1 Introduction

1.1 Motivation

Before going to study the course, let us try to motivate ourselves by asking the following five **Wh** questions.

1. **What** is Numerical Analysis ?
2. **Where** is Numerical Analysis used ?
3. **When** is Numerical Analysis required ?
4. **Who** requires Numerical Analysis ?
5. **Why** should we learn Numerical Analysis ?

And, finally we ask the following:

How to do Numerical Analysis ?

1. **What** is Numerical Analysis ?

   Numerical Analysis is a branch of mathematics that deals with

   (a) Developing approximation procedure (called Numerical Methods) to solve mathematical problems.

(b) Developing efficient algorithm to implement the above approximation procedure as computer codes (called Implementation).

(c) Analyzing error for the above approximation procedure (a bit of error analysis, convergence, stability).

Numerical Analysis is not about playing with numerical values in computer. It is about mathematical ideas, insights.

What are Numerical methods?

They are algorithms to compute:

(i) approximations of functions,

(ii) approximations of derivatives of functions at some point,

(iii) approximations of integrals of functions in some interval,

(iv) approximations of solution/s of linear/nonlinear single/system of algebraic/differential equation/s.

2. **Where** is Numerical Analysis used ?
(I will put a picture describing whole situation.)

3. **When** is Numerical Analysis required ?

In the following situation one can go for numerical approximations:

(a) The mathematical problem is known to have solution but does not have any method to obtain exact analytical expression.

(b) Mathematical problem is very difficult to handle exactly.

(c) There is a lack of enough data about the inputs for the mathematical problem and therefore not possible to use the available exact methods.

4. **Who** requires Numerical Analysis ?

Everyone working in the field of science and engineering field. For eg:

- Mathematicians,
- Statisticians,
- Physicists,
- Chemists,
- Engineers, etc.

5. **Why** should we learn Numerical Analysis ?

There are at least three reasons for which we need to learn and understand numerical methods.

(a) Learning different numerical methods and their analysis will make a person more familiar with the techniques of developing new numerical methods. This is important because when the available methods are not enough or not efficient for a specific problem to solve.

(b) In many situations, one has more methods for a given problem. Hence choosing an appropriate method is important for producing an accurate result in lesser time.

(c) With a good enough knowledge of numerical methods, one can use it properly and efficiently to get solution of problems and understand what is going wrong if the results obtained are not expected.

Through the rest of the course we will answer 'How to do numerical analysis?'

1.2 Syllabus we are going to cover

(Already I mentioned in the class. This portion will be filled.)

1.3 Prerequisite

(Already I mentioned in the class. This portion will be filled.)

2

Floating Point Approximation of Real Numbers and Error Analysis

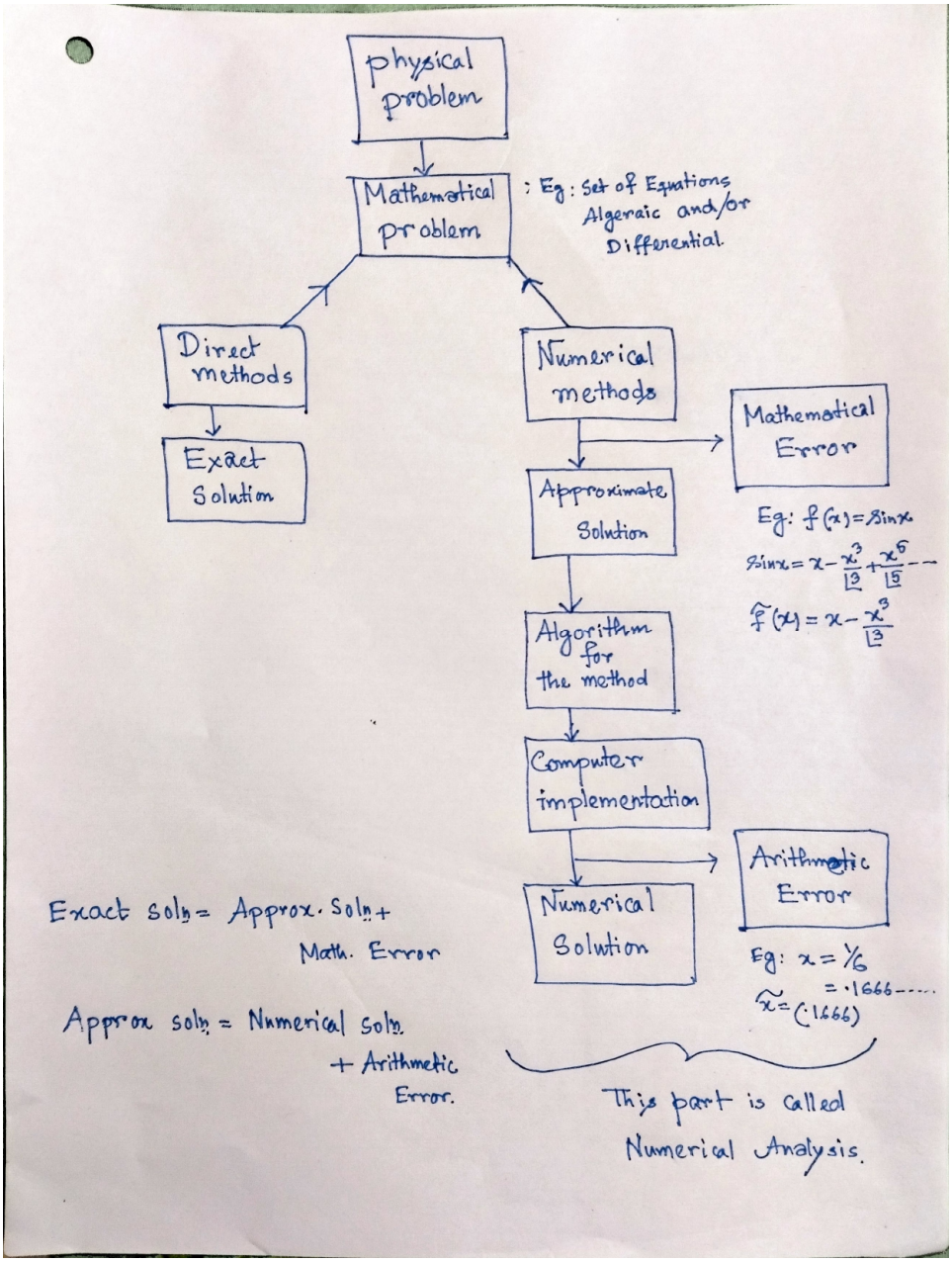
We have known that Numerical Analysis is all about

- Developing approximation procedure called numerical methods, to approximate a solution of a given Mathematical problem (whenever a solution exists),
- Developing algorithms and implement them as computer codes,
- Analyzing error in numerical approximation.

Therefore, the approximate solution, obtained by numerical method, will involve an error which is precisely the difference between the exact solution and the approximate solution. Thus, we have

Exact Solution = Approximate Solution + Error.

This error is called the **Mathematical** error.
In the next step while developing the algorithm and implementing them as computer code, we use a set of numerical values to evaluate the approximate solution obtained in numerical method. After evaluation we obtain a set of numerical values which is called the **numerical solution** to the given Mathematical problem. Due to memory restrictions, a computing device can store only a finite number of digits. Therefore, a real number cannot be stored exactly. Certain approximation needs to be done, and only an



approximate value of the given number is stored in the device. For further calculations, this approximate value is used instead of the exact value of the number. Hence, during the process of computation, the computer introduces a new error, called the **arithmetic error**. Then, we note that

Approximate Solution = Numerical Solution + Arithmetic Error.

Therefore, we have

Exact Solution = Numerical Solution + Mathematical Error + Arithmetic Error.

The Total Error is defined as

Total Error = Mathematical Error + Arithmetic Error.

The error (arithmetic error) involved in the numerical solution when compared to the exact solution can be worse than the mathematical error.
In this chapter, we introduce the floating-point representation of a real number and see two method to obtain floating-point approximation of a given real number. We introduce different types of errors that we come across in numerical analysis and their effects in the computation. At the end of this chapter, we will be familiar with the arithmetic errors, their effect on computed results and some ways to minimize this error in the computation.

2.1 Floating-Point Representation

Let $\beta \in \mathbb{N}$, with $\beta \geqslant 2$. Let $x \in \mathbb{R}$ be a real number. The real number x can be represented in base β as

$$x := (-1)^s \times (.d_1 d_2 \dots d_t d_{t+1} \dots)_\beta \times \beta^e, \tag{2.1.1}$$

where $d_i \in \{0, 1, \dots, \beta - 1\}$ for all $i \in \mathbb{N}$ with $d_1 \neq 0$ or $d_i = 0$, for all $i \in \mathbb{N}$, $s \in \{0, 1\}$, and $e \in \mathbb{Z}$. Here in above

- β is called the **base** or **radix**,
- s is called the **sign**,
- $(.d_1 d_2 \dots d_n d_{n+1} \dots)_\beta$ is called the **mantissa**,
- e is called the **exponent**,

of the given number x. The representation (2.1.1) of a real number is called the **floating-point representation**. The mantissa part can also be written as

$$(.d_1 d_2 \dots d_n d_{n+1} \dots)_\beta = \sum_{i=1}^{\infty} \frac{d_i}{\beta^i}. \tag{2.1.2}$$

We note that the right hand side of (2.1.2) is convergent series in $\mathbb{R}$ and hence is well defined.

Remark 2.1.1. *We note the following.*

- *When $\beta = 2$, the floating-point representation (2.1.1) is called the* **binary floating-point representation**.
- *When $\beta = 10$, the floating-point representation (2.1.1) is called the* **decimal floating-point representation**.

There are other representation of real numbers. For example **octal** *with $\beta = 8$,* **hexadecimal** *(used in ancient China) with $\beta = 16$,* **vigesimal** *(used in France) with $\beta = 20$ etc. In the course we will consider $\beta = 10$.*

2.1.1 Floating-Point Number System in a Computing Device

Due to memory restrictions, a computing device can store only a finite number of digits in the mantissa. In this section, we introduce the floating-point approximation and discuss how a given real number can be approximated.
A computing device stores a real number with only a finite number of digits in the mantissa. Although different computing devices have different ways of representing the numbers, here we introduce a mathematical form of this representation, which we will use throughout this course.

Definition 2.1.1 (*n*-**digit floating point number**). *Let $\beta \in \mathbb{N}$, with $\beta \geqslant 2$. An n-***digit floating point number** x_f *in* **base** β *is of the form*

$$x_f := (-1)^s \times (.d_1 d_2 \dots d_n)_\beta \times \beta^e \tag{2.1.3}$$

where $d_i \in \{0, 1, \dots, \beta - 1\}$ for all $i = 1, 2, \dots, n$, with $d_1 \neq 0$ or $d_i = 0$, for all $i = 1, 2, \dots, n$, $s \in \{0, 1\}$, $e \in \mathbb{Z}$ is an appropriate exponent. Here

$$(.d_1 d_2 \dots d_n)_\beta = \sum_{i=1}^{n} \frac{d_i}{\beta^i}. \tag{2.1.4}$$

The number of digits n in the mantissa is called the **precision** *or* **length** *of the floating-point number.*

Remark 2.1.2. *We note the following.*

- *When $\beta = 2$, the n-digit floating-point representation (2.1.1) is called the n-***digit binary floating-point representation**. *This is used in computer.*
- *When $\beta = 10$, the n-digit floating-point representation (2.1.1) is called the n-***digit decimal floating-point representation**. *It is used in daily life computation.*

Example 2.1.1. *The following are examples of real numbers in the decimal floating point representation.*

(i). The real number $x_f = 2347$ is represented in the decimal floating-point representation as

$$x_f = (-1)^0 \times .2347 \times 10^4.$$

In the above, we note that $\beta = 10, s = 0, d_1 = 2, d_2 = 3, d_3 = 4, d_4 = 7, e = 1$.

(ii). The real number $y_f = -0.000739$ is represented in the decimal floating-point representation as

$$y_f = (-1)^1 \times .739 \times 10^{-3}.$$

In the above, we note that $\beta = 10, s = 1, d_1 = 7, d_2 = 3, d_3 = 9, e = -3$.

Remark 2.1.3. *Any computing device has its own memory limitations in storing a real number. In terms of the floating-point representation, these limitations lead to the restrictions in the range of the exponent* (e) *and the number of digits in the mantissa* (n).

2.1.2 Characterization of Floating-Point Number System in a Computing Device

For a given computing device, there are integers $L, U, n \in \mathbb{Z}$ such that the exponent e is limited to a range

$$L \leqslant e \leqslant U,$$

and n is the precision. Therefore, a floating point number system, denoted by $\mathbb{F}$, is characterized by four integer $\beta, n \in \mathbb{N}$, $L, U \in \mathbb{Z}$ such that

- β= base or radix,
- n= precision,
- L=lower bound of the exponent range,
- U=upper bound of the exponent range.

Any floating point number $x_f \in \mathbb{F}$, can be represented as in (2.1.3).

2.1.3 Properties of $\mathbb{F}$

(i). The n-digit floating-point representation of a number is unique.

(ii). A floating point number system $\mathbb{F}$ is finite and discrete.

(iii). The total numbers in a given floating point number system is

$$2(\beta-1)\beta^{n-1}(U-L+1)+1.$$

Indeed, while counting the numbers, there are

- 2 choices of the sign s,
- $(\beta-1)$ number of choices of leading digit d_1,
- β number of choices of each of the remaining digits $d_2, d_2, \ldots d_n$,
- $(U-L+1)$ number of possible values of the exponent e.

Therefore, in total $2(\beta-1)\beta^{n-1}(U-L+1)$ numbers. Finally, 1 for number zero.

(iv). The mantissa $m(x_f) := (.d_1 d_2 \dots d_n)_\beta = \sum_{i=1}^{n} \frac{d_i}{\beta^i}$ satisfies

$$\frac{1}{\beta} \leqslant m(x_f) < 1.$$

(v). The smallest positive floating point number is

$$\text{UFL} = \beta^{L-1}.$$

Indeed, for the smallest positive floating point number

- the sign, $s = 0$,
- the leading digit, $d_1 = 1$,
- the remaining digits, $d_2 = d_2 = \cdots = d_n = 0$,
- the exponent, $e = L$.

Hence,

$$\text{UFL} = (.10\dots,0)_\beta \beta^L = \frac{1}{\beta}\beta^L = \beta^{L-1}.$$

It is also called Under Flow Level (UFL).

(vi). The largest positive floating point number is

$$\text{OFL} = (1-\beta^{-n})\beta^U.$$

Indeed, for the largest positive floating point number

- the sign, $s = 0$,
- all the digits, $d_1 = d_2 = d_2 = \cdots = d_n = (\beta - 1)$,
- the exponent, $e = U$.

Hence,

$$\begin{aligned}
\text{UFL} &= (.(\beta-1)(\beta-1)\dots(\beta-1))_\beta \beta^U, \\
&= \sum_{i=1}^{n} \frac{(\beta-1)}{\beta^i}\beta^U, \\
&= \frac{(\beta-1)}{\beta}\beta^U \left(1 + \frac{1}{\beta} + \cdots \frac{1}{\beta^{n-1}}\right), \\
&= \frac{(\beta-1)}{\beta}\beta^U \left(\frac{1-\left(\frac{1}{\beta}\right)^n}{1-\left(\frac{1}{\beta}\right)}\right), \\
&= \frac{(\beta-1)}{\beta}\beta^U \frac{(1-\beta^{-n})}{\frac{(\beta-1)}{\beta}},
\end{aligned}$$

$$= (1-\beta^{-n})\beta^{U}.$$

It is also called Under Flow Level (UFL).

(vii). All floating point number $x_f \in \mathbb{F}$ must satisfies

$$\text{UFL} \leqslant |x_f| \leqslant \text{OFL}.$$

(viii). Floating point numbers are not uniformly distributed throughout their range, but they are equally spaced only between successive powers of β.

(ix). If an arithmetic operation leads to a result in which a computed number x satisfies $|x| > \text{OFL}$, then, the calculation may terminate. Then error occurs in the computation. It is called an overflow error.

(x). If an arithmetic operation leads to a result in which a computed number x satisfies $|x| < \text{UFL}$, then, the the floating point representation of x (note that floating point representation of real number will be discussed in subsection 2.1.5) will be set to zero. This also makes error in the computation. It is called an underflow error.

In subsection 2.1.4, we introduce the concept of under and over flow of memory, which is a result of the restriction in the exponent. The restriction on the length of the mantissa is discussed in subsection 2.1.5.

2.1.4 Overflow and Underflow of Memory

- When the value of the exponent e in a floating-point number exceeds the maximum limit of the memory, we encounter the overflow of memory.
- When the value of the exponent e goes below the minimum of the range, then we encounter underflow.

During a computation,

- if some computed number x_f has an exponent $e > L$ (this also occurs if $|x_f| > \text{OFL}$) then we say, the memory **overflow** occurs.
- if $e < U$, (this also occurs if $|x_f| < \text{UFL}$) we say the memory **underflow** occurs.

and

Remark 2.1.4. *We note the following.*

(i).
- *In the case of overflow of memory in a floating-point number, a computer will usually produce meaningless results or simply prints the symbol* inf *or* NaN.

- *When a computation involves an undetermined quantity (like* $0\times\infty, \infty-\infty, 0/0$*), then the output of the computed value on a computer will be the symbol* NaN *(means 'not a number'). For instance, if* x_f *is a sufficiently large number that results in an overflow of memory when stored on a computing device, and* y_f *is another number that results in an underflow, then their product will be returned as* NaN.

(ii). On the other hand, we feel that the underflow is more serious than overflow in a computation. Because, when underflow occurs, a computer will simply consider the number as zero without any warning. The error is disguised in the computation and computer will not inform you about that whereas for the case of overflow computer will give you warning to modify the computer algorithm(code). If a computation involves such an underflow of memory, then there is a danger of having a large difference between the actual value and the computed value. However, by writing a separate subroutine, one can monitor and get a warning whenever an underflow occurs.

2.1.5 Floating-Point Approximation of Real Numbers

In general, a real number can have infinitely many digits, which a computing device cannot hold in its memory. Rather, each computing device will have its own limitation on the length of the mantissa. If a given real number has infinitely many digits in the mantissa of the floating-point form as in (2.1.1), then the computing device converts this number into an n-digit floating-point form as in (2.1.3). Such an approximation is called the **floating-point approximation** of a real number.
There are many ways to get floating-point approximation of a given real number. Here we introduce two types of floating-point approximations.

Definition 2.1.2 (Chopping)**.** *Let* $x \in \mathbb{R}$ *be a real number given in the floating-point representation* (2.1.1) *as*

$$x = (-1)^s \times (.d_1 d_2 \dots d_n d_{n+1} \dots)_\beta \times \beta^e. \tag{2.1.5}$$

The floating-point approximations of x *using* n**-digit chopping** *is given by*

$$fl(x) = (-1)^s \times (.d_1 d_2 \dots d_n)_\beta \times \beta^e. \tag{2.1.6}$$

The floating-point approximations of x *using* n**-digit rounding** *is given by*

$$fl(x) = \begin{cases} (-1)^s \times (.d_1 d_2 \dots d_n)_\beta \times \beta^e & , \quad 0 \leqslant d_{n+1} < \frac{\beta}{2}, \\ (-1)^s \times (.d_1 d_2 \dots (d_n+1))_\beta \times \beta^e & , \quad \frac{\beta}{2} \leqslant d_{n+1} < \beta, \end{cases} \tag{2.1.7}$$

where

$$(.d_1d_2\ldots(d_n+1))_\beta = (.d_1d_2\ldots(d_n)_\beta + (.\underbrace{00\ldots0}_{(n-1)\text{- times}}1)_\beta,$$
$$= \sum_{i=1}^{n}\frac{d_i}{\beta^i} + \frac{1}{\beta^n}.$$

Remark 2.1.5. *We note the following. If for a real number x, $d_{n+1} \geqslant \frac{\beta}{2}$ and $d_n = \beta - 1$, then in case of n-th digit rounding $d_n + 1$ in $fl(x)$ would be β. Therefore, we put $d_n = 0$ and add 1 to d_{n-1}. Again, if $d_{n-1} = \beta - 1$, we put $d_{n-1} = 0$ and add 1 to d_{n-2}. We continue further if required until all the digits are exhausted. It may happen that $d_1 = d_2 = \cdots = d_n = \beta - 1$ and $d_{n+1} \geqslant \frac{\beta}{2}$, then for $fl(x)$ we put $d_1 = 1$, all other $d_i = 0$ for $i = 2, 3, \ldots, n$ and increase the exponent by 1, if*

$$x = (-1)^s \times (.\underbrace{(\beta-1)(\beta-1)\ldots(\beta-1)}_{n\text{-times}}d_{n+1})_\beta \times \beta^e, \quad d_{n+1} \geqslant \frac{\beta}{2}$$

then in case of n-th digit rounding

$$fl(x) = (-1)^s \times (.1\ \underbrace{00\ldots0}_{(n-1)\text{-times}})_\beta \times \beta^{e+1}.$$

Example 2.1.2. *(i). The floating-point representation of π is given by*

$$\pi = (-1)^0 \times (.31415926\ldots)_{10} \times 10^1.$$

The floating-point approximation of π using 5-digit chopping is given by

$$fl(\pi) = (-1)^0 \times (.31415)_{10} \times 10^1.$$

The floating-point approximation of π using 5-digit rounding is given by

$$fl(\pi) = (-1)^0 \times (.31416)_{10} \times 10^1.$$

because $d_6 = 9 \geqslant 5$. So in $fl(\pi)$, $d_5 = 5 + 1 = 6$.

Remark 2.1.6. *Most of the modern processors, including Intel, uses IEEE 754 standard format. This format uses 52 bits in mantissa, (64-bit binary representation), 11 bits in exponent and 1 bit for sign. This representation is called the* **double precision**. *number. When we perform a computation without any floating-point approximation, we say that the computation is done using* **infinite precision** *(also called* **exact arithmetic**).

2.1.6 Arithmetic Operations of Floating-Point Approximations

In this subsection, we describe the procedure of performing arithmetic operations using n-digit rounding. The procedure of performing arithmetic operation using n-digit chopping can be done in a similar way.

Procedure of performing arithmetic operations

Let $\odot$ denote any one of the basic arithmetic operations $+$, $-$, $\times$, $\div$. Let $x, y \in \mathbb{R}$ be two real numbers. The process of computing $x \odot y$ using n-digit rounding is as follows.

- **Step 1:** First, we consider the n-digit rounding approximations $\text{fl}(x)$ and $\text{fl}(y)$ of the numbers x and y, respectively.
- **Step 2:**Then, we perform the calculation $\text{fl}(x) \odot \text{fl}(y)$ using exact arithmetic.
- **Step 3:** Finally, we consider the n-digit rounding approximations $\text{fl}(\text{fl}(x) \odot \text{fl}(y))$ of $\text{fl}(x) \odot \text{fl}(y)$.

The result from **Step 3** is the value of$x \odot y$ using n-digit rounding.

Example 2.1.3. *Let us consider the function $f : [0, \infty) \to \mathbb{R}$ given by*

$$f(x) = x(\sqrt{x+1} - \sqrt{x}), \quad x \in \mathbb{R}.$$

We evaluate $f(100000)$ using 6-digit rounding arithmetic. We have

$$f(100000) = 100000(\sqrt{100001} - \sqrt{100000})$$

We see that

$$\sqrt{100001} = 316.229347\cdots = .316229347\cdots \times 10^3,$$

and

$$\sqrt{100000} = 316.227766\cdots = .316227766\cdots \times 10^3.$$

Using 6-digit rounding we note that

$$fl(\sqrt{100001}) = .316229 \times 10^3, \quad \textit{and} \quad fl(\sqrt{100000}) = .316228 \times 10^3.$$

Then,

$$fl(\sqrt{100001}) - fl(\sqrt{100000}) = .000001 \times 10^3 = .1 \times 10^{-2}.$$

Hence,

$$fl(fl(\sqrt{100001}) - fl(\sqrt{100000})) = .1 \times 10^{-2}.$$

Again we see that $fl(100000) = .1 \times 10^6$. *Then,*

$$fl(100000) \times fl(fl(\sqrt{100001}) - fl(\sqrt{100000})) = .1 \times 10^6 \times .1 \times 10^{-2} = .01 \times 10^4 = .1 \times 10^3.$$

Finally, we have

$$fl(f(100000)) = .1 \times 10^3 = 100.$$

Similarly, using 6-digit chopping, the value of $fl(f(100000)) = 200$.

2.2 Types of Errors

The approximate representation of a real number obviously differs from the actual number, whose difference is called an **error**.

Definition 2.2.1 (Errors)**.** *(i). The* **error** *in a computed quantity is defined as*

$$\textbf{Error} = \textbf{True Value} - \textbf{Approximate Value}.$$

(ii). Absolute value of an error is called the **absolute error**.

(iii). The **relative error** *is a measure of the error in relation to the size of the true value as given by*

$$\textbf{Relative Error} = \frac{\textbf{Error}}{\textbf{True Value}}, \qquad \textit{Provided} \quad \textbf{True Value} \neq 0.$$

(iv). Absolute value of the relative error is called the **absolute relative error**.

(v). The **percentage error** *is defined as*

$$\textbf{Percentage Error} = 100 \times |\textbf{Relative Error}|.$$

Remark 2.2.1. *Let* x_A *denote the approximation to the real number* $x \in \mathbb{R}$. *We use the following notations.*

$$E(x_A) := Error(x_A) = x - x_A. \tag{2.2.1}$$

$$E_a(x_A) := Absolute\ Error(x_A) = |x - x_A|. \tag{2.2.2}$$

$$E_r(x_A) := Relative\ Error(x_A) = \frac{x - x_A}{x}, \quad x \neq 0. \tag{2.2.3}$$

$$E_{ar}(x_A) := Absolute\ Relative\ Error(x_A) = \left|\frac{x - x_A}{x}\right|, \quad x \neq 0. \tag{2.2.4}$$

The absolute error has to be understood more carefully because a relatively small difference between two large numbers can appear to be large, and a relatively large

difference between two small numbers can appear to be small. On the other hand, the relative error gives a percentage of the difference between two numbers, which is usually more meaningful as illustrated below.

Example 2.2.1.

2.3 Sources of Errors

- Mathematical modeling of a physical problem.
- Uncertainty in physical data or empirical measurements.
- Errors from previous computations.
- Blunders in arithmetic computations or in methods.
- **Machine Errors**.
- **Mathematical Truncation Error**.

2.4 Machine Errors

In this chapter we will be discussing Machine Errors.

2.4.1 Errors in Floating-point Approximations

With most of real numbers $x \in \mathbb{R}$, we have $\mathrm{fl}(x) \neq x$. We now measure the errors for floating for approximations of real numbers.

