北京科技大学
University of Science and Technology Beijing

Preface for Numerical Analysis

Liancun Zheng
School of Mathematics and Physics
University of Science and Technology Beijing
E-mail:liancunzheng@ustb.edu.cn

1. Introduction

Good morning, you are welcome. I am Liancun Zheng, Professor at School of Mathematics and Physics, University of Science and Technology Beijing.

It is my pleasure to meet you here and to give this class. This course, has begun in 2009, for International students of Master or Ph.D, whom major are in science and technology, engineering, computer, material, and so on. It will help you to solve numerical analysis and computational problems, make you powerful in future.

Due to COVID-19, some students are temporarily unable to attend classes. We shall give teaching both online and offline by helping of Tencent, RainClassroom and Wechat. I hope every one study hard, and have a happy life at USTB.

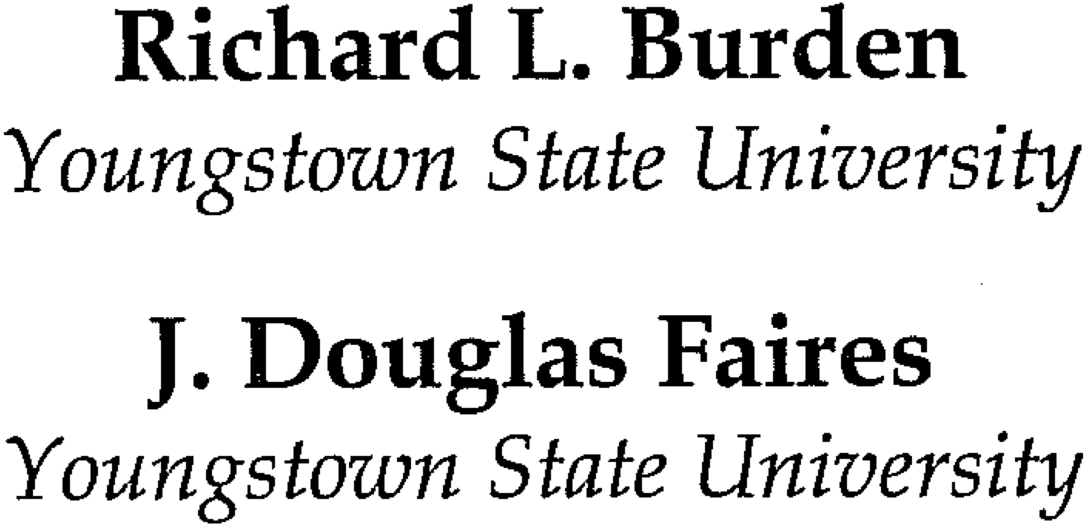
2. Some pictures with students, omitted here.

3. Preface for this course

During the past decades, giant needs for ever more sophisticated mathematical models and increasingly complex and extensive computer simulations have arisen. In this fashion, two indissociable activities, mathematical modeling and computer-simulation, have gained a major status in all aspects of science, technology, and industry.

Numerical analysis is here understood as the part of Mathematics that describes and analyzes all the numerical schemes that are used on computers; its objective consists in obtaining a clear, precise, and faithful, representation of all the "information" contained in a mathematical model. It is the natural extension of more classical tools, such as analytic solutions, special transforms, functional analysis, as well as stability and asymptotic analysis.

4.The text book: Numerical Analysis,Seventh(Eighth) Edition



Richard L. Burden

Youngstown State University

J. Douglas Faires

Youngstown State University

For students who are going to study this course

1. The grade of this course includes two parts:

   (1) homework or quiz, 40%(50%)

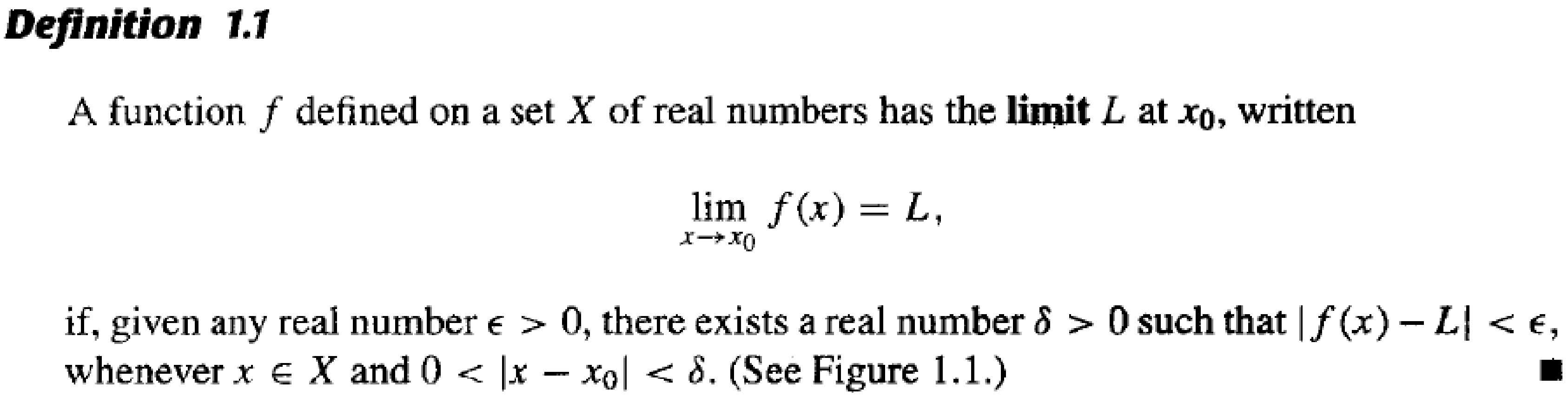
   (2) final examination, 60%(50%)

2. You need to register this course at web of graduate school , USTB. I can mark your studying only when you completed the registration.

