

Taylor Approximation

Taylor's theorem plays a vital role in numerical analysis. The theorem enables us to approximate a given sufficiently smooth function by a polynomial called the *Taylor polynomial* and provides us with an expression for the error involved in the approximation called the *truncation error* or the *remainder term*.

In Section 1.1, we first define Taylor's polynomial for a given function at a given point and prove Taylor's theorem. In Section 1.2, we define the truncation error in approximating a sufficiently smooth function at a point by Taylor's polynomial and introduce the concept of estimating the truncation error.

1.1 Taylor's Theorem

Let f be a real-valued function defined on an interval I . We say $f \in C^n(I)$ if f is n -times continuously differentiable at every point in I . Also, we say $f \in C^\infty(I)$ if f is continuously differentiable of any order at every point in I .

The most important result used very frequently in numerical analysis, especially in error analysis of numerical methods, is the Taylor's expansion of a function in a neighborhood of a point $a \in \mathbb{R}$. In this section, we define the Taylor's polynomial and prove an important theorem called the *Taylor's theorem*. The idea of the proof of this theorem is similar to the one used in proving the mean value theorem, where we construct a function and apply Rolle's theorem several times to it.

Definition 1.1.1 [Taylor's Polynomial].

Let f be n -times continuously differentiable at a given point a . The **Taylor's polynomial** of degree n for the function f about the point a , denoted by T_n , is defined by

$$T_n(x) = \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k, \quad x \in \mathbb{R}. \quad (1.1)$$

Theorem 1.1.2 [Taylor's Theorem].

Let f be $(n+1)$ -times continuously differentiable function on an open interval containing the points a and x . Then there exists a number ξ between a and x such that

$$f(x) = T_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1}, \quad (1.2)$$

where T_n is the Taylor's polynomial of degree n for f about the point a given by (1.1) and the second term on the right hand side is called the **remainder term**.

Proof.

Let us assume $x > a$ and prove the theorem. The proof is similar if $x < a$.

Define $g(t)$ by

$$g(t) = f(t) - T_n(t) - A(t-a)^{n+1}$$

and choose A so that $g(x) = 0$, which gives

$$A = \frac{f(x) - T_n(x)}{(x-a)^{n+1}}.$$

Note that

$$g^{(k)}(a) = 0 \text{ for } k = 0, 1, \dots, n.$$

Also, observe that the function g is continuous on $[a, x]$ and differentiable in (a, x) .

Apply Rolle's theorem to g on $[a, x]$ (after verifying all the hypotheses of Rolle's theorem) to get

$$a < c_1 < x \text{ satisfying } g'(c_1) = 0.$$

Again apply Rolle's theorem to g' on $[a, c_1]$ to get

$$a < c_2 < c_1 \text{ satisfying } g''(c_2) = 0.$$

In turn apply Rolle's theorem to $g^{(2)}, g^{(3)}, \dots, g^{(n)}$ on intervals $[a, c_2], [a, c_3], \dots, [a, c_n]$, respectively.

At the last step, we get

$$a < c_{n+1} < c_n \text{ satisfying } g^{(n+1)}(c_{n+1}) = 0.$$

But

$$g^{(n+1)}(c_{n+1}) = f^{(n+1)}(c_{n+1}) - A(n+1)!,$$

which gives

$$A = \frac{f^{(n+1)}(c_{n+1})}{(n+1)!}.$$

Equating both values of A , we get

$$f(x) = T_n(x) + \frac{f^{(n+1)}(c_{n+1})}{(n+1)!}(x-a)^{n+1}.$$

This completes the proof.

Observe that the mean value theorem (Theorem A.3.7) is a particular case of the Taylor's theorem.

Remark 1.1.3.

The representation (1.2) is called the *Taylor's formula* for the function f about the point a .

The Taylor's theorem helps us to obtain an approximate value of a sufficiently smooth function in a small neighborhood of a given point a when the value of f and all its derivatives up to a sufficient order is known at the point a . For instance, if we know $f(a), f'(a), \dots, f^{(n)}(a)$, and we seek an approximate value of $f(a+h)$ for some real number h , then the Taylor's theorem can be used to get

$$f(a+h) \approx f(a) + f'(a)h + \frac{f''(a)}{2!}h^2 + \dots + \frac{f^{(n)}(a)}{n!}h^n.$$

Note here that we have not added the remainder term and therefore used the

approximation symbol \approx . Observe that the remainder term

$$R_n(h) = \frac{f^{(n+1)}(\xi)}{(n+1)!} h^{n+1}$$

is not known since it involves the evaluation of $f^{(n+1)}$ at some unknown value ξ lying between a and $a+h$. Also, observe that as $h \rightarrow 0$, the remainder term approaches to zero, provided $f^{(n+1)}$ is bounded. This means that for smaller values of h , the Taylor's polynomial gives a good approximation of $f(a+h)$.

1.2 Truncation Error and its Estimate

The error involved in approximating $f(x)$ by the Taylor's polynomial $T_n(x)$ is given by $f(x) - T_n(x)$ and is called the **Truncation error**. Using (1.2), we can write the truncation error as

$$f(x) - T_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1},$$

for some ξ between x and a .

As remarked above, the truncation error cannot be obtained explicitly because it involves an unknown parameter ξ . However, we can estimate the truncation error under a smoothness condition on f .

Let f be an $(n+1)$ -times continuously differentiable function defined on a closed and bounded interval $I := [\alpha, \beta]$ (i.e., $f \in C^{n+1}(I)$). Then, there exists real number M_{n+1} such that

$$|f^{(n+1)}(\xi)| \leq M_{n+1}, \text{ for all } \xi \in I.$$

Therefore, for any fixed points $a, x \in I$, we can write

$$\left| \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1} \right| \leq \frac{M_{n+1}}{(n+1)!} |x-a|^{n+1}.$$

We can further get an estimate of the remainder term that is independent of x as

$$\left| \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1} \right| \leq \frac{M_{n+1}}{(n+1)!} |\beta-a|^{n+1},$$

which holds for all $x \in I$. Observe that the right hand side of the above estimate is a fixed number. We refer to such estimates as **truncation error estimates** or **remainder estimates**.

Example 1.2.1.

A second degree polynomial approximation to

$$f(x) = \sqrt{x+1}, \quad x \in [-1, \infty)$$

using the Taylor's formula about $a = 0$ is given by

$$f(x) \approx 1 + \frac{x}{2} - \frac{x^2}{8},$$

where the remainder term is neglected and hence what we obtained here is only an approximate representation of f .

The truncation error is obtained using the remainder term in the formula (1.2) with $n = 2$ and is given by

$$f(x) - T_2(x) = \frac{x^3}{16(\sqrt{1+\xi})^5},$$

for some point ξ between 0 and x .

Note that we cannot obtain a remainder estimate in the present example as f''' is not bounded in $[-1, \infty)$. However, for any $0 < \delta < 1$, if we restrict the domain of f to $[-\delta, \infty)$, then we can obtain the remainder estimate for a fixed $x \in [-\delta, \infty)$ as

$$|f(x) - T_2(x)| \leq \frac{|x|^3}{16(\sqrt{1-\delta})^5}.$$

Further, if we restrict the domain of f to $[-\delta, b]$ for some real number $b > 0$, then we get the remainder estimate independent of x as

$$|f(x) - T_2(x)| \leq \frac{b^3}{16(\sqrt{1-\delta})^5}.$$

1.3 Taylor's Series

Since the remainder term in (1.2) represents the error in Taylor's approximation of a given (sufficiently smooth) function, it is very important to write the remainder term in the form that gives the best possible information about the behaviour of the error. This is particularly important in numerical analysis where the interest is to understand how fast the error goes to zero as we improve the approximation by increasing the degree of the polynomial.

Definition 1.3.1 [Taylor's Series].

Let f be a C^∞ -function in a neighborhood of a point a . The power series

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k$$

is called the **Taylor's series** of f about the point a .

The question now is when this series converges and what is the limit of this series. These questions are answered in the following theorem¹ whose proof is omitted for this course.

Theorem 1.3.2.

Let $f \in C^\infty(I)$ and let $a \in I$. Assume that there exists an open neighborhood (interval) $N_a \subset I$ of the point a and there exists a constant M (may depend on a) such that

$$|f^{(k)}(x)| \leq M^k,$$

for all $x \in N_a$ and $k = 0, 1, 2, \dots$. Then for each $x \in I$, we have

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x-a)^k.$$

The following example illustrates the procedure for approximating a given function by its Taylor's polynomial and obtaining the corresponding truncation error.

Example 1.3.3.

As another example, let us approximate the function $f(x) = \cos(x)$ by a polynomial using Taylor's theorem about the point $a = 0$. First, let us take the Taylor's series expansion

$$\begin{aligned} f(x) &= \cos(0) - \sin(0)x - \frac{\cos(0)}{2!}x^2 + \frac{\sin(0)}{3!}x^3 + \dots \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}. \end{aligned}$$

Now, we truncate this infinite series to get an approximate representation of $f(x)$ in

¹Apostol, T. M., Mathematical Analysis, Narosa, 1974.

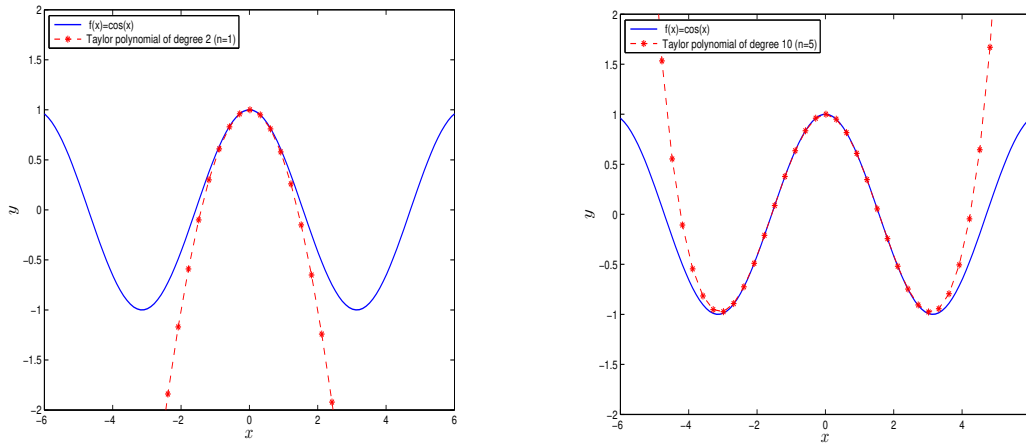


Figure 1.1: Comparison between the graph of $f(x) = \cos(x)$ and the Taylor polynomial of degree 2 and 10 about the point $a = 0$.

a sufficiently small neighborhood of $a = 0$ as

$$f(x) \approx \sum_{k=0}^n \frac{(-1)^k}{(2k)!} x^{2k},$$

which is the Taylor's polynomial of degree $2n$ for the function $f(x) = \cos(x)$ about the point $a = 0$. The truncated series is

$$f(x) - T_{2n}(x) = \sum_{k=n+1}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}.$$

Looking at the leading term of the truncated series, the truncation error can be taken as

$$f(x) - T_{2n}(x) = (-1)^{n+1} \frac{\cos(\xi)}{(2(n+1))!} x^{2(n+1)},$$

where ξ lies between 0 and x . It is important to observe in this particular example that for a given n , we get the Taylor polynomial of degree $2n$.