Theorem 2.4.1. *Let $x \in \mathbb{R}, x \neq 0$ be such that UFL $\leqslant |fl(x)| \leqslant$ OFL. Then,*

$$E_{ar}(fl(x)) \leqslant \begin{cases} \beta^{-n+1}, & \text{in case of chopping,} \\ \frac{1}{2}\beta^{-n+1}, & \text{in case of rounding.} \end{cases}$$

Proof. Let us assume that x has floating-point representation

$$x = (-1)^s \times (.d_1 d_2 \dots d_n d_{n+1} \dots)_\beta \times \beta^e = (-1)^s \beta^e \sum_{i=1}^{\infty} \frac{d_i}{\beta^i}.$$

First, We note that

$$|x| = \left| (-1)^s \beta^e \left(\sum_{i=1}^{\infty} \frac{d_i}{\beta^i} \right) \right| = \beta^e \sum_{i=1}^{\infty} \frac{d_i}{\beta^i} \geqslant \beta^e \frac{d_1}{\beta} \geqslant \beta^{e-1}. \tag{2.4.1}$$

We prove the result for two cases separately.

Case of chopping : In case of n-digit chopping

$$\text{fl}(x) = (-1)^s \times (.d_1 d_2 \dots d_n)_\beta \times \beta^e = (-1)^s \beta^e \sum_{i=1}^{n} \frac{d_i}{\beta^i}.$$

Hence,

$$\begin{aligned} E_a(\text{fl}(x)) = |x - \text{fl}(x)| &= \left| (-1)^s \beta^e \left(\sum_{i=1}^{\infty} \frac{d_i}{\beta^i} - \sum_{i=1}^{n} \frac{d_i}{\beta^i} \right) \right|, \\ &= \left| \beta^e \left(\sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} \right) \right|, \\ &= \beta^e \sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} \leqslant \beta^e \sum_{i=n+1}^{\infty} \frac{\beta - 1}{\beta^i}, \\ &= \beta^e (\beta - 1) \sum_{i=n+1}^{\infty} \frac{1}{\beta^i}, \\ &= \beta^e (\beta - 1) \frac{\frac{1}{\beta^{n+1}}}{1 - \frac{1}{\beta}} = \frac{\beta^e}{\beta^n} = \beta^{e-n}. \end{aligned}$$

Using (2.4.1), we have

$$E_{ar}(\text{fl}(x)) = \frac{|x - \text{fl}(x)|}{|x|} \leqslant \frac{\beta^{e-n}}{\beta^{e-1}} = \beta^{-n+1}.$$

Case of rounding : Sub-case: $0 \leqslant d_{n+1} < \frac{\beta}{2}$. We note that

$$\text{fl}(x) = (-1)^s \times (.d_1 d_2 \dots d_n)_\beta \times \beta^e = (-1)^s \beta^e \sum_{i=1}^{n} \frac{d_i}{\beta^i}.$$

Hence,

$$\begin{aligned} E_a(\text{fl}(x)) = |x - \text{fl}(x)| &= \left| (-1)^s \beta^e \left(\sum_{i=1}^{\infty} \frac{d_i}{\beta^i} - \sum_{i=1}^{n} \frac{d_i}{\beta^i} \right) \right|, \\ &= \left| \beta^e \left(\sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} \right) \right|, \\ &= \beta^e \sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} = \beta^e \left\{ \frac{d_{n+1}}{\beta^{n+1}} + \sum_{i=n+2}^{\infty} \frac{d_i}{\beta^i} \right\}, \\ &\leq \beta^e \left\{ \frac{\frac{\beta}{2} - 1}{\beta^{n+1}} + (\beta - 1) \sum_{i=n+2}^{\infty} \frac{1}{\beta^i} \right\} \\ &\left(\because 0 \leqslant d_{n+1} < \frac{\beta}{2} \implies 0 \leqslant d_{n+1} \leqslant \frac{\beta}{2} - 1, \quad \text{and} \quad d_i \leqslant \beta - 1 \right) \end{aligned}$$

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

Using (2.4.1), we have

$$
E_{ar}(\mathrm{fl}(x)) = \frac{|x-\mathrm{fl}(x)|}{|x|} \leqslant \frac{1}{2}\frac{\beta^{e-n}}{\beta^{e-1}} = \frac{1}{2}\beta^{-n+1}.
$$

Case of rounding : Sub-case: $\frac{\beta}{2} \leqslant d_{n+1} < \beta$. We note that

$$
\mathrm{fl}(x) = (-1)^s \times (.d_1 d_2 \ldots (d_n+1))_\beta \times \beta^e = (-1)^s \beta^e \left(\sum_{i=1}^{n} \frac{d_i}{\beta^i} + \frac{1}{\beta^n} \right).
$$

Hence,

$$
\begin{aligned}
E_a(\mathrm{fl}(x)) = |x-\mathrm{fl}(x)| &= \left| (-1)^s \beta^e \left(\sum_{i=1}^{\infty} \frac{d_i}{\beta^i} - \sum_{i=1}^{n} \frac{d_i}{\beta^i} - \frac{1}{\beta^n} \right) \right|, \\
&= \left| \beta^e \left(\sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} - \frac{1}{\beta^n} \right) \right|, \\
&\left(\because \sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} \leq (\beta-1) \sum_{i=n+1}^{\infty} \frac{1}{\beta^i} = (\beta-1)\frac{\frac{1}{\beta^{n+1}}}{1-\frac{1}{\beta}} = \frac{1}{\beta^n} \right) \\
&= \beta^e \left(\frac{1}{\beta^n} - \sum_{i=n+1}^{\infty} \frac{d_i}{\beta^i} \right) \\
&\leq \beta^e \left(\frac{1}{\beta^n} - \frac{d_{n+1}}{\beta^{n+1}} \right) \\
&\left(\because \frac{\beta}{2} \leqslant d_{n+1} \right) \\
&\leq \beta^e \left(\frac{1}{\beta^n} - \frac{\beta}{2\beta^{n+1}} \right) = \frac{1}{2}\beta^{e-n}.
\end{aligned}
$$

Finally, using (2.4.1), we have

$$
E_{ar}(\mathrm{fl}(x)) = \frac{|x-\mathrm{fl}(x)|}{|x|} \leqslant \frac{1}{2}\frac{\beta^{e-n}}{\beta^{e-1}} = \frac{1}{2}\beta^{-n+1}.
$$

This completes the proof. □

2.4.2 Accuracy of Floating-point Approximations

Unit round-off

We now introduce two measures that give a fairly precise idea of the possible accuracy in a floating-point approximation.

Definition 2.4.1 (Unit round-off (**u**)). *The* **unit round-off** *of a computer is denoted by* **u** *and is defined as*

$$\mathbf{u} = \sup\left\{E_{\mathbf{ar}}(fl(\mathbf{x})) : \mathbf{x} \in \mathbb{R},\ OFL \leq |\mathbf{x}| \leq UFL\right\}.$$

i.e., it is the maximum possible value of $E_{ar}(fl(x))$ *for any x satisfying the condition as in Theorem 2.4.1. By Theorem 2.4.1 we note that*

$$\mathbf{u} = \begin{cases} \beta^{-n+1}, & \text{in case of chopping,} \\ \frac{1}{2}\beta^{-n+1}, & \text{in case of rounding.} \end{cases}$$

Alternative characterization of unit round-off

Definition 2.4.2 (Machine Epsilon(e_{mach})). *The* **machine epsilon** *of a computer is denoted by* ε_{mach} *and is defined as*

$$\varepsilon_{mach} = \inf\left\{\delta \in \mathbb{F} : \delta > 0, \quad \text{and} \quad fl(1+\delta) > 1\right\}.$$

Theorem 2.4.2.

$$\varepsilon_{mach} = \begin{cases} \beta^{-n+1}, & \text{in case of chopping,} \\ \frac{1}{2}\beta^{-n+1}, & \text{in case of rounding.} \end{cases}$$

Proof. See book: Atkinson: Introduction ot Numerical Analysis, 2ed., p-15. □

Remark 2.4.1. *From the definition of* ε_{mach}*, we note that for any* $\delta \in \mathbb{F}$ *with* $0 < \delta < \varepsilon_{mach}$ *implies* $fl(1+\delta) = 1$*, i.e., in computer the number* $1+\delta$ *is identical with* 1.

Remark 2.4.2. *we note that* $\varepsilon_{mach} \neq UFL$*.* ε_{mach} *is determined by the precision i.e., the number of digits in the mantissa field of Floating-point system, whereas the UFL is determined by the number of lower bound in the exponent field.*

Example 2.4.1.

Maximal Accuracy

Definition 2.4.3 (Maximal Accuracy). *The maximal accuracy in a floating-point representation is denoted my* M_{acc} *and is defined as*

$$M_{acc} = \sup\left\{m \in \mathbb{Z} : m \geqslant 0, \quad \text{and} \quad fl(m) = m\right\}.$$

Remark 2.4.3. *From the definition of maximal accuracy, we note that for any* $fl(M_{acc}+1) \neq M_{acc}+1$.

Theorem 2.4.3.

$$M_{acc} = \beta^n.$$

Proof. See book: Atkinson: Introduction ot Numerical Analysis, 2ed., p-16. □

2.4.3 Loss of Significance

In place of relative error, we often use the concept of significant digits that is closely related to relative error.

Definition 2.4.4 (Significant digits). *Let us consider a floating-point number system with base* $\beta \in \mathbb{N}$ *with* $\beta \geqslant 2$. *For* $x \in \mathbb{R}$, *let* x_A *be an approximation of* x. *Let*

$$s = \sup\left\{k \in \mathbb{Z} : \beta^k \leqslant |x|\right\} \quad \text{and} \quad r = \sup\left\{t \in \mathbb{N} : |x - x_A| \leq \frac{1}{2}\beta^{s+1-t}\right\}$$

Then, we say x_A *has* r *significant digits to approximate* x.

Example 2.4.2. *We find significant digits for the following approximations.*

(i). $x = \frac{1}{3} = .3333\ldots$ *and* $x_A = .333$. *We note that*

$$10^{-1} = .1 < x, \quad \textit{hence,} \quad s = -1$$

and

$$\begin{aligned} |x - x_A| = (.0003333\ldots) &\leqslant (.3333\ldots) \times 10^{-3}, \\ &= \frac{1}{2} \times (.6666\ldots) \times 10^{-3}, \\ &< \frac{1}{2} \times 10^{-3} = \frac{1}{2} \times 10^{-1+1-3} \end{aligned}$$

Therefore, x_A *has* 3 *significant digits to approximate* x.

(ii). $x = 23.496$ *and* $x_A = 23.494$ *We note that*

$$10^1 < x, \quad \textit{hence,} \quad s = 1,$$

and

$$\begin{aligned} |x - x_A| = (.002) &= .2 \times 10^{-2}, \\ &= \frac{1}{2} \times .4 \times 10^{-2}, \end{aligned}$$

$$< \frac{1}{2} \times 10^{-2} = \frac{1}{2} \times 10^{1+1-4}$$

Therefore, x_A has 4 significant digits to approximate x.

(iii). $x = 0.02138$ *and* $x_A = 0.02144$ *We note that*

$$10^{-2} < x, \quad \textit{hence,} \quad s = -2,$$

and

$$\begin{aligned} |x - x_A| = (.00006) &= .6 \times 10^{-4}, \\ &= \frac{1}{2} \times 1.2 \times 10^{-4}, \\ &= \frac{1}{2} \times .12 \times 10^{-3}, \\ &< \frac{1}{2} \times 10^{-3} = \frac{1}{2} \times 10^{-2+1-2} \end{aligned}$$

Therefore, x_A has 2 significant digits to approximate x.

Remark 2.4.4. *(i). We note that*

$$E_{ar}(x_A) = \frac{|x - x_A|}{|x|} \leqslant \frac{1}{2} \frac{\beta^{s-r+1}}{\beta^s} = \frac{1}{2} \beta^{1-r}.$$

Hence, absolute relative error decreases with an increase in the number of significant digits.

(ii). Number of significant digits roughly measures the number of leading non-zero digits of x_A that are correct relative to the corresponding digits in the true value x. However, this is **not a precise way** *to get the number of significant digits as we have seen in the above examples.*

(iii). The role of significant digits in numerical calculations is very important in the sense that the loss of significant digits may result in drastic amplification of the relative error as illustrated in the following example.

Example 2.4.3 (Loss of Significance)**.** *(i). Let us consider two real numbers*

$$x = 7.6545428 = .76545428 \times 10^1, \quad y = 7.6544201 = .76544201 \times 10^1.$$

Let

$$x_A = 7.6545421 = .76545421 \times 10^1, \quad y_A = 7.6544200 = .76544200 \times 10^1$$

Lecture-6

Propagation of Error in Function evaluation

Let $f: \mathbb{R} \to \mathbb{R}$ be a function and $x_T \in \mathbb{R}$.
We are interested to find $f(x_T)$ in computer.

When we put x_T in a computer, first it generates an approximation say x_A due to its finite precision (or finite memory)

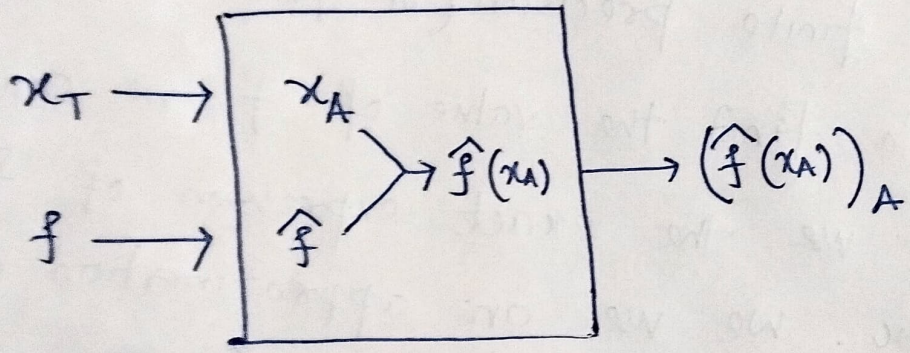
Again to find the value of f, we cannot always use the exact expression of f in computer. We use an approximation of f (which is sometime called a numerical algorithm) say $\hat{f}$.

Therefore, to evalnate $f(x_T)$, the computer uses the approximated value x_A of x_T and the approximated function $\hat{f}$ of f.

Hence, we suppose to recieve the value $\hat{f}(x_A)$ from the computer.

P-2

But again after using the algorithm $\hat{f}$, Computer does a final step approximation of $\hat{f}(x_A)$ and we finally recieve $\left(\hat{f}(x_A)\right)_A$.



Computer.

Therefore, the error for this output (computation)

$$Er = f(x_T) - \left(\hat{f}(x_A)\right)_A$$

$$= \underbrace{f(x_T) - f(x_A)}_{\text{Propagated Er}} + \underbrace{f(x_A) - \hat{f}(x_A)}_{\text{Truncation Er}}$$

$$+ \underbrace{\hat{f}(x_A) - \left(\hat{f}(x_A)\right)_A}_{\text{rounding/chopping Er.}}$$

Propagated Error: It is the difference between the values of the function f at the true value x_T and at the approximated value x_A of the input, i.e. $f(x_T) - f(x_A)$.

It is independent of the choice of an algorithm.

Truncation Error: It is the difference between the values of the function f and its approximation $\hat{f}$ at the same point x_A,

ie, $f(x_A) - \hat{f}(x_A)$

It depends on the choice of approximation or algorithm.

Rounding/chopping Error It is the difference between the result produced by a given algorithm using exact arithmatic and the result produced by the same algorithm using finite precision arithmatic,

ie, $\hat{f}(x_A) - (\hat{f}(x_A))_A$.

It depends on the machin's finite precision arithmatic.

P-4

Computational Error: It is the combination of Truncation error and Rounding error.

Error = Propagated Error + Computational Error

Computational Error = Truncation Error + Rounding Error.

Example: Eg① Let $f: \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \sin x, \quad x \in \mathbb{R}.$$

we want to evaluate $f(\pi/8)$ ie. $\sin(\pi/8)$ in computer.

Here $x_T = \pi/8$, let us assume computer uses an approximation of π as 3

$$\pi = 3.14159\ldots, \sim 3$$

$$x_T = \pi/8, \quad x_A = 3/8 = .375$$

Let we use the algorithm

$$\hat{f}(x) = x - \frac{x^3}{\lfloor 3}, \qquad \sin x = x - \frac{x^3}{\lfloor 3} + \frac{x^5}{\lfloor 5} \ldots$$

$$\therefore \hat{f}(3/8) = 3/8 - \frac{27}{512 \times 6} = 3/8 - \frac{9}{1024}$$

$\hat{f}(3/8) = .375 - .008780625 =$

Let us assume the computer uses 5-digit rounding

$\left(\hat{f}(3/8)\right)_A = .375 - .00878 = .38375$

Therefore the error in this computation

$$\text{Error} = f(\pi/8) - \left(\hat{f}(3/8)\right)_A$$

$$= \sin(\pi/8) - .38375$$

$$= \underbrace{\sin(\pi/8) - \sin(3/8)}_{\text{propagated}} + \overbrace{\sin(3/8) - \left(\frac{3}{8} - \frac{1}{\lfloor 3}\left(\frac{3}{8}\right)^3\right)}^{\text{Truncation}}$$

$$+ \underbrace{\left(\frac{3}{8} - \frac{1}{\lfloor 3}\left(\frac{3}{8}\right)^3\right) - .38375}_{\text{rounding}}$$

Eg(2) Let $f: \mathbb{R} \to \mathbb{R}$ be a twice differentiable fn. with $|f''(x)| \le M_0$ $|f''(x)| \le M_2$ for all $x \in \mathbb{R}$

and for some $M_0, M_2 > 0$.

We want to find $f'(x_T)$ at some pt. $x_T \in \mathbb{R}$ with the help of computer.

P-6 We know by analytical theory (Calculus)

$$f'(x_T) = \lim_{h \to 0} \frac{f(x_T+h) - f(x_T)}{h},$$

(as f is diffble so this limit exists).

we use the algorithm $\hat{f}'_h(x_T) = \frac{f(x_T+h) - f(x_T)}{h}$ for $h \neq 0$.

to evalute $f'(x_T)$.

we finally get $\left(\hat{f}'_h(x_A)\right)_A$. from the computer.

The error for the computation

$$\text{Error} = f'(x_T) - \left(\hat{f}'_h(x_A)\right)_A$$

$$= \underbrace{f'(x_T) - f'(x_A)}_{\text{propagated}} + \underbrace{f'(x_A) - \hat{f}'_h(x_A)}_{\text{Truncation}} + \underbrace{\hat{f}'_h(x_A) - \left(\hat{f}'_h(x_A)\right)_A}_{\text{Rounding.}}$$

T.E. = Truncation Error $= f'(x_A) - \hat{f}'_h(x_A)$

$$T.E. = f'(x_A) - \hat{f}_h'(x_A)$$

$$= f'(x_A) - \frac{f(x_A+h) - f(x_A)}{h}$$

$$= -\frac{h}{2} f''(\xi).$$

[Since f is twice differentiable fn, by Taylor's theorem

$$f(x_A+h) = f(x_A) + h f'(x_A) + \frac{h^2}{2} f''(\xi)$$

for some $\xi \in [x_A, x_A+h]$

or $\xi \in [x_A+h, x_A]$

$$\Rightarrow \frac{f(x_A+h) - f(x_A)}{h} - f'(x_A) = \frac{h}{2} f''(\xi).$$

$$\therefore |T.E| = \frac{1}{2}|h|\,|f''(\xi)| \leq \frac{M_2|h|}{2} \longrightarrow (1).$$

$(\because |f''(x)| \leq M_2 \ \forall x \in \mathbb{R})$.

R.E. = Rounding error

$$= \hat{f}_h'(x_A) - \left(\hat{f}_h'(x_A)\right)_A$$

$$|R.E.| = \left| \hat{f}_h'(x_A) - \left(\hat{f}_h'(x_A)\right)_A \right| \leq \varepsilon_{mach} \left| \hat{f}_h'(x_A) \right|$$

$(\because$ we proved $Er(fl(x)) \leq \varepsilon_{mach}.)$

P-8 Here $x = \widehat{f_h'}(x_A)$ and $fl(x) = \left(\widehat{f_h'}(x_A)\right)_A$.

$\therefore |R.E.| \leq \varepsilon_{mach} \left|\widehat{f_h'}(x_A)\right|$

$$= \varepsilon_{mach} \left| \frac{f(x_A+h) - f(x_A)}{h} \right|$$

$$\leq \varepsilon_{mach} \frac{\left(|f(x_A+h)| + |f(x_A)|\right)}{|h|}$$

$$\leq \frac{2\,\varepsilon_{mach}\, M_0}{|h|} \quad \left(\because |f(x)| \leq M_0 \ \forall x \in \mathbb{R}\right)$$

$\longrightarrow$ (2)

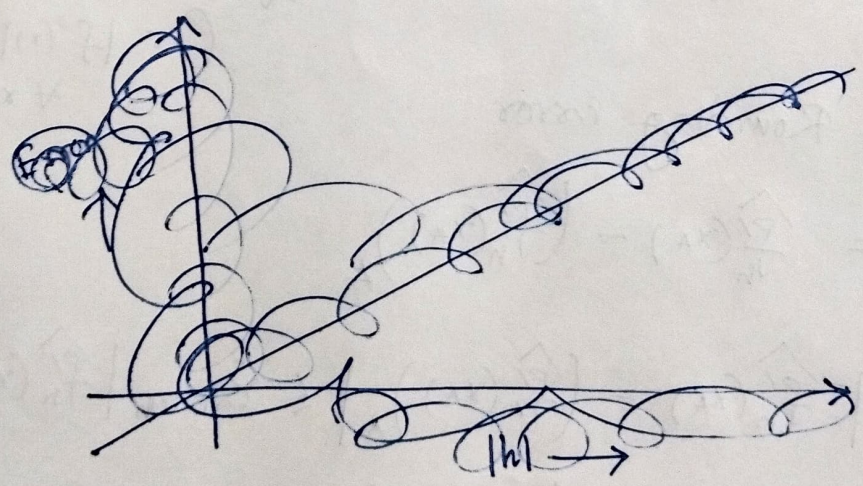
By (1) and (2) we have

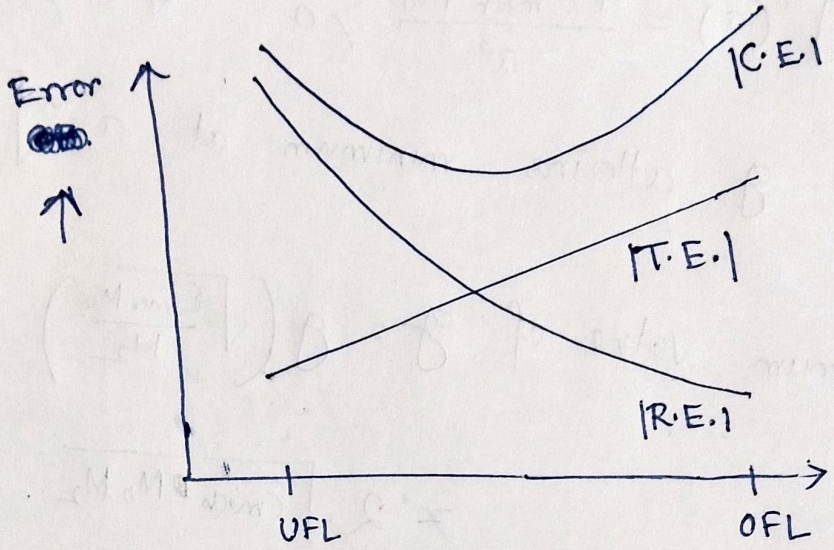
Computational Error = T.E. + R.E.

$$|\text{Computational error}| \leq |T.E.| + |R.E.|$$

$$\leq \frac{|h|\, M_2}{2} + \frac{2\,\varepsilon_{mach}\, M_0}{|h|}$$

$h \neq 0$.





The computational error is minimized when

$$|h| = \sqrt[2]{\frac{\varepsilon_{mach} M_0}{M_2}}$$

let $g : (0, \infty) \longrightarrow \mathbb{R}$ given by

$$g(a) = \frac{a M_2}{2} + \frac{2 \varepsilon_{mach} M_0}{a}, \quad a > 0$$

To find minimum of g

Diff. g with respect to a

$$g'(a) = M_2/2 - \frac{2 \varepsilon_{mach} M_0}{a^2}$$

$$g'(a) = 0 \Rightarrow M_2/2 = \frac{2 \varepsilon_{mach} M_0}{a^2}$$

$$\Rightarrow a^2 = 4 \varepsilon_{mach} M_0 / M_2$$

$$a = \sqrt[2]{\frac{\varepsilon_{mach} M_0}{M_2}}, \quad \text{as } a > 0.$$

Q-010

$$g''(a) = \frac{4\varepsilon_{max} M_0}{a^3} > 0. \quad \forall a > 0$$

Hence, g attains minimum at $a = 2\sqrt{\frac{\varepsilon_{max} M_0}{M_2}}$.

Minimum value of $g = g\left(\sqrt{\frac{\varepsilon_{max} M_0}{M_2}}\right)$

$$= 2\sqrt{\varepsilon_{max} M_0 M_2}.$$

~~The above example~~, we have seen evidence of the above example in Lab-assignment-2:

For the function $f(x) = \sin x$ at $x = 1$..

We now discuss propagated Error.

Let $f : \mathbb{R} \to \mathbb{R}$ be diffble with

$|f'(\xi_1)| \leq M_1$ for some $M_1 > 0$ and for all $x \in N_\delta(x_T) = \delta$-nbd ie $|x - x_T| < \delta$.

$E(f(x_A))$ = propagated Error $= f(x_T) - f(x_A)$

$= f'(\xi)(x_T - x_A)$ (by Lagrange Mean Value theorem)

for some $\xi \in (x_T, x_A)$ or (x_A, x_T).

Assume $f(x_T) \neq 0 \neq x_T$.

Absolute relative propagated error

$$Ear(f(x_A)) = \frac{|f(x_T) - f(x_A)|}{|f(x_T)|}$$

$$= \frac{|f'(\xi_1)|\,|x_T - x_A|}{|f(x_T)|}$$

$$= \frac{|f'(\xi_1)|}{|f(x_T)|} \frac{|x_T - x_A|}{|x_T|} |x_T|$$

$$= \frac{|f'(\xi_1)|\,|x_T|}{|f(x_T)|} Ear(x_A).$$

P-2

$$\therefore \quad E_{ar}(f(x_A)) = \frac{|f'(\xi)|\,|x_T|}{|f(x_T)|} \cdot E_{ar}(x_A)$$

$E_{ar}(f(x_A))$ — Forward Error

$E_{ar}(x_A)$ — Backward Error

$$E_{ar}(f(x_A)) \leq \frac{M_1 |x_T|}{|f(x_T)|} E_{ar}(x_A)$$

$\frac{M_1 |x_T|}{|f(x_T)|}$ — we want to understand this quantity and how does this affect the propagated Error.