Chapter 1 Mathematical Preliminaries

1.1 Review of Calculus

The concepts of *limit* and *continuity* of a function are fundamental to the study of calculus.



Definition 1.1

A function f defined on a set X of real numbers has the **limit** L at x_0, written

$$\lim_{x \to x_0} f(x) = L,$$

if, given any real number $\epsilon > 0$, there exists a real number $\delta > 0$ such that $|f(x) - L| < \epsilon$, whenever $x \in X$ and $0 < |x - x_0| < \delta$. (See Figure 1.1.) ■

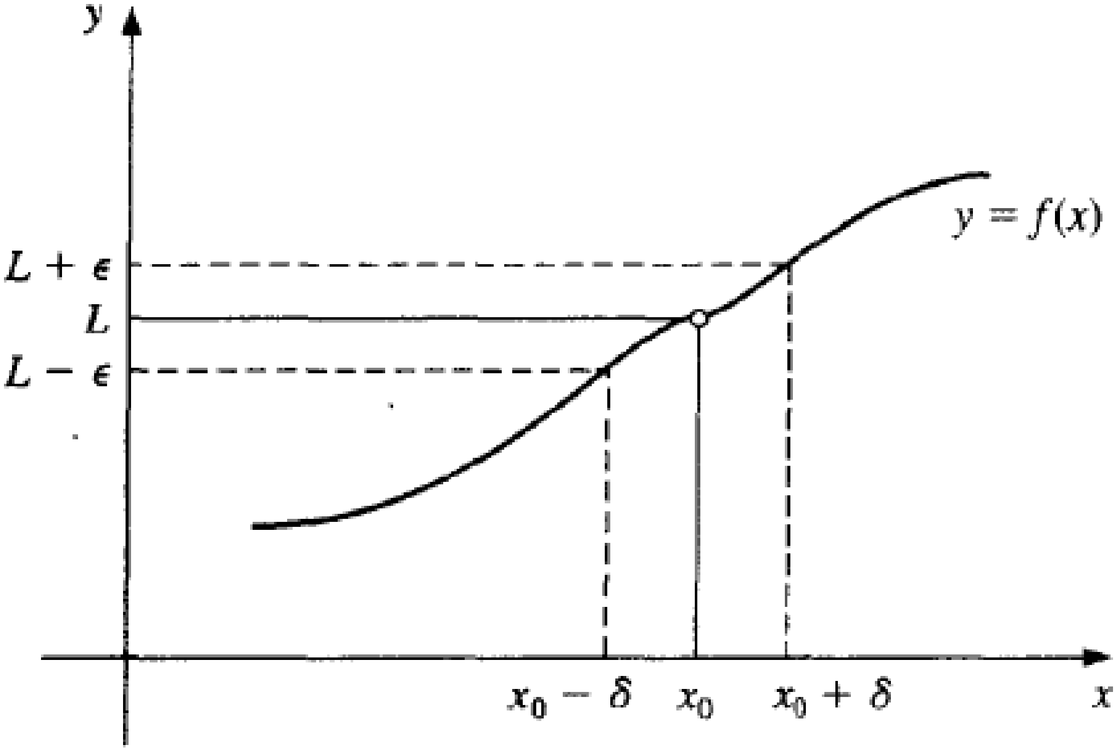


Figure 1.1

Defination 1.2

Let f be a function defined on a set X of real numbers and $x_0 \in X$. Then f is **continuous** at x_0 if

$$\lim_{x \to x_0} f(x) = f(x_0).$$

The function f is continuous on the set X if it is continuous at each number in X.

$C(X)$ denotes the set of all functions that are continuous on X. When X is an interval of the real line, the parentheses in this notation are omitted. For example, the set of all functions continuous on the closed interval $[a, b]$ is denoted $C[a, b]$.

The *limit of a sequence* of real or complex numbers is defined in a similar manner.