Figure 1.1 shows the comparison between the Taylor polynomial (red dot and dash line) of degree 2 ($n = 1$) and degree 10 ($n = 5$) for $f(x) = \cos(x)$ about $a = 0$ and the graph of $\cos(x)$ (blue solid line). We observe that for $n = 1$, Taylor polynomial gives a good approximation in a small neighborhood of $a = 0$. But sufficiently away from 0, this polynomial deviates significantly from the actual graph of $f(x) = \cos(x)$. Whereas, for $n = 5$, we get a good approximation in a relatively larger neighborhood of $a = 0$.

1.4 Exercises

1. Find the Taylor's polynomial of degree 2 for the function

$$f(x) = \sqrt{x+1}$$

about the point $a = 1$. Also find the remainder.

2. Use Taylor's polynomial of degree 2 about $a = 0$ to evaluate approximately the value of the function $f(x) = e^x$ at $x = 0.5$. Obtain the remainder $R_2(0.5)$ in terms of the unknown ξ and show that ξ lie in the interval $(0, 0.5)$ by explicitly computing it.
3. Obtain Taylor polynomial of degree 3 and 5 for the function $f(x) = \sin(x)$ about the point $a = 0$. Give the remainder term in both the cases.

Order of Convergence

In Section A.1, we defined convergent sequences and discussed some conditions under which a given sequence of real numbers converges. The definition and the discussed conditions never tell us how fast the sequence converges to the limit. Even if we know that a sequence of approximations converge to the exact one (limit), it is important in numerical analysis to know how fast the sequence of approximate values converge to the exact value. In this section, we introduce two important notations called *big Oh* and *little oh*, which are basic tools for the study of speed of convergence. We end this chapter by defining the notion of order of convergence.

2.1 Big Oh and Little oh Notations

The notions of big Oh and little oh are well understood through the following examples.

Example 2.1.1.

Consider the two sequences $\{n\}$ and $\{n^2\}$ both of which are unbounded and tend to infinity as $n \rightarrow \infty$. However we feel that the sequence $\{n\}$ grows ‘slowly’ compared to the sequence $\{n^2\}$.

Consider also the sequences $\{1/n\}$ and $\{1/n^2\}$ both of which decrease to zero as $n \rightarrow \infty$. However we feel that the sequence $\{1/n^2\}$ decreases more rapidly compared to the sequence $\{1/n\}$.

The above examples motivate us to develop tools that compare two sequences $\{a_n\}$ and $\{b_n\}$. Landau has introduced the concepts of Big Oh and Little oh for comparing two sequences that we will define below.

Definition 2.1.2 [Big Oh and Little oh].

Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers. Then

1. the sequence $\{a_n\}$ is said to be **Big Oh** of $\{b_n\}$, and write $a_n = O(b_n)$, if there exists a real number C and a natural number N such that

$$|a_n| \leq C |b_n|, \quad \text{for all } n \geq N.$$

2. the sequence $\{a_n\}$ is said to be **Little oh** (sometimes said to be **small oh**) of $\{b_n\}$, and write $a_n = o(b_n)$, if for every $\epsilon > 0$ there exists a natural number N such that

$$|a_n| \leq \epsilon |b_n|, \quad \text{for all } n \geq N.$$

Remark 2.1.3.

1. If $b_n \neq 0$ for every n , then we have $a_n = O(b_n)$ if and only if the sequence $\left\{\frac{a_n}{b_n}\right\}$ is bounded. That is, there exists a constant C such that

$$\left|\frac{a_n}{b_n}\right| \leq C.$$

2. If $b_n \neq 0$ for every n , then we have $a_n = o(b_n)$ if and only if the sequence $\left\{\frac{a_n}{b_n}\right\}$ converges to 0. That is,

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0.$$

3. For any pair of sequences $\{a_n\}$ and $\{b_n\}$ such that $a_n = o(b_n)$, it follows that $a_n = O(b_n)$.

The converse is not true. Consider the sequences $a_n = n$ and $b_n = 2n + 3$, for which $a_n = O(b_n)$ holds but $a_n = o(b_n)$ does not hold.

4. Let $\{a_n\}$ and $\{b_n\}$ be two sequences that converge to 0. Then $a_n = O(b_n)$ means the sequence $\{a_n\}$ tends to 0 at least as fast as the sequence $\{b_n\}$; and $a_n = o(b_n)$ means the sequence $\{a_n\}$ tends to 0 faster than the sequence $\{b_n\}$.

5. If the sequence $\{a_n\}$ is bounded, then $a_n = O(b_n)$ where $b_n = 1$ for every n . In this case $a_n = O(1)$.

The Big Oh and Little oh notations may be adapted to functions as follows.

Definition 2.1.4 [Big Oh and Little oh for Functions].

Let $x_0 \in \mathbb{R}$. Let f and g be functions defined in an interval containing x_0 . Then

1. the function f is said to be **Big Oh** of g as $x \rightarrow x_0$, and write $f(x) = O(g(x))$, if there exists a real number C and a real number $\delta > 0$ such that

$$|f(x)| \leq C |g(x)|, \quad \text{whenever } |x - x_0| \leq \delta.$$

2. the function f is said to be **Little Oh** (also, **Small oh**) of g as $x \rightarrow x_0$, and write $f(x) = o(g(x))$, if for every $\epsilon > 0$ there exists a real number δ such that

$$|f(x)| \leq \epsilon |g(x)|, \quad \text{whenever } |x - x_0| \leq \delta.$$

Observe that the remarks similar to the ones discussed in Remark 2.1.3 hold in the case of functions as well.

Example 2.1.5.

The Taylor's formula for $f(x) = \cos(x)$ about the point $a = 0$ is

$$\cos(x) = \sum_{k=0}^n \frac{(-1)^k}{(2k)!} x^{2k} + (-1)^{n+1} \frac{\cos(\xi)}{(2(n+1))!} x^{2(n+1)}$$

where ξ lies between x and 0.

Let us denote the remainder term (truncation error) as (for a fixed n)

$$g(x) = (-1)^{n+1} \frac{\cos(\xi)}{(2(n+1))!} x^{2(n+1)}.$$

Clearly, $g(x) \rightarrow 0$ as $x \rightarrow 0$. The question now is

‘How fast does $g(x) \rightarrow 0$ as $x \rightarrow 0$?’

The answer is

‘As fast as $x^{2(n+1)} \rightarrow 0$ as $x \rightarrow 0$.’

That is,

$$g(x) = O(x^{2(n+1)}) \text{ as } x \rightarrow 0.$$

In this case, we may say that the sequence

$$\left\{ \sum_{k=0}^n \frac{(-1)^k}{(2k)!} x^{2k} \right\}$$

converges to $\cos(x)$ with order $2(n+1)$.

2.2 Order of Convergence

In the oh notations, we compared two sequences (or two functions) to understand how fast a sequence approaches the limit. We can also measure the rapidity of a convergent sequence $\{a_n\}$ in approaching the limit by looking at how the terms of the sequence get closer to the limit as n increases.

Definition 2.2.1 [Order of Convergence].

Let $\{a_n\}$ be a sequence such that $\lim_{n \rightarrow \infty} a_n = a$.

1. We say that the order of convergence is atleast *linear* if there exists a constant $c < 1$ and a natural number N such that

$$|a_{n+1} - a| \leq c |a_n - a| \quad \text{for all } n \geq N.$$

2. We say that the order of convergence is atleast *superlinear* if there exists a sequence $\{\epsilon_n\}$ that converges to 0, and a natural number N such that

$$|a_{n+1} - a| \leq \epsilon_n |a_n - a| \quad \text{for all } n \geq N.$$

3. We say that the order of convergence is atleast *quadratic* if there exists a constant C (not necessarily less than 1), and a natural number N such that

$$|a_{n+1} - a| \leq C |a_n - a|^2 \quad \text{for all } n \geq N.$$

4. Let $\alpha \in \mathbb{R}_+$. We say that the order of convergence is atleast α if there exists a constant C (not necessarily less than 1), and a natural number N such that

$$|a_{n+1} - a| \leq C |a_n - a|^\alpha \quad \text{for all } n \geq N.$$

Remark 2.2.2.

Another equivalent way to define the order of convergence of a convergent sequence is as follows:

We say that a sequence $\{a_n\}$ converges to a limit a with order of convergence $\alpha > 0$ if there exists a constant $\lambda > 0$ such that

$$\lim_{n \rightarrow \infty} \frac{|a_{n+1} - a|}{|a_n - a|^\alpha} = \lambda.$$

The number λ is called the *asymptotic error constant*.

2.3 Exercises

1. Prove or disprove:

- i) $2n^2 + 3n + 4 = o(n)$ as $n \rightarrow \infty$.
- ii) $\frac{n+1}{n^2} = o(\frac{1}{n})$ as $n \rightarrow \infty$.
- iii) $\frac{n+1}{n^2} = O(\frac{1}{n})$ as $n \rightarrow \infty$.
- iv) $\frac{n+1}{\sqrt{n}} = o(1)$ as $n \rightarrow \infty$.
- v) $\frac{1}{\ln n} = o(\frac{1}{n})$ as $n \rightarrow \infty$.
- vi) $\frac{1}{n \ln n} = o(\frac{1}{n})$ as $n \rightarrow \infty$.
- vii) $\frac{e^n}{n^5} = O(\frac{1}{n})$ as $n \rightarrow \infty$.

2. Prove or disprove:

- i) $e^x - 1 = O(x^2)$ as $x \rightarrow 0$.
- ii) $x^{-2} = O(\cot x)$ as $x \rightarrow 0$.
- iii) $\cot x = o(x^{-1})$ as $x \rightarrow 0$.
- iv) For $r > 0$, $x^r = O(e^x)$ as $x \rightarrow \infty$.
- v) For $r > 0$, $\ln x = O(x^r)$ as $x \rightarrow \infty$.

Mathematical Preliminaries

This chapter reviews some of the concepts and results from calculus that are frequently used in this course. We recall important definitions and theorems, and outline proofs of certain theorems. The readers are assumed to be familiar with a first course in calculus.

In **Section A.1**, we introduce sequences of real numbers, recall the definition of convergence of a sequence, and state the sandwich theorem. Further, we discuss the concept of limit and continuity of a real function in **Section A.2**. We state the sandwich theorem for continuous functions and the intermediate value theorem. These theorems play vital role in convergence analysis and finding initial guesses to some iterative methods for solving nonlinear equations. In **Section A.3** we define the notion of derivative of a function, and prove Rolle's theorem and mean-value theorem for derivatives. The mean-value theorem for integration is discussed in **Section A.4**. These two theorems are crucially used in devising methods for numerical integration and differentiation.