For $f: \mathbb{R} \to \mathbb{R}$, $f(x_T) \neq 0 \neq x_T$

let we define $M_f(x_T, x_A) = \dfrac{\dfrac{|f(x_T) - f(x_A)|}{|f(x_T)|}}{\dfrac{|x_T - x_A|}{|x_T|}}$

$\forall\ x_T, x_A \in \mathbb{R}$.

= Error magnification factor.

Condition number of f at x_T is given by

$$K_f(x_T) = \inf_{\delta > 0} \sup_{0 < |x_T - x_A| < \delta} M_f(x_T, x_A) = \overline{\lim_{x_A \to x_T}} M_f(x_T, x_A).$$

Since, $0 \leq M_f(x_1, x_2) \leq \dfrac{M_1 |x_1|}{|f(x_1)|}$, $\forall\, x_1, x_2 \in \mathbb{R}$

$K_f(x_1)$ is a non-negative finite number.

If f is differentiable at $x = x_1$ we have

$$K_f(x_1) = \frac{|f'(x_1)||x_1|}{|f(x_1)|} \quad \forall\, x_1 \in \mathbb{R} \text{ with } f(x_1) \neq 0 = x_1.$$

Incase either $x_1 = 0$ or $f(x_1) = 0$,

we consider abeolute condition number

an condition number which is given by

$$K_f(x_1) = \inf_{\delta > 0} \sup_{0 < |x_1 - x_2| < \delta} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|}$$

$$= |f'(x_1)| \quad \text{if } f \text{ is differentiable at } x = x_1.$$

Let $\epsilon > 0$ be given. By definition of $K_f(x_1)$

There exists $\delta_1 > 0$ such that

$$\sup_{0 < |x_1 - x_2| < \delta_1} M_f(x_1, x_2) \leq K_f(x_1) + \epsilon.$$

P-4 Therefore $\forall\ x_A$ with $|x_A - x_T| < \delta_1$.

$$M_f(x_T, x_A) \leq K_f(x_T) + \varepsilon.$$

$$\Rightarrow \frac{\dfrac{|f(x_T) - f(x_A)|}{|f(x_T)|}}{\dfrac{|x_T - x_A|}{|x_T|}} \leq K_f(x_T) + \varepsilon$$

by defn of $M_f(x_T, x_A)$.

$$\Rightarrow \frac{|f(x_T) - f(x_A)|}{|f(x_T)|} \leq \left(K_f(x_T) + \varepsilon\right) \frac{|x_T - x_A|}{|x_T|}$$

ie. $$E_{ar}\left(f(x_A)\right) \leq \left(K_f(x_T) + \varepsilon\right) E_{ar}(x_A).$$

$$\approx K_f(x_T)\, E_{ar}(x_A).$$

(for $0 < \varepsilon < \varepsilon_{mach}\, K_f(x_T)$ in computer.)

Therefore, the propagated error is magnified with the factor of the condition number of the fn.

—.

Insensitive / well conditioned

A problem is said to be insensitive or well conditioned if a given change in the input data causes a reasonable change in the output.

In general this happens if

$$K_f(x_T) \leq 10$$

Sensitive / ill conditioned

A problem is said to be sensitive or ill-conditioned if the input data causes much larger change in the output data.

In general this occures if

$$K_f(x_T) > 10.$$

Eg: 1. $f : (0, \infty) \to \mathbb{R}$, $f(x) = \sqrt{x}$

$$K_f(x_T) = \tfrac{1}{2} \cdot \forall\, x_T > 0$$

2. $f : \mathbb{R} \to \mathbb{R}$ be diffble bijective and f^{-1} is also diffble.

Show that

$$K_{f^{-1}}(y_T) = \frac{1}{K_f(x_T)}$$

with $f(x_T) = y_T$.

We note that condition number is the property of the function itself. It does neither depend on the algorithm of approximation of f nor on the machine precision.

Therefore, larger condition number can cause severe error in computation even if

we choose best algorithm and very high precision computer.

We will see this kind of example in solving a system of linear equation.

Lecture - 8

We are interested to solve the $m \times n$ linear system of equations:

$$\textcircled{1} \leftarrow \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots\cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots\cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots\cdots + a_{mn}x_n = b_m \end{cases}$$

where, $m, n \in \mathbb{N}$, $a_{ij}, b_i \in \mathbb{R}$, $1 \le i \le m$, $1 \le j \le n$, are given and $x_j \in \mathbb{R}$, $1 \le j \le n$ are unknowns.

The above system can be written as

$$\textcircled{2} \leftarrow Ax = b \quad (\text{posed in } \mathbb{R}^m)$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in M_{m,n}(\mathbb{R}), \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m$$

and $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$.

Definition (Solution)

A solution of the system ①/② is an n-tuple

$$s = \begin{bmatrix} s_1 \\ s_2 \\ \\ \\ s_n \end{bmatrix} \in \mathbb{R}^n \quad \text{such that} \quad As = b.$$

Let $b \in \mathbb{R}^m$. Let us denote the set of all solutions of ①/② by S_b and is given by

$$S_b = \{ s \in \mathbb{R}^n \mid As = b \}. \text{ (solution set)}$$

Defn. The system ①/② is said to be consistent if $S_b \neq \phi$ i.e., $\exists$ a solution of ①/②, Otherwise it is said to be inconsistent.

Defn. The system ①/② is said to be homogeneous if $b = 0$, otherwise it is said to be non-homogeneous.

Remark 1 We know that $A0 = 0$. Therefore, $0 \in S_0 \neq \phi$. Hence, homogeneous system is always consistent.

Let $A \in M_{m,n}(\mathbb{R})$, $m, n \in \mathbb{N}$.

Therefore, $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is a linear map.

We use the following notations.

$$N(A) := \{ x \in \mathbb{R}^n \mid Ax = 0 \}.$$

$:=$ Kernel of A,

$:=$ S_0.

$$R(A) := \{ Ax \mid x \in \mathbb{R}^n \}$$

$:=$ Range of A.

We know (From linear algebra)

⊙ $N(A)$ is a subspace of $\mathbb{R}^n$.

let $\eta(A) :=$ dimension of $N(A)$

$:=$ nullity of A.

Hence, $0 \leq \eta(A) \leq n$. ⟶ (1)

⊙ $R(A)$ is a subspace of $\mathbb{R}^m$

let $r(A) :=$ dimension of $R(A)$

$:=$ rank of A

Hence, $0 \leq r(A) \leq m$. ⟶ (2)

Theorem 1 (Rank-Nullity)

For $A \in M_{m,n}(\mathbb{R})$, $r(A) + n(A) = n$, ie.

Rank of A + Nullity of A $= n$.

proof: (Friedberg: Linear Algebra, Hoffman & Kunze: linear Algebra).

From Rank-Nullity theorem and ① we also have

$$0 \leq r(A) \leq n \longrightarrow ③$$

Therefore ② and ③ implies

$$0 \leq r(A) \leq \min\{m, n\}.$$

We will use Rank-Nullity theorem to conclude few results regarding solvability of system

$$Ax = b \quad (\text{posed in } \mathbb{R}^m).$$

1. Homogeneous system

Let us recall the solution set for homogeneous system is denoted by S_0, i.e.

$$S_0 := \{ x \in \mathbb{R}^n \mid Ax = 0 \} := N(A).$$

Corollary (1) If $r(A) < n$, then system ①/② has infinite number of solutions.

proof: By rank-nullity theorem

$$\eta(A) = n - r(A) > 0 \quad (\because r(A) < n)$$

$\Rightarrow$ dimension of $N(A) \geq 1$

$\Rightarrow$ $\exists$ $s_0 \in N(A) = S_0$

$\Rightarrow$ $ts_0 \in N(A) = S_0$ (As $N(A)$ is a subspace of $\mathbb{R}^n$)

$\forall\ t \in \mathbb{R}$

Hence, S_0 contains infinite number of elements,

i.e., Cardinality of $S_0 := |S_0| = \infty$.

Corollary (2) For $A \in M_{m,n}(\mathbb{R})$,

if $m < n$, then the system ①/② has infinite number of solutions.

proof: we know (have shown above)

$$r(A) \leq \min\{m, n\} = m < n$$

$(\because\ m < n)$

Hence,

By corollary ①, the system ①/② has infinite number of solutions.

2. Non-Homogeneous system

For $A \in M_{m,n}(\mathbb{R})$, $m, n \in \mathbb{N}$.

Let $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix}$, $\alpha_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \\ a_{mi} \end{bmatrix} \in \mathbb{R}^m$

$\alpha_1 \quad \alpha_2 \quad \alpha_n$ Columns of A

$i = 1, 2, \cdots, n.$

we denote the subspace of $\mathbb{R}^m$ span by the vectors $\alpha_1, \alpha_2, \cdots, \alpha_n$ as

$$\text{Col}(A) := \text{span}\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$$
$$:= \text{Column space of } A.$$

$$\text{Colr}(A) := \text{dimension of } \text{Col}(A)$$
$$:= \text{Column rank of } A$$

Theorem 2 For $A \in M_{m,n}(\mathbb{R})$,

$R(A) = \text{Col}(A)$ and hence $r(A) = \text{colr}(A)$.

proof. To show $R(A) = \text{Col}(A)$

Let $y \in R(A) \Rightarrow \exists\ x \in \mathbb{R}^n$ such that

$$Ax = y. \longrightarrow (*)_1$$

As $x \in \mathbb{R}^n$, x can be expressed as

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$

for some $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$,

and $\{e_1, e_2, \ldots, e_n\} \subset \mathbb{R}^n$ is the standard basis of $\mathbb{R}^n$, ie.

$$e_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{at } j\text{th position.}$$

Then, $$Ax = x_1 Ae_1 + x_2 Ae_2 + \cdots + x_n Ae_n \longrightarrow (*)_2$$

(A is linear)

We observe that

$$Ae_1 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} = \alpha_1$$

$$Ae_2 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} = \alpha_2$$

So on $Ae_3 = \alpha_3 \ldots$ $Ae_j = \alpha_j$ $j = 1, 2, \ldots, n$.

$\longrightarrow (*)_3$

Hence, using $(*)_1, (*)_3$, from $(*)_2$ we have

$$y = x_1\alpha_1 + x_2\alpha_2 + \cdots + x_n\alpha_n$$

$\Rightarrow y \in \text{col}(A)$

$\Rightarrow R(A) \subset \text{col}(A)$. $\longrightarrow (*)_4$

To prove the other inclusion, let $z \in \text{col}(A)$.

$\exists$ $\omega_1, \omega_2, \ldots, \omega_n \in \mathbb{R}$ so that

$$z = \omega_1\alpha_1 + \omega_2\alpha_2 + \cdots + \omega_n\alpha_n$$

$$= \omega_1 Ae_1 + \omega_2 Ae_2 + \cdots + \omega_n Ae_n$$

$(\because Ae_j = \alpha_j \quad j = 1, 2, \ldots, n)$

$\Rightarrow z = A(\omega_1 e_1 + \omega_2 e_2 + \cdots + \omega_n e_n)$

$\Rightarrow z = A\omega$ $(\because A \text{ is linear})$

for $\omega = \omega_1 e_1 + \cdots + \omega_n e_n \in \mathbb{R}^n$.

$\Rightarrow z \in R(A)$.

$\Rightarrow \text{col}(A) \subset R(A)$. $\longrightarrow (*)_5$

$(*)_4$ and $(*)_5 \Rightarrow R(A) = \text{col}(A)$.

Hence, dimension of $R(A)$ = dimension of $\text{col}(A)$

$r(A) = \text{col}\,r(A)$.

———.

Remark 2

We note that for a given $b \in \mathbb{R}^m$, the system ①/② has a solution

iff there exists $s \in \mathbb{R}^n$ so that $As = b$

iff $b \in R(A)$

iff $b \in \text{col}(A)$ (by theorem 2 above).

On the other hand system ①/② can be written as

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = b$$

$$\Rightarrow \quad x_1 \alpha_1 + x_2 \alpha_2 + \cdots + x_n \alpha_n = b \longrightarrow ③.$$

Finding soln. of ①/② means find $x_1, x_2, \ldots, x_n \in \mathbb{R}$ so that $x_1 \alpha_1 + x_2 \alpha_2 + \cdots + x_n \alpha_n = b$.

Theorem 3 System ①/② is consistent iff

$$r(A) = r(A|b),$$

where $A|b = [A\ b]$ = Augmented matrix.
(b column is added)

proof: The system ①/② is consistent

iff $b \in R(A)$

iff $b \in Col(A)$ (by theorem.2)

iff $b \in span\{\alpha_1, \alpha_2, .., \alpha_n\}$ (α_i's are columns of A)

iff $span\{\alpha_1, \alpha_2, .., \alpha_n\} = span\{\alpha_1, \alpha_2, .., \alpha_n, b\}$.

iff $Col(A) = Col(A|b)$

iff $r(A) = r(A|b)$

Theorem 4 Let $b \in \mathbb{R}^m$ and $s_1 \in \mathbb{R}^n$ be a solution of $Ax = b$ (i.e. $\& As_1 = b$).

Then, $S_b := s_1 + S_0$, (they are equal as sets)

where $s_1 + S_0 := \{ s_1 + z \mid z \in S_0 \}$

$= \{ s_1 + z \mid Az = 0 \}$.

proof: To show, first, $S_b \subseteq s_1 + S_0$.

Let $\omega \in S_b$. Then, $A\omega = b$

Now, $A(\omega - s_1) = A\omega - As_1$

$= b - b$ ($\because As_1 = b$ given)

$= 0$

$\Rightarrow \omega - s_1 \in S_0$

$\Rightarrow \exists\ z_0 \in S_0$ so that $\omega - s_1 = z_0$

$\Rightarrow \omega = s_1 + z_0 \in s_1 + S_0$ ($\because z_0 \in S_0$)

$\therefore \Rightarrow S_b \subset s_1 + S_0$. $\longrightarrow (*)_7$

To prove $s_1 + S_0 \subseteq S_b$.

let $w_1 \in s_1 + S_0 \Rightarrow \exists\ z_1 \in S_0$ so that

$w_1 = s_1 + z_1$

$\Rightarrow Aw_1 = As_1 + Az_1 = b$ ($\because As_1 = b$, $Az_1 = 0$)

P-12

$\Rightarrow$ w_1 is a solution of $Ax = b$

$\Rightarrow$ $w_1 \in S_b$

$\therefore$ $\Rightarrow$ $s_1 + S_0 \subseteq S_b$. $\longrightarrow$ $(*)_8$

$(*)_7$ and $(*)_8$ $\Rightarrow$ $S_b = s_1 + S_0$.

Theorem 5 For every $b \in \mathbb{R}^m$, the system $Ax = b$ has solution. Then, $m \leq n$.

proof: $Ax = b$ has solution iff $b \in R(A)$

By hypothesis $\forall\ b \in \mathbb{R}^m$, $b \in R(A)$

$\Rightarrow$ $\mathbb{R}^m \subseteq R(A)$

But we know, $R(A) \subseteq \mathbb{R}^m$.

Hence, $R(A) = \mathbb{R}^m \Rightarrow r(A) = m$

Also by observation above $r(A) \leq n$

Hence, $m \leq n$.

Remark 3 For $A \in M_{m,n}(\mathbb{R})$, we have known that $r(A) \leq \min\{m,n\}$.

(i) $r(A) < \min\{m,n\}$

$\Rightarrow r(A) < m$ and $r(A) < n$

$\Rightarrow R(A) \subsetneq \mathbb{R}^m$ and $\eta(A) = n - r(A) > 0$

$\Rightarrow$ For a given $b \in \mathbb{R}^m$,

- if $b \notin R(A)$, $Ax = b$ has no solution
- if $b \in R(A)$, $Ax = b$ has infinite number of solution. As $b \in R(A)$ $\exists\ s_1 \in \mathbb{R}^n$, so that $As_1 = b$.

The soln. set is $S_b = s_1 + S_0$.

Now as $\eta(A) = n - r(A) > 0$

By Corollary 1, $|S_0| = \infty$

Hence, $|S_b| = \infty$.

(ii) $r(A) = \min\{m,n\} = m < n$

$\Rightarrow r(A) = m$ and $r(A) < n$

$\Rightarrow R(A) = \mathbb{R}^m$ and $\eta(A) = n - r(A) > 0$.

$\Rightarrow$ For given $b \in \mathbb{R}^m$, $Ax = b$ has solution.

and as $r(A) < n$ by Corollary 1, $|S_0| = \infty$. $Ax = b$ has infinite number of soln.

(iii) $r(A) = \min\{m,n\} = n < m$

$\Rightarrow r(A) = n$ and $r(A) < m$

$\Rightarrow R(A) \subsetneq \mathbb{R}^m$ and $\eta(A) = n - r(A) = 0$

$\Rightarrow$ For a given $b \in \mathbb{R}^m$

if $b \notin R(A)$, $Ax = b$ has no solution

if $b \in R(A)$, $Ax = b$ has unique solution

As $S_b = s_1 + S_0$,

as $\eta(A) = 0$, $|S_0| = 1$

actually $0 \in S_0$.

so $S_b = s_1 + \{0\}$.

(iv) $r(A) = \min\{m,n\} = m = n$.

$\Rightarrow r(A) = m$ and $r(A) = n$

$\Rightarrow R(A) = \mathbb{R}^m$ and $\eta(A) = n - r(A) = 0$.

$\Rightarrow$ for given $b \in \mathbb{R}^m = \mathbb{R}^n$, the system $Ax = b$ has unique solution.

We are interested in case (iv)

Theorem 6

Let $A \in M_{n,n}(\mathbb{R}) := M_n(\mathbb{R})$

The followings are equivalent

(i) $Ax = 0$ has only trivial solution $(x = 0)$

(ii) for all $b \in \mathbb{R}^n$, $Ax = b$ has unique soln.

(iii) A is invertible.

We conclude: The system $Ax = b$

— uniquely solvable if A is invertible (non-singular)

— has infinitely many solution if A is singular and $b \in R(A) = Col(A)$

— has no solution if A is singular and $b \notin R(A) = Col(A)$.

Lecture - 9

We want to solve the equation

$$Ax = b \quad (\text{posed in } \mathbb{R}^m), \longrightarrow (1)$$

where $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$ is the unknown and

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in M_{m,n}(\mathbb{R}) \quad \text{and}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m \quad \text{be given,} \quad n, m \in \mathbb{N}.$$

We will consider the case $n = m = r(A)$ ie $\det A \neq 0$.

The system (1) has unique solution.

For other cases: Either the system has no soln. or infinite no. of solution.

For infinite soln case, choose a particular one putting extra constraint

For non existence case, Find $s \in \mathbb{R}^n$ such that $\|As - b\|_2 = \inf_{x \in \mathbb{R}^n} \|Ax - b\|_2$

Involves pseudoinverse

SVD / Polar D.

$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_2^2}$,

P-2 Method to Find solutions

Direct Method

- Gaussian elimination
- Factorization
- Gauss-Jordan

Iterative method

- Gauss-Jacobi
- Gauss-Seidel
- SOR
- residual corrector or iterative refinement

Gaussian Elimination

Motivation:

A = Diagonal $A = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$.

Eqns are

$a_{11}x_1 = b_1$

$a_{22}x_2 = b_2$

$a_{33}x_3 = b_3$

Soln is given by $x_i = b_i/a_{ii}$, $i=1,2,3$

why $a_{ii} \neq 0$? ($\because \det A \neq 0$)

For general n, $x_i = b_i/a_{ii}$ $i=1,\dots,n$. $i=1,2,3$

A = Lower triangular.

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Eqns

$a_{11}x_1 = b_1$

$a_{21}x_1 + a_{22}x_2 = b_2$

$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3$

$$x_1 = b_1/a_{11}$$

$$x_2 = \frac{1}{a_{22}}\left(b_2 - a_{21}x_1\right)$$

$$x_3 = \frac{1}{a_{33}}\left(b_3 - a_{31}x_1 - a_{32}x_2\right)$$

Why $a_{ii} \neq 0$ $i=1,2,3$

For general n, $x_1 = b_1/a_{11}$

$$x_j = \frac{1}{a_{jj}}\left(b_j - \sum_{k=1}^{j-1} a_{jk}x_k\right)$$

$j = 2, \ldots, n$

This method is called Forward substitution.

A = Upper triangular

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix},$$

Eqns

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$

$$a_{22}x_2 + a_{23}x_3 = b_2$$

$$a_{33}x_3 = b_3$$

$$x_3 = b_3/a_{33}$$

$$x_2 = \frac{1}{a_{22}}\left(b_2 - a_{23}x_3\right)$$

$$x_1 = \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3\right)$$

General $x_n = \frac{b_n}{a_{nn}}$, $x_j = \frac{1}{a_{jj}}\left(b_j - \sum_{k=i+1}^{n} a_{jk}x_k\right)$, $k = n-1, n-2, \ldots 1$

P-4

Gaussian Elimination method:

Converting the matrix A into an upper triangular matrix by elementary row operations. (It does not change the solution set)

Then using backward substitution we get the result.

Let us consider the 3×3 system.

$$a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 = b_1^{(1)} \longrightarrow E_1$$

$$a_{21}^{(1)} x_1 + a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 = b_2^{(1)} \longrightarrow E_2$$

$$a_{31}^{(1)} x_1 + a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 = b_3^{(1)} \longrightarrow E_3$$

To make it upper triangular form, we need to eliminate x_1 from eqn.s E_2, E_3.

Recall elementary Row operations

$$A = \begin{bmatrix} — R_1 — \\ — R_2 — \\ \cdots \\ — R_n — \end{bmatrix}$$

type-I. λR_i , $\lambda \neq 0$

-II $R_i \longleftrightarrow R_k$ $i \neq k$

-III $R_i \longrightarrow R_i + \lambda R_k$ $\lambda \neq 0$

Step-1: If $a_{11}^{(1)} \neq 0$

$$E_2 \longrightarrow E_2 - \frac{a_{21}^{(1)}}{a_{11}^{(1)}} E_1 = E_2 - m_{21} E_1, \quad m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}}$$

$$E_3 \longrightarrow E_3 - \frac{a_{31}^{(1)}}{a_{11}^{(1)}} E_1 = E_2 - m_{31} E_1, \quad m_{31} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}}$$

The new system:

$$a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 = b_1^{(1)}$$

$$0 + a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 = b_2^{(2)}$$

$$0 + a_{32}^{(2)} x_2 + a_{33}^{(2)} = b_3^{(2)}$$

and $a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)}$, $b_i^{(2)} = b_i^{(1)} - m_{i1} b_1^{(1)}$, $i,j = 2,3$

Step-2: If $a_{22}^{(2)} \neq 0$:

$$E_3 \longrightarrow E_3 - \frac{a_{32}^{(2)}}{a_{22}^{(2)}} E_2 = E_3 - m_{32} E_2$$

$$m_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}}$$

The new system

$$a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 = b_1^{(1)}$$

$$0 + a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 = b_2^{(2)}$$

$$0 + 0 + a_{33}^{(3)} x_3 = b_3^{(3)}$$

$a_{33}^{(3)} = a_{33}^{(2)} - m_{32} \times a_{23}^{(2)}$

$b_3^{(3)} = b_3^{(2)} - m_{32} b_2^{(2)}$

P-6 Use back word Substitution

$$x_3 = \frac{b_3^{(3)}}{a_{33}^{(3)}}$$

$$x_2 = \frac{1}{a_{22}^{(2)}}\left(b_2^{(2)} - a_{23}^{(2)} x_3\right)$$

$$x_1 = \frac{1}{a_{11}^{(1)}}\left(b_1^{(1)} - a_{12}^{(1)} x_2 - a_{13}^{(1)} x_3\right)$$

Note one thing

$$A = A^{(1)} \longrightarrow A^{(2)} \longrightarrow A^{(3)} = U.$$

$$A^{(2)} = L_1 A^{(1)}$$

$$A^{(3)} = L_2 A^{(2)}$$

$$I \xrightarrow[E_3 \to E_3 - m_{31}E_1]{E_2 \to E_2 - m_{21}E_1} L_1$$

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{bmatrix}$$

$$I \xrightarrow{E_3 \to E_3 - m_{32}E_2} L_2$$

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{bmatrix}$$

$$\therefore \quad U = A^{(3)} = L_2 A^{(2)} = L_2 L_1 A^{(1)} = L_2 L_1 A.$$

$$\therefore \quad A = L_1^{-1} L_2^{-1} U = LU$$

$$L = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix}$$

This is called LU decomposition of A

Example:

$$2x_1 + x_2 + 3x_3 = 2$$

$$4x_1 + 4x_2 + 7x_3 = 4$$

$$2x_1 + 5x_2 + 9x_3 = 8$$

$$A^{(1)} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{bmatrix} \qquad b^{(1)} = \begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix}$$

$$[A^{(1)} \mid b^{(1)}]$$

$$\xrightarrow[R_3 \to R_3 - R_1]{R_2 \to R_2 - 2R_1} \begin{bmatrix} 2 & 1 & 3 \\ 0 & 2 & 1 \\ 0 & 4 & 6 \end{bmatrix} \qquad b^{(2)} = \begin{bmatrix} 2 \\ 0 \\ 6 \end{bmatrix}$$

$A^{(2)}$

P-8

$$\xrightarrow{R_3 \to R_3 - 2R_2} \underbrace{\begin{bmatrix} 2 & 1 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 4 \end{bmatrix}}_{A^{(3)} = U} \quad \overbrace{\begin{bmatrix} 2 \\ 0 \\ 6 \end{bmatrix}}^{b^{(3)}}$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

$$x_3 = \frac{1}{4} \times 6 = 3/2$$

$$x_2 = 1/2\,(0 - 3/2) = -3/4$$

$$x_1 = 1/2\,(2 + 3/4 - 3 \times 3/2) = -7/8.$$

Qn. Is it possible to do Gaussian elimination for every matrix even $\det A \neq 0$.

Ans: No. We state the criterion.

But for $\det A \neq 0$, we wants to solve. what is the remedy then?

LU decomposition is not unique in general.

State: Existence and uniqueness of LU decomposition.

~~For glutist~~ ~~comment here.~~

LU - decomposition of a square matrix

Defn. (LU factorization / decomposition)

A matrix $A \in M_n(\mathbb{R})$ is said to have LU factorization / decomposition if there exists a lower triangular matrix L and an upper triangular matrix U such that

$$A = LU$$

Existence of LU decomposition

Theorem 1 (Horn & Johnson: Matrix Analysis)

Let $A \in M_n(\mathbb{R})$ with $\det A \neq 0$.

If all the leading principal submatrices $A[1,2,\ldots,k]$, $k = 1,2,\ldots,n$ are non-singular, then, A has an LU decomposition with both L and U are non-singular.

Eg: $A = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 3 & 4 \\ 2 & 1 & 3 \end{bmatrix}$

$A[1] = 1$, $\det A[1] = 1$

$A[1,2] = \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix}$, $\det A[1,2] = 0$

$A[1,2,3] = A$, $\det A[1,2,3] = -9 \neq 0$.