Definition 1.3

Let $\{x_n\}_{n=1}^{\infty}$ be an infinite sequence of real or complex numbers. The sequence $\{x_n\}_{n=1}^{\infty}$ has the **limit** x (**converges to** x) if, for any $\epsilon > 0$, there exists a positive integer $N(\epsilon)$ such that $|x_n - x| < \epsilon$, whenever $n > N(\epsilon)$. The notation

$$\lim_{n \to \infty} x_n = x, \quad \text{or} \quad x_n \to x \quad \text{as} \quad n \to \infty,$$

means that the sequence $\{x_n\}_{n=1}^{\infty}$ converges to x. ■

(Intermediate Value Theorem)

If $f \in C[a, b]$ and K is any number between $f(a)$ and $f(b)$, then there exists a number c in (a, b) for which $f(c) = K$. ■

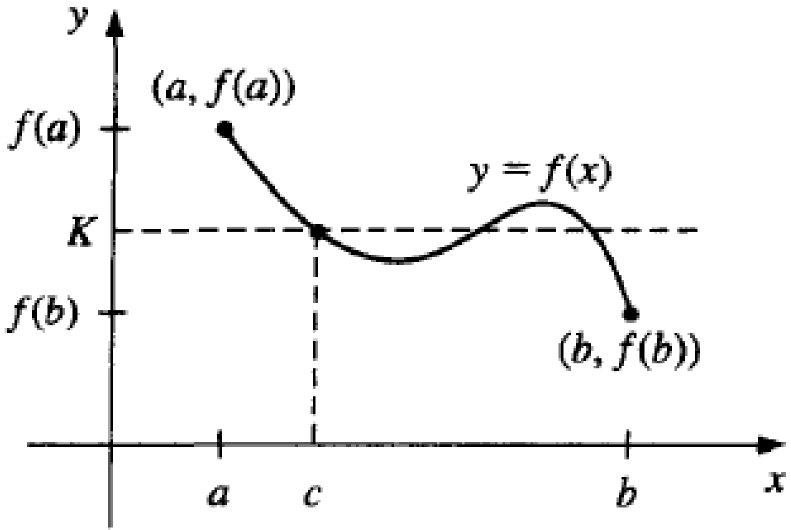


Figure 1.3

To show that $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a solution in the interval $[0, 1]$, consider $f(x) = x^5 - 2x^3 + 3x^2 - 1$. Since

$$f(0) = -1 < 0 < 1 = f(1)$$

and f is continuous, the Intermediate Value Theorem implies that a number x exists with $0 < x < 1$, for which $x^5 - 2x^3 + 3x^2 - 1 = 0$. ■

Theorem 1.4

If f is a function defined on a set X of real numbers and $x_0 \in X$, then the following statements are equivalent:

a. f is continuous at x_0;

b. If $\{x_n\}_{n=1}^{\infty}$ is any sequence in X converging to x_0, then $\lim_{n\to\infty} f(x_n) = f(x_0)$.

Definition 1.5

Let f be a function defined in an open interval containing x_0. The function f is **differentiable** at x_0 if

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists. The number $f'(x_0)$ is called the **derivative** of f at x_0. A function that has a derivative at each number in a set X is **differentiable** on X.

The derivative of f at x_0 is the slope of the tangent line to the graph of f at $(x_0, f(x_0))$, as shown in Figure 1.2. ■

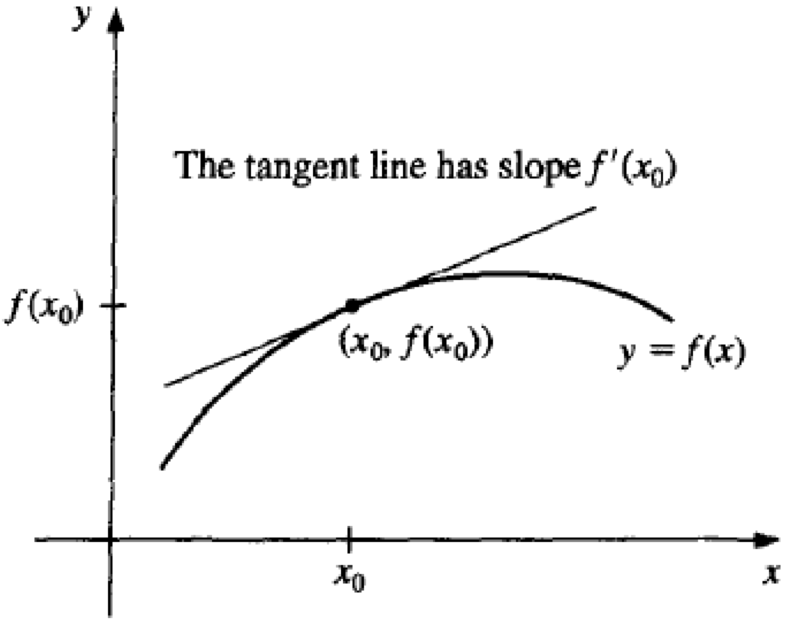


Figure 1.2

Theorem 1.6 If the function f is differentiable at x_0, then f is continuous at x_0.

The set of all functions that have n continuous derivatives on X is denoted $C^n(X)$, and the set of functions that have derivatives of all orders on X is denoted $C^\infty(X)$. Polynomial, rational, trigonometric, exponential, and logarithmic functions are in $C^\infty(X)$, where X consists of all numbers for which the functions are defined. When X is an interval of the real line, we will again omit the parentheses in this notation.

The next theorems are of fundamental importance in deriving methods for error estimation. The proofs of these theorems and the other unreferenced results in this section can be found in any standard calculus text.