A.1 Sequences of Real Numbers

Let $a, b \in \mathbb{R}$ be such that $a < b$. We use the notations $[a, b]$ and (a, b) for the closed and the open intervals, respectively, and are defined by

$$[a, b] = \{ x \in \mathbb{R} : a \leq x \leq b \}$$

and

$$(a, b) = \{ x \in \mathbb{R} : a < x < b \}.$$

Definition A.1.1 [Sequence].

A *sequence* of real numbers is an ordered list of real numbers

$$a_1, a_2, \dots, a_n, a_{n+1}, \dots$$

In other words, a *sequence* is a function that associates the real number a_n for each natural number n . The notation $\{a_n\}$ is used to denote the sequence, *i.e.*

$$\{a_n\} := a_1, a_2, \dots, a_n, a_{n+1}, \dots$$

Note

We may also use the notation $\{a_n\}_{n=1}^{\infty}$ to denote a sequence. This notation is particularly useful when we define a sequence where the index starts from some other integer, say, with $n = 0$ instead of $n = 1$. If we simply use the notation $\{a_n\}$, it means $n = 1, 2, \dots$

The concept of convergence of a sequence plays an important role in the numerical analysis. For instance, an iterative procedure is often constructed to approximate a solution x of an equation. This procedure produces a sequence of approximations, which is expected to converge to the exact solution x .

Definition A.1.2 [Convergence of a Sequence].

Let $\{a_n\}$ be a sequence of real numbers and let a be a real number. The sequence $\{a_n\}$ is said to *converge* to a , written as

$$\lim_{n \rightarrow \infty} a_n = a \quad (\text{or } a_n \rightarrow a \text{ as } n \rightarrow \infty),$$

if for every $\epsilon > 0$, there exists a natural number N such that

$$|a_n - a| < \epsilon \quad \text{whenever } n \geq N.$$

The real number a is called the *limit* of the sequence $\{a_n\}$.

The following result is very useful in computing the limit of a sequence sandwiched between two sequences having a common limit.

Theorem A.1.3 [Sandwich Theorem].

Let $\{a_n\}, \{b_n\}, \{c_n\}$ be sequences of real numbers such that

1. there exists an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, the sequences satisfy the inequalities $a_n \leq b_n \leq c_n$ and
2. $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} c_n = a$, for some real number a .

Then the sequence $\{b_n\}$ also converges and $\lim_{n \rightarrow \infty} b_n = a$.

Definition A.1.4 [Bounded Sequence].

A sequence $\{a_n\}$ is said to be a **bounded sequence** if there exists a real number M such that

$$|a_n| \leq M \quad \text{for every } n \in \mathbb{N}.$$

Note that a bounded sequence, in general, need not converge, for instance $\{(-1)^n\}$. However, it is always possible to find a convergent subsequence.

Theorem A.1.5 [Bolzano-Weierstrass theorem].

Every bounded sequence $\{a_n\}$ has a convergent subsequence $\{a_{n_k}\}$.

Definition A.1.6 [Monotonic Sequences].

A sequence $\{a_n\}$ of real numbers is said to be

1. an **increasing sequence** if $a_n \leq a_{n+1}$, for every $n \in \mathbb{N}$.
2. a **strictly increasing sequence** if $a_n < a_{n+1}$, for every $n \in \mathbb{N}$.
3. a **decreasing sequence** if $a_n \geq a_{n+1}$, for every $n \in \mathbb{N}$.
4. a **strictly decreasing sequence** if $a_n > a_{n+1}$, for every $n \in \mathbb{N}$.

A sequence $\{a_n\}$ is said to be a (strictly) **monotonic sequence** if it is either (strictly) increasing or (strictly) decreasing.

The following theorem shows the advantage of working with monotonic sequences.

Theorem A.1.7.

Bounded monotonic sequences always converge.

The following result gives the algebra of limits of sequences.

Theorem A.1.8.

Let $\{a_n\}$ and $\{b_n\}$ be two sequences. Assume that $\lim_{n \rightarrow \infty} a_n$ and $\lim_{n \rightarrow \infty} b_n$ exist. Then

1. $\lim_{n \rightarrow \infty} (a_n + b_n) = \lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n$.
2. $\lim_{n \rightarrow \infty} c a_n = c \lim_{n \rightarrow \infty} a_n$, for any number c .
3. $\lim_{n \rightarrow \infty} a_n b_n = \lim_{n \rightarrow \infty} a_n \lim_{n \rightarrow \infty} b_n$.
4. $\lim_{n \rightarrow \infty} \frac{1}{a_n} = \frac{1}{\lim_{n \rightarrow \infty} a_n}$, provided $\lim_{n \rightarrow \infty} a_n \neq 0$.

A.2 Limits and Continuity

In the previous section, we introduced the concept of limit for a sequences of real numbers. We now define the “limit” in the context of functions.

Definition A.2.1 [Limit of a Function].

1. Let f be a function defined on the left side (or both sides) of a , except possibly at a itself. Then, we say “the **left-hand limit** of $f(x)$ as x approaches a , equals l ” and denote

$$\lim_{x \rightarrow a^-} f(x) = l,$$

if we can make the values of $f(x)$ arbitrarily close to l (as close to l as we like) by taking x to be sufficiently close to a and x less than a .

2. Let f be a function defined on the right side (or both sides) of a , except possibly at a itself. Then, we say “the **right-hand limit** of $f(x)$ as x approaches a , equals r ” and denote

$$\lim_{x \rightarrow a^+} f(x) = r,$$

if we can make the values of $f(x)$ arbitrarily close to r (as close to r as we like) by taking x to be sufficiently close to a and x greater than a .

3. Let f be a function defined on both sides of a , except possibly at a itself. Then,

we say “the **limit** of $f(x)$ as x approaches a , equals L ” and denote

$$\lim_{x \rightarrow a} f(x) = L,$$

if we can make the values of $f(x)$ arbitrarily close to L (as close to L as we like) by taking x to be sufficiently close to a (on either side of a) but not equal to a .

Remark A.2.2.

Note that in each of the above definitions the value of the function f at the point a does not play any role. In fact, the function f need not be defined at a .

In the previous section, we have seen some limit laws in the context of sequences. Similar limit laws also hold for limits of functions. We have the following result, often referred to as “the limit laws” or as “algebra of limits”.

Theorem A.2.3.

Let f, g be two functions defined on both sides of a , except possibly at a itself. Assume that $\lim_{x \rightarrow a} f(x)$ and $\lim_{x \rightarrow a} g(x)$ exist. Then

1. $\lim_{x \rightarrow a} (f(x) + g(x)) = \lim_{x \rightarrow a} f(x) + \lim_{x \rightarrow a} g(x).$
2. $\lim_{x \rightarrow a} c f(x) = c \lim_{x \rightarrow a} f(x),$ for any number $c.$
3. $\lim_{x \rightarrow a} f(x)g(x) = \lim_{x \rightarrow a} f(x) \lim_{x \rightarrow a} g(x).$
4. $\lim_{x \rightarrow a} \frac{1}{g(x)} = \frac{1}{\lim_{x \rightarrow a} g(x)},$ provided $\lim_{x \rightarrow a} g(x) \neq 0.$

Remark A.2.4.

Polynomials, rational functions, all trigonometric functions wherever they are defined, have property called direct substitution property:

$$\lim_{x \rightarrow a} f(x) = f(a).$$

The following theorem is often used to estimate limits of functions.

Theorem A.2.5.

If $f(x) \leq g(x)$ when x is in an interval containing a (except possibly at a) and the limits of f and g both exist as x approaches a , then

$$\lim_{x \rightarrow a} f(x) \leq \lim_{x \rightarrow a} g(x).$$

The following theorem is often used to compute limits of functions.

Theorem A.2.6 [Sandwich Theorem].

Let f , g , and h be given functions such that

1. $f(x) \leq g(x) \leq h(x)$ when x is in an interval containing a (except possibly at a) and
2. $\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} h(x) = L$,

then

$$\lim_{x \rightarrow a} g(x) = L.$$

We will now give a rigorous definition of the limit of a function. Similar definitions can be written down for left-hand and right-hand limits of functions.

Definition A.2.7.

Let f be a function defined on some open interval that contains a , except possibly at a itself. Then we say that the *limit* of $f(x)$ as x approaches a is L and we write

$$\lim_{x \rightarrow a} f(x) = L.$$

if for every $\epsilon > 0$ there is a number $\delta > 0$ such that

$$|f(x) - L| < \epsilon \text{ whenever } 0 < |x - a| < \delta.$$

Definition A.2.8 [Continuity].

A function f is

1. *continuous from the right* at a if

$$\lim_{x \rightarrow a+} f(x) = f(a).$$

2. *continuous from the left* at a if

$$\lim_{x \rightarrow a^-} f(x) = f(a).$$

3. *continuous* at a if

$$\lim_{x \rightarrow a} f(x) = f(a).$$

A function f is said to be *continuous on an open interval* if f is continuous at every point in the interval. If f is defined on a closed interval $[a, b]$, then f is said to be continuous at a if f is continuous from the right at a and similarly, f is said to be continuous at b if f is continuous from left at b .

Remark A.2.9.

Note that the definition for continuity of a function f at a , means the following three conditions are satisfied:

1. The function f must be defined at a . i.e., a is in the domain of f ,
2. $\lim_{x \rightarrow a} f(x)$ exists, and
3. $\lim_{x \rightarrow a} f(x) = f(a)$.

Equivalently, for any given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x) - f(a)| < \epsilon \text{ whenever } |x - a| < \delta.$$

Theorem A.2.10.

If f and g are continuous at a , then the functions $f + g$, $f - g$, cg (c is a constant), fg , f/g (provided $g(a) \neq 0$), $f \circ g$ (composition of f and g , whenever it makes sense) are all continuous.

Thus polynomials, rational functions, trigonometric functions are all continuous on their respective domains.

Theorem A.2.11 [Intermediate Value Theorem].

Suppose that f is continuous on the closed interval $[a, b]$ and let N be any number between $f(a)$ and $f(b)$, where $f(a) \neq f(b)$. Then there exists a point $c \in (a, b)$ such that

$$f(c) = N.$$

A.3 Differentiation

We next give the basic definition of a differentiable function.

Definition A.3.1 [Derivative].

The **derivative** of a function f at a point a , denoted by $f'(a)$, is defined as

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}, \quad (1.1)$$

if this limit exists. We say f is **differentiable** at a . A function f is said to be **differentiable** on (c, d) if f is differentiable at every point in (c, d) .

There are alternate ways of defining the derivative of a function, which leads to different difference formulae.

Remark A.3.2.

The derivative of a function f at a point $x = a$ can also be given by

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a) - f(a-h)}{h}, \quad (1.2)$$

and

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a-h)}{2h}, \quad (1.3)$$

provided the limits exist.

If we write $x = a + h$, then $h = x - a$ and $h \rightarrow 0$ if and only if $x \rightarrow a$. Thus, formula (1.1) can equivalently be written as

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}.$$

Remark A.3.3 [Interpretation].