Remark 1 LU decomposition of a matrix is not unique. As we can write

$$A = LU = LD \cdot D^{-1}U = \tilde{L}\tilde{U}$$ where $\tilde{L} = LD$, $\tilde{U} = D^{-1}U$

for any non-singular diagonal matrix D.

Eg:
$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\begin{pmatrix} 1/a & 0 \\ 0 & 1/b \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}$$

for any $a, b \in \mathbb{R}$, $a, b \neq 0$

$$= \underbrace{\begin{pmatrix} a & 0 \\ 3a & b \end{pmatrix}}_{\tilde{L}}\underbrace{\begin{pmatrix} 1/a & 2/a \\ 0 & -2/b \end{pmatrix}}_{\tilde{U}}$$

We have plenty of matrices $\tilde{L}, \tilde{U}$ for different non-zero real numbers a and b.

Theorem 2 (LDU-decomposition) [Horn and Johnson: Matrix Analysis]

Let $A \in M_n(\mathbb{R})$ be non-singular ($\det A \neq 0$).

If all the leading principal submatrices

$A[1,2,\ldots,k]$, $k = 1,2,\ldots,n$ are non-singular,

then the matrix A can be factored as

$$A = LDU,$$

where L = unit lower triangular

U = unit upper triangular

D = diagonal matrix $= \operatorname{diag}(d_1, d_2, \ldots, d_n)$,

$$d_1 = a_{11}$$

$$d_i = \frac{\det(A[1,2,\ldots,i])}{\det(A[1,2,\ldots,i-1])}, \quad i = 1,2,\ldots,n$$

Moreover, the factors L, U, D are unique.

Eg:
$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 4 \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 1 & 1/2 & 3/2 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}}_{U}.$$

P-4

There are three special types of LU factorization.

(i) Doolittle's $A = LU$, L = unit lower triangular

(ii) Crout's $A = LU$, U = unit upper triangular

(iii) Cholesky's $A = LL^T$.

Defn. (Doolittle's factorization)

A matrix $A \in M_n(\mathbb{R})$ is said to have a Doolittle's factorization if there exist a lower triangular matrix L with all diagonal elements 1 and an upper triangular matrix U such that $A = LU$

Defn. (Crout's factorization)

A matrix $A \in M_n(\mathbb{R})$ is said to have a Crout's factorization if there exist a lower triangular matrix L and an upper triangular matrix U with all diagonal elements 1 such that

$$A = LU.$$

Eg(1) Doolittle's method

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12}+u_{22} & l_{21}u_{13}+u_{23} \\ l_{31}u_{11} & l_{31}u_{12}+l_{32}u_{22} & l_{31}u_{13}+l_{32}u_{23}+u_{33} \end{pmatrix}$$

Comparing:

1st row: $\begin{cases} u_{11} = 2 \\ u_{12} = 1 \\ u_{13} = 3 \end{cases}$

1st Column: $\begin{cases} l_{21}u_{11} = 4 \Rightarrow l_{21} \times 2 = 4 \Rightarrow l_{21} = 2 \\ l_{31}u_{11} = 2 \Rightarrow l_{31} \times 2 = 2 \Rightarrow l_{31} = 1 \end{cases}$

2nd row: $\begin{cases} l_{21}u_{12} + u_{22} = 4 \Rightarrow 2 \times 1 + u_{22} = 4 \Rightarrow u_{22} = 2 \\ l_{21}u_{13} + u_{23} = 7 \Rightarrow 2 \times 3 + u_{23} = 7 \Rightarrow u_{23} = 1 \end{cases}$

2nd Column: $\{ l_{31}u_{12} + l_{32}u_{22} = 5 \Rightarrow 1 \times 1 + l_{32} \times 2 = 5 \Rightarrow l_{32} = 2$

3rd row: $\{ l_{31}u_{13} + l_{32}u_{23} + u_{33} = 9 \Rightarrow 1 \times 3 + 2 \times 1 + u_{33} = 9$

$\Rightarrow u_{33} = 4$

Finally we have

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 4 \end{pmatrix}$$

Eg(2) (Crout's method)

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12}+l_{22} & l_{21}u_{13}+l_{22}u_{23} \\ l_{31} & l_{31}u_{12}+l_{32} & l_{31}u_{13}+l_{32}u_{23}+l_{33} \end{pmatrix}$$

Comparing.

1st column:
$$\begin{cases} l_{11} = 2 \\ l_{21} = 4 \\ l_{31} = 2 \end{cases}$$

1st row:
$$\begin{cases} l_{11}u_{12} = 1 \Rightarrow 2 \times u_{12} = 1 \Rightarrow u_{12} = 1/2 \\ l_{11}u_{13} = 3 \Rightarrow 2 \times u_{13} = 3 \Rightarrow u_{13} = 3/2 \end{cases}$$

2nd Column:
$$\begin{cases} l_{21}u_{12} + l_{22} = 4 \Rightarrow 4 \times 1/2 + l_{22} = 4 \Rightarrow l_{22} = 2 \\ l_{31}u_{12} + l_{32} = 5 \Rightarrow 2 \times 1/2 + l_{32} = 5 \Rightarrow l_{32} = 4 \end{cases}$$

2nd row: $\left\{ l_{21}u_{13} + l_{22}u_{23} = 7 \right. \Rightarrow 4 \times \frac{3}{2} + 2 \times u_{23} = 7$

$\Rightarrow u_{23} = 1/2$

3rd column: $\left\{ l_{31}u_{13} + l_{32}u_{23} + l_{33} = 9 \right.$

$\Rightarrow 2 \times 3/2 + 4 \times 1/2 + l_{33} = 9$

$\Rightarrow l_{33} = 4$

Finally

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 4 & 2 & 0 \\ 2 & 4 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

General Algorithm for Doolittle's factorization

$A = (a_{ij})_{n \times n} = LU$. $L = (L_{ij})_{n \times n}$

$U = (U_{ij})_{n \times n}$

$$L_{ij} = \begin{cases} 0, & 1 \le i < j \le n \\ 1, & 1 \le i = j \le n \\ L_{ij}, & 1 \le j < i \le n \end{cases}$$

$$U_{ij} = \begin{cases} 0, & 1 \le j < i \le n \\ U_{ij}, & 1 \le i \le j \le n \end{cases}$$

P-8

$$A = LU$$

$$\Rightarrow \quad a_{ij} = (LU)_{ij}$$

$$= \sum_{k=1}^{n} L_{ik} U_{kj}$$

$$= \sum_{k=1}^{i-1} L_{ik} U_{kj} + L_{ii} U_{ij} + \sum_{k=i+1}^{n} L_{ik} U_{kj}$$

$$\sum_{k=i+1}^{n} L_{ik} U_{kj} = 0$$

$$\boxed{a_{ij} = \sum_{k=1}^{i-1} L_{ik} U_{kj} + L_{ii} U_{ij}} \qquad (\text{[illegible]})$$

1st row: $i=1$, $\quad a_{1j} = U_{1j}$, $j = 1, 2, \cdots, n$

1st column: $j=1$ $\quad a_{i1} = \sum_{k=1}^{i-1} L_{ik} U_{k1} + U_{i1}$

$i=1$, [illegible] $U_{11} = a_{11}$

$i = 2, 3, \ldots, n$

$$a_{i1} = L_{ii} U_{11} \qquad (\because U_{k1} = 0,\ U_{i1} = 0 \text{ for } i, k \geq 2)$$

$$\Rightarrow L_{i1} = a_{i1} / U_{11}$$

as $U_{11} = a_{11} \neq 0$.

For $i = 2, 3, \ldots, n-1$

$$u_{ii} = a_{ii} - \sum_{k=1}^{i-1} L_{ik} U_{ki}$$

For $j = i+1, \ldots, n$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} L_{ik} U_{kj} \quad \text{(ith row)} \quad (\because L_{ii} = 1)$$

$$l_{ji} = \frac{1}{u_{ii}} \left(a_{ji} - \sum_{k=1}^{i-1} L_{jk} U_{ki} \right) \quad \text{(ith column)}$$

Finally $i = n$,

$$u_{nn} = a_{nn} - \sum_{k=1}^{n-1} L_{nk} U_{kn}.$$

General Algorithm for Crout's factorization

$A = (a_{ij})_{n \times n}$ $\quad L = (L_{ij})_{n \times n}$, $\quad U = (U_{ij})_{n \times n}$

$$L_{ij} = \begin{cases} 0, & 1 \le i < j \le n \\ L_{ij}, & 1 \le j \le i \le n \end{cases}$$

$$U_{ij} = \begin{cases} 0, & 1 \le j < i \le n, \\ 1, & 1 \le i = j = n, \\ U_{ij}, & 1 \le i < j \le n. \end{cases}$$

P-10

$$a_{ij} = (LU)_{ij} = \sum_{k=1}^{n} L_{ik} U_{kj}$$

$$= \sum_{k=1}^{i-1} L_{ik} U_{kj} + L_{ii} U_{ij} + 0$$

For $j=1$, $a_{i1} = \sum_{k=1}^{i-1} L_{ik} U_{k1} + L_{ii} U_{i1}$ (1st column)

$i=1 \Rightarrow a_{11} = L_{11} U_{11} \Rightarrow L_{11} = a_{11}$ $(\because U_{11}=1)$

$i=2,\ldots,n$, $a_{i1} = L_{i1} U_{11}$, $L_{i1} = a_{i1}$ (1st row)

For $j = 2, \ldots, n-1$

$$L_{jj} = a_{jj} - \sum_{k=1}^{j-1} L_{jk} U_{kj}$$

$i = j+1, \ldots, n$

$$L_{ij} = a_{ij} - \sum_{k=1}^{j-1} L_{ik} U_{kj}, \quad (j\text{th-Column.})$$

$$U_{ji} = \frac{1}{l_{jj}} \left(a_{ji} - \sum_{k=1}^{j-1} L_{jk} U_{ki} \right) \quad (j\text{th row})$$

Finally, $j = n$

$$l_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk} U_{kn}$$

Lecture - 11

Definition (Symmetric matrix)

A matrix $A \in M_n(\mathbb{R})$ is said to be symmetric matrix if $A = A^T$ (Transpose of A)

Defn. (positive semi-definite)

A symmetric matrix $A \in M_n(\mathbb{R})$ is said to be positive semi-definite if

$$\forall\ x \in \mathbb{R}^n, \quad x^T A x \geq 0$$

Defn. (positive definite)

A symmetric matrix $A \in M_n(\mathbb{R})$ is said to be positive definite if

$$\forall\ x \in \mathbb{R}^n - \{0\}, \quad x^T A x > 0$$

Defn. (Submatrices)

Let $k \in \mathbb{N}$, $1 \leq k \leq m$ and $l \in \mathbb{N}$, $1 \leq l \leq n$

Let $\alpha \subseteq \{1, 2, .., m\}$ with $|\alpha| = k$ (Cardinality of α)

and $\beta \subseteq \{1, 2, .., n\}$ with $|\beta| = l$ (" of β)

Let $A = (a_{ij}) \in M_{m,n}(\mathbb{R})$.

A submatrices of A of order $k \times l$ with indices α and β is a matrix $A[\alpha, \beta] \in M_{k,l}(\mathbb{R})$.

P-2 such that

$$A[\alpha,\beta] = (a_{ij})_{i\in\alpha, j\in\beta}.$$

If $|\alpha| = |\beta| = k$, then $A[\alpha,\beta] \in M_k(\mathbb{R})$ is a square matrix. Determinant of $A[\alpha,\beta]$ is called a minor.

① If $\alpha = \beta$ (in that case ofcourse $k \le \min\{m,n\}$)

Then the matrix $A[\alpha] = A[\alpha,\alpha] = (a_{ij})_{i,j\in\alpha} \in M_k(\mathbb{R})$ is called a principal submatrix of order k and determinant of $A[\alpha]$ is called a principal minor.

② If $\alpha = \beta = \{1, 2, .., k\}$. Then $A[\alpha] = A[\alpha,\alpha]$ is called leading principal submatrix of order k and determinant $A[\alpha]$ is called the leading principal minor.

Eg:
$$A=\begin{bmatrix}1 & 2 & 3 & 4 & 5\\ 0 & 9 & 8 & 7 & 6\\ 1 & 3 & 5 & 7 & 9\end{bmatrix}_{3\times 5}$$

$\alpha=\{2,3\}$, $\beta=\{1,2,5\}$.

$$A[\alpha,\beta]=\begin{bmatrix}a_{21} & a_{22} & a_{25}\\ a_{31} & a_{32} & a_{35}\end{bmatrix}=\begin{bmatrix}0 & 9 & 6\\ 1 & 3 & 9\end{bmatrix}$$

a submatrix of order 2×3 with indices $\{2,3\}$, $\{1,2,5\}$.

$\alpha_1=\{1,3\}$, $=\alpha=\{1,3\}=\beta_1$

$$A[\alpha_1]=\begin{bmatrix}a_{11} & a_{13}\\ a_{31} & a_{33}\end{bmatrix}=\begin{bmatrix}1 & 3\\ 1 & 5\end{bmatrix}$$

a principal submatrix of order 2

$\alpha_2=\{1,2\}=\beta_2$

$$A[\alpha_2]=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}=\begin{bmatrix}1 & 2\\ 0 & 9\end{bmatrix}$$

(leading principal of order 2)

$\alpha_3=\{1,2,3\}=\beta_3$

$$A[\alpha_3]=\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}=\begin{bmatrix}1 & 2 & 3\\ 0 & 9 & 8\\ 1 & 3 & 5\end{bmatrix}$$

(leading principal of order 3)

P-4

Definition (cholesky factorization)

A matrix $A \in M_n(\mathbb{R})$ is said to have a cholesky decomposition if there exists a lower triangular matrix $L \in M_n(\mathbb{R})$ such that

$$A = LL^T$$

---

we note that if A has a cholesky decomposition then A is necessarily symmetric and positive-semidefinite. Furthermore if A is non-singular, then A is positive-definite.

$$A = LL^T \Rightarrow A^T = (LL^T)^T = (L^T)^T L^T = LL^T = A.$$

For $x \in \mathbb{R}^n$.

$$x^T Ax = x^T LL^T x = (L^T x)^T L^T x = y^T y \geq 0, \quad y = L^T x$$

Theorem 1

A matrix $A \in M_n(\mathbb{R})$ is positive definite iff A has a cholesky factorization. $A = LL^T$ with L = lower triangular with non-zero diagonal elements.

proof: (postponed)

Lemma 1 Let $A \in M_n(\mathbb{R})$ be a symmetric matrix.

The following are equivalent

(i) A is positive definite

(ii) All the principal minors of the matrix A are positive

(iii) All the eigen-values of the matrix A are positive.

Algorithm to find cholesky decomposition

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

$\det A[1] = 4 > 0$

$\det A[1,2] = 16 - 1 = 15 > 0$

$\det A[1,2,3] = 4(12-4) - 1(2+3) + 1(-2-4)$

$= 32 - 5 - 6 = 21 > 0.$

Note $A = A^T$.

$\Rightarrow$ A is positive definite.

$$\begin{bmatrix} 4 & -1 & 1 \\ -1 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

P-6

$$\begin{bmatrix} 4 & -1 & 1 \\ -1 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{21}l_{11} & l_{21}^2+l_{22}^2 & l_{21}l_{31}+l_{22}l_{32} \\ l_{31}l_{11} & l_{21}l_{31}+l_{22}l_{32} & l_{31}^2+l_{32}^2+l_{33}^2 \end{bmatrix}$$

Comparing row wise (Columnwise also can be done)

$l_{11}^2 = 4 \Rightarrow l_{11} = 2$ (one could also take $l_{11} = -2$ and proceed)

$l_{11}l_{21} = -1 \Rightarrow l_{21} = -1/2$

$l_{11}l_{31} = 1 \Rightarrow l_{31} = 1/2$

$l_{21}^2 + l_{22}^2 = 4 \Rightarrow 1/4 + l_{22}^2 = 4 \Rightarrow l_{22}^2 = 4 - 1/4 = 15/4$

$l_{22} = \sqrt{15}/2$ (one could also take $l_{22} = -\sqrt{15}/2$)

$l_{21}l_{31} + l_{22}l_{32} = 2$

$-1/2 \times 1/2 + \sqrt{15}/2\, l_{32} = 2 \Rightarrow \sqrt{15}/2\, l_{32} = 2 + 1/4 = 9/4$

$\Rightarrow l_{32} = \frac{9}{2\sqrt{15}}$

$l_{31}^2 + l_{32}^2 + l_{33}^2 = 3$

$1/4 + \frac{81}{60} + l_{33}^2 = 3 \Rightarrow l_{33}^2 = 3 - 1/4 - \frac{27}{20}$

$= \frac{60-5-27}{20} = \frac{28}{20} = 7/5$

$l_{33} = \sqrt{7/5}$ (-ve can also be taken)

$$L = \begin{pmatrix} 2 & 0 & 0 \\ -1/2 & \sqrt{15}/2 & 0 \\ 1/2 & \frac{9}{2\sqrt{15}} & \sqrt{7}/5 \end{pmatrix}$$

General Algorithm

$A = LL^T \Rightarrow A = a_{ij}$ $\quad L = \begin{cases} 0, & 1 \le i < j \le n \\ L_{ij} & 1 \le j \le i \le n \end{cases}$

$$a_{ij} = (LL^T)_{ij} = \sum_{k=1}^{n} L_{ik} L^T_{kj}$$

$$= \sum_{k=1}^{n} L_{ik} L_{jk} \qquad (\because L^T_{kj} = L_{jk})$$

$$= \sum_{k=1}^{i \wedge j} L_{ik} L_{jk} \qquad \left(\begin{array}{l} \because L_{ik} = 0 \text{ if } k > i \\ L_{jk} = 0 \text{ if } k > j \\ i \wedge j = \min\{i, j\} \end{array}\right)$$

$$\boxed{a_{ij} = \sum_{k=1}^{i \wedge j} L_{ik} L_{jk}}$$

For $i = 1$, and $j = 1$ $\quad a_{11} = \sum_{k=1}^{1} L_{1k} L_{1k} = L_{11}^2$

$$L_{11} = \sqrt{a_{11}} \qquad (\because a_{11} > 0, \ \det A[1] = a_{11} > 0)$$

$j = 2, \cdots, n$

$$a_{1j} = \sum_{k=1}^{1 \wedge j} L_{1k} L_{jk} = L_{11} L_{j1}$$

$$\Rightarrow L_{j1} = a_{1j} / L_{11} = \frac{a_{1j}}{\sqrt{a_{11}}}.$$

P-8

For $i = 2, 3, \ldots, n-1$

$$a_{ij} = \sum_{k=1}^{i \wedge j} L_{ik} L_{jk}$$

For $j = i$, $a_{ii} = \sum_{k=1}^{i} L_{ik} L_{ik}$

$$= \sum_{k=1}^{i-1} L_{ik}^2 + L_{ii}^2$$

$$\Rightarrow L_{ii}^2 = a_{ii} - \sum_{k=1}^{i-1} L_{ik}^2$$

$$\Rightarrow L_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} L_{ik}^2 \right)^{1/2}$$

For $j = i+1, \ldots, n$

$$a_{ij} = \sum_{k=1}^{i \wedge j} L_{ik} L_{jk} = \sum_{k=1}^{i} L_{ik} L_{jk}$$

$$= \sum_{k=1}^{i-1} L_{ik} L_{jk} + L_{ii} L_{ji}$$

$$\Rightarrow L_{ii} L_{ji} = a_{ij} - \sum_{k=1}^{i-1} L_{ik} L_{jk}$$

$$\Rightarrow L_{ji} = \left(a_{ij} - \sum_{k=1}^{i-1} L_{ik} L_{jk} \right) \Big/ L_{ii}$$

For $i = n$, $j = n$ $\quad a_{nn} = \sum_{k=1}^{n} L_{nk} L_{nk} = \sum_{k=1}^{n-1} L_{nk}^2 + L_{nn}^2$

$$\Rightarrow L_{nn} = \left(a_{nn} - \sum_{k=1}^{n-1} L_{nk}^2 \right)^{1/2}.$$

Finally

$i=j=1$, $L_{11} = \sqrt{a_{11}}$

For $i=1$, $j=2,..,n$ $L_{j1} = a_{j1}/L_{11}$

For $i=2,..,n-2$, $j=i$, $L_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} L_{ik}^2\right)^{1/2}$

$j=i+1,..,n$

$$L_{ji} = \left(a_{ij} - \sum_{k=1}^{i-1} L_{ik}L_{jk}\right)/L_{ii}$$

For $i=j=n$, $L_{nn} = \left(a_{nn} - \sum_{k=1}^{n-1} L_{nk}^2\right)^{1/2}$.

Tridiagonal system

A matrix of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & & & 0 \\ a_{21} & a_{22} & a_{23} & & & 0 \\ 0 & a_{32} & a_{33} & a_{34} & & \\ 0 & 0 & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & & & 0 & a_{n,n-1} & a_{nn} \end{bmatrix}$$

ie. $a_{ij} = 0$, for $|i-j| \geq 2$.

P-10

We now show a LU decomposition (Crout's) of a tridiagonal matrix (we assume that LU decomposition is possible)

A has at most $n+2(n-1) = 3n-2$ non-zero entries

One can show that the Crout's decomposition of A is of the form.

$$A = LU \quad \text{where}$$

$$L = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \ddots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & L_{n,n-1} \; L_{n,n} \end{bmatrix} \qquad L_{ij} = \begin{cases} 0, & i<j \\ & i>j+1 \end{cases}$$

and

$$U = \begin{bmatrix} 1 & U_{12} & 0 & \cdots & 0 \\ 0 & 1 & U_{23} & \ddots & 0 \\ \vdots & \ddots & 1 & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ 0 & & & 0 & 1 \end{bmatrix} \qquad \begin{array}{l} U_{ij} = 0, \; i>j \\ \qquad\; i<j-1 \\ U_{ii} = 1 \end{array}$$

$$A = LU, \quad a_{ij} = (LU)_{ij} = \sum_{k=1}^{n} L_{ik} U_{kj}$$

$$a_{ij} = \sum_{k=1}^{i-1} L_{ik} U_{kj} + L_{ii} U_{ij}$$

$$= \sum_{k=1}^{(i-1) \wedge j} L_{ik} U_{kj} + L_{ii} U_{ij}$$

$$a_{11} = L_{11}$$

$$a_{i,i-1} = L_{i,i-1} \quad \text{for} \quad i = 2, 3, \ldots, n$$

$$a_{ii} = L_{i,i-1} U_{i-1,i} + L_{ii}, \quad i = 2, 3, \ldots, n$$

$$a_{i,i+1} = L_{ii} U_{i,i+1} \quad \text{for} \quad i = 2, 3, \ldots, n-1$$

Eg:

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 & 0 \\ L_{21} & L_{22} & 0 & 0 \\ 0 & L_{32} & L_{33} & 0 \\ 0 & 0 & L_{43} & L_{44} \end{bmatrix} \begin{bmatrix} 1 & U_{12} & 0 & 0 \\ 0 & 1 & U_{23} & 0 \\ 0 & 0 & 1 & U_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

b. [illegible]

$$= \begin{bmatrix} L_{11} & L_{11}U_{12} & 0 & 0 \\ L_{21} & L_{21}U_{12}+L_{22} & L_{22}U_{23} & 0 \\ 0 & L_{32} & L_{32}U_{23}+L_{33} & L_{33}U_{34} \\ 0 & 0 & L_{43} & L_{43}U_{34}+L_{44} \end{bmatrix}$$

Complete it

Pivoting

We note that in Gaussian elimination after $(k-1)$th step (for $1 \le k \le n-1$), the transformed matrix takes the form

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & & \ddots & & & \\ 0 & 0 & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & a_{k+1,k}^{(k)} & & \vdots \\ & & & \vdots & & \\ 0 & 0 & & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix}$$

where $a_{ll}^{(l)} \neq 0$, $l = 1, \ldots, k-1$. are called pivot elements.

If $a_{kk}^{(k)} = 0$, then we cannot proceed to define the multiplier m_{ik} to eliminate the entries in the kth column below $a_{kk}^{(k)}$. the usual Gaussian elimination is stopped.

We need to modify the method

The idea is to scan the kth column of $A^{(k)}$ from the kth row through nth row for

the first non-zero entry.

If $a_{pk}^{(k)} \neq 0$ with $k+1 \leq p \leq n$ (smallest), then

we do the row interchanges

$$R_k \longleftrightarrow R_p$$

and the process is continued.

This method is called pivoting.

To reduce round off error, it is sometime require to perform row interchanges even if the pivot elements are not zero.

(i) If $a_{kk}^{(k)}$ is small in magnitude compare to $a_{jk}^{(k)}$, $k \leq j \leq n$, the the magnitude of the multiplier

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

will be very large. compare to other elements.

Hence, round-off error is incorporated in the computation of one of the terms

$$a_{il}^{(k+1)} = a_{il}^{(k)} - m_{ik} a_{kl}^{(k)}.$$

P-14

(ii) When performing the backward substitution

$$x_k = \left(b_k^{(k)} - \sum_{j=k+1}^{n} a_{kj}^{(k)} \right) \Big/ a_{kk}^{(k)}$$

with a very smaller value of $a_{kk}^{(k)}$ compare to the numerator, any error in the numerator might be increased because of the division by $a_{kk}^{(k)}$.

Eg: Consider the system

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Gaussian elimination

$$\begin{pmatrix} \varepsilon & 1 \\ 0 & 1-1/\varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2-1/\varepsilon \end{pmatrix}$$

$$x_2 = \frac{2-1/\varepsilon}{1-1/\varepsilon}, \quad x_1 = \frac{1}{\varepsilon}(1-x_2) = \frac{1}{1-\varepsilon}$$

If the computer has 5-digit rounding approximation let $\varepsilon = 10^{-6}$. Then, $2-1/\varepsilon = 2-10^6 = -.999998$

$$1-1/\varepsilon = 1-10^6 = -.999999$$

$$fl(2-1/\varepsilon) = -10^6 = fl(1-1/\varepsilon) \Rightarrow x_2 = 1 \text{ and } x_1 = 0.$$

But $x_2 \approx 1$ and $x_1 \approx 1$ (Computer commits large error).

Partial pivoting

To avoid this problem, pivoting is performed by $a_{pq}^{(k)}$ with larger in magnitude.

Select an element in the same column that is below the diagonal and has the largest absolute value.

Let p be the smallest integer with $k \le p \le n$ such that

$$|a_{pk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|$$

if $p \neq k$, perform $R_k \longleftrightarrow R_p$ (row interchanges)

This method is called partial pivoting.

Complete pivoting

(For better accuracy but much time required)

Pivoting can incorporate the interchange of both rows and columns

choose smallest integere p, q with

$k \le p \le n$, $k \le q \le n$.

