A Concise Introduction to Numerical Analysis

Lecture Notes

Mohammad Sabawi

Department of Mathematics
College of Education for Women
Tikrit University

Email: mohammad.sabawi@tu.edu.iq

01 October 2021

Contents

Preface

The aim of these class notes is to cover the necessary materials in a standard numerical analysis course and it is not intended to add to the plethora of Numerical Analysis texts. We tried our best to write these notes in concise, clear and accessible way, to make them more attractive to the readers. These lecture notes cover the basic and fundamental concepts and principles in numerical analysis and it is not a comprehensive introduction to numerical analysis. We emphasise in these notes on the mathematical principles via explaining them by the aid of numerical software **MATLAB**. The prerequisite material for this course are a course in **Calculus**, **Linear Algebra** and **Differential Equations**. A basic knowledge in MATLAB is helpful but it is not necessary. There is a glut of numerical software nowadays, among these we chose to use MATLAB because of its wide capabilities in scientific computing.

The notes contain sufficient material for a full year of study and can be covered in two courses for undergraduate mathematics and engineering students.These notes consist of eight chapters cover the basic and fundamental topics in numerical analysis. Each chapter contains some relevant examples to illustrate the concepts and ideas introduced in the chapter and ends with a set of exercises address the topics covered in each chapter.

Chapter 1

Introduction

1.1 Numerical Analysis: An Introduction

Numerical analysis is a branch of mathematics studies the methods and algorithms which used for solving a variety of problems in different areas of todays life such as *mathematics*, *physics*, *engineering*, *medicine* and *social* and *life sciences*. The main objective of numerical analysis is investigation finding new mathematical approaches for approximating the underlying problems, and also development of the current algorithms and numerical schemes to make them more efficient and reliable. The advent of computers revolutionise numerical analysis and nowadays with parallel and super computers the numerical computations became more easier compared with the past where solving simple problems take a long time, much effort and require hard work. In principle, numerical analysis mainly focuses on the ideas of *stability*, *convergence*, *accuracy*, *consistency* and *error analysis*. In the literature numerical analysis also known as **scientific computing**, **scientific computation**, **numerics**, **computational mathematics** and **numerical mathematics**. Numerical analysis can be divided into the following fields:

1. **Numerical Solutions of Linear Algebraic Equations**.
2. **Numerical Solutions of Nonlinear Algebraic Equations**.
3. **Interpolation and Extrapolation**.
4. **Approximation Theory and Curve Fitting**.
5. **Numerical Differentiation**.
6. **Numerical Integration**.

7. **Numerical Optimisation**.
8. **Numerical Solutions of Eigenvalue Problems**.
9. **Numerical Solutions of Ordinary Differential Equations**.
10. **Numerical Solutions of Partial Differential Equations**.
11. **Numerical Solutions of Integral Equations**.
12. **Numerical Modelling**.
13. **Numerical Simulation**.

Numerical analysis is dated back to the Babylonians works in approximating the square root of 2. During this long journey of evolution many many scientists contributed to its development and progress among these we just name a few such as **Lagrange**, **Gauss**, **Newton**, **Euler**, **Legendre** and **Simpson**.

1.2 Numbers Representation in Computer

Human beings do arithmetic in their daily life using the *decimal (base* 10*) number system*. Nowadays, most computers use *binary (base* 2*) number system*. We enter the information to computers using the decimal system but computers translate them to the binary system by using the *machine language*.

Definition 1 (**Scientific Notation**). *Let k be a real number, then k can be written in the following form*

$$k = m \times 10^n,$$

where m is any real number and the exponent n is an integer. This notation is called the ***scientific notation*** *or* ***scientific form*** *and sometimes referred to as* ***standard form****.*

Example 1. *Write the following numbers in scientific notation:*

1. 0.00000834.

2. 25.45879.

3. 3400000.

4. 33.

5. $2,300,000,000$.

6. 2.718282.

Solution:

1. $0.00000834 = 8.34 \times 10^{-6}$.
2. $25.45879 = 2.545879 \times 10^{1}$.
3. $3400000 = 3.4 \times 10^{6}$.
4. $33 = 3.3 \times 10^{1}$.
5. 2.3×10^{9}.
6. $2.718282 = 2.718282 \times 10^{0}$.