Take the graph of f , draw the line joining the points $(a, f(a))$ and $(x, f(x))$. Take its slope and take the limit of these slopes as $x \rightarrow a$. Then the point $(x, f(x))$ tends to $(a, f(a))$. The limit is nothing but the slope of the tangent line at $(a, f(a))$ to the curve $y = f(x)$. This geometric interpretation will be very useful in describing the Newton-Raphson method in the context of solving nonlinear equations.

Theorem A.3.4.

If f is differentiable at a , then f is continuous at a .

Proof.

Using the identities

$$f(x) = \frac{f(x) - f(a)}{x - a}(x - a) + f(a)$$

and taking limit as $x \rightarrow a$ yields the desired result.

The converse of Theorem A.3.4 is not true. For, the function $f(x) = |x|$ is continuous at $x = 0$ but is not differentiable there.

Theorem A.3.5.

Suppose f is differentiable at a . Then there exists a function ϕ such that

$$f(x) = f(a) + (x - a)f'(a) + (x - a)\phi(x),$$

and $\lim_{x \rightarrow a} \phi(x) = 0$.

Proof.

Define ϕ by

$$\phi(x) = \frac{f(x) - f(a)}{x - a} - f'(a).$$

Since f is differentiable at a , the result follows on taking limits on both sides of the last equation as $x \rightarrow a$.

Theorem A.3.6 [Rolle's Theorem].

Let f be a function that satisfies the following three hypotheses:

1. f is continuous on the closed interval $[a, b]$.
2. f is differentiable on the open interval (a, b) .
3. $f(a) = f(b)$.

Then there is a number c in the open interval (a, b) such that $f'(c) = 0$.

Proof.

If f is a constant function *i.e.*, $f(x) = f(a)$ for every $x \in [a, b]$, clearly such a c exists. If f is not a constant, then at least one of the following holds.

Case 1: The graph of f goes above the line $y = f(a)$ *i.e.*, $f(x) > f(a)$ for some $x \in (a, b)$.

Case 2: The graph of f goes below the line $y = f(a)$ *i.e.*, $f(x) < f(a)$ for some $x \in (a, b)$.

In case (1), *i.e.*, if the graph of f goes above the line $y = f(a)$, then the global maximum cannot be at a or b . Therefore, it must lie in the open interval (a, b) . Denote that point by c . That is, global maximum on $[a, b]$ is actually a local maximum, and hence $f'(c) = 0$. A similar argument can be given in case (2), to show the existence of local minimum in (a, b) . This completes the proof of Rolle's theorem.

The mean value theorem plays an important role in obtaining error estimates for certain numerical methods.

Theorem A.3.7 [Mean Value Theorem].

Let f be a function that satisfies the following hypotheses:

1. f is continuous on the closed interval $[a, b]$.
2. f is differentiable on the open interval (a, b) .

Then there is a number c in the open interval (a, b) such that

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

or, equivalently,

$$f(b) - f(a) = f'(c)(b - a).$$

Proof.

Define ϕ on $[a, b]$ by

$$\phi(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

By Rolle's theorem there exists a $c \in (a, b)$ such that $\phi'(c) = 0$ and hence the proof is complete.

A.4 Integration

In Theorem A.3.7, we have discussed the mean value property for the derivative of a function. We now discuss the mean value theorems for integration.

Theorem A.4.1 [Mean Value Theorem for Integrals].

If f is continuous on $[a, b]$, then there exists a number c in $[a, b]$ such that

$$\int_a^b f(x) dx = f(c)(b - a).$$

Proof.

Let m and M be minimum and maximum values of f in the interval $[a, b]$, respectively. Then,

$$m(b - a) \leq \int_a^b f(x) dx \leq M(b - a).$$

Since f is continuous, the result follows from the intermediate value theorem.

Recall the average value of a function f on the interval $[a, b]$ is defined by

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

Observe that the first mean value theorem for integrals asserts that the average of an integrable function f on an interval $[a, b]$ belongs to the range of the function f .

Remark A.4.2 [Interpretation].

Let f be a function on $[a, b]$ with $f > 0$. Draw the graph of f and find the area under the graph lying between the ordinates $x = a$ and $x = b$. Also, look at a rectangle with base as the interval $[a, b]$ with height $f(c)$ and compute its area. Both values are the same.

The Theorem A.4.1 is often referred to as the *first mean value theorem for integrals*. We now state the second mean value theorem for integrals, which is a general form of Theorem A.4.1.

Theorem A.4.3 [Second Mean Value Theorem for Integrals].

Let f and g be continuous on $[a, b]$, and let $g(x) \geq 0$ for all $x \in \mathbb{R}$. Then there exists a number $c \in [a, b]$ such that

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

The proof of this theorem is left as an exercise.

A.5 Exercises

Sequences of Real Numbers

1. Let L be a real number and let $\{a_n\}$ be a sequence of real numbers. If there exists a positive integer N such that

$$|a_n - L| \leq \mu |a_{n-1} - L|,$$

for all $n \geq N$ and for some fixed $\mu \in (0, 1)$, then show that $a_n \rightarrow L$ as $n \rightarrow \infty$.

2. Consider the sequences $\{a_n\}$ and $\{b_n\}$, where

$$a_n = \frac{1}{n}, \quad b_n = \frac{1}{n^2}, \quad n = 1, 2, \dots$$

Clearly, both the sequences converge to zero. For the given $\epsilon = 10^{-2}$, obtain the smallest positive integers N_a and N_b such that

$$|a_n| < \epsilon \text{ whenever } n \geq N_a, \text{ and } |b_n| < \epsilon \text{ whenever } n \geq N_b.$$

For any $\epsilon > 0$, show that $N_a > N_b$.

3. Let $\{x_n\}$ and $\{y_n\}$ be two sequences such that $x_n, y_n \in [a, b]$ and $x_n < y_n$ for each $n = 1, 2, \dots$. If $x_n \rightarrow b$ as $n \rightarrow \infty$, then show that the sequence $\{y_n\}$ converges. Find the limit of the sequence $\{y_n\}$.
4. Let $I_n = \left[\frac{n-2}{2n}, \frac{n+2}{2n} \right]$, $n = 1, 2, \dots$ and $\{a_n\}$ be a sequence with a_n is chosen arbitrarily in I_n for each $n = 1, 2, \dots$. Show that $a_n \rightarrow \frac{1}{2}$ as $n \rightarrow \infty$.

Limits and Continuity

5. Let f be a real-valued function such that $f(x) \geq \sin(x)$ for all $x \in \mathbb{R}$. If $\lim_{x \rightarrow 0} f(x) = L$ exists, then show that $L \geq 0$.

6. Let P and Q be polynomials. Find

$$\lim_{x \rightarrow \infty} \frac{P(x)}{Q(x)} \quad \text{and} \quad \lim_{x \rightarrow 0} \frac{P(x)}{Q(x)}$$

in each of the following cases.

- i) The degree of P is less than the degree of Q .
- ii) The degree of P is greater than the degree of Q .
- iii) The degree of P is equal to the degree of Q .

7. Show that the equation $\sin x + x^2 = 1$ has at least one solution in the interval $[0, 1]$.

8. Let $f(x)$ be continuous on $[a, b]$, let x_1, \dots, x_n be points in $[a, b]$, and let g_1, \dots, g_n be real numbers having same sign. Show that

$$\sum_{i=1}^n f(x_i)g_i = f(\xi) \sum_{i=1}^n g_i, \quad \text{for some } \xi \in [a, b].$$

9. Let $f : [0, 1] \rightarrow [0, 1]$ be a continuous function. Prove that the equation $f(x) = x$ has at least one solution lying in the interval $[0, 1]$ (Note: A solution of this equation is called a *fixed point* of the function f).

10. Show that the equation $f(x) = x$, where

$$f(x) = \sin\left(\frac{\pi x + 1}{2}\right), \quad x \in [-1, 1]$$

has at least one solution in $[-1, 1]$.

Differentiation

11. Let $c \in (a, b)$ and $f : (a, b) \rightarrow \mathbb{R}$ be differentiable at c . If c is a local extremum (maximum or minimum) of f , then show that $f'(c) = 0$.

12. Suppose f is differentiable in an open interval (a, b) . Prove the following statements

- (a) If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is non-decreasing.
- (b) If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant.
- (c) If $f'(x) \leq 0$ for all $x \in (a, b)$, then f is non-increasing.

13. Let $f : [a, b] \rightarrow \mathbb{R}$ be given by $f(x) = x^2$. Find a point c specified by the mean-value theorem for derivatives. Verify that this point lies in the interval (a, b) .

Integration

14. Let $g : [0, 1] \rightarrow \mathbb{R}$ be a continuous function. Show that there exists a $c \in (0, 1)$ such that

$$\int_0^1 x^2(1-x)^2 g(x) dx = \frac{1}{30} g(\xi).$$

15. If n is a positive integer, show that

$$\int_{\sqrt{n\pi}}^{\sqrt{(n+1)\pi}} \sin(t^2) dt = \frac{(-1)^n}{c},$$

where $\sqrt{n\pi} \leq c \leq \sqrt{(n+1)\pi}$.

Arithmetic Error Analysis

Numerical analysis deals with developing methods, called *numerical methods*, to approximate a solution of a given Mathematical problem (whenever a solution exists). The approximate solution obtained by a method involves an error, which we call the *mathematical error*, and is precisely the difference between the exact solution and the approximate solution. Thus, we have

$$\text{Exact Solution} = \text{Approximate Solution} + \text{Mathematical Error}.$$

The study of numerical methods is incomplete if we don't develop algorithms and implement the algorithms as computer codes. The outcome of the computer code is a set of numerical values to the approximate solution obtained using a numerical method. Such a set of numerical values is called the *numerical solution* to the given Mathematical problem. During the process of computation, the computer introduces a new error, called the *arithmetic error*, and we have

$$\text{Approximate Solution} = \text{Numerical Solution} + \text{Arithmetic Error}.$$

The error involved in the numerical solution when compared to the exact solution can be worse than the mathematical error and is now given by

$$\text{Exact Solution} = \text{Numerical Solution} + \text{Mathematical Error} + \text{Arithmetic Error}.$$

The *Total Error* is defined as

$$\text{Total Error} = \text{Mathematical Error} + \text{Arithmetic Error}.$$

A digital calculating device can hold only a finite number of digits because of memory restrictions. Therefore, a number cannot be stored exactly. Certain approximation needs to be done, and only an approximate value of the given number will finally be stored in the device. For further calculations, this approximate value is used instead of the exact value of the number. This is the source of arithmetic error.

In this chapter, we introduce the floating-point representation of a real number and illustrate a few ways to obtain floating-point approximation of a given real number. We further 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.

3.1 Floating-Point Representation

Let $\beta \in \mathbb{N}$ and $\beta \geq 2$. Any real number can be represented exactly in **base** β as

$$(-1)^s \times (.d_1 d_2 \cdots d_n d_{n+1} \cdots)_\beta \times \beta^e, \quad (3.1)$$

where $d_i \in \{0, 1, \dots, \beta - 1\}$ with $d_1 \neq 0$ or $d_1 = d_2 = d_3 = \cdots = 0$, $s = 0$ or 1 , and an appropriate integer e called the **exponent**. Here

$$(.d_1 d_2 \cdots d_n d_{n+1} \cdots)_\beta = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_n}{\beta^n} + \frac{d_{n+1}}{\beta^{n+1}} + \cdots \quad (3.2)$$

is a β -fraction called the **mantissa**, s is called the **sign** and the number β is called the **radix**. The representation (3.1) of a real number is called the **floating-point representation**.

