(x)dx &= \int_{x_i}^{x_{i+1}} f(x,y(x))dx \\ y(x_{i+1}) - y(x_i) &= \int_{x_i}^{x_{i+1}} f(x,y(x))dx \\ &\approx hf\left(x_{i+1}, y\left(x_{i+1}\right)\right) \\ &\Rightarrow y\left(x_{i+1}\right) = y\left(x_i\right) + hf\left(x_{i+1}, y_{i+1}\right). \\ &\Rightarrow y_{i+1} \approx y_i + hf\left(x_{i+1}, y_{i+1}\right) \end{aligned}$$

2.

$$y'\left(x_{i+1}\right) \approx \frac{y\left(x_{i+1}\right) - y\left(x_i\right)}{h} \quad \text{backward divided-difference}$$

$$\Rightarrow y\left(x_{i+1}\right) \approx y\left(x_i\right) + hy'\left(x_{i+1}\right) = y\left(x_i\right) + hf\left(x_{i+1}, y\left(x_{i+1}\right)\right)$$

$$y_{i+1} = y_i + hf\left(x_{i+1}, y_{i+1}\right)$$

$$\begin{aligned} R_{i+1} &= y\left(x_{i+1}\right) - y_{i+1} \\ &= y\left(x_{i+1}\right) - \left[y\left(x_i\right) + hf\left(x_{i+1}, y_{i+1}\right)\right] \\ &= -\frac{h^2}{2} y''\left(\xi_i\right) \\ &= O\left(h^2\right) \end{aligned}$$

Proposition 9.5

1. *The local truncation error of Implicit Euler's Method is* $O(h^2)$
2. *The global truncation error of Implicit Euler's Method is* $O(h)$
3. *The accuracy of Implicit Euler's Method is* 1 *order.*

9.2.3 Trapezoidal rule

Proposition 9.6

The iterative scheme of Trapezoidal Method is

$$y_{i+1} = y_i + \frac{h}{2}\left[f\left(x_i, y_i\right) + f\left(x_{i+1}, y_{i+1}\right)\right]$$

Proof

$$
\begin{aligned}
y'(x)=f(x,y) \Rightarrow \int_{x_i}^{x_{i+1}} y'(x)dx &= \int_{x_i}^{x_{i+1}} f(x,y(x))dx \\
y\left(x_{i+1}\right)-y\left(x_i\right) &= \int_{x_i}^{x_{i+1}} f(x,y(x))dx \\
&\approx \frac{h}{2}\left[f\left(x_i, y\left(x_i\right)\right)+f\left(x_{i+1}, y\left(x_{i+1}\right)\right]\right. \\
y\left(x_{i+1}\right) &\approx y\left(x_i\right)+\frac{h}{2}\left[f\left(x_i, y\left(x_i\right)\right)+f\left(x_{i+1}, y\left(x_{i+1}\right)\right)\right] \\
y_{i+1} &= y_i+\frac{h}{2}\left[f\left(x_i, y_i\right)+f\left(x_{i+1}, y_{i+1}\right)\right]
\end{aligned}
$$

Proposition 9.7

1. *The local truncation error of Trapezoidal method is* $O(h^3)$
2. *The global truncation error of Trapezoidal method is* $O(h^2)$
3. *The accuracy of Trapezoidal method is* 2 *order.*

Proof

$$
R_{i+1}=y\left(x_{i+1}\right)-y_{i+1}=-\frac{h^3}{12} y'''\left(\xi_i\right)=O\left(h^3\right)
$$

9.2.4 Midpoint rule

Proposition 9.8

The iterative scheme of Midpoint rule is

$$
y_{i+2}=y_i+2 h f\left(x_{i+1}, y_{i+1}\right)
$$

Proof

$$
\begin{aligned}
y'\left(x_{i+1}\right) &\approx \frac{y\left(x_{i+2}\right)-y\left(x_i\right)}{2 h} \\
y\left(x_{i+2}\right) &\approx y\left(x_i\right)+2 h y'\left(x_{i+1}\right) \\
y_{i+2} &= y_i+2 h f\left(x_{i+1}, y_{i+1}\right) \qquad \text{double -step}
\end{aligned}
$$

Proposition 9.9

the iterative scheme of Modified Midpoint rule can also be represented as

$$
y_{i+1}=y_i+h f\left(x_i+\frac{h}{2}, y_i+\frac{h}{2} f\left(x_i, y_i\right)\right)
$$