P-16

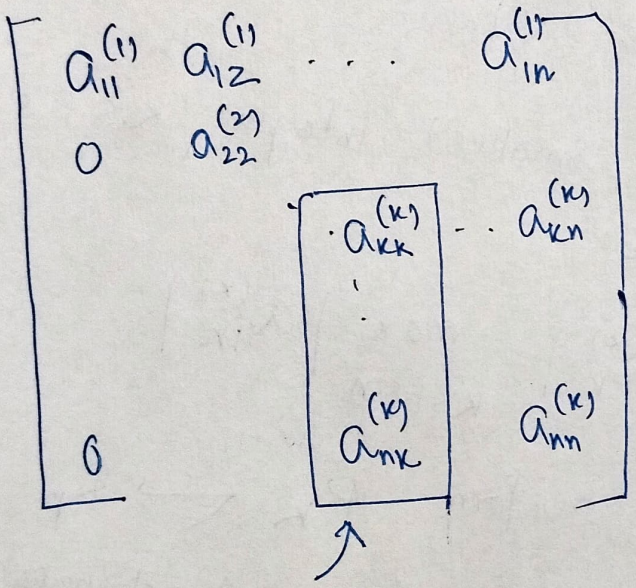
so that

$$|a_{pq}^{(k)}| = \max_{\substack{k \le i \le n \\ k \le j \le n}} |a_{ij}^{(k)}|$$

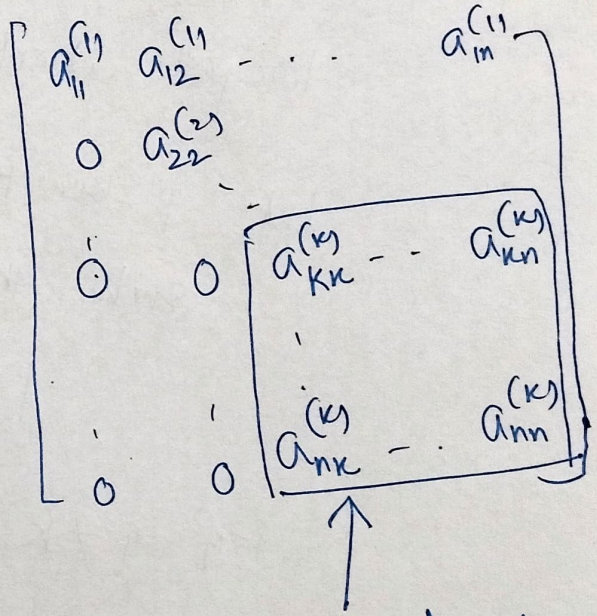
If $p \neq k$, perform $R_k \longleftrightarrow R_p$ and

if $q \neq k$ " $C_k \longleftrightarrow C_p$

(Column interchange)



Compare elements here for partial pivoting



Compare elements here for Complete pivoting.

Lecture-12

Let V be a vector space over the field $\mathbb{R}$.

Defn. (Norm)

A norm on V is a function $\|\cdot\| : V \longrightarrow \mathbb{R}$ satisfying

(i) $\|x\| \geq 0$, $\forall\, x \in V$ and $\|x\| = 0$ iff $x = 0$

(ii) $\|\alpha x\| = |\alpha| \|x\|$ $\forall\, \alpha \in \mathbb{R}$, $x \in V$, (Homogeneity)

(iii) $\|x+y\| \leq \|x\| + \|y\|$, $\forall\, x, y \in V$ (Triangle inequality)

On the vector space V, we already have an algebraic structure given by vector addition and scalar multiplication.

A norm on V introduces a topological structure on V and this topology on V is induced by a metric $d : V \times V \longrightarrow \mathbb{R}$, given by

$$d(x,y) = \|x - y\|, \quad x, y \in V \qquad \text{①}$$

One can show that, d defined in ①, satisfies all the defining properties of a metric.

P-2 This metric d has two special properties.

1. (Translation invariance)

$$d(x+z, y+z) = \|(x+z) - (y+z)\|$$

$$= \|x - y\|$$

$$= d(x, y), \quad \forall\ x, y, z \in V.$$

2. (Convexity of open balls in V)

Defn. A subset $C \subset V$ is said to be convex if $w, z \in C$ and $\alpha \in [0,1]$, then, $\alpha w + (1-\alpha) z \in C$.

Let $r > 0$, $x_0 \in V$. Open ball of radius r and centred at x_0 is given by

$$B_r(x_0) = \{ y \in V \mid \|y - x_0\| < r \}.$$

⊙ $B_r(x_0)$ is convex.

Let $w, z \in B_r(x_0)$ and $\alpha \in [0, 1]$.

To show. $\alpha w + (1-\alpha) z \in B_r(x_0)$.

$$\|(\alpha w + (1-\alpha) z) - x_0\| = \|\alpha w + (1-\alpha) z - (\alpha x_0 + (1-\alpha) x_0)\|$$

$$= \|\alpha (w - x_0) + (1-\alpha)(z - x_0)\|$$

$$\|\alpha\omega + (1-\alpha)z - x_0\| \leq \|\alpha(\omega - x_0)\| + \|(1-\alpha)(z-x_0)\|$$

(using Triangle inequality)

$$= |\alpha| \|\omega - x_0\| + |1-\alpha| \|z - x_0\|$$

$$< \alpha r + (1-\alpha) r = r$$

($\because \alpha \in [0,1]$, $\alpha \geq 0$ and $1-\alpha \geq 0$)

$$\left(\begin{array}{l} \omega \in B_r(x_0) \Rightarrow \|\omega - x_0\| < r \\ z \in B_r(x_0) \Rightarrow \|z - x_0\| < r \end{array} \right)$$

$\therefore \|\alpha\omega + (1-\alpha)z - x_0\| < r$

$\Rightarrow \alpha\omega + (1-\alpha)z \in B_r(x_0)$.

Hence, $B_r(x_0)$ is Convex.

Defn. (Convergence of a sequence in V.)

A sequence $\{x_n\}_{n \in \mathbb{N}} \subset V$ is said to converge to $x \in V$ if

$$\lim_{n \to \infty} \|x_n - x\| = 0.$$

i.e. given any $\varepsilon > 0$, $\exists\ N \in \mathbb{N}$ such that

$\forall\ n \geq N \Rightarrow \|x_n - x\| < \varepsilon$.

P-4 On the product space $V \times V$, let us consider a metric $d_1 : (V \times V) \times (V \times V) \longrightarrow \mathbb{R}$ given by

$$d_1((x_1, y_1), (x_2, y_2)) = \|x_1 - x_2\| + \|y_1 - y_2\|$$

$$\forall\ (x_1, y_1), (x_2, y_2) \in V \times V.$$

d_1 induces a topology on $V \times V$.

[one could also consider other metrics on $V \times V$

for example: $1 \leq p < \infty$

$$d_p((x_1, y_1), (x_2, y_2)) = \left(\|x_1 - x_2\|^p + \|y_1 - y_2\|^p\right)^{1/p}$$

$$d_\infty((x_1, y_1), (x_2, y_2)) = \max\left\{\|x_1 - x_2\|, \|y_1 - y_2\|\right\}.$$

All these metrices generates the same topology on $V \times V$]

Similarly on $\mathbb{R} \times V$, let us considere a metric $\tilde{d}_1 : (\mathbb{R} \times V) \times (\mathbb{R} \times V) \longrightarrow \mathbb{R}$ given by

$$\tilde{d}_1((\alpha_1, x_1), (\alpha_2, x_2)) = |\alpha_1 - \alpha_2| + \|x_1 - x_2\|.$$

$$\forall\ (\alpha_1, x_1), (\alpha_2, x_2) \in \mathbb{R} \times V.$$

$\tilde{d}_1$ induces a topology on $\mathbb{R} \times V$.

On V we have two algebraic operations called vector addition and scalar multiplication.

- $a : V \times V \longrightarrow V$

$$(x,y) \longmapsto a(x,y) = x+y \qquad \text{(vector addition)}$$

$$m : \mathbb{R} \times V \longrightarrow V$$

$$(\alpha, x) \longmapsto m(\alpha, x) = \alpha x \qquad \text{(scalar multiplication)}$$

⊙ $a : V \times V \longrightarrow V$ is continuous.

We use sequential approach to show continuity of a.

Let $\{(x_n, y_n)\}_{n \in \mathbb{N}} \subset V \times V$ be a sequence such that $(x_n, y_n) \longrightarrow (x,y)$ in $V \times V$ as $n \to \infty$

i.e., $d_1((x_n, y_n), (x,y)) \longrightarrow 0$ as $n \to \infty$

i.e., $\|x_n - x\| + \|y_n - y\| \longrightarrow 0$ as $n \to \infty$. $\longrightarrow (*)_1$

To show $a(x_n, y_n) \longrightarrow a(x,y)$ in V as $n \to \infty$.

$$d(a(x_n,y_n), a(x,y)) = \|a(x_n,y_n) - a(x,y)\|$$

$$= \|(x_n + y_n) - (x+y)\|$$

$$= \|(x_n - x) + (y_n - y)\|$$

$$\leq \|x_n - x\| + \|y_n - y\| \qquad \text{(by Triangle inequality)}$$

$$\longrightarrow 0 \text{ as } n \to \infty \text{ (by } (*)_1\text{)}$$

P-6

Hence, $d(a(x_n,y_n), a(x,y)) \longrightarrow 0$ as $n \to \infty$

ie $a(x_n,y_n) \longrightarrow a(x,y)$ as $n \to \infty$.

So a is continuous.

⊙ m is continuous

Let $\{(\alpha_n, x_n)\}_n \subset \mathbb{R} \times V$ be a sequence such that

$(\alpha_n, x_n) \longrightarrow (\alpha, x)$ in $\mathbb{R} \times V$ as $n \to \infty$

To show $m(\alpha_n, x_n) \longrightarrow m(\alpha, x)$ in V as $n \to \infty$.

$$d(m(\alpha_n, x_n), m(\alpha, x)) = \|(\alpha_n, x_n) - (\alpha, x)\|$$

$$= \|\alpha_n x_n - \alpha x\|$$

$$= \|\alpha_n (x_n - x) + (\alpha_n - \alpha) x\|$$

$$\leq \|\alpha_n (x_n - x)\| + \|(\alpha_n - \alpha) x\|$$

$$\leq |\alpha_n| \|x_n - x\| + |\alpha_n - \alpha| \|x\|$$

$$\leq M \|x_n - x\| + |\alpha_n - \alpha| \|x\|$$

$(\because |\alpha_n - \alpha| + \|x_n - x\| \to 0$ as $n \to \infty$

$|\alpha_n - \alpha| \longrightarrow 0$ as $n \to \infty$

$\Rightarrow \{\alpha_n\}$ is bdd $\Rightarrow \exists\ M > 0$ such that $|\alpha_n| \leq M \ \forall\ n \in \mathbb{N}$)

$$d(m(\alpha_n, x_n), m(\alpha, x)) \leq \max\{M, \|x\|\} (|\alpha_n - \alpha| + \|x_n - x\|)$$

$$= \max\{M, \|x\|\} \tilde{d}_1((\alpha_n, x_n), (\alpha, x))$$

$$\longrightarrow 0 \quad \text{as} \quad n \to \infty$$

$$(\because (\alpha_n, x_n) \to (\alpha, x) \text{ in } \mathbb{R} \times V)$$

Hence, $m(\alpha_n, x_n) \longrightarrow m(\alpha, x)$ as $n \to \infty$.

$\therefore$ m is continuous.

① The norm ie. $g : V \to \mathbb{R}$, $g(x) = \|x\|$ is also continuous (even Lipschitz)

proof: Let $x, y \in V$. By triangle inequality

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|$$

$$\|x\| - \|y\| \leq \|x - y\| \longrightarrow (*)_2$$

Again $\|y\| = \|y - x + x\| \leq \|y - x\| + \|x\|$.

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

$$\Rightarrow \quad -\|x - y\| \leq \|x\| - \|y\| \longrightarrow (*)_3$$

$(*)_2$ and $(*)_3 \Rightarrow$ $-\|x - y\| \leq \|x\| - \|y\| \leq \|x - y\|$

ie. $|\|x\| - \|y\|| \leq \|x - y\|$

ie $|g(x) - g(y)| \leq \|x - y\|$.

P-8 $\therefore$ g is Lipschitz continuous with Lipschitz constant $= 1$.

Defn (Normed linear space)

A normed linear space is a vector space V endowed with a norm.

The metric topology induced by the norm is called its norm topology.

$\odot$ V becomes a topological vector space.

Defn (Banach space)

A normed linear space is said to be a Banach space if it is complete under the norm topology, i.e., every Cauchy sequence is convergent.

Ex: $\mathbb{R}^N = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \;\middle|\; x_i \in \mathbb{R},\ i = 1, 2, \ldots, N \right\}$

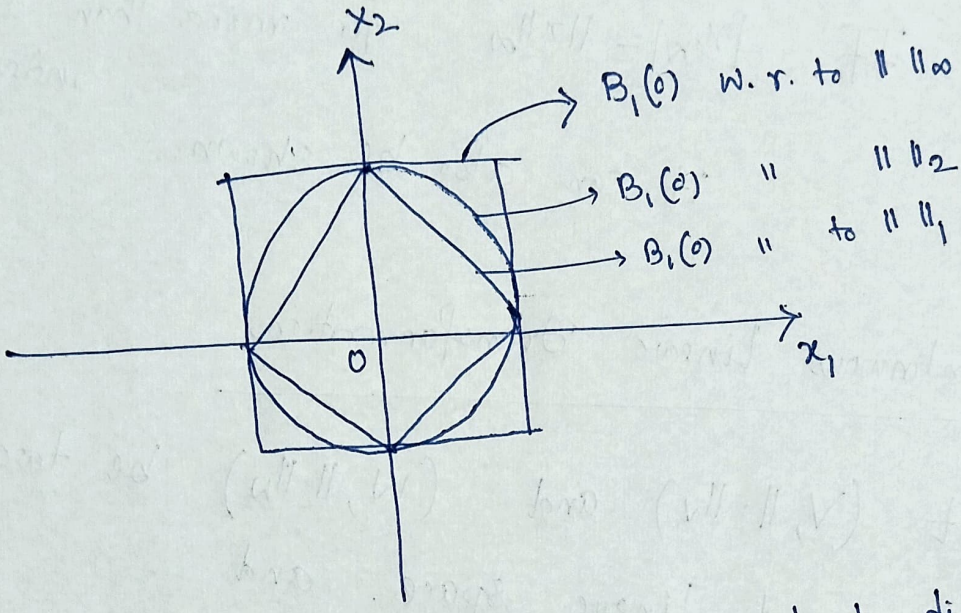
Let $1 \le p \le \infty$ be fixed.

On $\mathbb{R}^N$, we defined norm

$$\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p} \qquad \forall\, x \in \mathbb{R}^N$$

$$\|x\|_\infty = \max_{1 \le i \le N} |x_i| \qquad \forall\, x \in \mathbb{R}^N.$$

In practical example we consider $p = 1, 2, \infty$.



Unit ball $B_1(0)$ in $\mathbb{R}^2$ with respect to different norms $\|\ \|_p$

Explanation: $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$. for fixed $x \in \mathbb{R}^N$.

$$\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p\right)^{1/p}$$

Assume $\exists\ k \in \{1, 2, \cdots, N\}$ so that $|x_k| = \max_{1\le i\le N} |x_i|$

and $|x_k| > |x_i|$ for $i = 1, 2 \cdots, N$, $i \neq k$.

Then, $$\|x\|_p = \left(\sum_{\substack{i=1 \\ i\neq k}}^{N} |x_i|^p + |x_k|^p\right)^{1/p}$$

$$= \left(\sum_{\substack{i=1 \\ i\neq k}}^{N} \frac{|x_i|^p}{|x_k|^p} + 1\right)^{1/p} |x_k|$$

letting $p\to\infty$ $\lim_{p\to\infty} \|x\|_p = |x_k| = \|x\|_\infty$.

P-10

If $|x_k| = \|x\|_\infty$ for more than 1 index k

it can also be shown.

Continuous Linear Transformations

Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be two normed linear space. and

$T: V \longrightarrow W$ be a linear operator

$(\because \quad T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$

$\forall \ \alpha, \beta \in \mathbb{R}, \ x, y \in V)$

Defn. (Continuous lin. operator)

T is said to be continuous linear operator if it is continuous as a function between the topological space V and W (endowed with their norm topologies).

Defn. (Bounded subset)

A subset S of a normed linear space is bounded if it is contained in a ball, i.e.,

$\exists \ r > 0$ such that $S \subset B_r(0)$.

Proposition 1

Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be two normed linear space and $T: V \to W$ be a linear operator. The followings are equivalent

(i) T is continuous

(ii) T is continuous at 0

(iii) $\exists\ c > 0$ such that

$$\|Tx\|_W \leq c\,\|x\|_V \quad \forall\ x \in V$$

(iv) $T(B_1(0))$ is a bounded subset in W,

where $B_1(0) = \{ x \in V \mid \|x\|_V \leq 1 \}$.

(closed unit ball)

proof: (i) $\Leftrightarrow$ (ii)

IF T is continuous, then it is continuous at $x=0$.

conversely, let T be continuous at 0.

let $x \in V$. and $\{x_n\}_n \subset V$ be such that

$$x_n \to x \text{ in } V \text{ as } n \to \infty.$$

$$\Rightarrow \|x_n - x\|_V \to 0 \text{ as } n \to \infty.$$

let $z_n = x_n - x$, $n \in \mathbb{N}$, so $z_n \to 0$ in V as $n \to \infty$

P-12 As T is continuous at 0

$$Tz_n \longrightarrow T0 = 0 \quad \text{in } W \text{ as } n \to \infty$$

$$\Rightarrow \|Tz_n\|_W \longrightarrow 0 \text{ as } n \to \infty$$

$$\Rightarrow \|T(x_n - x)\|_W \longrightarrow 0 \text{ as } n \to \infty$$

$$\Rightarrow \|Tx_n - Tx\|_W \longrightarrow 0 \text{ as } n \to \infty$$

$$\Rightarrow Tx_n \longrightarrow Tx \text{ in } W \text{ as } n \to \infty.$$

Hence, T is continuous.

(ii) $\Longleftrightarrow$ (iii)

let T be continuous at 0. Using ε-δ defn. of continuity for $\varepsilon = 1$, $\exists\ \delta > 0$ such that

$$\forall\ x \in V,\ \|x\|_V < \delta \Rightarrow \|Tx\|_W < 1 \longrightarrow (*)_4$$

let $x \in V$. Take $y = \begin{cases} 0, & \text{if } x = 0 \\ \dfrac{\delta x}{2\|x\|_V}, & \text{if } x \neq 0. \end{cases}$

Then $\|y\|_V = \begin{cases} 0, & \text{if } x = 0 \\ \delta/2, & x \neq 0 \end{cases}$

So $\|y\|_V < \delta$

Hence, by $(*)_4$ $\|Ty\|_W < 1$

$$\Rightarrow \|T\left(\frac{\delta x}{2\|x\|_V}\right)\|_W < 1$$

$$\Rightarrow \frac{\delta}{2\|x\|_V}\|T(x)\|_W < 1$$

$\Rightarrow \|T(x)\|_W < \frac{2}{\delta}\|x\|_V. \quad \forall x \in V.$

Hence (iii) holds with $K = 2/\delta$.

Conversely: let $x_n \to 0$ in V as $n \to \infty$.

(ii) $\Rightarrow \|T(x_n)\|_W \leq K\|x_n\|_V \to 0$ as $n \to \infty$

$\Rightarrow T(x_n) \to 0$ in W as $n \to \infty$

$\Rightarrow T$ is continuous.

(iii) $\Leftrightarrow$ (iv)

From (iii) $\|Tx\|_W \leq K \quad \forall x \in B_1(0)$.

$(\because \|x\|_V \leq 1)$

$\Rightarrow T(B_1(0)) \subseteq B_K(0)$ in W.

So $T(B_1(0))$ is bounded in W.

Conversely, let $T(B_1(0))$ is bdd in W.

$\exists\ K > 0$ such that $\|T(x)\|_W \leq K$

$\forall x \in B_1(0)$.

Let $x \in V$, Set $y = \begin{cases} 0, & \text{if } x = 0 \\ \frac{x}{\|x\|_V}, & \text{if } x \neq 0. \end{cases}$

So $\|y\|_V = \begin{cases} 0, & x = 0 \\ 1, & x \neq 0 \end{cases} \Rightarrow y \in B_1(0)$.

$\Rightarrow \|T(y)\|_W \leq K$

$$\left\| T\left(\frac{x}{\|x\|_V}\right) \right\|_W \leq K \qquad x \neq 0.$$

$$\|T(x)\|_W \leq K \|x\|_V \qquad \forall\ x \neq 0$$

and $T(0) = 0$.

So $\|T(x)\|_W \leq K\|x\|_V \quad \forall\ x \in V.$

⊙ Set of all continuous linear operator $= B(V, W)$

⊙ Continuous linear operators are also known as bounded linear transformations.

we defined

$$\|T\| = \sup_{\|x\|_V \leq 1} \|Tx\|_W. \longrightarrow (2).$$

By proposition 1. (iv) $\|T\|$ is finite.

proposition 2

$$\|T\| = \sup_{\|x\|_V = 1} \|Tx\|_W = \sup_{x \neq 0} \frac{\|Tx\|_W}{\|x\|_V}.$$

proof: Let us denote

$$a = \sup_{\|x\|_V} \|Tx\|_W \quad , \quad b = \sup_{x \neq 0} \frac{\|Tx\|_W}{\|x\|_V}.$$

$$U_1 = \left\{ \|Tx\|_W, \mid \|x\|_V = 1 \right\}, \quad U_2 = \left\{ \frac{\|Tx\|_W}{\|x\|_V} \mid x \neq 0 \right\}.$$

We note that $\|Tx\|_W = \frac{\|Tx\|_W}{\|x\|_V} \in U_2$ if $\|x\|_V = 1$

$\Rightarrow \quad U_1 \subseteq U_2$

$x \neq 0$

Again $\frac{\|Tx\|_W}{\|x\|_V} = \left\|T\left(\frac{x}{\|x\|_V}\right)\right\|_W \in U_1$

or $\left\|\frac{x}{\|x\|_V}\right\|_V = 1$

$\Rightarrow \quad U_2 \subseteq U_1$

So $\quad U_1 = U_2$

$a = \sup U_1 = \sup U_2 = b. \longrightarrow (*)_5$

As $\{x \mid \|x\|_V = 1\} \subset B_1(0)$

$$\sup_{\|x\|_V=1} \|Tx\|_W \leq \sup_{\|x\|_V \leq 1} \|Tx\|_W$$

$$a \leq \|T\|. \longrightarrow (*)_6$$

By defn. of b, $\frac{\|T(x)\|_W}{\|x\|_V} \leq b \quad \forall\, x \neq 0$

$\Rightarrow \quad \|T(x)\|_W \leq b\|x\|_V \quad \forall\, x \neq 0$

$\longrightarrow (*)_6.$

$(*)_6$ is also true for $x = 0$. (both sides are 0)

So $\|Tx\|_W \leq b\|x\|_V \quad \forall\, x \in V.$

In particular if $\|x\|_V \leq 1$, $\|Tx\|_W \leq b.$, for all $x \in B_1(0)$

$\sup_{\|x\|_V \leq 1} \|Tx\|_W \leq b \Rightarrow \|T\| \leq b.$

$\longrightarrow (*)_7$

P-16 $(*)_5$, $(*)_6$ and $(*)_7$

$\Rightarrow \quad \alpha = \|T\| = b.$

Corollary 1 $\forall \; x \in V$

$$\|Tx\|_W \leq \|T\| \, \|x\|_V .$$

---

proposition 3 $\|T\| = \sup_{x \neq 0} \frac{\|Tx\|_W}{\|x\|_V}$, defines a norm on $\mathcal{B}(V,W)$.

If W is Banach then $\mathcal{B}(V,W)$ is also Banach.

Defn. Two norms defined on a same vector space are said to be equivalent if the topologies induced by these two norms coincide. Equivalently two norms $\| \; \|$ and $\| \; \|_*$ on V is said to be equivalent if $\exists$ constants $C_1, C_2 > 0$ such that

$$C_1 \|x\| \leq \|x\|_* \leq C_2 \|x\| \quad \forall \; x \in V.$$

Proposition 4

Any two norms on a finite dimensional vector space are equivalent.

Corollary

Any finite dimensional normed linear space is complete. In particular, any finite dimensional subspace of a normed linear space is closed.

Proposition 5

A normed linear space $(V, \|\cdot\|_V)$ is finite dimensional iff the closed unit ball in V is compact.

⊙ $x \in \mathbb{R}^N$ Equivalence of $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$ in $\mathbb{R}^N$.

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_N|$$

$$\|x\|_\infty = \max_{1 \le i \le N} |x_i|$$

let $\|x\|_\infty = |x_k|$ for some $k \in \{1, 2, \ldots, N\}$.

So $\|x\|_\infty = |x_k| \le |x_1| + \cdots + |x_N| = \|x\|_1$.

Again $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_N| \le |x_k| + |x_k| + \cdots + |x_k|$

$= N|x_k| = N\|x\|_\infty$.

Hence,

$$\|x\|_\infty \le \|x\|_1 \le N\|x\|_\infty . \longrightarrow (*)_6$$

$$\|x\|_2^2 = |x_1|^2 + |x_2|^2 + \cdots + |x_N|^2$$

$$\le |x_k|^2 + |x_k|^2 + \cdots + |x_k|^2$$

$$= N|x_k|^2 = N\|x\|_\infty^2$$

$$\Rightarrow \|x\|_2 \le \sqrt{N}\,\|x\|_\infty .$$

Also $|x_k|^2 \le |x_1|^2 + |x_2|^2 + \cdots + |x_N|^2 = \|x\|_2^2$

$$\Rightarrow \|x\|_\infty \le \|x\|_2$$

Hence $\|x\|_\infty \le \|x\|_2 \le \sqrt{N}\,\|x\|_\infty \longrightarrow (*)_7$

$(*)_6$ and $(*)_7$ gives

$$\|x\|_1 \le N\|x\|_\infty \le N\|x\|_2 \Rightarrow \frac{1}{N}\|x\|_1 \le \|x\|_2$$

$$\Rightarrow \frac{1}{N}\|x\|_1 \le \|x\|_2 \le \sqrt{N}\,\|x\|_\infty$$

(using $(*)_7$)

Also, $\|x\|_\infty \le \|x\|_2 \le \sqrt{N}\,\|x\|_\infty \le \sqrt{N}\,\|x\|_1$

$(\because \|x\|_\infty \le \|x\|_1)$.

$$\therefore \|x\|_\infty \le \|x\|_2 \le \sqrt{N}\,\|x\|_1$$

Lecture-13

Let us recall on $B(V,W)$ we defined norm

$$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|_W}{\|x\|_V}, \quad T \in B(V,W)$$

(1) ←

This norm is called operator norm subordinate to the vector norm on V and W.