1.2.1 Floating-Point Numbers

In the decimal system any real number $a \neq 0$ can be written in the **normalised decimal floating-point form** in the following way

$$a = \pm 0.d_1 d_2 d_3 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n,\ 1 \leq d_1 \leq 9,\ 0 \leq d_i \leq 9, \tag{1.1}$$

for each $i = 2, \cdots$, and n is an integer called the **exponent** (n can be positive, negative or zero). In computers we use a finite number of digits in representing the numbers and we obtain the following form

$$b = \pm 0.d_1 d_2 d_3 \cdots d_k \times 10^n,\ 1 \leq d_1 \leq 9,\ 0 \leq d_i \leq 9, \tag{1.2}$$

for each $i = 2, \cdots, k$. These numbers are called **k-digit decimal machine numbers**.

Also, the normalised floating-point decimal representation of the number $a \neq 0$ can be written in other way as

$$a = \pm r \times 10^n,\ (\frac{1}{10} \leq r < 1), \tag{1.3}$$

the number r is called the **normalised mantissa**.

The floating-point representation in binary number system can be defined by the same way as in the decimal number system. If $a \neq 0$, it can be represented as

$$a = \pm p \times 2^m,\ (\frac{1}{2} \leq p < 1), \tag{1.4}$$

where $p = (0.b_1 b_2 b_3 \cdots)_2$, $b_1 = 1$.

1.3 Errors

Occurrence of error is unavoidable in the field of scientific computing. Instead, numerical analysts try to investigate the possible and best ways to minimise the error. The study of the error and how to estimate and minimise it are the fundamental issues in *error analysis.*

1.3.1 Error Analysis

In numerical analysis we approximate the exact solution of the problem by using numerical method and consequently an error is committed. The numerical error is the difference between the exact solution and the approximate solution.

Definition 2 (**Numerical Error**). *Let x be the exact solution of the underlying problem or a true value and x^* its approximate solution or approximate value, then the error (denoted by e) in solving this problem or in approximating the value of x is*

$$e = x - x^*. \tag{1.5}$$

1.3.2 Sources of Error in Numerical Computations

- **Blunders (Gross Errors)** These errors also called **humans errors**, and are caused by humans mistakes and oversight and can be minimised by taking care during scientific investigations. These errors will add to the total error of the underlying problem and can significantly affect the accuracy of solution.
- **Modelling Errors** These errors arise during the modelling process when scientists ignore effecting factors in the model to simplify the problem. Also, these errors known as **formulation errors**.
- **Data Uncertainty** These errors are due to the uncertainty of the physical problem data and also known as **data errors** or **noise**.
- **Discretisation Errors** Computers represent a function of continuous variable by a number of discrete values. Also, scientists approximate and replace complex continuous problems by discrete ones and this results in **discretisation errors**.
- **Loss of Significance** This phenomenon occurs when subtracting two nearly equal numbers and can be avoided by using some mathematical

tricks such as algebraic manipulation. It is also known as **subtractive cancellation**, **catastrophic cancellation** or **loss of significant digits**.

- **Rounding Errors** Computers represent numbers in finite number of digits and hence some quantities cannot be represented exactly. The error caused by replacing a number a by its closest machine number is called the **roundoff error** or **round-off error** and the process is called **correct rounding**. This type of error happens when a true value of a real number x sometimes not stored or saved exactly due to the limited fixed precision of computer's representation.

- **Chopping Errors** These errors occur when chopping a number with infinite digits or a number with $k+1$ digits and replaced it by a $k-$digits number.

- **Truncation Errors** These errors arise when replacing complicated mathematical expressions by simple and elementary mathematical formulas. As an example of truncation error approximating a complicated function with truncated Taylor series. We will discuss truncation error in detailed way later.

1.3.3 Floating Point Representation

1.3.4 Absolute and Relative Errors

Definition 3 (**Absolute Error**). *The absolute error $\hat{e}$ of the error e is defined as the absolute value of the error e*

$$\hat{e} = |x - x^*|. \tag{1.6}$$

Definition 4 (**Relative Error**). *The relative error $\tilde{e}$ of the error e is defined as the ratio between the absolute error $\hat{e}$ and the absolute value of the true value x*

$$\tilde{e} = \frac{\hat{e}}{|x|} = \frac{|x - x^*|}{|x|}, \; x \neq 0. \tag{1.7}$$

Example 2. *Let $x = 3.141592653589793$ is the value of the constant ratio π correct to 15 decimal places and $x^* = 3.14159265$ be an approximation of x. Compute the following quantities:*

a. The error.

b. The absolute error.

c. The relative error.