Proof

$$
\left\{\begin{array}{l}
y\left(x_i+\frac{h}{2}\right) \approx y_i+\frac{h}{2} f\left(x_i, y_i\right) \\
y\left(x_{i+1}\right) \approx y_i+h f\left(x_i+\frac{h}{2}, y\left(x_i+\frac{h}{2}\right)\right)
\end{array}\right.
$$

Proposition 9.10

1. *The local truncation error of Midpoint method is* $O(h^3)$
2. *The global truncation error of Midpoint method is* $O(h^2)$
3. *The accuracy of Midpoint method is* 2 *order.*

9.2.5 Modified Euler's method(Predictor-Corrector method)

Proposition 9.11 (Iterative scheme of Modified Euler's method)

$$\begin{cases} \overline{y_{i+1}} = y_i + hf\left(x_i, y_i\right) & \textit{Euler method} \\ y_{i+1} = y_i + \frac{h}{2}\left[f\left(x_i, y_2\right) + f\left(x_{i+1}, \overline{y_{i+1}}\right)\right] & \textit{Trapezoidal method} \end{cases}$$

Thus one of the iterative schemes of Modified Euler's method is

$$y_{i+1} = y_i + \frac{h}{2}\left[f\left(x_i, y_i\right) + f\left(x_{i+1}, y_i + hf\left(x_i, y_i\right)\right)\right]$$

Proposition 9.12 (Predictor-Corrector Scheme)

$$\begin{cases} y_{i+1}^p = y_i + hf\left(x_i, y_i\right) \\ y_{i+1}^c = y_{i+} + hf\left(x_{i+1}, y_{i+1}^p\right) \\ y_{i+1} = \frac{1}{2}\left(y_{i+1}^p + y_{i+1}^c\right) \end{cases}$$

Remark These two iterative methods are essentially the same.

Proposition 9.13

1. *The local truncation error of Modified Euler's method is* $O(h^3)$
2. *The global truncation error of Modified Euler's method is* $O(h^2)$
3. *The accuracy of Modified Euler's method is* 2 *order.*

Example 9.2 Use predictor-corrector method to approximate

$$\begin{cases} \frac{du}{dt} = u - \frac{2t}{u}, t \in [0.14], \quad h = 0.1 \\ u(0) = 1 \end{cases}$$

9.3 Runge-kutta method

Proposition 9.14 (Runge - kutta scheme)

Runge - kutta scheme is as follow

$$\begin{cases} y_{i+1} = y_i + h\left[\lambda_1 k_1 + \lambda_2 k_2\right] \\ k_1 = f\left(x_i, y_i\right) = y'\left(x_i\right) \\ k_2 = f\left(x_i + ph, y_i + phk_1\right) \end{cases}$$

determine λ_1, λ_2, p to satisfy accuracy 2 or local truncation error $O\left(h^3\right)$

suppose $y\left(x_i\right) = y_i \quad R_{i+1} = y\left(x_{i+1}\right) - y_{i+1}$

Taylor expansion

$$\begin{aligned} f(x+h, y+k) =& f(x, y) + \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) f(x, y) + \frac{\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2}{2!} f(x, y) + \ldots \\ &+ \frac{\left(h\frac{\partial}{\partial x} + k\frac{d}{\partial y}\right)^n}{n!} f(x, y) + \frac{\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^{n+1}}{(n+1)!} f(x + \theta h, y + \theta k) \end{aligned}$$

$$
\begin{aligned}
k_2 &= f\left(x_i + ph, y_i + phk_1\right) \\
&= f\left(x_i, y_i\right) + phf_x\left(x_i, y_i\right) + phk_1 f_y\left(x_i, y_i\right) + O\left(h^3\right) \\
&= f\left(x_i, y_i\right) + phf''\left(x_i, y_i\right) + O\left(h^3\right)
\end{aligned}
$$

$$
\begin{aligned}
y_{i+1} &= y_i + h\left[\lambda_1 k_1 + \lambda_2 k_2\right] \\
&= y_i + h\left[\lambda_1 y\left(x_i\right) + \lambda_2 y\left(x_i\right) + \lambda_2 phy''\left(x_i\right) + \lambda_2 O\left(h^2\right)\right] \\
&= y_i + \left(\lambda_1 + \lambda_2\right) hy'\left(x_i\right) + \lambda_2 ph^2 y''\left(x_i\right) \cdot O\left(h^3\right)
\end{aligned}
$$

$$
\begin{aligned}
y\left(x_{i+1}\right) &= y\left(x_i\right) + hy'\left(x_i\right) + \frac{h^2}{2} y''\left(x_i\right) + O\left(h^3\right) \\
R_{i+1} &= y\left(x_{i+1}\right) - y_{i+1} = O\left(h^3\right) \\
&\Leftrightarrow \begin{cases} \lambda_1 + \lambda_2 = 1 \\ \lambda_2 p = \frac{1}{2} \end{cases}
\end{aligned}
$$

Remark Modified Euler's method, is also 2nd order Runge-kutta scheme.