Theorem 1.7

(Rolle's Theorem)

Suppose $f \in C[a, b]$ and f is differentiable on (a, b). If $f(a) = f(b)$, then a number c in (a, b) exists with $f'(c) = 0$. (See Figure 1.3.) ■

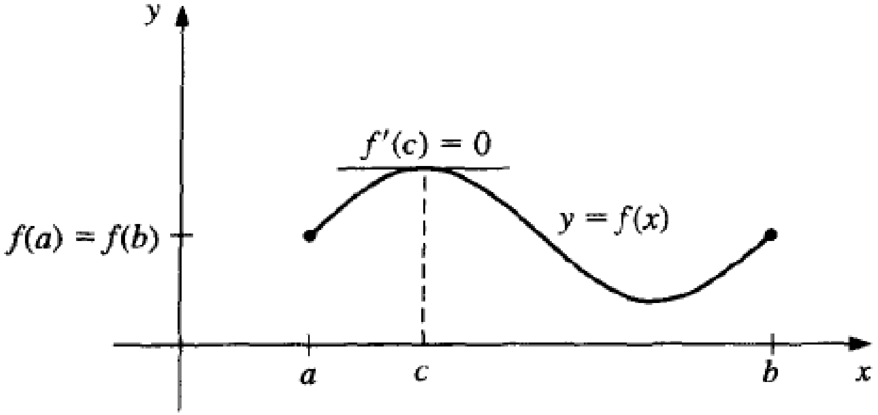


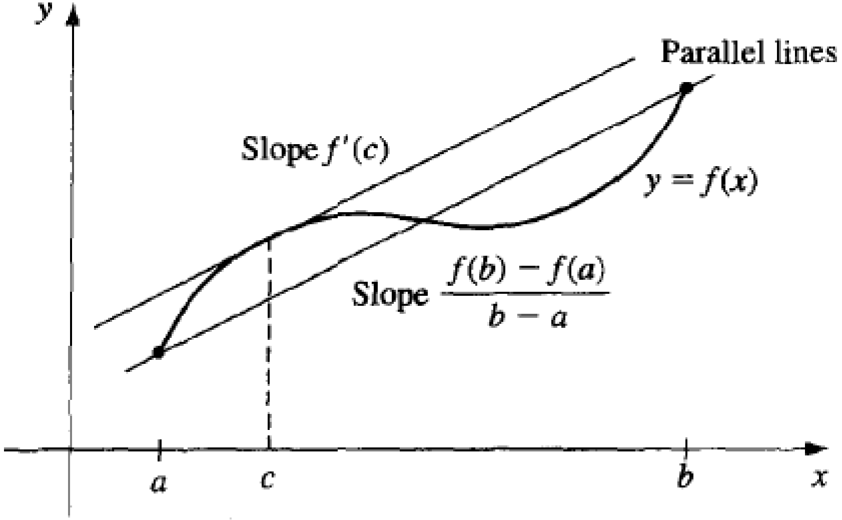
Figure 1.3

Theorem 1.8

(Mean Value Theorem)

If $f \in C[a, b]$ and f is differentiable on (a, b), then a number c in (a, b) exists with

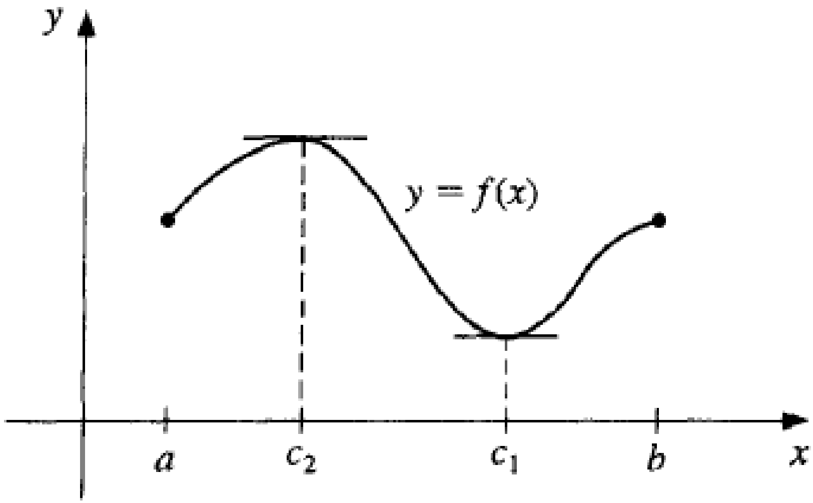
$$f'(c) = \frac{f(b) - f(a)}{b - a}. \quad \text{(See Figure 1.4.)}$$



Theorem 1.9

(Extreme Value Theorem)

If $f \in C[a, b]$, then $c_1, c_2 \in [a, b]$ exist with $f(c_1) \leq f(x) \leq f(c_2)$, for all $x \in [a, b]$. In addition, if f is differentiable on (a, b), then the numbers c_1 and c_2 occur either at the endpoints of $[a, b]$ or where f' is zero. (See Figure 1.5.) ∎