Proposition 1

This norm satisfies

(i) $\|Tx\|_W \leq \|T\| \|x\|_V \quad \forall x \in V$

(ii) for $I \in B(V,V)$ (identity operator)

$$\|I\| = 1$$

(iii) $A \in B(V,W)$, $B \in B(W,Z)$

Then, $AB \in B(V,Z)$ and

$$\|AB\| \leq \|A\| \|B\|.$$

proof: (i) Already shown. Follows from (1).

(ii) $$\|I\| = \sup_{x \neq 0} \frac{\|Ix\|_V}{\|x\|_V} = 1.$$

(iii) $$\|ABx\|_Z \leq \|A\| \|Bx\|_Z \quad \text{(by (i))}$$

$$\leq \|A\| \|B\| \|x\|_V \quad \text{(by (i))}$$

$\forall x \in V$.

P-2

For $x \neq 0$ $\quad \dfrac{\|ABx\|_Z}{\|x\|_V} \leq \|A\| \|B\|$

$\Rightarrow \sup_{x \neq 0} \dfrac{\|ABx\|_Z}{\|x\|_V} \leq \|A\| \|B\|$

$\Rightarrow \|AB\| \leq \|A\| \|B\|.$

In Numerical Analysis we will be taking both V and W to be finite dimensional vector space over $\mathbb{R}$. Let $\dim V = n$, $\dim W = m$

From theory of linear Algebra

V is isomorphic to $\mathbb{R}^n$ and

W is " to $\mathbb{R}^m$.

and $\mathcal{B}(V, W) = M_{m,n}(\mathbb{R})$

Now onwards we will consider $A : \mathbb{R}^n \to \mathbb{R}^m$ linear operator and $A \in M_{m,n}(\mathbb{R}^n)$.

Lecture-13

Let $\delta = \{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subset V$ be a basis of V

and $\gamma = \{\beta_1, \ldots, \beta_m\} \subset W$ be a " of W

$T \in B(V, W)$. Then T has representation

$$[T]_\delta^\gamma \in M_{m,n}(\mathbb{R}).$$

Matrix Norm subordinate to $\|\cdot\|_1$ norm

$$\|A\|_1 = \sup_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1}, \quad x \in \mathbb{R}^n.$$

we show that $\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$

proof: For any $x \in \mathbb{R}^n$, with respect to the standard basis $\{e_1, e_2, \ldots, e_n\} \subset \mathbb{R}^n$ we can write

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$

$$Ax = x_1 Ae_1 + x_2 Ae_2 + \cdots + x_n Ae_n$$

($\because A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ linear).

P-4 Considering $\|\ \|_1$ on $\mathbb{R}^m$

$$\|Ax\|_1 = \|x_1 Ae_1 + x_2 Ae_2 + \cdots + x_n Ae_n\|_1$$

$$\leq |x_1|\,\|Ae_1\|_1 + |x_2|\,\|Ae_2\|_1 + \cdots + |x_n|\,\|Ae_n\|_1$$

(Triangle inequality)

$$\leq |x_1|\,\|Ae_k\|_1 + |x_2|\,\|Ae_k\|_1 + \cdots + |x_n|\,\|Ae_k\|_1$$

where $k \in \{1, 2, \ldots, n\}$ such that

$$\|Ae_k\| = \max_{1 \leq j \leq n} \|Ae_j\| \longrightarrow (*)_1$$

$$= \|Ae_k\|_1 \left(|x_1| + |x_2| + \cdots + |x_n|\right)$$

$$= \|Ae_k\|_1\, \|x\|_1$$

$$\Rightarrow \quad \|Ax\|_1 \leq \|Ae_k\|_1\, \|x\|_1$$

For $x \neq 0$ $\quad \dfrac{\|Ax\|_1}{\|x\|_1} \leq \|Ae_k\| = \max_{1 \leq j \leq n} \|Ae_j\|$

$$\Rightarrow \sup_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} \leq \max_{1 \leq j \leq n} \|Ae_j\|.$$

$$\Rightarrow \|A\|_1 \leq \max_{1 \leq j \leq n} \|Ae_j\|.$$

We note that for $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$

$$Ae_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix} \in \mathbb{R}^m \quad j\text{th column of } A$$

$$\|Ae_j\|_1 = \sum_{i=1}^{m} |a_{ij}|$$

Hence, $\|A\|_1 \leq \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|$ $\longrightarrow (*)_2$

Again or

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

$$= \|Ae_k\|_1 \quad (\text{by } (*)_1)$$

$$\leq \|A\|_1 \, \|e_k\|_1$$

$$= \|A\|_1 \quad (\because \|e_k\|_1 = 1)$$

$\longrightarrow (*)_3$

$(*)_2$ and $(*)_3 \Rightarrow$ $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|$

This is called maximum Column-sum.

P-6

Matrix norm Subordinate to $\|\ \|_\infty$ norm

$$\|A\|_\infty = \sup_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}.$$

We show that $\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$

proof: (Exercise).

Matrix norm subordinate to $\|\ \|_2$ norm

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

We show that $\|A\|_2 = \sqrt{\max_{1 \le i \le r} \lambda_i}$,

where r = rank of A

$\lambda_1, \ldots, \lambda_r$ are eigen values of A^TA

(These are called singular values).

proof: (Exercise)

Condition number of a matrix

Recall the definition of condition number of a differentiable function

$f : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$

$$K_f(x) = \begin{cases} \dfrac{|f'(x)|\,|x|}{|f(x)|}, & x \neq 0 \neq f(x). \\ |f'(x)|, & \text{otherwise.} \end{cases}$$

Generalizing to the fn. $f : \mathbb{R}^n \to \mathbb{R}^m$

differentiable

$$K_f(x) = \begin{cases} \dfrac{\|x\|_{\mathbb{R}^n}\, \|f'(x)\|_{M_{m,n}(\mathbb{R})}}{\|f(x)\|_{\mathbb{R}^m}}, & x \neq 0 \neq f(x) \\ \|f'(x)\|_{M_{m,n}(\mathbb{R})}, & \text{otherwise.} \end{cases}$$

$f'(x) \in M_{m,n}(\mathbb{R})$ Jacobian matrix.

For $f : \mathbb{R}^n \to \mathbb{R}^m$, $f(x) = Ax$, $A \in M_{m,n}(\mathbb{R})$.

P-8

$$K_A(x) = \|A\| \frac{\|x\|_{\mathbb{R}^n}}{\|Ax\|_{\mathbb{R}^m}}, \quad x \neq 0 \neq Ax.$$

$$\left[\because f'(x) = A\right]$$

We are interested in error analysis in finding solution of $Ax = b$.

x is unknown here.

Therefore $K_A(x)$ is not known quantity.

Hence, we consider another quantity which is independent of x (we will see later if the most possible error magnifying factor)

Define

$$K_A = \sup_{\substack{x \neq 0 \\ Ax \neq 0}} K_A(x)$$

$$= \sup_{\substack{x \neq 0 \\ Ax \neq 0}} \frac{\|A\| \|x\|_{\mathbb{R}^n}}{\|Ax\|_{\mathbb{R}^m}}$$

If A is invertible and $A \in M_n(\mathbb{R})$.

$$\boxed{K_A = \|A\| \|A^{-1}\|.}$$

This is called condition number of an invertible matrix A.

Lecture-14

For a matrix $A \in M_n(\mathbb{R})$ with $\det A \neq 0$, condition number of A is given by

$$K(A) = \|A\| \|A^{-1}\|,$$

where $\|\cdot\|$ denotes a matrix norm subordinate to a vector norm of $\mathbb{R}^n$.

Remark 1

$K(A)$ depends on the subordinate vector norm of $\mathbb{R}^n$.

We have seen that for $1 \leq p \leq \infty$, $x \in \mathbb{R}^n$

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, & 1 \leq p < \infty \\ \max_{1 \leq i \leq n} |x_i|, & p = \infty. \end{cases}$$

$\|x\|_p$ is a vector norm in $\mathbb{R}^n$.

The matrix norm subordinate to the vector norm $\|\cdot\|_p$ of $\mathbb{R}^n$ is denoted by and given by

$$\|A\|_{*,p} = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \quad 1 \leq p \leq \infty,$$

$A \in M_n(\mathbb{R})$.

P-2 Let us denote the corresponding condition number of a matrix $A \in M_n(\mathbb{R})$, $\det A \neq 0$ by $K(A)_p$ and is given by

$$K(A)_p = \|A\|_{*,p} \|A^{-1}\|_{*,p} \quad , \quad 1 \leq p \leq \infty.$$

Eg:(1)

$$A = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & 1 \\ 3 & -1 & 4 \end{bmatrix}$$

$$\|A\|_{*,1} = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}| = \max\{6, 2, 6\} = 6$$

$$\|A\|_{*,\infty} = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| = \max\{4, 2, 8\} = 8$$

$$A^{-1} = \begin{bmatrix} \frac{1}{2} & \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{5}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$\|A^{-1}\|_{*,1} = \frac{9}{2}$$

$$\|A^{-1}\|_{*,\infty} = \frac{7}{2}$$

$$K(A)_1 = \|A\|_{*,1} \|A^{-1}\|_{*,1} = 6 \times \frac{9}{2} = 27$$

$$K(A)_\infty = \|A\|_{*,\infty} \|A^{-1}\|_{*,\infty} = 8 \times \frac{7}{2} = 28.$$

Lecture - 14

Properties of $K(A)$

(i) $K(I) = 1$, $I \in M_n(\mathbb{R})$ is the identity matrix.

We note that for any matrix norm $\|\ \|$ subordinate to a vector norm of $\mathbb{R}^n$

$$\|I\| = \sup_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \sup_{x \neq 0} \frac{\|x\|}{\|x\|} = 1$$

Hence, $K(I) = \|I\| \|I^{-1}\| = 1$.

(ii) $K(A) \geq 1$

As A is invertible, $AA^{-1} = I$

So, $\|I\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\| = K(A)$

[using the property of a matrix norm subordinate to a vector norm]

$\Rightarrow K(A) \geq 1$

$(\because \|I\| = 1)$

$(\because \|AB\| \leq \|A\| \|B\|)$

(iii) For any $c \in \mathbb{R} \setminus \{0\}$, $A \in M_n(\mathbb{R})$, $\det A \neq 0$

$$K(cA) = K(A),$$

P-4

$$K(cA) = \|cA\| \, \|(cA)^{-1}\|$$

$$= |c| \, \|A\| \cdot |c^{-1}| \, \|A^{-1}\|$$

$$= \|A\| \, \|A^{-1}\| = K(A).$$

(iv) $K(A) = K(A^{-1})$. This follows from the definition of $K(A)$.

(v) For $D \in M_n(\mathbb{R})$, $D = \text{diag}(d_1, d_2, \ldots, d_n)$

$d_i \neq 0$, $i = 1, 2, \ldots, n$

$$K(D)_p = \max_{1 \le i \le n} |d_i| \Big/ \min_{1 \le i \le n} |d_i| \quad , \quad 1 \le p \le \infty.$$

For $1 \le p \le \infty$, we know that

$$\|D\|_{*,p} = \max_{1 \le i \le n} |d_i|$$

As $D^{-1} = \text{diag}(1/d_1, 1/d_2, \ldots, 1/d_n)$

$$\|D^{-1}\|_{*,p} = \max_{1 \le i \le n} \left|\frac{1}{d_i}\right| = \frac{1}{\min_{1 \le i \le n} |d_i|}$$

Hence, $K(D)_p = \|D\|_{*,p} \, \|D^{-1}\|_{*,p} = \dfrac{\max_{1 \le i \le n} |d_i|}{\min_{1 \le i \le n} |d_i|}$

(VI) For $A \in M_n(\mathbb{R})$, $\det A \neq 0$, A normal,

$$K(A)_2 = \frac{\max\limits_{1\le i\le n} |\lambda_i|}{\min\limits_{1\le i\le n} |\lambda_i|}, \qquad \lambda_i \in \rho(A)$$

eigen values of A.

Recall: $A \in M_n(\mathbb{R})$ is normal if

$$AA^T = A^TA$$

$$\|A\|_{*,2} = \sqrt{\max_{\lambda\in\rho(A^TA)}} = \sqrt{\max_{\lambda\in\rho(A^TA)} |\lambda|}$$

$$= \sqrt{\max_{\lambda\in\rho(A)} |\lambda|^2} \qquad (\because A \text{ is normal})$$

$$= \max_{\lambda\in\rho(A)} |\lambda|$$

$$= \max_{1\le i\le n} |\lambda_i|, \qquad \lambda_i \in \rho(A).$$

Also, $\|A^{-1}\|_{*,2} = \max\limits_{1\le i\le n} |1/\lambda_i| = \dfrac{1}{\min\limits_{1\le i\le n} |\lambda_i|}$

Hence, $K(A_*)_2 = \dfrac{\max\limits_{1\le i\le n} |\lambda_i|}{\min\limits_{1\le i\le n} |\lambda_i|}$

Theorem 1 (Gastinel)

Let $A \in M_n(\mathbb{R})$, $\det A \neq 0$ and $\|\ \|$ be any matrix norm subordinate to a vector norm of $\mathbb{R}^n$. Then,

$$\frac{1}{K(A)} = \min \left\{ \frac{\|A-B\|}{\|A\|} \;\middle|\; B \in M_n(\mathbb{R}),\ \det B = 0 \right\}$$

proof: Let $B \in M_n(\mathbb{R})$ with $\det B = 0$.

Then, there exists $x_0 \in \mathbb{R}^n$, $x_0 \neq 0$ such that

$$Bx_0 = 0$$

Now, $x_0 = A^{-1}Ax_0$ $\quad (\because A^{-1}A = I)$

$$\|x_0\| = \|A^{-1}Ax_0\| \leq \|A^{-1}\|\,\|Ax_0\|$$

$$= \|A^{-1}\|\,\|Ax_0 - Bx_0\| \quad (\because Bx_0 = 0)$$

$$\leq \|A^{-1}\|\,\|A-B\|\,\|x_0\|$$

$$\Rightarrow \quad 1 \leq \|A^{-1}\|\,\|A-B\| \quad (\because x_0 \neq 0,\ \|x_0\| \neq 0)$$

$$\Rightarrow \frac{1}{\|A^{-1}\|} \leq \|A-B\|$$

$$\Rightarrow \frac{1}{K(A)} = \frac{1}{\|A\|\,\|A^{-1}\|} \leq \frac{\|A-B\|}{\|A\|}.$$

Since $B \in M_n(\mathbb{R})$ is arbitrary singular matrix,

$$\frac{1}{K(A)} \leq \inf_{\substack{B \in M_n(\mathbb{R}) \\ \det B = 0}} \frac{\|A-B\|}{\|A\|}$$

We need to find $\tilde{B} \in M_n(\mathbb{R})$, $\det \tilde{B} = 0$ such that

$$\frac{\|A-\tilde{B}\|}{\|A\|} = \frac{1}{K(A)}.$$

We know $\|A^{-1}\| = \sup_{x \neq 0} \frac{\|A^{-1}x\|}{\|x\|} = \sup_{\|y\|=1} \|A^{-1}y\|$

As $A^{-1}: \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is continuous, and the set $\{ y \in \mathbb{R}^n \mid \|y\| = 1 \}$ is a closed bdd set (hence it is compact),

$\exists\ y_0 \in \mathbb{R}^n$, $\|y_0\| = 1$ such that

$$\|A^{-1}y_0\| = \|A^{-1}\|.$$

Also, there exist $\omega \in \mathbb{R}^n$, so that

$$\omega^T A^{-1} y_0 = 1 \quad \text{and} \quad \|\omega^T\| = \frac{1}{\|A^{-1}y_0\|}$$

(Existence of such vector is given by Hahn-Banach theorem)

P-8

Choose $\tilde{B} = A - y_0 \omega^T$

we note that $\tilde{B}(A^{-1}y_0) = (A - y_0\omega^T)A^{-1}y_0$

$= y_0 - y_0 \omega^T A^{-1} y_0$

$= y_0 - y_0$

$= 0 \quad (\because \omega^T A^{-1} y_0 = 1)$

So, $\tilde{B}$ is singular.

$\|A - \tilde{B}\| = \|y_0 \omega^T\| \quad (\because \tilde{B} = A - y_0\omega^T)$

$= \sup_{x \neq 0} \frac{\|y_0 \omega^T x\|}{\|x\|}$

$= \sup_{x \neq 0} \frac{\|y_0\| |\omega^T x|}{\|x\|} \quad (\because \omega^T x \in \mathbb{R})$

$\|y_0 \omega^T x\| = \|y_0\| |\omega^T x|$

$= \|y_0\| \sup_{x \neq 0} \frac{|\omega^T x|}{\|x\|}$

$= \|y_0\| \, \|\omega^T\|$

$= \frac{1}{\|A^{-1}y_0\|} \quad (\because \|y_0\| = 1, \ \|\omega^T\| = \frac{1}{\|A^{-1}y_0\|})$

$= \frac{1}{\|A^{-1}\|} \quad (\because \|A^{-1}y_0\| = \|A^{-1}\|)$

$\Rightarrow \frac{\|A - \tilde{B}\|}{\|A\|} = \frac{1}{\|A^{-1}\|\|A\|} = \frac{1}{K(A)}$

Hence, $\frac{1}{K(A)} = \min \left\{ \frac{\|A - B\|}{\|A\|} \mid B \in M_n(\mathbb{R}), \det B = 0 \right\}$.

Remark 2

$K(A)$ measures how close the matrix A to a singular matrix.

Fix a matrix $B \in M_n(\mathbb{R})$, $\det B = 0$.

Above, theorem says

$$\frac{\|A\|}{\|A-B\|} \leq K(A)$$

If $A \to B$, then, $K(A) \to \infty$, i.e.

if A is nearly a singular matrix, then condition number will be very large.

And we have seen $K(A) \geq 1$.

Eg: (2) Let $A = \alpha I$, $\alpha > 0$.

$\det A = \alpha^n$.

If $0 < \alpha < 1$, $\det A \to 0$ as $n \to \infty$.

If $\alpha > 1$, $\det A \to \infty$ as $n \to \infty$

In computer for $0 < \alpha < 1$, and for large enough dimension n, $\det A = 0$ even if A is non-singular.

P-10 We note that

$$K(A) = K(\alpha I) = K(I) = 1.$$

We observe that to determine non-singularity of a matrix in a computer it is more appropriate to consider the quantity Condition number than determinant.

Error Analysis

Let $x \in \mathbb{R}^n$ be the solution of the system

$$Ax = b \quad , \quad b \neq 0. \leftarrow (1)$$

Let $\hat{A}$ be a perturbation (an approximation) of A and $\hat{b}$ be a " (an ") of b.

Let $\hat{x} \in \mathbb{R}^n$ be the solution of $\hat{A}\hat{x} = \hat{b}$ ——→ (2)

(As $\det A \neq 0$ and set of nonsingular matrices are open in $M_n(\mathbb{R})$, for small enough δ if $\|A - \hat{A}\| < \delta$, $\hat{A}$ is also non-singular. So, system (2) has unique soln.).

Let us denote $\delta x = x - \hat{x} = E(\hat{x})$

$\delta A = A - \hat{A} = E(\hat{A})$

$\delta b = b - \hat{b} = E(\hat{b})$.

We find an estimate of the absolute relative error $E_{ar}(\hat{x}) = \frac{\|x - \hat{x}\|}{\|x\|} = \frac{\|\delta x\|}{\|x\|}$

in terms of $E_{ar}(\hat{A}) = \frac{\|\delta A\|}{\|A\|}$ and

$$E_{ar}(\hat{b}) = \frac{\|\delta b\|}{\|b\|}.$$

As $\hat{x}$ satisfies $\hat{A}\hat{x} = \hat{b}$

So $(A - \delta A)(x - \delta x) = b - \delta b$ $\left(\because \hat{A} = A - \delta A,\ \hat{b} = b - \delta b,\ \hat{x} = x - \delta x\right)$

$\Rightarrow (A - \delta A)x - (A - \delta A)\delta x = b - \delta b$

$\Rightarrow Ax - \delta Ax - (A - \delta A)\delta x = b - \delta b$

$\Rightarrow \delta Ax + (A - \delta A)\delta x = \delta b \quad (\because Ax = b)$

$\Rightarrow (A - \delta A)\delta x = -\delta Ax + \delta b$

$\Rightarrow A(I - A^{-1}\delta A)\delta x = -\delta Ax + \delta b.$

$\Rightarrow (I - A^{-1}\delta A)\delta x = -A^{-1}(\delta Ax - \delta b).$ ⟶ ③

P-12

Assume that $\|\delta A\| < \frac{1}{\|A^{-1}\|}$.

Hence, $\|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| < 1$

Therefore $(I - A^{-1}\delta A)^{-1}$ exists and

$$\|(I - A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|} \longrightarrow (4)$$

From (3) we get

$$\delta x = -(I - A^{-1}\delta A)^{-1} A^{-1}(\delta A x - \delta b)$$

$$\Rightarrow \|\delta x\| \leq \|(I - A^{-1}\delta A)^{-1}\| \|A^{-1}\| \|\delta A x - \delta b\|$$

$$\leq \frac{1}{1 - \|A^{-1}\delta A\|} \|A^{-1}\| (\|\delta A x\| + \|\delta b\|)$$

(using (4))

$$\leq \frac{1}{1 - \|A^{-1}\|\|\delta A\|} \cdot \|A^{-1}\| (\|\delta A\|\|x\| + \|\delta b\|)$$

$$\left[\because \|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| \right.$$
$$\left. 0 < 1 - \|A^{-1}\|\|\delta A\| \leq 1 - \|A^{-1}\delta A\| \right.$$
$$\left. \frac{1}{1 - \|A^{-1}\delta A\|} \leq \frac{1}{1 - \|A^{-1}\|\|\delta A\|} \right]$$

$$\Rightarrow \quad \|\delta x\| \leq \frac{\|A\|\|A^{-1}\|}{1 - \|A^{-1}\|\|A\| \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|}\|x\| + \frac{\|\delta b\|}{\|A\|} \right)$$

$$\Rightarrow \quad \frac{\|\delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\|x\|} \right)$$

$$\leq \frac{K(A)}{1 - K(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

$$\left(\because \quad b = Ax \right.$$

$$\|b\| \leq \|A\|\|x\|$$

$$\left. \frac{1}{\|A\|\|x\|} \leq \frac{1}{\|b\|} \right)$$

Finally. $$\boxed{E_{ar}(\hat{x}) \leq \frac{K(A)}{1 - K(A) E_{ar}(\hat{A})} \left(E_{ar}(\hat{A}) + E_{ar}(\hat{b}) \right).}$$

If $\delta b = 0$, then $E_{ar}(\hat{b}) = 0$ and

$$E_{ar}(\hat{x}) \leq \frac{K(A) E_{ar}(\hat{A})}{1 - K(A) E_{ar}(\hat{A})}$$

If $\delta A = 0$, then $E_{ar}(\hat{A}) = 0$ and

$$E_{ar}(\hat{x}) \leq K(A) E_{ar}(\hat{b}).$$

Again if $\delta b = 0$, we also have

$$\hat{A}\hat{x} = b \qquad (\because \delta b = 0,\ \hat{b} = b)$$

$$(A - \delta A)\hat{x} = b$$

$$\Rightarrow A\hat{x} - \delta A\hat{x} = b$$

$$\Rightarrow A(x - \delta x) - \delta A\hat{x} = b$$

$$\Rightarrow Ax - A\delta x - \delta A\hat{x} = b$$

$$A\delta x = -\delta A\hat{x} \qquad (\because Ax = b)$$

$$\Rightarrow \delta x = -A^{-1}\delta A\hat{x}$$

$$\Rightarrow \|\delta x\| \leq \|A^{-1}\| \|\delta A\| \|\hat{x}\|$$

$$\Rightarrow \frac{\|\delta x\|}{\|\hat{x}\|} \leq \|A^{-1}\| \|\delta A\| = \|A^{-1}\| \|A\| \frac{\|\delta A\|}{\|A\|}$$

$$\Rightarrow \boxed{\frac{\|\delta x\|}{\|\hat{x}\|} \leq K(A)\, E_{ar}(\hat{A})}$$

Lecture - 14

Residual

One way to verify a pt. $\hat{x} \in \mathbb{R}^n$ is a solution of $Ax = b$ is to substitute it in the equation. Since most of the cases, we get an approximate solution $\hat{x}$ of $Ax = b$.

There is an error, let $E(\hat{x}) = x - \hat{x}$.

Defn. (residual)

The residual of an approximate solution $\hat{x}$ of the system $Ax = b$ is given by

$$r = r(\hat{x}) = b - A\hat{x}.$$

we note that $r = b - A\hat{x} = Ax - A\hat{x}$

$$= AE(\hat{x}).$$

we now find an estimate of the relative error $E_{ar}(\hat{x})$ in terms of residual.

we have, $r = b - A\hat{x} = AE(\hat{x}) \longrightarrow (5)$

$$\Rightarrow \|r\| = \|AE(\hat{x})\| \leq \|A\| \, \|E(\hat{x})\|$$

$$\Rightarrow \frac{\|r\|}{\|A\|} \leq \|E(\hat{x})\| \longrightarrow (6).$$

Again from ⑤

$$E(\hat{x}) = A^{-1} r$$

$$\|E(\hat{x})\| \leq \|A^{-1}\| \|r\| \longrightarrow ⑦$$

⑥ and ⑦ $\Rightarrow$ $\dfrac{\|r\|}{\|A\|} \leq \|E(\hat{x})\| \leq \|A^{-1}\| \|r\|$

$$\Rightarrow \frac{\|r\|}{\|A\|\|x\|} \leq \frac{\|E(\hat{x})\|}{\|x\|} \leq \frac{\|A^{-1}\|\|r\|}{\|x\|} \quad (\because x \neq 0)$$

$$\Rightarrow \frac{\|r\|}{\|A\|\|A^{-1}\|} \cdot \frac{\|A^{-1}\|}{\|x\|} \leq E_{ar}(\hat{x}) \leq \frac{\|A\|\|A^{-1}\|}{\|A\|\|x\|} \cdot \|r\|$$

$$\Rightarrow \frac{\|r\|}{K(A)} \cdot \frac{\|A^{-1}\|}{\|x\|} \leq E_{ar}(\hat{x}) \leq \frac{K(A)}{\|A\|\|x\|} \cdot \|r\|$$

$\longrightarrow ⑧$

$[\because K(A) = \|A\|\|A^{-1}\|]$

We know that $b = Ax \Rightarrow \|b\| \leq \|A\|\|x\|$

$$\Rightarrow \frac{1}{\|A\|\|x\|} \leq \frac{1}{\|b\|} \longrightarrow ⑨$$

Also $x = A^{-1}b \Rightarrow \|x\| \leq \|A^{-1}\|\|b\|$

$$\Rightarrow \frac{1}{\|b\|} \leq \frac{\|A^{-1}\|}{\|x\|} \longrightarrow ⑩$$

using ⑨ and ⑩, we obtain from ⑧ that

$$\Rightarrow \frac{\|r\|}{\|b\| K(A)} \leq E_{ar}(\hat{x}) \leq K(A) \frac{\|r\|}{\|b\|}.$$

$$\Rightarrow \boxed{\frac{E_{ar}(r)}{K(A)} \leq E_{ar}(\hat{x}) \leq K(A)\, E_{ar}(r),} \quad \longrightarrow (11)$$

where we use the notation

$$E_{ar}(r) = \frac{\|r\|}{\|b\|} = \frac{\|b - A\hat{x}\|}{\|b\|}$$

= absolute relative residual

If $K(A)$ is close to 1 sufficiently, the ~~relative~~ absolute relative error $E_{ar}(\hat{x})$ and absolute relative residual are of same size. In this case the absolute relative residual taken as measure of absolute relative error.