Remark 3.1.1.

When $\beta = 2$, the floating-point representation (3.1) is called the **binary floating-point representation** and when $\beta = 10$, it is called the **decimal floating-point representation**.

Note

Throughout this course, we always take $\beta = 10$.

3.1.1 Floating-Point Approximation

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.

Although different computing devices have different ways of representing numbers, here we introduce a mathematical form of this representation, which we will use throughout this course.

Definition 3.1.2 [*n*-Digit Floating-point Number].

Let $\beta \in \mathbb{N}$ and $\beta \geq 2$. An *n-digit floating-point number* in *base* β is of the form

$$(-1)^s \times (.d_1 d_2 \cdots d_n)_\beta \times \beta^e \quad (3.3)$$

where

$$(.d_1 d_2 \cdots d_n)_\beta = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_n}{\beta^n} \quad (3.4)$$

where $d_i \in \{0, 1, \dots, \beta - 1\}$ with $d_1 \neq 0$ or $d_1 = d_2 = d_3 = \cdots = 0$, $s = 0$ or 1 , and an appropriate exponent e .

Remark 3.1.3.

When $\beta = 2$, the *n*-digit floating-point representation (3.3) is called the *n-digit binary floating-point representation* and when $\beta = 10$, it is called the *n-digit decimal floating-point representation*.

Example 3.1.4.

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

1. The real number $x = 6.238$ is represented in the decimal floating-point representation as

$$6.238 = (-1)^0 \times 0.6238 \times 10^1,$$

in which case, we have $s = 0$, $\beta = 10$, $e = 1$, $d_1 = 6$, $d_2 = 2$, $d_3 = 3$ and $d_4 = 8$.

2. The real number $x = -0.0014$ is represented in the decimal floating-point representation as

$$x = (-1)^1 \times 0.14 \times 10^{-2}.$$

Here $s = 1$, $\beta = 10$, $e = -2$, $d_1 = 1$ and $d_2 = 4$.

Remark 3.1.5.

The floating-point representation of the number $1/3$ is

$$\frac{1}{3} = 0.33333 \dots = (-1)^0 \times (0.33333 \dots)_{10} \times 10^0.$$

An n -digit decimal floating-point representation of this number has to contain only n digits in its mantissa. Therefore, $1/3$ cannot be represented as an n -digit floating-point number.

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 number of digits in the mantissa (n) and the range of the exponent (e). In Section 3.1.2, 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 Section 3.1.3.

3.1.2 Underflow and Overflow 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, whereas when this value goes below the minimum of the range, then we encounter underflow. Thus, for a given computing device, there are integers m and M such that the exponent e is limited to a range

$$m \leq e \leq M. \quad (3.5)$$

During the calculation, if some computed number has an exponent $e > M$ then we say, the memory **overflow** occurs and if $e < m$, we say the memory **underflow** occurs.

Remark 3.1.6.

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 your 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 is a sufficiently large number that results in an overflow of memory when stored on a computing device, and x is another number that results in an underflow, then their product will be returned as **NaN**.

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. However, by writing a separate subroutine, one can monitor and get a warning whenever an underflow occurs.

Example 3.1.7 [Overflow].

Run the following MATLAB code on a computer with 32-bit intel processor:

```
i=308.25471;
fprintf('%f %f\n',i,10^i);
i=308.25472;
fprintf('%f %f\n',i,10^i);
```

We see that the first print command shows a meaningful (but very large) number, whereas the second print command simply prints `inf`. This is due to the overflow of memory while representing a very large real number.

Also try running the following code on the MATLAB:

```
i=308.25471;
fprintf('%f %f\n',i,10^i/10^i);
i=308.25472;
fprintf('%f %f\n',i,10^i/10^i);
```

The output will be

```
308.254710  1.000000
308.254720  NaN
```

If your computer is not showing `inf` for `i = 308.25472`, try increasing the value of `i` till you get `inf`.

Example 3.1.8 [Underflow].

Run the following MATLAB code on a computer with 32-bit intel processor:

```
j=-323.6;
if(10^j>0)
    fprintf('The given number is greater than zero\n');
elseif (10^j==0)
    fprintf('The given number is equal to zero\n');
else
    fprintf('The given number is less than zero\n');
end
```

The output will be

The given number is greater than zero

When the value of j is further reduced slightly as shown in the following program

```
j=-323.64;
if(10^j>0)
    fprintf('The given number is greater than zero\n');
elseif (10^j==0)
    fprintf('The given number is equal to zero\n');
else
    fprintf('The given number is less than zero\n');
end
```

the output shows

The given number is equal to zero

If your computer is not showing the above output, try decreasing the value of j till you get the above output.

In this example, we see that the number $10^{-323.64}$ is recognized as zero by the computer. This is due to the underflow of memory. Note that multiplying any large number by this number will give zero as answer. 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.

3.1.3 Chopping and Rounding a Number

The number of digits in the mantissa, as given in Definition 3.1.2, is called the *precision* or *length* of the floating-point number. 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 (or sufficiently large number of) digits in the mantissa of the floating-point form as in (3.1), then the computing device converts this number into an n -digit floating-point form as in (3.3). 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 approximation.

Definition 3.1.9 [Chopped and Rounded Numbers].

Let x be a real number given in the floating-point representation (3.1) as

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

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

$$\text{fl}(x) = (-1)^s \times (.d_1 d_2 \cdots d_n)_\beta \times \beta^e. \quad (3.6)$$

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

$$\text{fl}(x) = \begin{cases} (-1)^s \times (.d_1 d_2 \cdots d_n)_\beta \times \beta^e & , \quad 0 \leq d_{n+1} < \frac{\beta}{2} \\ (-1)^s \times (.d_1 d_2 \cdots (d_n + 1))_\beta \times \beta^e & , \quad \frac{\beta}{2} \leq d_{n+1} < \beta \end{cases}, \quad (3.7)$$

where

$$(-1)^s \times (.d_1 d_2 \cdots (d_n + 1))_\beta \times \beta^e := (-1)^s \times \left((.d_1 d_2 \cdots d_n)_\beta + (. \underbrace{00 \cdots 0}_{(n-1)\text{-times}} 1)_\beta \right) \times \beta^e.$$

Note

As already mentioned, throughout this course, we always take $\beta = 10$. **Also, we do not assume any restriction on the exponent $e \in \mathbb{Z}$.**

Example 3.1.10.

The floating-point representation of π is given by

$$\pi = (-1)^0 \times (.31415926 \cdots) \times 10^1.$$

The floating-point approximation of π using *five-digit chopping* is

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

which is equal to 3.1415. Since the sixth digit of the mantissa in the floating-point representation of π is a 9, the floating-point approximation of π using *five-digit rounding* is given by

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

which is equal to 3.1416.

Remark 3.1.11.

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*).

3.1.4 Arithmetic Using n -Digit Rounding and Chopping

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 is done in a similar way.

Let \odot denote any one of the basic arithmetic operations '+', '-', '×' and '÷'. Let x and y be real numbers. The process of computing $x \odot y$ using n -digit rounding is as follows.

Step 1: Get the n -digit rounding approximation $\text{fl}(x)$ and $\text{fl}(y)$ of the numbers x and y , respectively.

Step 2: Perform the calculation $\text{fl}(x) \odot \text{fl}(y)$ using exact arithmetic.

Step 3: Get the n -digit rounding approximation $\text{fl}(\text{fl}(x) \odot \text{fl}(y))$ 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 3.1.12.

Consider the function

$$f(x) = x(\sqrt{x+1} - \sqrt{x}).$$

Let us evaluate $f(100000)$ using a six-digit rounding. We have

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

The evaluation of $\sqrt{100001}$ using six-digit rounding is as follows.

$$\begin{aligned} \sqrt{100001} &\approx 316.229347 \\ &= 0.316229347 \times 10^3. \end{aligned}$$

The six-digit rounded approximation of 0.316229347×10^3 is given by 0.316229×10^3 . Therefore,

$$\text{fl}(\sqrt{100001}) = 0.316229 \times 10^3.$$

Similarly,

$$\text{fl}(\sqrt{100000}) = 0.316228 \times 10^3.$$

The six-digit rounded approximation of the difference between these two numbers is

$$\text{fl}\left(\text{fl}(\sqrt{100001}) - \text{fl}(\sqrt{100000})\right) = 0.1 \times 10^{-2}.$$

Finally, we have

$$\begin{aligned} \text{fl}(100000) \times (0.1 \times 10^{-2}) &= (0.1 \times 10^6) \times (0.1 \times 10^{-2}) \\ &= 0.1 \times 10^3. \end{aligned}$$

Thus, the value of $f(100000)$ using six-digit rounding is 100. Similarly, we can see that using six-digit chopping, the value of $f(100000)$ is 200.

Definition 3.1.13 [Machine Epsilon].

The *machine epsilon* of a computer is the smallest positive floating-point number δ such that

$$\text{fl}(1 + \delta) > 1.$$

For any floating-point number $\hat{\delta} < \delta$, we have $\text{fl}(1 + \hat{\delta}) = 1$, and $1 + \hat{\delta}$ and 1 are identical within the computer's arithmetic.

Remark 3.1.14.

From Example 3.1.8, it is clear that the machine epsilon for a 32-bit intel processor lies between the numbers $10^{-323.64}$ and $10^{-323.6}$. It is possible to get the exact value of this number, but it is no way useful in our present course, and so we will not attempt to do this here.

3.2 Types of Errors

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

Definition 3.2.1 [Errors].

1. The **error** in a computed quantity is defined as

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

2. Absolute value of an error is called the **absolute error**.

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

$$\text{Relative Error} = \frac{\text{Error}}{\text{True Value}}.$$

Here, we assume that the true value is non-zero.

4. The **percentage error** is defined as

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

Remark 3.2.2.

Let x_A denote the approximation to the real number x . We use the following notations:

$$E(x_A) := \text{Error}(x_A) = x - x_A. \quad (3.8)$$

$$E_a(x_A) := \text{Absolute Error}(x_A) = |E(x_A)| \quad (3.9)$$

$$E_r(x_A) := \text{Relative Error}(x_A) = \frac{E(x_A)}{x}, \quad x \neq 0. \quad (3.10)$$

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 3.2.3.

Let $x = 100000$, $x_A = 99999$, $y = 1$ and $y_A = 1/2$. We have

$$E_a(x_A) = 1, \quad E_a(y_A) = \frac{1}{2}.$$

Although $E_a(x_A) > E_a(y_A)$, we have

$$E_r(x_A) = 10^{-5}, \quad E_r(y_A) = \frac{1}{2}.$$

Hence, in terms of percentage error, x_A has only $10^{-3}\%$ error when compared to x whereas y_A has 50% error when compared to y .