Definition 1.10

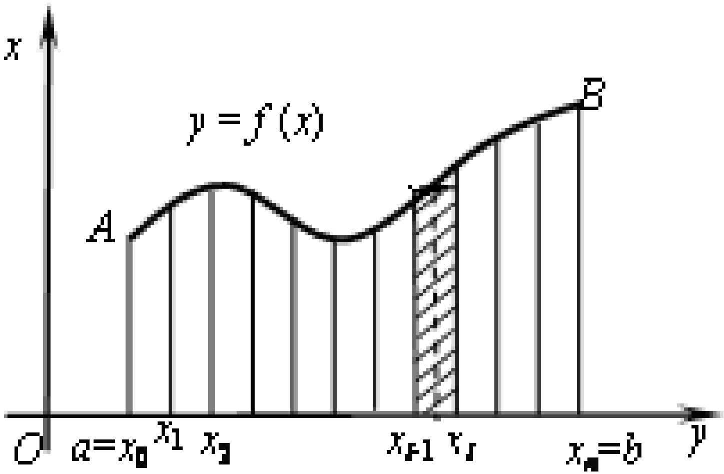
The **Riemann integral** of the function f on the interval $[a, b]$ is the following limit, provided it exists:

$$\int_a^b f(x)\,dx = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(z_i)\,\Delta x_i,$$

where the numbers $x_0, x_1, \ldots, x_n$ satisfy $a = x_0 \le x_1 \le \cdots \le x_n = b$, and where $\Delta x_i = x_i - x_{i-1}$, for each $i = 1, 2, \ldots, n$, and z_i is arbitrarily chosen in the interval $[x_{i-1}, x_i]$. ■

Every continuous function f on $[a, b]$ is Riemann integrable on $[a, b]$. This permits us to choose, for computational convenience, the points x_i to be equally spaced in $[a, b]$, and for each $i = 1, 2, \ldots, n$, to choose $z_i = x_i$. In this case,

$$\int_a^b f(x)\,dx = \lim_{n \to \infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i),$$


x
$y=f(x)$
B
A
O
$a=x_0$
x_1
x_2
x_{i-1}
x_i
$x_n=b$
y

Mean Value Theorem for Integrals

Suppose $f \in C[a,b]$, then there, at least, exists a point , such that

$$\int_a^b f(x)\mathrm{d}x = f(\xi)(b-a), \qquad \xi \in [a,b]$$

Proof: Since $f(x) \in [a, b]$, then there exist a minimum *m* and a maximum *M*,

Such that, $m \le f(x) \le M$, integrating from *a* to *b*, obtain

$$m(b-a) \le \int_a^b f(x)\mathrm{d}x \le M(b-a),$$

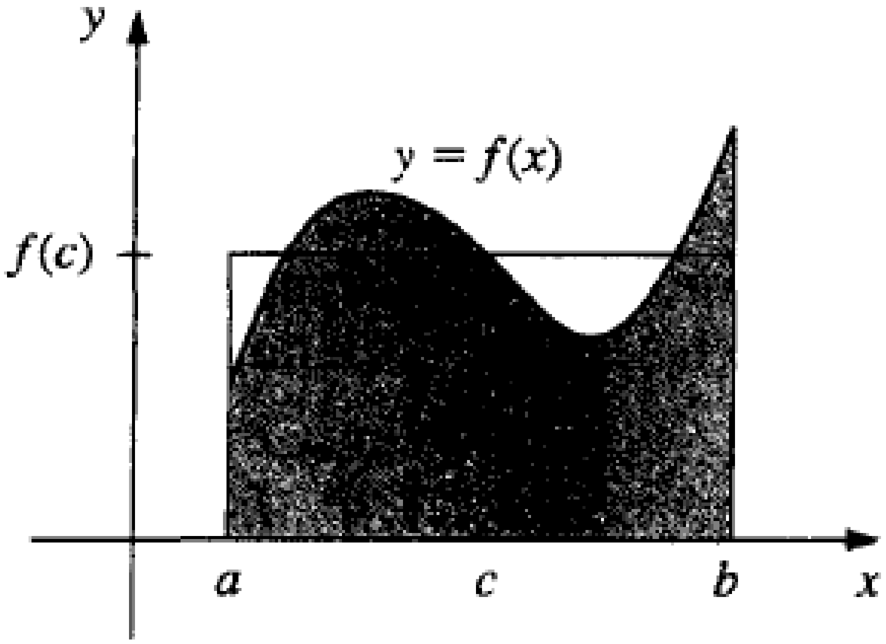
then,

$$m \le \frac{\int_a^b f(x)\mathrm{d}x}{(b-a)} \le M,$$

In view of intermediate theorem, there exists,at least, a $\xi \in [a, b]$, such that

$$f(\xi) = \frac{\int_a^b f(x)\mathrm{d}x}{(b-a)}.$$

$$\int_a^b f(x)\mathrm{d}x = f(\xi)(b-a), \qquad \xi \in [a,b]$$



Theorem 1.11

(Weighted Mean Value Theorem for Integrals)
Suppose $f \in C[a, b]$, the Riemann integral of g exists on $[a, b]$, and $g(x)$ does not change sign on $[a, b]$. Then there exists a number c in (a, b) with

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx. \qquad \blacksquare$$

When $g(x) \equiv 1$, Theorem 1.11 is the usual Mean Value Theorem for Integrals. It gives the **average value** of the function f over the interval $[a, b]$ as

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

(See Figure 1.10.)