Eg: $A = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix}$ $\quad \|A^{-1}\| = \begin{bmatrix} 101/4 & -99/4 \\ -99/4 & 101/4 \end{bmatrix}$

$\|A\|_\infty = 2,$ $\quad \|A^{-1}\|_\infty = 50$

$K(A)_\infty = 2 \times 50 = 100.$

$$\frac{E_{ar}(r)}{100} \leq E_{ar}(\hat{x}) \leq 100\, E_{ar}(r).$$

P-18

Eg: Consider the system

$$1.01x_1 + 0.99x_2 = 2$$

$$0.99x_1 + 1.01x_2 = 2$$

Exact solution $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Let $\hat{x} = \begin{pmatrix} 1.01 \\ 1.01 \end{pmatrix}$ be an approximate soln.

$$E(\hat{x}) = \begin{pmatrix} -0.01 \\ -0.01 \end{pmatrix} \quad \begin{pmatrix} -0.02 \\ -0.02 \end{pmatrix} = r(\hat{x})$$

Let $\hat{y} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ be another approximate soln.

$$E(\hat{y}) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad r(\hat{y}) = \begin{pmatrix} -0.02 \\ 0.02 \end{pmatrix}$$

~~For $\hat{x}$~~ For both the approximation

$$\|r(\hat{x})\|_p = \|r(\hat{y})\|_p$$

but for $\hat{y}$, the error $\|E(\hat{y})\|_p$ is larger

compare to $\|E(\hat{x})\|_p$.

Iterative methods to solve $Ax = b$

Problem: Find approximate solution to $Ax = b$, where $A \in M_n(\mathbb{R})$ has following properties

(i) It is a large system, ie. n is very big, for example $n = O(10^6)$

(ii) A is sparse with a large percent of 0 entries

(iii) A is structured (ie. the product Ax for given $x \in \mathbb{R}^n$ can be computed efficiently)

Idea: Avoiding computing A^{-1}. Perform the "cheap" operation Ax.

We start with "Gauss - Jacobi" iteration

P-2

Want to solve:

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

Rewrite it:

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

provided $a_{ii} \neq 0 \quad \forall \; i = 1, 2, \ldots, n$

(Idea is to write the system in a fixed point form by keeping all the diagonal term on the LHS and other term on the RHS of the system).

In a compact form

$$x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j\right), \qquad i = 1, 2, \ldots, n$$

Gauss-Jacobi iterations

⊙ **Initialization**: Choose a starting point

$$x^0 = \begin{pmatrix} x_1^0 \\ x_2^0 \\ \vdots \\ x_n^0 \end{pmatrix}.$$

⊙ **Iteration steps**: For $K = 0, 1, 2, \ldots$ until the stop criteria

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

$$x_i^{K+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^K \right) \quad \leftarrow (1)$$

end

end

⊙ **Choices of starting vector x^0**:

(i) A vector with all entries 1: $x_i^0 = 1$ i.e.

$$x^0 = \begin{pmatrix} 1 \\ 1 \\ \\ 1 \end{pmatrix}$$

(ii) A vector " " " 0: $x^0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

(iii) The load vector $x^0 = b$

(iv) For diagonal dominant matrix $x_i^0 = b_i / a_{ii}$, $i = 1, 2, \ldots, n$ (best choice).

P-4

(v) Since this is a linear system, if the iteration converges, it converges for any choice of initial vector. So anything works.

⊙ stopping criteria:

Small Error:

(i) For given $0 < \varepsilon << 1$,

$$\|x^k - x^{k-1}\| \lesssim \varepsilon \quad \text{for some vector norm.}$$

Small relative error:

(ii) For given, $0 < \varepsilon << 1$,

$$\frac{\|x^k - x^{k-1}\|}{\|x^k\|} \lesssim \varepsilon \quad , \text{ if } x^k \neq 0$$

(iii) Small residual: $r^k = b - Ax^k$

$$\|r^k\| \lesssim \varepsilon.$$

(iv) Maximum number of iteration reached

(v) Any combinations of the above.

Advantages / Disad.

- We need to create 2 vectors for computation, x^k and x^{k+1}, need more space.
- Non-sequential: Good for parellel computing.
- Convergence is slow.

Improvement of Grauss-Jacobi iteration: Grauss-Seidal

Recall the Grauss-Jacobi iteration

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

i.e.

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

we update the value of this part with recent ((k+1)th step) value.

i.e. we take x_j^{k+1} for x_j^k

for $1 \leq j \leq i-1$.

P-6 Hence we have Gauss-Seidal iterations:

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

Gauss - Seidal iterations

⊙ Initialization: As in Gauss - Jacobi

⊙ Iteration steps:

For $K = 0, 1, 2, \ldots$ until the stop criteria

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

② ← $$x_i^{K+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{K+1} - \sum_{j=i+1}^{n} a_{ij} x_j^{K} \right)$$

end

end

⊙ choices of starting vector x^0 and

⊙ stoping criteria as in Gauss - Jacobi

Advantages / Disadvantages:

⊙ Require only one vector for both x^K and x^{K+1}, saves memory space

⊙ Not good for parallel computing: (sequential method)

⊙ Converges faster than Gauss - Jacobi.

Eg: (1)

$$2x_1 - x_2 = 0$$
$$-x_1 + 2x_2 - x_3 = 1$$
$$-x_2 + 2x_3 = 2$$

We note that exact soln. $x = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$

Choose initial vector $x^0 = \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}$, $x_i^0 = b_i / a_{ii}$, $i = 1, 2, 3$.

① Gauss - Jacobi iteration:

$$\begin{cases} x_1^{k+1} = \frac{1}{2} x_2^k \\ x_2^{k+1} = \frac{1}{2} + \frac{1}{2} x_1^k + \frac{1}{2} x_3^k \\ x_3^{k+1} = 1 + \frac{1}{2} x_2^k \end{cases}$$

We compute some iterations

$$x^1 = \begin{pmatrix} 0.25 \\ 1 \\ 1.25 \end{pmatrix}$$

$$x^2 = \begin{pmatrix} 0.5 \\ 1.25 \\ 1.5 \end{pmatrix}$$

$$x^3 = \begin{pmatrix} 0.625 \\ 1.5 \\ 1.625 \end{pmatrix}$$

P-8

Observation:

⊙ It seems the iteration is converging.
We need to run more steps.

⊙ Slow convergence rate.

⊙ **Gauss-Seidal iteration:**

$$\begin{cases} x_1^{k+1} = \frac{1}{2} x_2^k \\ x_2^{k+1} = \frac{1}{2} + \frac{1}{2} x_1^{k+1} + \frac{1}{2} x_3^k \\ x_3^{k+1} = 1 + \frac{1}{2} x_2^{k+1} \end{cases}$$

$$x^1 = \begin{pmatrix} 0.25 \\ 1.125 \\ 1.5625 \end{pmatrix}, \quad x^2 = \begin{pmatrix} 0.5625 \\ 1.5625 \\ 1.7813 \end{pmatrix},$$

Observation

⊙ Converges a bit faster than Gauss-Jacobi iterations.

We observe that the Gauss-Jacobi and Gauss-Seidal iteration is fixed and independent of problem or parameter. These two iterations are rigid in some sense. There is no way to make some adjustment (depending on the problem) to get better result. This is the disadvantage of the above two methods.

We now consider another method which has a parameter for some adjustment.

A version based on Gauss-Seidal. (SOR (Successive Over Relaxation)

$$x_i^{k+1} = (1-\omega)\, x_i^k + \omega \frac{1}{a_{ii}} \left(b_i - \underbrace{\sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k}_{\text{G-S.}} \right)$$

(3) ←

$\omega \in \mathbb{R}$.

(later we will see if SOR converges then $0 < \omega < 2$).

ω: Relaxation parameter.

P-10

For convergence $0 < \omega < 2$

- $\omega = 1$: Gauss-Seidal
- $0 < \omega < 1$: Under relaxation
- $1 < \omega < 2$: Over relaxation

Eg: (2) For the previous example: Take $\omega = 1.2$

SOR iteration:

$$\begin{cases} x_1^{k+1} = -0.2x_1^k + 0.6x_2^k \\ x_2^{k+1} = -0.2x_2^k + 0.6(1 + x_1^{k+1} + x_3^k) \\ x_3^{k+1} = -0.2x_3^k + 0.6(2 + x_2^{k+1}) \end{cases}$$

$$x^0 = \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}, \quad x^1 = \begin{pmatrix} 0.3 \\ 1.28 \\ 1.708 \end{pmatrix}, \quad x^2 = \begin{pmatrix} 0.708 \\ 1.8290 \\ 1.9442 \end{pmatrix}$$

Observation: Faster convergence than both Gauss-Jacobi and Gauss-Seidal.

Remark: We could also consider verson depend on Gauss-Jacobi (i.e $x^{k+1} = (1-\omega)x^k + \omega \cdot (G\text{-}J)$

We want to write all three above methods (iterative) in a standard form.

All three methods can be written in a compact way as: For iterations $k = 0, 1, 2, \ldots$

$$\text{G-J}: \quad x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^k \right)$$

$$\text{G-S}: \quad x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right)$$

$$\text{SOR}: \quad x_i^{k+1} = (1-\omega) x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right)$$

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

For defining the methods we have assumed

$a_{ii} \neq 0$ for $i = 1, 2, \ldots, n$.

Basic idea of all three method above is:

1. To solve the system $Ax = b$. $\longrightarrow$ (1)

Write the system in the form

$$x = Mx + y \quad \longleftarrow (2)$$

for some $M \in M_n(\mathbb{R})$ and $y \in \mathbb{R}^n$ given.

Equation ② is called fixed point formulation of the system.

Qn: why?

Ans: Let us define a function $F: \mathbb{R}^n \to \mathbb{R}^n$ by

$$F(x) = Mx + y \quad , x \in \mathbb{R}^n,$$

$M \in M_n(\mathbb{R})$ and $y \in \mathbb{R}^n$ are given.

To find a soln. of ① means to find a pt $s \in \mathbb{R}^n$ so that $As = b$,

or equivalently, to find a pt. $s \in \mathbb{R}^n$ so that

$$s = Ms + b.$$

or, equivalently to find a pt. $s \in \mathbb{R}^n$ so that

$$F(s) = s \quad (\text{ie. a fixed pt. of } F).$$

Therefore, our problem will be solved if we could find a fixed pt of the function F.

---

2. To find a fixed pt. of F, we use iteration. We choose an initial condition $x^0 \in \mathbb{R}^n$ and generate a sequence $\{x^k\}_{k \in \mathbb{N} \cup \{0\}}$ as

$$x^{k+1} = F(x^k) = Mx^k + y, \quad k = 0, 1, \ldots \longrightarrow (3)$$

3. Show that the seqn. $\{x^k\}_{k \in \mathbb{N} \cup \{0\}} \subset \mathbb{R}^n$ is convergent say $\lim_{k \to \infty} x^k = s$ (say) $\in \mathbb{R}^n$.

Assume F is continuous, then letting $k \to \infty$ in (3)

we have $\lim_{k \to \infty} x^{k+1} = \lim_{k \to \infty} F(x^k)$

$$s = F(s) = Ms + y \longrightarrow (4)$$

This gives a solution of (1).

We now discuss under what condition on M any y

The seqn. $\{x^k\}_{k \in \mathbb{N} \cup \{0\}}$ defined in ③ converges

Defn. (Spectral Radius)

Let $M \in M_n(\mathbb{R})$, the spectral radius of A is denoted by $\rho(M)$ and is defined by

$$\rho(M) = \sup \{ |\lambda| \mid \lambda \in \mathbb{C}, \det(\lambda I - M) = 0 \}$$

$$= \max \{ |\lambda| \mid \lambda \text{ is an eigen-value of } M \text{ in } \mathbb{C} \}.$$

Let us recall two results from the supplimentary notes:

Theorem-1 (theorem 7.9, page-490)

Let $M \in M_n(\mathbb{R})$. $\lim_{k \to \infty} M^k = 0$ in $M_n(\mathbb{R})$ iff $\rho(M) < 1$.

Theorem-2 (theorem 7.10, page-491)

Let $M \in M_n(\mathbb{R})$. The series $\sum_{j=0}^{\infty} M^j$ convergent in $M_n(\mathbb{R})$ iff $\rho(M) < 1$.

In that case $\sum_{j=0}^{\infty} M^j = (I - M)^{-1}$.

Errors and convergence analysis of iterative methods.

We have known that, all the iterative method discussed above can be put in the form

$$x^{k+1} = Mx^k + y, \quad k \in \mathbb{N} \cup \{0\}. \quad \leftarrow (1)$$

Theorem 13 For any $x^0 \in \mathbb{R}^n$, the sequence $\{x^k\}_{k=0}^{\infty}$ defined by (1) above converges to the unique solution of $x = Mx + y$ (which in turn soln of $Ax = b$) if and only if $\rho(M) < 1$.

proof: Let $\rho(M) < 1$. By Theorem 1 and 2 above, we know that $\lim_{k\to\infty} M^k = 0$ in $M_n(\mathbb{R})$. $\longrightarrow (2)$ (in any matrix norm)

also, the sequence $S_k = \sum_{j=0}^{k-1} M^j$ converges and $\longrightarrow (3)$

$\lim_{k\to\infty} S_k = (I - M)^{-1}$ in $M_n(\mathbb{R})$. $\longrightarrow (4)$.

Now, from (1) $x^k = Mx^{k-1} + y$ for $k \geq 1$

$$= M(Mx^{k-2} + y) + y$$

$$= M^2 x^{k-2} + My + y$$

$$= M^k x^0 + (M^{k-1} + M^{k-2} + \cdots + I)y.$$

Hence, $x^k = M^k x^0 + S_k y$ (S_k is defined in ③)

By ② and ④ we have as $k \to \infty$.

$\lim_{k\to\infty} x^k$ exists and say $\lim_{k\to\infty} x^k = \tilde{x} \in \mathbb{R}^n$.

$$\tilde{x} := \lim_{k\to\infty} x^k = \lim_{k\to\infty} M^k x^0 + \lim_{k\to\infty} S_k y$$

$$= (I - M)^{-1} y.$$

Hence, $\tilde{x}$ satisfies $(I-M)\tilde{x} = y$

$$\tilde{x} = M\tilde{x} + y.$$

[Note: In the ~~class~~ class room, we showed $\{x^k\}$ is Cauchy, and hence concluded $\lim_{k\to\infty} x^k$ exists.

Here, I have given another proof].

Conversely: let $\{x^k\}_{k \in \mathbb{N}\cup\{0\}}$ be defined in ① Converges to the unique soln $\tilde{x}$ of ②

To show $\rho(M) < 1$.

~~Let $\rho(M) \geq 1$. There exists~~ For any $x^0 \in \mathbb{R}^n$,

Hence, $x^{k+1} = Mx^k + y$, $\tilde{x} = M\tilde{x} + y$

$$\Rightarrow x^{k+1} - \tilde{x} = M(x^k - \tilde{x}) = M^2(x^{k-1} - \tilde{x})$$

$$= M^{k+1}(x^0 - \tilde{x})$$

As $\lim_{k\to\infty} x^{k+1} = \tilde{x}$, $\lim_{k\to\infty} M^{k+1}(x^0 - \tilde{x}) = 0 \Rightarrow \lim_{k\to\infty} M^k = 0$ in $M_n(\mathbb{R})$

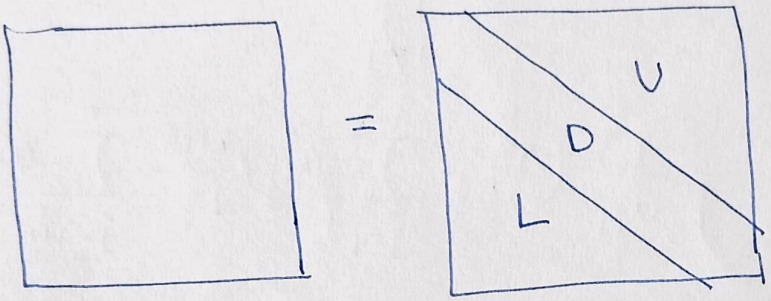
$\forall\, x^0 \in \mathbb{R}^n \Rightarrow \rho(M) < 1$ (by theorem-1).

Lecture-16-17

We now see that all three methods above are written is the form $x^{k+1} = Mx^k + y$

Splitting of the matrix A:

$$A = L + D + U$$



$A = (a_{ij}) \in M_n(\mathbb{R})$. Then,

$$L = \begin{bmatrix} 0 & \cdots & 0 \\ a_{21} & \ddots & \vdots \\ a_{n1} & a_{n,n-1} & 0 \end{bmatrix} \quad \text{i.e.} \quad L_{ij} = \begin{cases} 0, & \text{if } 1 \le i \le j \le n \\ a_{ij}, & \text{if } 1 \le j < i \le n \end{cases}$$

$$D = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & \\ & & 0 \\ 0 & \cdots & 0 \; a_{nn} \end{bmatrix} \qquad D = \operatorname{diag}(a_{11}, a_{12}, \ldots, a_{nn})$$

$$U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & & \\ & & 0 & a_{n-1,n} \\ 0 & & \cdots & 0 \end{bmatrix} \qquad U_{ij} = \begin{cases} a_{ij}, & \text{if } 1 \le i < j \le n \\ 0, & \text{if } 1 \le j \le i \le n \end{cases}$$

Eg: $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, $L = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{bmatrix}$, $D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}$, $U = \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix}$

Now, we have $Ax = b$

$\Rightarrow (L+D+U)x = b$

$\Rightarrow Lx + Dx + Ux = b$. $\longrightarrow$ ① ②

G-J iteration: $Dx^{K+1} = b - Lx^K - Ux^K$

so, $x^{K+1} = D^{-1}(b - (L+U)x^K)$

$x^{K+1} = D^{-1}b - D^{-1}(L+U)x^K$

[Assuming $a_{ii} \neq 0 \ \forall i = 1, 2, \ldots, n$
so $D = \text{diag}(a_{11}, \ldots, a_{nn})$
is invertible]

$\leftarrow$ ③ $\boxed{x^{K+1} = M_J x^K + y_J}$ where

$M_J = -D^{-1}(L+U) \in M_n(\mathbb{R})$, $y_J = D^{-1}b \in \mathbb{R}^n$

G-S iteration: $Dx^{K+1} + Lx^{K+1} = b - Ux^K$

$\Rightarrow (D+L)x^{K+1} = b - Ux^K$

$\Rightarrow x^{K+1} = (D+L)^{-1}b - (D+L)^{-1}Ux^K$

$\leftarrow$ ④ $\Rightarrow \boxed{x^{K+1} = M_S x^K + y_S}$

where $M_S = -(D+L)^{-1}U$, $y_S = (D+L)^{-1}b \in \mathbb{R}^n$
$\in M_n(\mathbb{R})$.

[Note $\det(D+L) = \det D \neq 0$
so $(D+L)$ is invertible]

Lecture-16-17

SOR: $x^{k+1} = (1-\omega)x^k + \omega \cdot G\text{-}S$

$$x^{k+1} = (1-\omega)x^k + \omega D^{-1}(b - Lx^{k+1} - Ux^k)$$

$$\Rightarrow Dx^{k+1} = (1-\omega)Dx^k + \omega b - \omega Lx^{k+1} - \omega Ux^k$$

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

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

$$\Rightarrow \boxed{x^{k+1} = M_{SOR}x^k + y_{SOR}} \quad \text{where.}$$

(5) ←

$$M_{SOR} = (D+\omega L)^{-1}[(1-\omega)D - \omega U] \in M_n(\mathbb{R})$$

$$y_{SOR} = \omega(D+\omega L)^{-1}b \in \mathbb{R}^n$$

Advantages of this representation is that it helps us to do convergence analysis of the iteration scheme.

Theorem (4) Let $A \in M_n(\mathbb{R})$ be strictly diagonal dominant.

Then, G-J and G-S converges for any $x^0 \in \mathbb{R}^n$.

proof: It is suffices to show that

$$\rho(M_J) < 1 \quad \text{and} \quad \rho(M_S) < 1.$$

Then by theorem (3), the result follows.

$\rho(M_J) < 1$:

Let λ be an eigen value of M_J with corresponding eigen vector v. we assume $\|v\|_\infty = 1$.

Let $k \in \{1, 2, .., n\}$ be such that

$$|v_k| = \|v\|_\infty = \max_{1 \le i \le n} |v_i| = 1.$$

So $|v_i| \le 1 \quad \forall \quad 1 \le i \le n$.

Now, $M_J v = \lambda v \Rightarrow D^{-1}(L+U)v = \lambda v$, $\left[\because M_J = D^{-1}(L+U)\right]$

$\Rightarrow (L+U)v = \lambda D v$

For k-th component $\lambda (Dv)_k = ((L+U)v)_k$

$$\Rightarrow \lambda a_{kk} v_k = \sum_{j=1}^{n} (L+U)_{kj} v_j$$

$$\Rightarrow \lambda a_{kk} v_k = \sum_{j=1}^{k-1} a_{kj} v_j + \sum_{j=k+1}^{n} a_{kj} v_j$$

$$\left[\because L_{ij} = \begin{cases} a_{ij}, & 1 \le j < i \le n \\ 0, & 1 \le i \le j \le n \end{cases} \right.$$

$$\left. v_j = \begin{cases} 0, & 1 \le j \le i \le n \\ a_{ij}, & 1 \le i < j \le n \end{cases} \right]$$

$$\Rightarrow |\lambda a_{kk} v_k| \le \sum_{j=1}^{k-1} |a_{kj} v_j| + \sum_{j=k+1}^{n} |a_{kj} v_j|$$

$$\Rightarrow |\lambda| |a_{kk}| |v_k| \le \sum_{j=1}^{k-1} |a_{kj}| |v_j| + \sum_{j=k+1}^{n} |a_{kj}| |v_j|$$

$$\Rightarrow |\lambda| |a_{kk}| \le \sum_{j=1}^{k-1} |a_{kj}| + \sum_{j=k+1}^{n} |a_{kj}|$$

$(\because |v_k| = 1$
$|v_j| \le 1, \ j \ne k)$

$$\Rightarrow |\lambda| |a_{kk}| \le \sum_{\substack{j=1 \\ j \ne k}}^{n} |a_{kj}| < |a_{kk}|$$

(by definition of strict diagonal dominant)

$$\Rightarrow |\lambda| < 1 \Rightarrow \rho(M_J) < 1$$

$(\because$ Eigen values of M_J are finite in number).

P-12

To show $\rho(M_S) < 1$

Again as before, for eigen value λ, let v be an eigen vector with $\|v\|_\infty = 1$.

Choosing $1 \le K \le n$ as $|v_K| = \|v\|_\infty = \max_{1 \le i \le n} |v_i| = 1$.

Now, $M_S v = \lambda v$

$\Rightarrow (D+L)^{-1} U v = \lambda v \quad (\because M_S = (D+L)^{-1} U)$

$\Rightarrow Uv = \lambda (D+L) v$

Consider the K-th component $(Uv)_K = \lambda ((D+L)v)_K$

$$\Rightarrow \sum_{j=1}^{n} U_{Kj} v_K = \lambda a_{KK} v_K + \lambda \sum_{j=1}^{n} L_{Kj} v$$

$$\Rightarrow \sum_{j=K+1}^{n} a_{Kj} v_K = \lambda a_{KK} v_K + \lambda \sum_{j=1}^{K-1} a_{Kj} v_j$$

$$\Rightarrow -\lambda a_{KK} v_K = -\sum_{j=K+1}^{n} a_{Kj} v_j + \lambda \sum_{j=1}^{K-1} a_{Kj} v_j$$

$$\Rightarrow |\lambda||a_{KK}||v_K| \le \sum_{j=K+1}^{n} |a_{Kj}||v_j| + |\lambda| \sum_{j=1}^{K-1} |a_{Kj}||v_j|$$

$$\Rightarrow |\lambda||a_{KK}| \le \sum_{j=K+1}^{n} |a_{Kj}| + |\lambda| \sum_{j=1}^{K-1} |a_{Kj}| \quad (\because |v_j| \le 1)$$

$$\Rightarrow |\lambda| \left(|a_{KK}| - \sum_{j=1}^{K-1} |a_{Kj}| \right) \le \sum_{j=K+1}^{n} |a_{Kj}| \longrightarrow (10)$$

Since $|a_{kk}| > \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| = \sum_{j=1}^{k-1} |a_{kj}| + \sum_{j=k+1}^{n} |a_{kj}|$

$$\Rightarrow |a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}| > \sum_{j=k+1}^{n} |a_{kj}| > 0.$$

$$\Rightarrow 1 > \sum_{j=k+1}^{n} |a_{kj}| \Big/ \left(|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}|\right) \longrightarrow (11)$$

from (10), we have

$$|\lambda| \leq \frac{\sum_{j=k+1}^{n} |a_{kj}|}{\left(|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}|\right)} < 1 \quad (\text{using (11)})$$

$$\Rightarrow |\lambda| < 1 \Rightarrow \rho(M_S) < 1.$$

This finishes the proof of the theorem.

Theorem (5) For any $\omega \in \mathbb{R}$, $\rho(M_{SOR}) \geq |1-\omega|$.

Hence, if SOR method converges, then $0 < \omega < 2$.

proof We first compute $\det(M_{SOR})$.

Let us recall, $M_{SOR} = (D+\omega L)^{-1}[(1-\omega)D - \omega U]$

$$\det M_{SOR} = \det\left\{(D+\omega L)^{-1}[(1-\omega)D - \omega U]\right\}$$

$$= \det\left\{(D+\omega L)^{-1}\right\} \det\left\{(1-\omega)D - \omega U\right\}$$

$$= \frac{1}{\det(D+\omega L)} \det\left\{(1-\omega)D - \omega U\right\}.$$

[since, $L_{ii} = 0$, diagonal elements of $(D+\omega L)_{ii} = (D+\omega L)_{ii}$

$= D_{ii}$

Similarly, $U_{ii} = 0$

diagonal elements of $(1-\omega)D - \omega U = ((1-\omega)D - \omega U)_{ii}$

$= (1-\omega)D_{ii}$]

Hence, $\det M_{SOR} = \frac{1}{D_{11}D_{22}\cdots D_{nn}} \times (1-\omega)^n D_{11}D_{22}\cdots D_{nn}$

$$= (1-\omega)^n.$$

$\Rightarrow \lambda_1 \cdots \lambda_n = (1-\omega)^n$ (let $\lambda_1, \lambda_2 \ldots, \lambda_n$'s are eigen values of M_{SOR}).