The errors defined above are between a given number and its approximate value. Quite often we also approximate a given function by another function that can be handled more easily. For instance, a sufficiently differentiable function can be approximated using Taylor's theorem (Theorem 1.1.2). The error between the function value and the value obtained from the corresponding Taylor's polynomial is defined as *truncation error* as defined in Section 1.2.

3.3 Loss of Significance

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

Definition 3.3.1 [Significant β -Digits].

Let β be a radix and $x \neq 0$. If x_A is an approximation to x , then we say that x_A approximates x to r *significant β -digits* if r is the largest non-negative integer such that

$$\frac{|x - x_A|}{|x|} \leq \frac{1}{2}\beta^{-r+1}. \quad (3.11)$$

We also say x_A *has r significant β -digits in x* .

Note

When $\beta = 10$, we refer significant 10-digits by significant digits.

Example 3.3.2.

1. For $x = 1/3$, the approximate number $x_A = 0.333$ has three significant digits, since

$$\frac{|x - x_A|}{|x|} = 0.001 < 0.005 = 0.5 \times 10^{-2}.$$

Thus, $r = 3$.

2. For $x = 0.02138$, the approximate number $x_A = 0.02144$ has three significant digits, since

$$\frac{|x - x_A|}{|x|} \approx 0.0028 < 0.005 = 0.5 \times 10^{-2}.$$

Thus, $r = 3$.

3. For $x = 0.02132$, the approximate number $x_A = 0.02144$ has two significant

digits, since

$$\frac{|x - x_A|}{|x|} \approx 0.0056 < 0.05 = 0.5 \times 10^{-1}.$$

Thus, $r = 2$.

4. For $x = 0.02138$, the approximate number $x_A = 0.02149$ has two significant digits, since

$$\frac{|x - x_A|}{|x|} \approx 0.0051 < 0.05 = 0.5 \times 10^{-1}.$$

Thus, $r = 2$.

5. For $x = 0.02108$, the approximate number $x_A = 0.0211$ has three significant digits, since

$$\frac{|x - x_A|}{|x|} \approx 0.0009 < 0.005 = 0.5 \times 10^{-2}.$$

Thus, $r = 3$.

6. For $x = 0.02108$, the approximate number $x_A = 0.02104$ has three significant digits, since

$$\frac{|x - x_A|}{|x|} \approx 0.0019 < 0.005 = 0.5 \times 10^{-2}.$$

Thus, $r = 3$.

Remark 3.3.3.

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 it is evident from the above examples.

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 3.3.4.

Let us consider two real numbers

$$x = 7.6545428 = 0.76545428 \times 10^1 \text{ and } y = 7.6544201 = 0.76544201 \times 10^1.$$

The numbers

$$x_A = 7.6545421 = 0.76545421 \times 10^1 \text{ and } y_A = 7.6544200 = 0.76544200 \times 10^1$$

are approximations to x and y , correct to seven and eight significant digits, respectively. The exact difference between x_A and y_A is

$$z_A = x_A - y_A = 0.12210000 \times 10^{-3}$$

and the exact difference between x and y is

$$z = x - y = 0.12270000 \times 10^{-3}.$$

Therefore,

$$\frac{|z - z_A|}{|z|} \approx 0.0049 < 0.5 \times 10^{-2}$$

and hence z_A has only three significant digits with respect to z . Thus, we started with two approximate numbers x_A and y_A which are correct to seven and eight significant digits with respect to x and y , respectively. But their difference z_A has only three significant digits with respect to z . Hence, there is a loss of significant digits in the process of subtraction. A simple calculation shows that

$$E_r(z_A) \approx 53581 \times E_r(x_A).$$

Similarly, we have

$$E_r(z_A) \approx 375067 \times E_r(y_A).$$

Loss of significant digits is therefore dangerous. The loss of significant digits in the process of calculation is referred to as *loss of significance*.

The loss of significant digits results in accumulation of error during evaluation of a function value at a given point as illustrated in the following example.

Example 3.3.5.

Consider the function

$$f(x) = x(\sqrt{x+1} - \sqrt{x}).$$

From Example 3.1.12, the value of $f(100000)$ using six-digit rounding is 100, whereas the true value is 158.113. There is a drastic error in the value of the function, which is due to the loss of significant digits. It is evident that as x increases, the terms $\sqrt{x+1}$ and \sqrt{x} comes closer to each other and therefore loss of significance in their computed value increases.

Such a loss of significance can be avoided by rewriting the given expression of f in such a way that subtraction of near-by non-negative numbers is avoided. For instance,

we can rewrite the expression of the function f as

$$f(x) = \frac{x}{\sqrt{x+1} + \sqrt{x}}.$$

With this new form of f , we obtain $f(100000) = 158.114000$ using six-digit rounding.

Example 3.3.6.

Consider evaluating the function

$$f(x) = 1 - \cos x$$

near $x = 0$. Since $\cos x \approx 1$ for x near zero, there will be loss of significance in the process of evaluating $f(x)$ for x near zero. So, we have to use an alternative formula for $f(x)$ such as

$$\begin{aligned} f(x) &= 1 - \cos x \\ &= \frac{1 - \cos^2 x}{1 + \cos x} \\ &= \frac{\sin^2 x}{1 + \cos x} \end{aligned}$$

which can be evaluated quite accurately for small x .

Remark 3.3.7.

Unlike the above examples, we may not be able to always write an equivalent formula of a given function to avoid loss of significance in the evaluation. In such cases, we have to go for a suitable approximation of the given function by other functions, for instance Taylor's polynomial of desired degree, that do not involve loss of significance.

3.4 Propagation of Relative Error in Arithmetic Operations

Once an error is committed, it affects subsequent results as this error propagates through subsequent calculations. We first study how the results are affected by using approximate numbers instead of actual numbers and then will take up the effect of errors on function evaluation in the next section.

Let x_A and y_A denote the approximate numbers used in the calculation, and let x_T and y_T be the corresponding true values. We will now see how relative error propagates with the four basic arithmetic operations.

3.4.1 Addition and Subtraction

Let $x_T = x_A + \epsilon$ and $y_T = y_A + \eta$ be positive real numbers. The relative error $E_r(x_A \pm y_A)$ is given by

$$\begin{aligned} E_r(x_A \pm y_A) &= \frac{(x_T \pm y_T) - (x_A \pm y_A)}{x_T \pm y_T} \\ &= \frac{(x_T \pm y_T) - (x_T - \epsilon \pm (y_T - \eta))}{x_T \pm y_T} \end{aligned}$$

Upon simplification, we get

$$E_r(x_A \pm y_A) = \frac{\epsilon \pm \eta}{x_T \pm y_T}. \quad (3.12)$$

The above expression shows that there can be a drastic increase in the relative error during subtraction of two approximate numbers whenever $x_T \approx y_T$ as we have witnessed in **Example 3.3.4** and **Example 3.3.5**. On the other hand, it is easy to see from (3.12) that

$$|E_r(x_A + y_A)| \leq |E_r(x_A)| + |E_r(y_A)|,$$

which shows that the relative error propagates slowly in addition. Note that such an inequality in the case of subtraction is not possible.

3.4.2 Multiplication

The relative error $E_r(x_A \times y_A)$ is given by

$$\begin{aligned} E_r(x_A \times y_A) &= \frac{(x_T \times y_T) - (x_A \times y_A)}{x_T \times y_T} \\ &= \frac{(x_T \times y_T) - ((x_T - \epsilon) \times (y_T - \eta))}{x_T \times y_T} \\ &= \frac{\eta x_T + \epsilon y_T - \epsilon \eta}{x_T \times y_T} \\ &= \frac{\epsilon}{x_T} + \frac{\eta}{y_T} - \left(\frac{\epsilon}{x_T}\right) \left(\frac{\eta}{y_T}\right) \end{aligned}$$

Thus, we have

$$E_r(x_A \times y_A) = E_r(x_A) + E_r(y_A) - E_r(x_A)E_r(y_A). \quad (3.13)$$

Taking modulus on both sides, we get

$$|E_r(x_A \times y_A)| \leq |E_r(x_A)| + |E_r(y_A)| + |E_r(x_A)| |E_r(y_A)|$$

Note that when $|E_r(x_A)|$ and $|E_r(y_A)|$ are very small, then their product is negligible when compared to $|E_r(x_A)| + |E_r(y_A)|$. Therefore, the above inequality reduces to

$$|E_r(x_A \times y_A)| \lesssim |E_r(x_A)| + |E_r(y_A)|,$$

which shows that the relative error propagates slowly in multiplication.

3.4.3 Division

The relative error $E_r(x_A/y_A)$ is given by

$$\begin{aligned}
 E_r(x_A/y_A) &= \frac{(x_T/y_T) - (x_A/y_A)}{x_T/y_T} \\
 &= \frac{x_T(y_T - \eta) - y_T(x_T - \epsilon)}{x_T(y_T - \eta)} \\
 &= \frac{\epsilon y_T - \eta x_T}{x_T(y_T - \eta)} \\
 &= \frac{y_T}{y_T - \eta} (E_r(x_A) - E_r(y_A))
 \end{aligned}$$

Thus, we have

$$E_r(x_A/y_A) = \frac{1}{1 - E_r(y_A)} (E_r(x_A) - E_r(y_A)). \quad (3.14)$$

The above expression shows that the relative error increases drastically during division whenever $E_r(y_A) \approx 1$. This means that y_A has 100% error when compared to y , which is very unlikely because we always expect the relative error to be very small, ie., very close to zero. In this case the right hand side is approximately equal to $E_r(x_A) - E_r(y_A)$. Hence, we have

$$|E_r(x_A/y_A)| \lesssim |E_r(x_A) - E_r(y_A)| \leq |E_r(x_A)| + |E_r(y_A)|,$$

which shows that the relative error propagates slowly in division.

3.4.4 Total Error

In Subsection 3.1.4, we discussed the procedure of performing arithmetic operations using n -digit floating-point approximation. The computed value $\text{fl}(\text{fl}(x) \odot \text{fl}(y))$ involves an error (when compared to the exact value $x \odot y$) which comprises of

1. Error in $\text{fl}(x)$ and $\text{fl}(y)$ due to n -digit rounding or chopping of x and y , respectively, and
2. Error in $\text{fl}(\text{fl}(x) \odot \text{fl}(y))$ due to n -digit rounding or chopping of the number $\text{fl}(x) \odot \text{fl}(y)$.

The **total error** is defined as

$$\begin{aligned}
 (x \odot y) - \text{fl}(\text{fl}(x) \odot \text{fl}(y)) &= \{(x \odot y) - (\text{fl}(x) \odot \text{fl}(y))\} \\
 &\quad + \{(\text{fl}(x) \odot \text{fl}(y)) - \text{fl}(\text{fl}(x) \odot \text{fl}(y))\},
 \end{aligned}$$

in which the first term on the right hand side is called the **propagated error** and the second term is called the **floating-point error**. The **relative total error** is obtained by dividing both sides of the above expression by $x \odot y$.