Proof: Assume g(x) ≥0 on [a, b], and $f \in C[a,b]$, then there exist a minimum m and a maximum M, such that,

$$mg(x) \le f(x)g(x) \le Mg(x),$$

integrating from a to b, obtain

$$m\int_a^b g(x)\mathrm{d}x \le \int_a^b f(x)g(x)\mathrm{d}x \le M\int_a^b g(x)\mathrm{d}x,$$

If, $\int_a^b g(x)\mathrm{d}x=0$, $\quad then,$ $\int_a^b f(x)g(x)\mathrm{d}x = 0 \;\; for\ any\ x \in [a,b],$

If, $\int_a^b g(x)\mathrm{d}x > 0$, $\quad then,$ $\quad m \le \dfrac{\int_a^b f(x)g(x)\mathrm{d}x}{\int_a^b g(x)\mathrm{d}x} \le M,$

In view of intermediate theorem, there exists, at least, a $\xi \in [a, b]$, such that

$$f(\xi) = \frac{\int_a^b f(x)g(x)\mathrm{d}x}{\int_a^b g(x)\mathrm{d}x}, \text{ i.e.,} \int_a^b f(x)g(x)\mathrm{d}x = f(\xi)\int_a^b g(x)\mathrm{d}x$$

Theorem 1.12

(Generalized Rolle's Theorem)

Suppose $f \in C[a, b]$ is n times differentiable on (a, b). If $f(x)$ is zero at the $n+1$ distinct numbers $x_0, \ldots, x_n$ in $[a, b]$, then a number c in (a, b) exists with $f^{(n)}(c) = 0$. ■

The next theorem is the Intermediate Value Theorem. Although its statement seems reasonable, the proof is beyond the scope of the usual calculus course. It can, however, be found in most analysis texts (see, for example, [Fu, p. 67]).

Theorem 1.14 (Taylor's Theorem)

Suppose $f \in C^n[a, b]$, that $f^{(n+1)}$ exists on $[a, b]$, and $x_0 \in [a, b]$. For every $x \in [a, b]$, there exists a number $\xi(x)$ between x_0 and x with

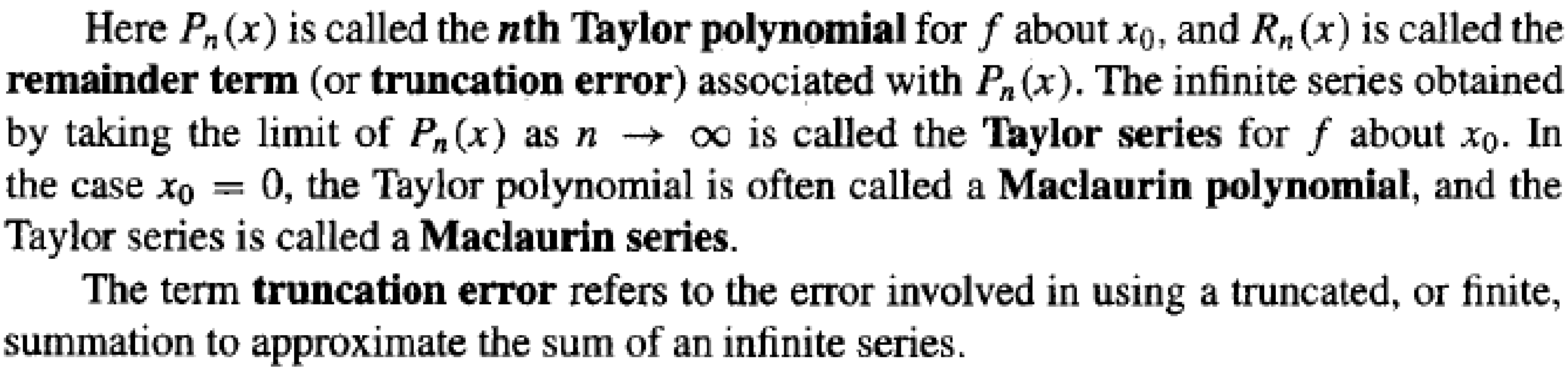
$$f(x) = P_n(x) + R_n(x),$$

where

$$\begin{aligned} P_n(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \end{aligned}$$