Proof $\lambda_1 = \lambda_2 = \frac{1}{2}, p = 1$

$$
\begin{cases}
y_{i+1} = y_i + h\left[\dfrac{1}{2} k_1 + \dfrac{1}{2} k_2\right] \\
k_1 = f\left(x_i, y_i\right) \\
k_2 = f\left(x_i + h, y_i + hk_1\right)
\end{cases}
\Leftrightarrow y_{i+1} = y_i + \frac{h}{2}\left[f\left(x_i, y_i\right) + f\left(x_{i+1}, y_i + hf\left(x_i, y_i\right)\right)\right]
$$

Remark Midpoint method is a 2nd order Runge-kutta method

Proof $\lambda_2 = 1 \quad \lambda_1 = 0. \quad p = \frac{1}{2}$

$$
\begin{cases}
y_{i+1} = y_i + hk_2 \\
k_1 = f\left(x_i, y_i\right) \\
k_2 = f\left(x_i + \dfrac{1}{2} h, y_i + \dfrac{1}{2} hk_1\right)
\end{cases}
\Leftrightarrow y_{i+1} = y_i + hf\left(x_i + \frac{1}{2} h, \quad y_i + \frac{1}{2} hf\left(x_i, y_i\right)\right)
$$

Proposition 9.15 (4 order Runge-kutta scheme)

$$
\begin{cases}
y_{i+1} = y_i + \dfrac{h}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right] \\
k_1 = f\left(x_i, y_i\right) \\
k_2 = f\left(x_i + \dfrac{h}{2}, y_i + \dfrac{h}{2} k_1\right) \\
k_3 = f\left(x_i + \dfrac{h}{2}, y_i + \dfrac{h}{2} k_2\right) \\
k_4 = f\left(x_i + h, y_i + hk_3\right)
\end{cases}
$$

9.4 Convergence of methods

Definition 9.3

A one step method is said to be convergent with respect to the differential equation it approximates if

$$\lim_{h\to 0}\max_{1\leqslant i\leqslant n}\left|y_i-y\left(x_i\right)\right|=0$$

♣

Definition 9.4

A one-step method is stable with the results depend continuously on the initial data. ♣

Example 9.3 show Euler's method for $\begin{cases} y'=\lambda y & x\in[0,1] \\ y(0)=y_0 \end{cases}$ is convergent

Proof The exact solution $y(x)=y_0e^{\lambda x}, \quad y\left(x_i\right)=y_0e^{\lambda x_i}$

By Euler's method

$$\begin{aligned}
y_{i+1} &= y_i+hf\left(x_i,y_i\right) \\
&= y_i+h\lambda y_i \\
&= (1+\lambda h)y_i \\
y_1 &= (1+\lambda h)y_0 \\
y_2 &= (1+\lambda h)y_1=(1+\lambda h)^2y_0 \\
y_i &= (1+\lambda h)^i y_0=(1+\lambda h)^{\frac{x_i}{n}}y_0 \\
&= \left((1+\lambda h)^{\frac{1}{\lambda h}}\right)^{\lambda x_i}y_0 \\
&\to y_0e^{\lambda x_i} \quad h\to 0
\end{aligned}$$

Explicit Euler's method

$$\begin{aligned}
&y_{i+1}=y_i+h\lambda y_i=(1+h\lambda)y_i=(1+\bar{h})y_i=(1+\bar{h})^{i+1}y_0 \\
&\varepsilon_0=y_0-\overline{y_0}\Rightarrow\varepsilon_{i+1}=y_{i+1}-\overline{y_{i+1}}=(1+\bar{h})^{i+1}\varepsilon_0 \quad \Rightarrow|1+\bar{h}|<1
\end{aligned}$$

Implicit Euler's method.

$$\begin{aligned}
y_{i+1}=y_i+h\lambda y_{i+1}\Rightarrow y_{i+1}=\frac{1}{1-\bar{h}}y_i=\left(\frac{1}{1-\bar{h}}\right)^{i+1}y_0 \\
\varepsilon_0=y_0-\overline{y_0}\Rightarrow\varepsilon_{i+1}=\left(\frac{1}{1-\bar{h}}\right)^{i+1}\varepsilon_0 \quad \Rightarrow|1-\bar{h}|>1
\end{aligned}$$