Example 3.4.1.

Consider evaluating the integral

$$I_n = \int_0^1 \frac{x^n}{x+5} dx, \quad \text{for } n = 0, 1, \dots, 30.$$

The value of I_n can be obtained in two different iterative processes, namely,

1. The *forward iteration* for evaluating I_n is given by

$$I_n = \frac{1}{n} - 5I_{n-1}, \quad I_0 = \ln(6/5).$$

2. The *backward iteration* for evaluating I_{n-1} is given by

$$I_{n-1} = \frac{1}{5n} - \frac{1}{5}I_n, \quad I_{30} = 0.54046330 \times 10^{-2}.$$

The following table shows the computed value of I_n using both iterative formulas along with the exact value. The numbers are computed using MATLAB using double precision arithmetic and the final answer is rounded to 6 digits after the decimal point.

n	Forward Iteration	Backward Iteration	Exact Value
1	0.088392	0.088392	0.088392
5	0.028468	0.028468	0.028468
10	0.015368	0.015368	0.015368
15	0.010522	0.010521	0.010521
20	0.004243	0.007998	0.007998
25	11.740469	0.006450	0.006450
30	-36668.803026	Not Computed	0.005405

Clearly the backward iteration gives exact value up to the number of digits shown, whereas forward iteration tends to increase error and give entirely wrong values. This is due to the propagation of error from one iteration to the next iteration. In forward iteration, the total error from one iteration is magnified by a factor of 5 at the next iteration. In backward iteration, the total error from one iteration is divided by 5 at the next iteration. Thus, in this example, with each iteration, the total error tends to increase rapidly in the forward iteration and tends to increase very slowly in the backward iteration.

3.5 Propagation of Relative Error in Function Evaluation

For a given function $f : \mathbb{R} \rightarrow \mathbb{R}$, consider evaluating $f(x)$ at an approximate value x_A rather than at x . The question is **how well does $f(x_A)$ approximate $f(x)$** ? To answer this question, we compare $E_r(f(x_A))$ with $E_r(x_A)$.

Assume that f is a C^1 function. Using the mean-value theorem, we get

$$f(x) - f(x_A) = f'(\xi)(x - x_A),$$

where ξ is an unknown point between x and x_A . The relative error in $f(x_A)$ when compared to $f(x)$ is given by

$$E_r(f(x_A)) = \frac{f'(\xi)}{f(x)}(x - x_A).$$

Thus, we have

$$E_r(f(x_A)) = \left(\frac{f'(\xi)}{f(x)} x \right) E_r(x_A). \quad (3.15)$$

Since x_A and x are assumed to be very close to each other and ξ lies between x and x_A , we may make the approximation

$$f(x) - f(x_A) \approx f'(x)(x - x_A).$$

In view of (3.15), we have

$$E_r(f(x_A)) \approx \left(\frac{f'(x)}{f(x)} x \right) E_r(x_A). \quad (3.16)$$

The expression inside the bracket on the right hand side of (3.16) is the amplification factor for the relative error in $f(x_A)$ in terms of the relative error in x_A . Thus, this expression plays an important role in understanding the propagation of relative error in evaluating the function value $f(x)$ and hence motivates the following definition.

Definition 3.5.1 [Condition Number of a Function].

The **condition number** of a continuously differentiable function f at a point $x = c$ is given by

$$\left| \frac{f'(c)}{f(c)} c \right|. \quad (3.17)$$

The condition number of a function at a point $x = c$ can be used to decide whether the evaluation of the function at $x = c$ is well-conditioned or ill-conditioned depending on whether this condition number is smaller or larger as we approach this point. It is not possible to decide a priori how large the condition number should be to say that the function evaluation is ill-conditioned and it depends on the circumstances in which we are working.

Definition 3.5.2 [Well-Conditioned and Ill-Conditioned].

The process of evaluating a continuously differentiable function f at a point $x = c$ is said to be *well-conditioned* if the condition number

$$\left| \frac{f'(c)}{f(c)} c \right|$$

at c is ‘small’. The process of evaluating a function at $x = c$ is said to be *ill-conditioned* if it is not well-conditioned.

In the following example, we illustrate a well-conditioned computation.

Example 3.5.3.

Consider the function

$$f(x) = \sqrt{x},$$

for all $x \in [0, \infty)$. Then

$$f'(x) = \frac{1}{2\sqrt{x}}, \text{ for all } x \in [0, \infty).$$

The condition number of f is

$$\left| \frac{f'(x)}{f(x)} x \right| = \frac{1}{2}, \text{ for all } x \in [0, \infty),$$

which shows that taking square roots is a well-conditioned process.

From (3.16), we have

$$|E_r(f(x_A))| \approx \frac{1}{2} |E_r(x_A)|.$$

Thus, $E_r(f(x_A))$ is more closer to zero than $E_r(x_A)$.

We next illustrate an ill-conditioned computation.

Example 3.5.4.

Consider the function

$$f(x) = \frac{10}{1 - x^2},$$

for all $x \in \mathbb{R}$. Then

$$f'(x) = \frac{20x}{(1 - x^2)^2},$$

so that

$$\begin{aligned} \left| \frac{f'(x)}{f(x)} x \right| &= \left| \frac{\left(\frac{20x}{(1 - x^2)^2} \right) x}{\frac{10}{(1 - x^2)}} \right| \\ &= \frac{2x^2}{|1 - x^2|} \end{aligned}$$

and this number can be quite large for $|x|$ near 1. Thus, for x near 1 or -1, the process of evaluating this function is ill-conditioned.

The above two examples gives us a feeling that if the process of evaluating a function is well-conditioned, then we tend to get less propagating relative error. But, this is not true in general as shown in the following example.

Example 3.5.5.

Consider the function

$$f(x) = \sqrt{x+1} - \sqrt{x}, \text{ for all } x \in (0, \infty).$$

For all $x \in (0, \infty)$, the condition number of this function is

$$\begin{aligned} \left| \frac{f'(x)}{f(x)} x \right| &= \frac{1}{2} \left| \frac{\left(\frac{1}{\sqrt{x+1}} - \frac{1}{\sqrt{x}} \right)}{\sqrt{x+1} - \sqrt{x}} x \right| \\ &= \frac{1}{2} \frac{x}{\sqrt{x+1}\sqrt{x}} \\ &\leq \frac{1}{2}, \end{aligned} \tag{3.18}$$

which shows that the process of evaluating f is well-conditioned for all $x \in (0, \infty)$. But, if we calculate $f(12345)$ using six-digit rounding, we find

$$\begin{aligned} f(12345) &= \sqrt{12346} - \sqrt{12345} \\ &= 111.113 - 111.108 \\ &= 0.005, \end{aligned}$$

while, actually, $f(12345) = 0.00450003262627751 \dots$. The calculated answer has 10% error.

The above example shows that a well-conditioned process of evaluating a function at a point is not enough to ensure the accuracy in the corresponding computed value. We need to check for the stability of the computation, which we discuss in the following subsection.

3.5.1 Stable and Unstable Computations

Suppose there are n steps to evaluate a function $f(x)$ at a point $x = c$. Then the total process of evaluating this function is said to have *instability* if at least one of the n steps is ill-conditioned. If all the steps are well-conditioned, then the process is said to be *stable*.

Example 3.5.6.

We continue the discussion in Example 3.5.5 and check the stability in evaluating the function f . Let us analyze the computational process. The function f consists of the following four computational steps in evaluating the value of f at $x = x_0$:

$$x_1 := x_0 + 1, \quad x_2 := \sqrt{x_1}, \quad x_3 := \sqrt{x_0}, \quad x_4 := x_2 - x_3.$$

Now consider the last two steps where we already computed x_2 and now going to compute x_3 and finally evaluate the function

$$f_4(t) := x_2 - t.$$

At this step, the condition number for f_4 is given by

$$\left| \frac{f'_4(t)}{f_4(t)} t \right| = \left| \frac{t}{x_2 - t} \right|.$$

Thus, f_4 is ill-conditioned when t approaches x_2 . Therefore, the above process of evaluating the function $f(x)$ is *unstable*.

Let us rewrite the same function $f(x)$ as

$$\tilde{f}(x) = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

The computational process of evaluating \tilde{f} at $x = x_0$ is

$$x_1 := x_0 + 1, \quad x_2 := \sqrt{x_1}, \quad x_3 := \sqrt{x_0}, \quad x_4 := x_2 + x_3, \quad x_5 := 1/x_4.$$

It is easy to verify that the condition number of each of the above steps is well-conditioned. For instance, the last step defines

$$\tilde{f}_5(t) = \frac{1}{x_2 + t},$$

and the condition number of this function is approximately,

$$\left| \frac{\tilde{f}'_5(x)}{\tilde{f}_5(x)} x \right| = \left| \frac{t}{x_2 + t} \right| \approx \frac{1}{2}$$

for t sufficiently close to x_2 . Therefore, this process of evaluating $\tilde{f}(x)$ is stable. Recall from **Example 3.3.5** that the above expression gives a more accurate value for sufficiently large x .

Remark 3.5.7.

As discussed in Remark 3.3.7, we may not be lucky all the time to come out with an alternate expression that lead to stable evaluation for any given function when the original expression leads to unstable evaluation. In such situations, we have to compromise and go for a suitable approximation with other functions with stable evaluation process. For instance, we may try approximating the given function with its Taylor's polynomial, if possible.

3.6 Exercises

Floating-Point Approximation

1. Let X be a sufficiently large number which result in an overflow of memory on a computing device. Let x be a sufficiently small number which result in underflow of memory on the same computing device. Then give the output of the following operations:
 (i) $x \times X$ (ii) $3 \times X$ (iii) $3 \times x$ (iv) x/X (v) X/x .
2. In the following problems, show all the steps involved in the computation.
 - i) Using 5-digit rounding, compute $37654 + 25.874 - 37679$.
 - ii) Let $a = 0.00456$, $b = 0.123$, $c = -0.128$. Using 3-digit rounding, compute $(a + b) + c$, and $a + (b + c)$. What is your conclusion?
 - iii) Let $a = 2$, $b = -0.6$, $c = 0.602$. Using 3-digit rounding, compute $a \times (b + c)$, and $(a \times b) + (a \times c)$. What is your conclusion?
3. To find the mid-point of an interval $[a, b]$, the formula $\frac{a+b}{2}$ is often used. Compute the mid-point of the interval $[0.982, 0.987]$ using 3-digit chopping. On the number line represent all the three points. What do you observe? Now use the more geometric formula $a + \frac{b-a}{2}$ to compute the mid-point, once again using 3-digit chopping. What do you observe this time? Why is the second formula more geometric?
4. Consider a computing device having exponents e in the range $m \leq e \leq M$, $m, M \in \mathbb{Z}$. If the device uses n -digit rounding binary floating-point arithmetic, then show that $\delta = 2^{-n}$ is the machine epsilon when $n \leq |m| + 1$.