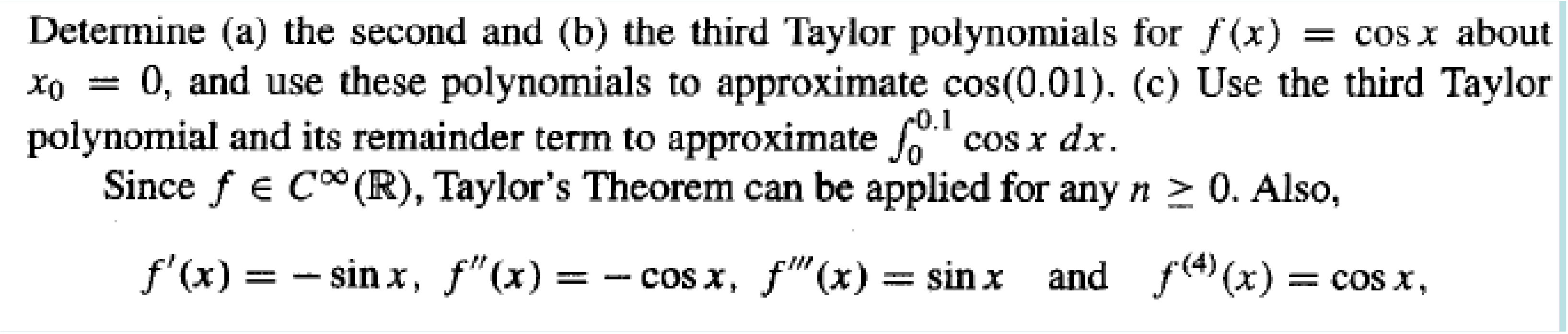
and

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}. \qquad \blacksquare$$



Here $P_n(x)$ is called the **nth Taylor polynomial** for f about x_0, and $R_n(x)$ is called the **remainder term** (or **truncation error**) associated with $P_n(x)$. The infinite series obtained by taking the limit of $P_n(x)$ as $n \to \infty$ is called the **Taylor series** for f about x_0. In the case $x_0 = 0$, the Taylor polynomial is often called a **Maclaurin polynomial**, and the Taylor series is called a **Maclaurin series.**

The term **truncation error** refers to the error involved in using a truncated, or finite, summation to approximate the sum of an infinite series.



Determine (a) the second and (b) the third Taylor polynomials for $f(x) = \cos x$ about $x_0 = 0$, and use these polynomials to approximate cos(0.01). (c) Use the third Taylor polynomial and its remainder term to approximate $\int_0^{0.1} \cos x \, dx$.

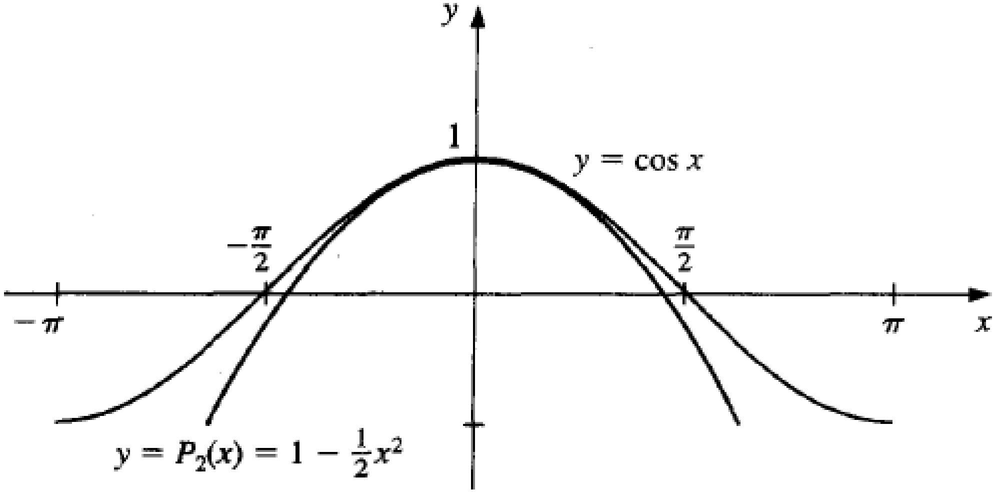
Since $f \in C^\infty(\mathbb{R})$, Taylor's Theorem can be applied for any $n \geq 0$. Also,

$$f'(x) = -\sin x, \quad f''(x) = -\cos x, \quad f'''(x) = \sin x \quad \text{and} \quad f^{(4)}(x) = \cos x,$$

Determine (a) the second and (b) the third Taylor polynomials for $f(x) = \cos x$ about $x_0 = 0$, and use these polynomials to approximate $\cos(0.01)$. (c) Use the third Taylor polynomial and its remainder term to approximate $\int_0^{0.1} \cos x \, dx$.

Since $f \in C^\infty(\mathbb{R})$, Taylor's Theorem can be applied for any $n \geq 0$. Also,

$$f'(x) = -\sin x, \quad f''(x) = -\cos x, \quad f'''(x) = \sin x \quad \text{and} \quad f^{(4)}(x) = \cos x,$$



1.2 Roundoff Errors and Computer Arithmetic

In our traditional mathematical world we permit numbers with an infinite number of digits. The arithmetic we use in this world defines $\sqrt{3}$ as that unique positive number that when multiplied by itself produces the integer 3. In the computational world, however, each representable number has only a fixed, finite number of digits. This means, for example, that only rational numbers—and not even all of these—can be represented exactly. Since $\sqrt{3}$ is not rational, it is given an approximate representation, one whose square will not be precisely 3, although it will likely be sufficiently close to 3 to be acceptable in most situations. In most cases, then, this machine arithmetic is satisfactory and passes without notice or concern, but at times problems arise because of this discrepancy.

Roundoff error is produced when a calculator or computer is used to perform real-number calculations. It occurs because the arithmetic performed in a machine involves numbers with only a finite number of digits, with the result that calculations are performed with only approximate representations of the actual numbers. In a typical computer, only a relatively small subset of the real number system is used for the representation of all the real numbers. This subset contains only rational numbers, both positive and negative, and stores the fractional part, together with an exponential part.