Solution:

a. The error

$$\begin{aligned} e = x - x^* &= 3.141592653589793 - 3.14159265 = 3.589792907376932e - 09 \\ &= 3.589792907376932 \times 10^{-9} = 0.000000003589792907376932. \end{aligned}$$

b. The absolute error

$$\hat{e} = |x - x^*| = |3.141592653589793 - 3.14159265| = 3.589792907376932e - 09.$$

c. The relative error

$$\begin{aligned} \tilde{e} &= \frac{\hat{e}}{|x|} = \frac{|x - x^*|}{|x|} = \frac{3.141592653589793 - 3.14159265}{3.141592653589793} \\ &= \frac{3.589792907376932e - 09}{3.141592653589793} = 1.142666571770530e - 09. \end{aligned}$$

Example 3. *Approximate the following decimal numbers to three significant digits by using rounding and chopping rules:*

1. $x_1 = 1.34579.$
2. $x_2 = 1.34679.$
3. $x_3 = 1.34479.$
4. $x_4 = 3.34379.$
5. $x_5 = 2.34579.$
6. $x_6 = 0.54387.$

Solution:

(i) ***Rounding:***

- (a) $x_1 = 1.35.$
- (b) $x_2 = 1.35.$
- (c) $x_3 = 1.34.$
- (d) $x_4 = 3.34.$
- (e) $x_5 = 2.35.$
- (f) $x_6 = 0.544.$

(ii) ***Chopping:***

- (a) $x_1 = 1.34.$
- (b) $x_2 = 1.34.$
- (c) $x_3 = 1.34.$
- (d) $x_4 = 3.34.$
- (e) $x_5 = 2.34.$
- (f) $x_6 = 0.543.$

1.4 Stable and Unstable Computations: Conditioning

Stability is one of the most important characteristics in any efficient and robust numerical scheme.

Definition 5 (**Numerical Stability**). *The numerical algorithm or process is called **stable** if the final result is relatively not affected by the perturbations during computation process. In other words, the numerical method or technique is stable if small changes in the initial conditions or initial data will produce small changes in outputs or final results. Otherwise it is called **unstable**.*

The **stability** notion is analogous and closely related to the notion of **conditioning**.

Definition 6 (**Conditioning**). ***Conditioning** is a measure of how **sensitive** the output to small changes in the input data. In literature **conditioning** is also called **sensitivity**.*

- *The problem is called **well-conditioned** or **insensitive** if small changes in the input data lead to small changes in the output data.*
- *The problem is called **ill-conditioned** or **sensitive** if small changes in the input data lead to big changes in the output data.*

Definition 7 (**Condition Number of a Function**). *If f is a differentiable function at x in its domain then the **condition number** of f at x is*

$$cond(f(x)) = \frac{|xf'(x)|}{|f(x)|},\ f(x) \neq 0. \tag{1.8}$$

Note: Condition number of a function f at x in its domain sometimes denoted by $C_f(x)$.

Definition 8 (**Condition Number of a Matrix**). *If A is a non-singular $n \times m$ matrix, the **condition number** of A is defined by*

$$cond(A) = \|A\|\|A^{-1}\|, \tag{1.9}$$

where

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}, \tag{1.10}$$

and x is a $m \times 1$ column vector.

Definition 9 (**Well-Posed Problem**). *The problem is* ***well-posed*** *if satisfies the following three conditions:*

a. The solution exists.

b. The solution is unique.

c. The solution depends continuously on problem data.

Otherwise, the problem is called ***ill-posed****.*

Remark 1. *Note that:*

1. *The problem is* ***ill-posed*** *or* ***sensitive*** *if cond* $\gg 1$.
2. *The problem is* ***well-posed*** *or* ***insensitive*** *if cond* < 1.

Example 4. *Find the condition number of the function* $f(x) = \sqrt{x}$.

Solution:

$$f(x) = \sqrt{x} \implies f'(x) = \frac{1}{2\sqrt{x}},\ x \neq 0,$$

implies that

$$cond(f(x)) = \frac{|xf'(x)|}{|f(x)|} = \frac{|\frac{x}{2\sqrt{x}}|}{|\sqrt{x}|} = \frac{1}{2}.$$

This indicates that the small changes in the input data lead to changes in the output data of half size the changes in the input data.