$\Rightarrow \quad |\lambda_1| \cdots |\lambda_n| = |1-\omega|^n \longrightarrow (12)$

If $|\lambda_i| < |1-\omega| \quad \forall\ i=1,\ldots,n \Rightarrow |\lambda_1||\lambda_2|\cdots|\lambda_n| < |1-\omega|^n$,

which contradicts (12).

Hence, $\exists\ 1 \le k \le n$ so that $|\lambda_k| \ge |1-\omega|$.

$\Rightarrow \quad \rho(M_{SOR}) \ge |\lambda_k| \ge |1-\omega|$.

$\Rightarrow \quad \rho(M_{SOR}) \ge |1-\omega|$. $\longrightarrow (13)$

For the 2nd part, if SOR method converges then

$\rho(M_{SOR}) < 1$ (by theorem (3))

now (13) $\Rightarrow \quad |1-\omega| < 1 \quad \Rightarrow -1 < 1-\omega < 1$

$\Rightarrow \quad -2 < -\omega < 0$

$\Rightarrow \quad 0 < \omega < 2$.

(proved)

Theorem (6) (Ostrowski-Reich)

If $A \in M_n(\mathbb{R})$ be positive definite matrix and $0 < \omega < 2$, then the SOR method converges for any choice of initial approximation x^0.

Theorem (7) (see: Ortega, Numerical analysis)

If $A \in M_n(\mathbb{R})$ be positive definite and tridiagonal, then $\rho(M_S) = (\rho(M_J))^2 < 1$ and the optimal choice of ω for SOR is $\omega_{opt} = \dfrac{2}{1+\sqrt{1-\rho(M_J)^2}}$. Hence $\rho(M_{SOR}) = \omega - 1$.

Let $\mathbb{K}$ be a field and $A \in M_n(\mathbb{K})$.

In our case we consider $\mathbb{K}$ to be $\mathbb{R}$ or $\mathbb{C}$.

Defn (Eigen value / Eigen vector)

A non-zero vector $v \in \mathbb{K}^n$ is called an eigen vector of A if there exists $\lambda \in \mathbb{K}$ such that $Av = \lambda v$ in $\mathbb{K}^n$.

In this case, $\lambda \in \mathbb{K}$ is called an eigen value of A corresponding to the eigen vector v.

$(\lambda, v) \in \mathbb{K} \times \mathbb{K}^n$ is called an eigen pair.

Remark 1

Let $v \in \mathbb{K}^n$ be an eigen vector of A. Then, αv for $\alpha \neq 0$, $\alpha \in \mathbb{K}$ is also an eigen vector of A. Indeed, we see that:

$$A(\alpha v) = \alpha Av = \alpha \lambda v = \lambda(\alpha v).$$

Defn (Characteristic polynomial)

The polynomial $C_A(t) = \det(tI - A)$ in t is called the characteristic polynomial of A.

Remark 2

We know that from Linear Algebra,

$\lambda \in \mathbb{K}$ is an eigen value of A if and only if

$C_A(\lambda) = \det(\lambda I - A) = 0$ in $\mathbb{K}$.

Indeed, $\lambda \in \mathbb{K}$ is an eigen value of A.

$\Leftrightarrow \exists\ v \in \mathbb{K}^n,\ v \neq 0$ such that $Av = \lambda v$ in $\mathbb{K}^n$

ie $(\lambda I - A)v = 0$ in $\mathbb{K}^n$

$\Leftrightarrow \det(\lambda I - A) = 0$ in $\mathbb{K}$.

Eg (1). $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} \in M_2(\mathbb{R})$

$$C_A(t) = \det \begin{bmatrix} t-1 & -1 \\ -4 & t-1 \end{bmatrix} = (t-1)^2 - 4 = t^2 - 2t - 3$$

$$= (t-3)(t+1)$$

Two eigen values $\lambda_1 = 3$, $\lambda_2 = -1$

Eg: (2) $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in M_2(\mathbb{R})$

$$C_A(t) = \det \begin{bmatrix} t & -1 \\ -1 & t \end{bmatrix} = t^2 + 1$$

A has no eigen values in $\mathbb{R}$.

But if we consider $A \in M_n(\mathbb{C})$.

Then, A has eigen values. $\lambda_1 = i$, $\lambda_2 = -i$.

Remark 3 Let (λ, v) be an eigen pair of A.

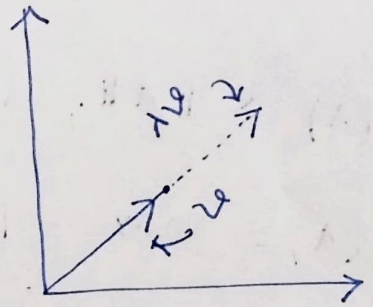
Then, $Av = \lambda v$

This eqn. says, in the direction of the vector v A acts as like scalar multiplication with scalar λ.

Geometrically: (Let $\mathbb{K} = \mathbb{R}$).

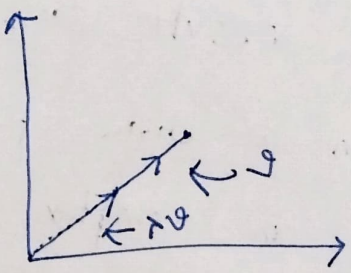
1. $\lambda > 1$, $A \in M_2(\mathbb{R})$

It magnifies the vector v.

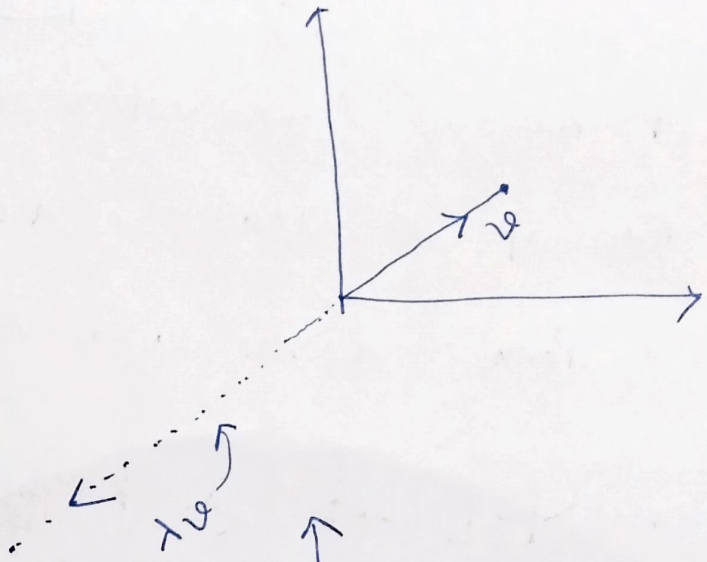


2. $0 < \lambda < 1$,

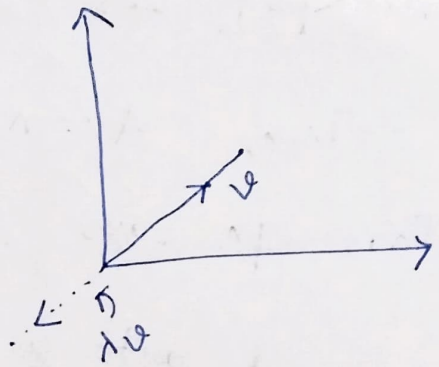
It shrinks the vector



3. $\lambda < -1$



4. $-1 < \lambda < 0$

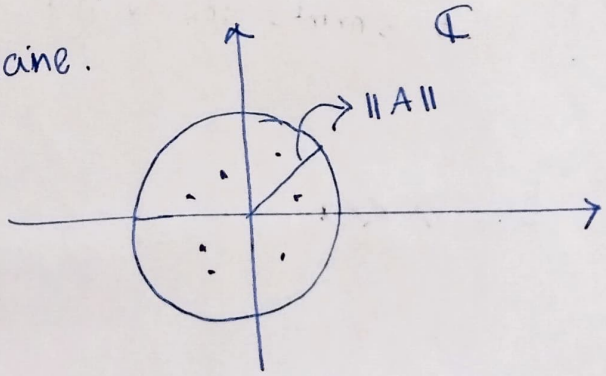


Let us recall, that for any matrix norm $\|\ \|$ in $M_n(\mathbb{C})$ subordinate to a vector norm in $\mathbb{C}^n$,

Then, $\rho(A) \leq \|A\|$; (Theorem 7.8, page-489, Supple note).

where $\rho(A) = \max \{ |\lambda| \mid \lambda \text{ is an eigenvalue of } A \text{ in } \mathbb{C} \}$.

The above result says eigen values of A lies in a disk of radius $\|A\|$, centred at 0 in complex plane.



we want provide better estimate for the location of the eigen values of A.

Defn For $A \in M_n(\mathbb{C})$, $A = (a_{ij})$

let $$R_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n$$

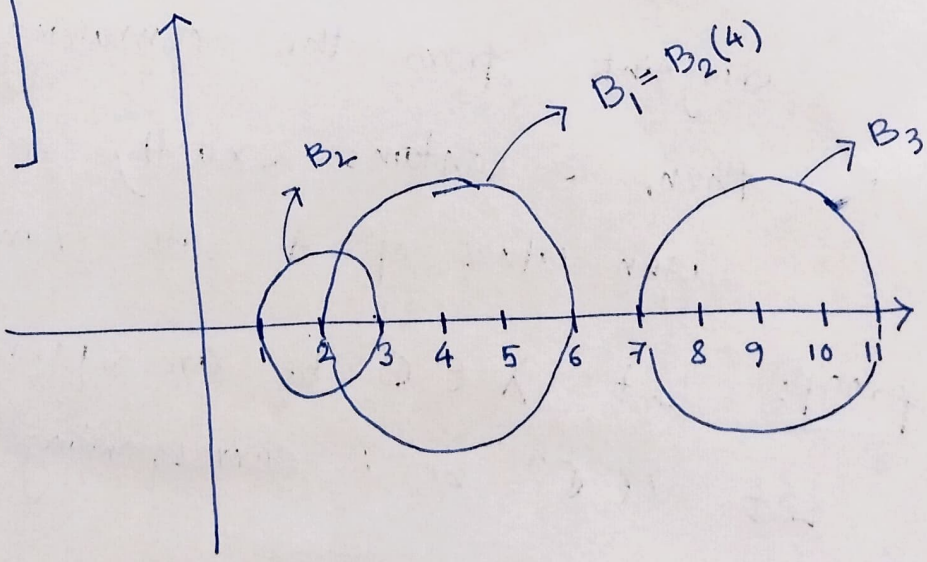
and $$C_j(A) = \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|, \quad j = 1, 2, \ldots, n$$

For $i = 1, 2, \ldots, n$, ith Gerschgorin disk is defined by

$$B_i := B_{R_i(A)}(a_{ii})$$

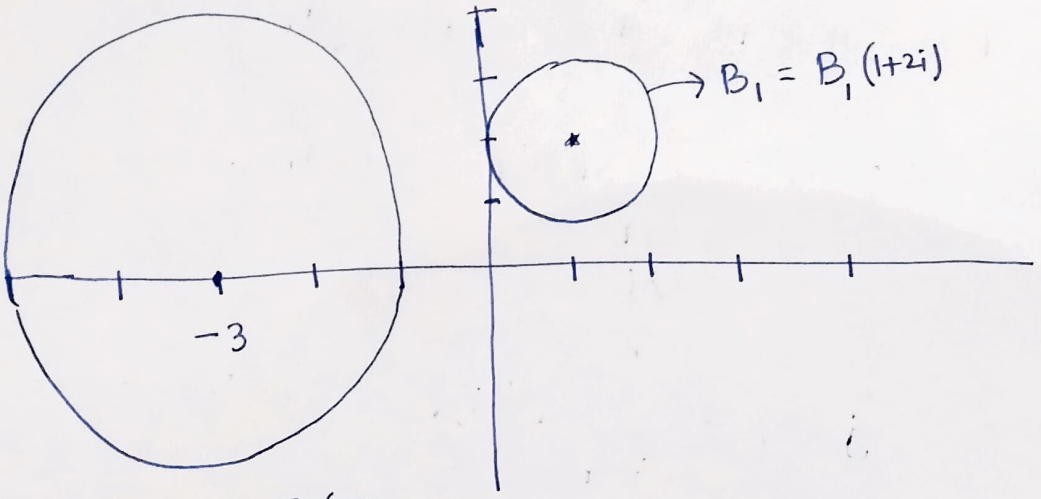
$$:= \{ z \in \mathbb{C} \mid |z - a_{ii}| \leq R_i(A) \}$$

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{bmatrix}$$



P-6

$$A = \begin{bmatrix} 1+2i & 1 \\ 2i & -3 \end{bmatrix}$$



$B_2 = B_2(-3)$.

Theorem 1 (Gerschgorin Circle theorem)

Let $A \in M_n(\mathbb{C})$. Then, every eigen value of A is contained in one of Gerschgorin disks. Moreover, if m $(1 \le m \le n)$ of these disks form a connected subset S, which is disjoint from the remaining $(n-m)$ disks, then, S contains exactly m of the eigen-values of A with counting multiplicities.

proof: Let $\lambda \in \mathbb{C}$ be an eigen value of A. Let $v \in \mathbb{C}^n$ be a corresponding eigen vector

with components $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{C}^n$.

Then, $Av = \lambda v$ in $\mathbb{C}^n$. $\longrightarrow (*)_1$

Let $k \in \{1, 2, \ldots, n\}$ be such that

$$|v_k| = \max_{1 \le j \le n} |v_j| \longrightarrow (*)_2$$

we note that as $v \neq 0 \Rightarrow |v_k| \neq 0$.

Then, the k-th component of $(*)_1$ (both sides)

$$(Av)_k = (\lambda v)_k$$

$$\sum_{j=1}^{n} a_{kj} v_j = \lambda v_k$$

$$\Rightarrow \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj} v_j + a_{kk} v_k = \lambda v_k$$

$$\Rightarrow \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj} v_j = (\lambda - a_{kk}) v_k$$

$$\Rightarrow |(\lambda - a_{kk}) v_k| = \left| \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj} v_j \right| \le \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}||v_j|$$

P-8

$$\Rightarrow |\lambda - a_{kk}||v_k| \leq \left(\sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|\right)|v_k| \quad (\text{by } *)_2$$

$$\Rightarrow |\lambda - a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| = R_k(A) \quad (\because |v_k| \neq 0)$$

$$\Rightarrow |\lambda - a_{kk}| \leq R_k(A)$$

$$\Rightarrow \lambda \in B_k.$$

proof of 2nd part is given in Atkinson, page- 588- 590, Theorem 9.1.

Theorem 2 Let $A = (a_{ij}) \in M_n(\mathbb{C})$ and

$$D_j = \{z \in \mathbb{C} \mid |z - a_{jj}| \leq C_j(A)\} \text{ For } j = 1, 2, \ldots, n$$

Then, every eigen value of A is contained in one of the disk D_j.

proof: Since, both A and A^T have same eigen values and characteristic polynomials. Result will follow, consider Theorem 1 with A replaced by A^T.

Power method

The power method is an iterative technique, which is used to determine the dominant eigen value of a matrix – that is eigen value with largest magnitude. By modifying the method slightly, it can also be used to determine other eigen values. This method produces not only an eigen value but also an eigen vector.

Def. (Dominant Eigen-value)

An eigen value of $A \in M_n(\mathbb{C})$ is said to be a dominant eigen value of A if

$$|\lambda| = \rho(A).$$

Remark 4

1. If a dominant eigen-value of a matrix is equal to zero, then all its eigen-values are zero and an eigen vector can be found by solving the linear system $Ax = 0$.
2. Let $\lambda_1, \lambda_2 \cdots, \lambda_n$ be eigen values of A repeated according to their algebraic multiplicities.

We can rearrange and rename them as

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

i.e. λ_1 is a dominant eigen value of A.

3. A matrix may have unique dominant eigen value or more than one dominant eigen value.

Moreover, if dominant eigen value is unique, the corresponding algebraic and geometric multiplicities could be more than one and also both algebraic and geometric multiplicities may not be the same.

Eg: (i) $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 3 \\ 0 & 0 & -1 \end{pmatrix}$

Eigen values are $1, -2, -1$

unique dominant eigen value $= -2$

$|-2| > |-1| = |1|$.

$$\begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 1 \\ 0 & 0 & -2 \end{bmatrix}$$

Eigen values $1, 2, -2$

dominant eigen values $2, -2$

not unique.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

unique dominant eigen value $= 2$

geometric multiplicity of $2 = 1$

Algebraic " of $2 = 2$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

unique dominant eigen value $= 2$

Algebraic multiplicity of $2 = 2$

geometric " of $2 = 2$.

Description of the method.

Assumption : Let $A \in M_n(\mathbb{R})$ has real eigen values $\lambda_1, \lambda_2, \ldots, \lambda_n$ (repeated according to their multiplicities)

(1). Eigen values are rearranged as

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

A has unique dominant eigen-value with algebraic and geometric multiplicities is equal to one (simple).

(2) The matrix A is diagonalizable.

ie. There exists a basis of $\mathbb{R}^n$ consisting of eigen vectors of A.

Let us denote the eigen pairs $\{(\lambda_i, v_i)\}_{i=1}^{n}$ where $\{v_1, v_2, \ldots, v_n\} \subset \mathbb{R}^n$ forms a basis of $\mathbb{R}^n$.

Method: Let $x^0 \in \mathbb{R}^n$ be a vector arbitrary so that

$$x^0 = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n \quad \text{with } \alpha_1 \neq 0$$

$\alpha_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$.

Then,

$$x^1 = Ax^0 = \alpha_1 Av_1 + \alpha_2 Av_2 + \cdots + \alpha_n Av_n$$

$$x^1 = \alpha_1 \lambda_1 v_1 + \alpha_2 \lambda_2 v_2 + \cdots + \alpha_n \lambda_n v_n$$

$$x^2 = Ax^1 = \alpha_1 \lambda_1^2 v_1 + \alpha_2 \lambda_2^2 v_2 + \cdots + \alpha_n \lambda_n^2 v_n$$

For $m \in \mathbb{N}$

$$x^m = A^m x^0 = \alpha_1 \lambda_1^m v_1 + \alpha_2 \lambda_2^m v_2 + \cdots + \alpha_n \lambda_n^m v_n.$$

$$x^m = \lambda_1^m \left[\alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j \right]$$

We note that $\left|\frac{\lambda_j}{\lambda_1}\right| < 1$ for $j = 2, \cdots, n$

Hence, $\left(\frac{\lambda_j}{\lambda_1}\right) \to 0$ as $m \to \infty$.

Hence, $x^m \approx \lambda_1^m \alpha_1 v_1$ provided $\alpha_1 \neq 0$.

But it may happen if $|\lambda_1| < 1$ small enough

so that for large $m \in \mathbb{N}$

$\lambda_1^m = 0$ in Computer

or if $|\lambda_1| > 1$ large enough so that

for large $m \in \mathbb{N}$, $\lambda_1^m = \text{Inf}$ in computer.

Therefore, iteration x^m will not be very practical for computing.

In practice, the approximate is normalized at each iteration.

P-14 **To find approximate eigen vector**

We define for $m \geq 1$

$$y^m = A x^{m-1}, \quad x^m = \frac{y^m}{\|y^m\|_\infty},$$

Let us elaborate.

$$y^1 = Ax^0, \quad x^1 = \frac{y^1}{\|y^1\|_\infty} = \frac{Ax^0}{\|Ax^0\|_\infty}$$

$$y^2 = Ax^1, \quad x^2 = \frac{y^2}{\|y^2\|_\infty} = \frac{Ax^1}{\|Ax^1\|_\infty} = \frac{A^2x^0}{\|Ax^0\|_\infty} \times \frac{\|Ax^0\|_\infty}{\|A^2x^0\|_\infty}$$

$$= \frac{A^2x^0}{\|A^2x^0\|_\infty}$$

Inductively, $y^m = Ax^{m-1}, \quad x^m = \frac{A^m x^0}{\|A^m x^0\|_\infty}$

As, $A^m x^0 = \alpha_1 \lambda_1^m v_1 + \alpha_2 \lambda_2^m v_2 + \cdots + \alpha_n \lambda_n^m v_n$

$$A^m x^0 = \lambda_1^m \left[\alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j \right]$$

and

$$\|A^m x^0\|_\infty = \left\| \lambda_1^m \left[\alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j \right] \right\|_\infty$$

$$= |\lambda_1|^m \left\| \alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j \right\|_\infty$$

Hence, $$x^m = \frac{\lambda_1^m \left(\alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j\right)}{|\lambda_1|^m \left\| \alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j \right\|_\infty}$$

$$x^m = \left(\frac{\lambda_1}{|\lambda_1|}\right)^m \cdot \frac{\left(\alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j\right)}{\left\| \alpha_1 v_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_j \right\|_\infty} \longrightarrow (*)_3$$

As $\left|\frac{\lambda_j}{\lambda_1}\right| < 1$ for $j = 2, 3, \cdots, n$

$$\left(\frac{\lambda_j}{\lambda_1}\right)^m \longrightarrow \infty \quad \text{as} \quad m \longrightarrow \infty$$

and $$\frac{\lambda_1}{|\lambda_1|} = \text{sign}(\lambda_1) = \pm 1$$

letting $m \longrightarrow \infty$, in $(*)_3$ we note that

$$x^m \longrightarrow \begin{cases} \dfrac{\alpha_1 v_1}{\|\alpha_1 v_1\|_\infty} & \text{or if } \lambda_1 > 0 \\ \text{oscillates between} & \\ \pm \dfrac{\alpha_1 v_1}{\|\alpha_1 v_1\|_\infty} & \text{if } \lambda_1 < 0. \end{cases}$$

$$\frac{\alpha_1}{|\alpha_1|} = \text{sign}(\alpha_1) = \pm 1$$

Therefore, as $m \to \infty$

$$x^m \longrightarrow \begin{cases} + \dfrac{v_1}{\|v_1\|_\infty} & \text{if } \lambda_1 > 0,\ \alpha_1 > 0. \\ - \dfrac{v_1}{\|v_1\|_\infty} & \text{if } \lambda_1 > 0,\ \alpha_1 < 0 \\ \text{Oscillates between } \pm \dfrac{v}{\|v_1\|_\infty} & \text{if } \lambda_1 < 0. \end{cases}$$

~~For approximate eigen vat~~

Therefore for large m, x^m provides an eigen vector (scaling with $\|\ \|_\infty$ norm).

To find approximate eigen value.

Let $k \in \{1, 2, .., n\}$ be such that

$v_{1k} \neq 0$ where $v_1 = \begin{pmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1n} \end{pmatrix} \in \mathbb{R}^n$.

Define for $m \geq 1$

$$\sigma_m = \frac{(y^m)_k}{(x^{m-1})_k}$$

$\longrightarrow$ (*)

$(y^m)_k$ = kth component of vector y^m

$(x^{m-1})_k$ = kth component of vector x^{m-1}.

We note that , $x^{m-1} = \frac{A^{m-1}x^0}{\|A^{m-1}x^0\|_\infty}$.

$y^m = Ax^{m-1}$.

$$y^m = \frac{AA^{m-1}x^0}{\|A^{m-1}x^0\|_\infty} = \frac{A^m x^0}{\|A^{m-1}x^0\|_\infty}.$$

So, $$(x^{m-1})_k = \frac{1}{\|A^{m-1}x^0\|_\infty}(A^{m-1}x^0)_k$$

$$= \frac{\lambda_1^{m-1}}{\|A^{m-1}x^0\|}\left(\alpha_1 v_{1k} + \sum_{j=2}^{n}\left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_{jk}\right)$$

$$(y^m)_k = \frac{1}{\|A^{m-1}x^0\|_\infty}(A^m x^0)_k$$

$$= \frac{\lambda_1^m}{\|A^{m-1}x^0\|_\infty}\left(\alpha_1 v_{1k} + \sum_{j=2}^{n}\left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_{jk}\right)$$

Hence, From (*)/ we have

$$\sigma_m = \frac{(y^m)_k}{(x^{m-1})_k} = \lambda_1 \frac{\left(\alpha_1 v_{1k} + \sum_{j=2}^{n}\left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_{jk}\right)}{\left(\alpha_1 v_{1k} + \sum_{j=2}^{n}\left(\frac{\lambda_j}{\lambda_1}\right)^m \alpha_j v_{jk}\right)}$$

Letting $m \to \infty$,

$$\sigma_m \longrightarrow \frac{\lambda_1(\alpha_1 v_{1k})}{\alpha_1 v_{1k}} = \lambda_1 \quad [\because \alpha_1 \neq 0 \neq v_{1k}]$$

P-18 Therefore for large m, σ_m gives an approximation of the eigen value λ_1.

Disadvantages of power method

1. The power method requires that the matrix has only one dominant eigen-value but apriori this information is not known.
2. Apriori it is not known that the matrix is diagonalizable
3. Apriori it is not known how to choose the initial vector x^0 so that $\alpha_1 \neq 0$.

Remark 5

1. Power can be modified to find eigen value and eigen vector in case of single dominant eigen value of geometric multiplicity greater than 1.
2. The power method (modified) can be used if there is multiple dominant eigen value, but the algorithm is more complicated.

3. The power method can also be used to determine eigen values other than dominant eigen value. This process is called deflation method.

Theorem 3

Let $(\lambda_i, v_i)_{i=1}^{n}$ be an eigen pairs of A with λ_1 has multiplicity 1.

Let $x \in \mathbb{R}^n$ such that $x^T v_1 = 1$.

Then, the matrix $B = A - \lambda_1 v_1 x^T \in M_n(\mathbb{R})$ has eigen values $0, \lambda_2, \ldots, \lambda_n$ with associated eigen vectors $v_1, w_2, \ldots, w_n$, where w_j, $j = 2, \ldots, n$ are given by

$$v_j = (\lambda_j - \lambda_1) w_j + \lambda_1 x^T w_j v_1$$

$j = 2, 3, \ldots, n$

proof: Wilkinson: The algebraic eigen value problem, p-596.

There are many choices of the vector x.

(i) Wielandt deflation: $x = \frac{1}{\lambda_1 v_{1k}} (a_{k1}, \ldots, a_{kn})^T$, where

P-20 $k \in \{1, 2, .., n\}$ is such that $v_{1k} \neq 0$

$(a_{k1}, \ldots, a_{kn}) \rightarrow k$ th row of A.

(ii) Hotelling deflation:

$$x = \frac{v_1}{v_1^T v_1}.$$

4. The sequence x^m converges to an eigen vector corresponding to the dominant eigen value λ_1 if there exists unique index $k \in \{1, 2, .., n\}$ such that

$$|v_{1k}| = \|v_1\|_2 .$$