The use of binary digits tends to conceal the computational difficulties that occur when a finite collection of machine numbers is used to represent all the real numbers. To examine these problems, we now assume, for simplicity, that machine numbers are represented in the normalized *decimal* floating-point form

$$\pm 0.d_1 d_2 \ldots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad \text{and} \quad 0 \le d_i \le 9,$$

for each $i = 2, \ldots, k$. Numbers of this form are called k-digit *decimal machine numbers*.

Any positive real number within the numerical range of the machine can be normalized to the form

$$y = 0.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \ldots \times 10^n .$$

The floating-point form of y, denoted $fl(y)$, is obtained by terminating the mantissa of y at k decimal digits. There are two ways of performing this termination. One method, called **chopping**, is to simply chop off the digits $d_{k+1} d_{k+2} \ldots$ to obtain

$$fl(y) = 0.d_1 d_2 \ldots d_k \times 10^n .$$

The other method, called **rounding**, adds $5 \times 10^{n-(k+1)}$ to y and then chops the result to obtain a number of the form

$$fl(y) = 0.\delta_1 \delta_2 \ldots \delta_k \times 10^n .$$

So, when rounding, if $d_{k+1} \geq 5$, we add 1 to d_k to obtain $fl(y)$; that is, we *round up*. When $d_{k+1} < 5$, we merely chop off all but the first k digits; so we *round down*. If we round down, then $\delta_i = d_i$, for each $i = 1, 2, \ldots, k$. However, if we round up, the digits might change.

The number π has an infinite decimal expansion of the form $\pi = 3.14159265\ldots$. Written in normalized decimal form, we have

$$\pi = 0.314159265\ldots \times 10^1.$$

The floating-point form of π using five-digit chopping is

$$fl(\pi) = 0.31415 \times 10^1 = 3.1415.$$

Since the sixth digit of the decimal expansion of π is a 9, the floating-point form of π using five-digit rounding is

$$fl(\pi) = (0.31415 + 0.00001) \times 10^1 = 3.1416. \qquad \blacksquare$$

The error that results from replacing a number with its floating-point form is called **roundoff error** (regardless of whether the rounding or chopping method is used). The following definition describes two methods for measuring approximation errors.

Definition 1.15

If p^* is an approximation to p, the **absolute error** is $|p - p^*|$, and the **relative error** is $\dfrac{|p - p^*|}{|p|}$, provided that $p \neq 0$. ■

Consider the absolute and relative errors in representing p by p^* in the following example.

a. If $p = 0.3000 \times 10^1$ and $p^* = 0.3100 \times 10^1$, the absolute error is 0.1, and the relative error is $0.333\overline{3} \times 10^{-1}$.

b. If $p = 0.3000 \times 10^{-3}$ and $p^* = 0.3100 \times 10^{-3}$, the absolute error is 0.1×10^{-4}, and the relative error is $0.333\overline{3} \times 10^{-1}$.

c. If $p = 0.3000 \times 10^4$ and $p^* = 0.3100 \times 10^4$, the absolute error is 0.1×10^3, and the relative error is again $0.333\overline{3} \times 10^{-1}$.

This example shows that the same relative error, $0.333\overline{3} \times 10^{-1}$, occurs for widely varying absolute errors. As a measure of accuracy, the absolute error can be misleading and the relative error more meaningful since the relative error takes into consideration the size of the value. ■

Definition 1.16

The number p^* is said to approximate p to t **significant digits** (or figures) if t is the largest nonnegative integer for which

$$\frac{|p - p^*|}{|p|} < 5 \times 10^{-t}.$$ ■

Throughout the text we will be examining approximation procedures, called *algorithms*, involving sequences of calculations. An **algorithm** is a procedure that describes, in an unambiguous manner, a finite sequence of steps to be performed in a specified order. The object of the algorithm is to implement a procedure to solve a problem or approximate a solution to the problem.

We use a **pseudocode** to describe the algorithms. This pseudocode specifies the form of the input to be supplied and the form of the desired output. Not all numerical procedures give satisfactory output for arbitrarily chosen input. As a consequence, a stopping technique independent of the numerical technique is incorporated into each algorithm to avoid infinite loops.

The Nth Taylor polynomial for $f(x) = \ln x$ expanded about $x_0 = 1$ is

$$P_N(x) = \sum_{i=1}^{N} \frac{(-1)^{i+1}}{i}(x-1)^i,$$

and the value of ln 1.5 to eight decimal places is 0.40546511. Suppose we want to compute the minimal value of N required for

$$|\ln 1.5 - P_N(1.5)| < 10^{-5}$$

without using the Taylor polynomial remainder term. From calculus we know that if $\sum_{n=1}^{\infty} a_n$ is an alternating series with limit A whose terms decrease in magnitude, then A and the Nth partial sum $A_N = \sum_{n=1}^{N} a_n$ differ by less than the magnitude of the $(N+1)$st term; that is,

$$|A - A_N| \leq |a_{N+1}|.$$

The following algorithm uses this bound.