Types of Errors

5. If $\text{fl}(x)$ is the approximation of a real number x in a computing device, and ϵ is the corresponding relative error, then show that $\text{fl}(x) = (1 - \epsilon)x$.
6. Let x , y and z be real numbers whose floating point approximations in a computing device coincide with x , y and z respectively. Show that the relative error in computing $x(y + z)$ equals $\epsilon_1 + \epsilon_2 - \epsilon_1\epsilon_2$, where $\epsilon_1 = E_r(\text{fl}(y + z))$ and $\epsilon_2 = E_r(\text{fl}(x \times \text{fl}(y + z)))$.
7. Let $\epsilon = E_r(\text{fl}(x))$. Show that
 - i) $|\epsilon| \leq 10^{-n+1}$ if the computing device uses n -digit (decimal) chopping.
 - ii) $|\epsilon| \leq \frac{1}{2}10^{-n+1}$ if the computing device uses n -digit (decimal) rounding.
 - iii) Can the equality hold in the above inequalities?

8. Let the approximation $\sin x \approx x$ be used on the interval $[-\delta, \delta]$ where $\delta > 0$. Show that for all $x \in [-\delta, \delta]$,

$$|\sin x - x| < \frac{x^3}{6}.$$

Find a $\delta > 0$ such that the inequality

$$|\sin x - x| < \frac{1}{2}10^{-6},$$

holds for all $x \in [-\delta, \delta]$.

9. Let $x_A = 3.14$ and $y_A = 2.651$ be obtained from x_T and y_T using 4-digit rounding. Find the smallest interval that contains
 (i) x_T (ii) y_T (iii) $x_T + y_T$ (iv) $x_T - y_T$ (v) $x_T \times y_T$ (vi) x_T/y_T .
10. The ideal gas law is given by $PV = nRT$ where R is the gas constant. We are interested in knowing the value of T for which $P = V = n = 1$. If R is known only approximately as $R_A = 8.3143$ with an absolute error at most 0.12×10^{-2} . What is the relative error in the computation of T that results in using R_A instead of R ?

Loss of Significance and Propagation of Error

11. Obtain the number of significant digits in x_A when compared to x for the following cases:
 (i) $x = 451.01$ and $x_A = 451.023$ (ii) $x = -0.04518$ and $x_A = -0.045113$
 (iii) $x = 23.4604$ and $x_A = 23.4213$.
12. Instead of using the true values $x_T = 0.71456371$ and $y_T = 0.71456238$ in calculating $z_T = x_T - y_T (= 0.133 \times 10^{-5})$, if we use the approximate values $x_A = 0.71456414$ and $y_A = 0.71456103$, and calculate $z_A = x_A - y_A (= 0.311 \times 10^{-5})$, then find the loss of significant digits in the process of calculating z_A when compared to the significant digits in x_A .
13. Let $x < 0 < y$ be such that the approximate numbers x_A and y_A has seven and nine significant digits with x and y respectively. Show that $z_A := x_A - y_A$ has at least six significant digits when compared to $z := x - y$.

Propagation of Relative Error

14. Let $x = 0.65385$ and $y = 0.93263$. Obtain the total error in obtaining the product of these two numbers using 3-digit rounding.

15. Find the condition number at a point $x = c$ for the following functions
(i) $f(x) = x^2$, (ii) $g(x) = \pi^x$, (iii) $h(x) = b^x$.
16. Let x_T be a real number. Let $x_A = 2.5$ be an approximate value of x_T with an absolute error at most 0.01. The function $f(x) = x^3$ is evaluated at $x = x_A$ instead of $x = x_T$. Estimate the resulting absolute error.
17. Is the process of computing the function $f(x) = (e^x - 1)/x$ stable or unstable for $x \approx 0$? Justify your answer.
18. Show that the process of evaluating the function

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

for $x \approx 0$ is unstable. Suggest an alternate formula for evaluating f for $x \approx 0$, and check if the computation process is stable using this new formula.

19. Check for stability of computing the function

$$f(x) = \left(x + \frac{1}{3}\right) - \left(x - \frac{1}{3}\right)$$

for large values of x .

20. Check for stability of computing the function

$$g(x) = \frac{\left(3 + \frac{x^2}{3}\right) - \left(3 - \frac{x^2}{3}\right)}{x^2}$$

for values of x very close to 0.

21. Check for stability of computing the function

$$h(x) = \frac{\sin^2 x}{1 - \cos^2 x}$$

for values of x very close to 0.

22. Compute the values of the function

$$\frac{\sin x}{\sqrt{1 - \sin^2 x}}$$

using your calculator (without simplifying the given formula) at $x = 89.9, 89.95, 89.99$ degrees. Also compute the values of the function $\tan x$ at these values of x and compare the number of significant digits.

Tutorial Sheet 1

Mathematical Preliminaries

1. Let L be a real number and let $\{a_n\}$ be a sequence of real numbers. If there exists a positive integer N and a $\mu \in (0, 1)$ such that

$$|a_n - L| \leq \mu |a_{n-1} - L|$$

holds for all $n \geq N$, then show that $a_n \rightarrow L$ as $n \rightarrow \infty$.

2. Show that the equation $\sin x + x^2 = 1$ has at least one solution in the interval $[0, 1]$.

3. Let $f(x)$ be a continuous function on $[a, b]$, let x_1, \dots, x_n be points in $[a, b]$, and let g_1, \dots, g_n be non-positive real numbers. Then show that

$$\sum_{i=1}^n f(x_i)g_i = f(\xi) \sum_{i=1}^n g_i, \quad \text{for some } \xi \in [a, b].$$

4. Let $f : [a, b] \rightarrow [a, b]$ be a continuous function. Prove that the equation $f(x) = x$ has at least one solution lying in the interval $[a, b]$ (Note: A solution of this equation is called a *fixed point* of the function f). Further if $\max_{x \in [a, b]} |f'(x)| < 1$, then show that the equation $f(x) = x$ has a unique solution in $[a, b]$.

5. Let g be a continuously differentiable function (C^1 function) such that the equation $g(x) = 0$ has at least n distinct roots. Show that the equation $g'(x) = 0$ has at least $n - 1$ distinct roots.

6. Evaluate an approximate value of the function $f(x) = e^{x^2}$ at $x = 1$ using $T_2(x)$ about the point $a = 0$. Obtain the remainder $R_2(1)$ in terms of some unknown real number ξ . Compute (approximately) a possible value of ξ .

7. For every $x \in \mathbb{R}$, show that there exists a $\xi_x \in \mathbb{R}$

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{\cos(\xi_x)}{6!}x^6.$$

8. Determine the best value of $\alpha \in \mathbb{R}$ in the equation

$$\tan^{-1} x = x + O(x^\alpha) \text{ as } x \rightarrow 0$$

9. Let $F : [0, 1] \rightarrow [0, 1]$ be a differentiable function. Let a sequence $\{x_n\}$ defined inductively by $x_{n+1} = F(x_n)$ be such that $\lim_{n \rightarrow \infty} x_n$ exists and denote the limit by x . Further assume that $F'(x) = 0$. Show that

$$x_{n+2} - x_{n+1} = o(x_{n+1} - x_n) \text{ as } n \rightarrow \infty.$$

10. Prove or disprove:

$$(i) \frac{n+1}{n^2} = O\left(\frac{1}{n}\right) \text{ as } n \rightarrow \infty \quad (ii) \frac{1}{\ln n} = o\left(\frac{1}{n}\right) \text{ as } n \rightarrow \infty$$

Tutorial Sheet 2

Arithmetic Error Analysis

1. Let X be a sufficiently large number which result in an overflow of memory on a computing device. Let x be a sufficiently small number which result in underflow of memory on the same computing device. Then give the output of the following operations:
 (i) $x \times X$ (ii) $3 \times X$ (iii) $3 \times x$ (iv) x/X (v) X/x .
2. In this question, computations are done on a computer which uses 3-digit chopping arithmetic.
 - i) Compute the mid-point of the interval $[0.982, 0.987]$. Show all the steps of the computation.
 - ii) Give a different way of computing the mid-point of $[0.982, 0.987]$, so that the mid-point lies in the given interval $[0.982, 0.987]$. Show all the steps of the computation. Explain the differences between both these computations.
3. In the following problems, show all the steps involved in the computation.
 - i) Using 5-digit rounding, compute $37654 + 25.874 - 37679$.
 - ii) Let $a = 0.00456$, $b = 0.123$, $c = -0.128$. Using 3-digit rounding, compute $(a + b) + c$, and $a + (b + c)$. What is your conclusion?
 - iii) Let $a = 2$, $b = -0.6$, $c = 0.602$. Using 3-digit rounding, compute $a \times (b + c)$, and $(a \times b) + (a \times c)$. What is your conclusion?
4. Consider a computing device having exponents e in the range $m \leq e \leq M$, $m, M \in \mathbb{Z}$. Let n be an integer such that $n \leq |m| + 1$.
 - i) If the device uses n -digit rounding binary floating-point arithmetic, then show that $\delta = 2^{-n}$ is the machine epsilon.
 - ii) What is the machine epsilon of the device if it uses n -digit rounding decimal floating-point arithmetic? Justify your answer.
5. Consider a computing device that uses n -digit chopping (decimal) arithmetic. Let $\text{fl}(x)$ denote the floating-point approximation of a positive real number x in this device. Prove

$$\left| \frac{x - \text{fl}(x)}{x} \right| \leq 10^{-n+1}.$$
6. The ideal gas law is given by $PV = nRT$ where R is the gas constant. We are interested in knowing the value of T for which $P = V = n = 1$. If R is known only approximately as $R_A = 8.3143$ with an absolute error at most 0.12×10^{-2} . Obtain an upper bound for the absolute relative error in the computation of T that results in using R_A instead of R ?

7. Let $x_A = 3.14$ and $y_A = 2.651$ be obtained from x_T and y_T using 4-digit rounding. Find the smallest interval that contains
 - (i) x_T
 - (ii) y_T
 - (iii) $x_T + y_T$
 - (iv) $x_T - y_T$
 - (v) $x_T \times y_T$
 - (vi) x_T / y_T .
8. Obtain the number of significant digits of $x_A = 0.025678$ present in $x = 0.025611$.
9. Instead of using the true values $x_T = 0.62457371$ and $y_T = 0.62457238$ in calculating $z_T = x_T - y_T$, if we use the approximate values $x_A = 0.62451251$ and $y_A = 0.62458125$, and calculate $z_A = x_A - y_A$, then find the loss of significant digits in the process of calculating z_A when compared to the significant digits in x_A .
10. Let $x_A = 0.04078$ has exactly 3 significant digits with respect to the real number x_T . Find the smallest interval in which x_T lies.
11. Given $x = 0.75371$ and $y = -0.49572$. Let the product $x * y$ be computed using 3-digit rounding floating-point arithmetic. What is the absolute value of the total error? [Give the final answer with at least 6-digits after decimal places]
12. For small values of x , the approximation $\sin x \approx x$ is often used. Obtain a range of values of x for which the approximation gives an absolute error of at most $\frac{1}{2} \times 10^{-6}$.
13. Is the process of computing the value of the function $f(x) = (e^x - 1)/x$ stable or unstable for $x \approx 0$? Justify your answer.