Example 5. *Let*

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0.5 & 3 \\ 0.1 & 1 & 0.3 \end{bmatrix},$$

the inverse of A can be computed by using MATLAB command ***inv(A)*** *to obtain*

$$A^{-1} = \begin{bmatrix} 4.7500 & -2.1667 & 5.8333 \\ 0.5000 & -0.3333 & 1.6667 \\ -3.2500 & 1.8333 & -4.1667 \end{bmatrix}.$$

Also, the condition number of A and its inverse can be computed using MATLAB commands ***cond(A)*** *and* ***cond(inv(A))*** *to have* $cond(A) = 37.8704$ *and* $cond(A^{-1}) = 37.8704$. *We notice that the matrix A and its inverse have the same condition numbers.*

Definition 10 (**Accuracy**). *It is a measure of closeness of the approximate solution to the exact solution.*

Definition 11 (**Precision**). *It is a measure of closeness of the two or more measurements to each other.*

Remark 2. *Note that the accuracy and precision are different and they are not related. The problem maybe very accurate but imprecise and vice versa.*

1.5 Convergence and Order of Approximation

Convergence of the numerical solution to the analytical solution is one of the important characteristic in any good and reliable numerical scheme.

Definition 12 (**Convergence of a Sequence**). *Let $\{a_n\}_{n=1}^{\infty}$ be an infinite sequence of real numbers. This sequence is said to be* ***convergent*** *to a real number a (has a* ***limit*** *at a) if, for any $\epsilon > 0$ there exists a positive integer $N(\epsilon)$ such that*

$$|a_n - a| < \epsilon, \text{ whenever } n > N(\epsilon). \tag{1.11}$$

Otherwise it is called a ***divergent*** *sequence, a is called the* ***limit of the sequence*** *a_n. Other commonly used notations for convergence are:*

$$\lim_{n\to\infty} a_n = a \quad or \quad a_n \to a \;\; as \;\; n, \quad or \quad \lim_{n\to\infty}(a_n - a) = 0, \tag{1.12}$$

this means that the sequence $\{a_n\}_{n=1}^{\infty}$ ***converges*** *to a otherwise it* ***diverges****.*

Definition 13 (**Order of Convergence**). *Let the sequence $\{a_n\}_{n=1}^{\infty}$ converges to a and set $e_n = a_n - a$ for any $n > 0$. If two positive constants M and q exist, such that*

$$\lim_{n\to\infty} \frac{|a_{n+1} - a|}{|a_n - a|^q} = \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^q} = M, \tag{1.13}$$

then the sequence $\{a_n\}_{n=1}^{\infty}$ is to be convergent to a with the ***order of convergence*** *q, the number M is called the* ***asymptotic error constant****.*

If $q = 1$, the convergence is called ***linear****.*
If $q = 2$, the convergence is called ***quadratic****.*

If $q = 3$, the convergence is called ***cubic****.*

Note that the convergence gets more rapid as q gets larger and larger.

Example 6. *Consider the sequence $\{\frac{1}{n}\}_{n=1}^{\infty}$, where n is a positive integer. Observe that $\frac{1}{n} \to 0$ as $n \to \infty$, it follows that*

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

Definition 14 (Order of Approximation $O(h^n)$). *The function $f(h)$ is said to be* ***big Oh*** *of the function $g(h)$, if two real constants c, and C exist such that*

$$|f(h)| \leq C|g(h)| \quad \text{whenever } h < c, \tag{1.14}$$

and denoted by $f(h) = O(g(h))$. The order of approximation is used to determine the rate at which a function grows.

Example 7. *Consider the functions $f(x) = x + 1$ and $g(x) = x^2$, where $x \geq 1$. Observe that $x \leq x^2$ and $1 \leq x^2$ for $x \geq 1$, hence $f(x) = x + 1 \leq 2x^2 = 2g(x)$ for $x \geq 1$. Consequently, $f(x) = O(g(x))$.*

Exercises

Exercise 1. *Write the following numbers in scientific form:*

1. 23.123.
2. $30,000,000$.
3. 0.000001573.
4. 39776444.
5. -345.386443.
6. -23000000.

Exercise 2. *Evaluate error, absolute error and relative error of the following values and their approximations:*

1. $x = 1,000,000,\ x^* = 999,999$.
2. $y = 0.00012887765,\ y^* = 0.00012897766$.
3. $z = 9776.96544,\ z^* = 9775.66544$.

Exercise 3. *Approximate the following numbers to four digits using rounding and chopping:*

1. 1.98876.
2. 33.87654.
3. 8.98879.
4. 2.88778.

Exercise 4. *Compute the condition number of the following functions:*

1. $f(x) = \cos(x)$.
2. $f(x) = \cos^{-1}(x)$.

.