METU – Department of Engineering Science

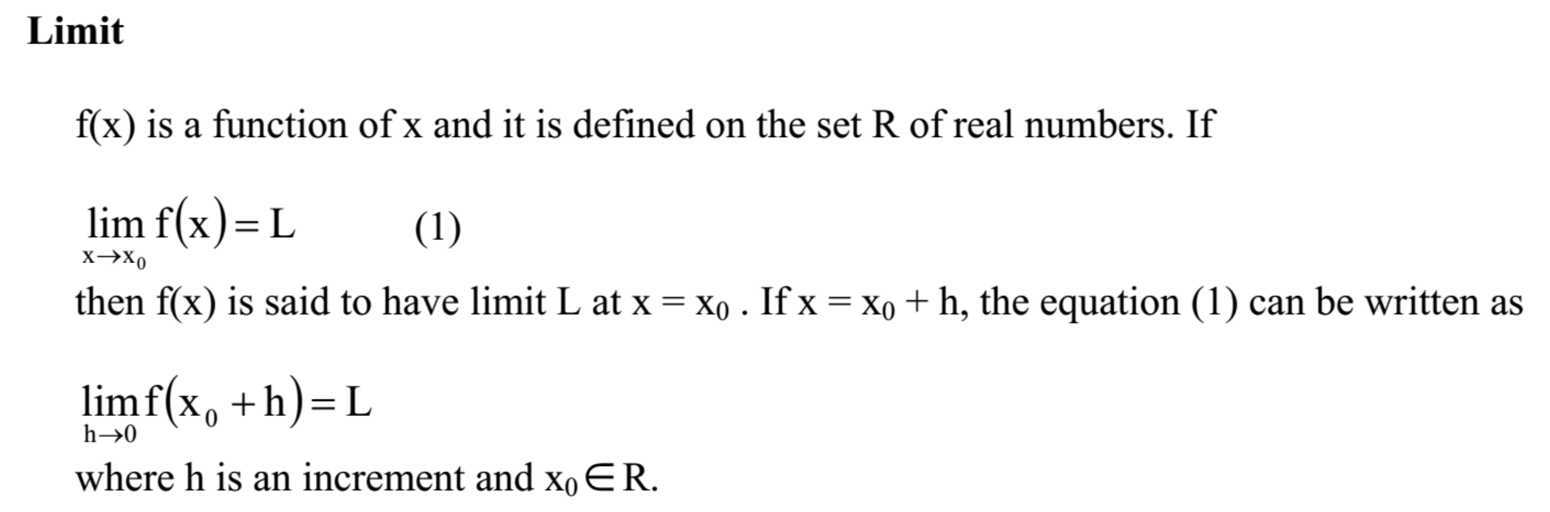
ES 361
COMPUTING METHODS IN ENGINEERING

Chapter 1
Introduction to Computing Methods

Dr. Berat Can CENGİZ

28.02.24

Mathematical Preliminaries



Limit

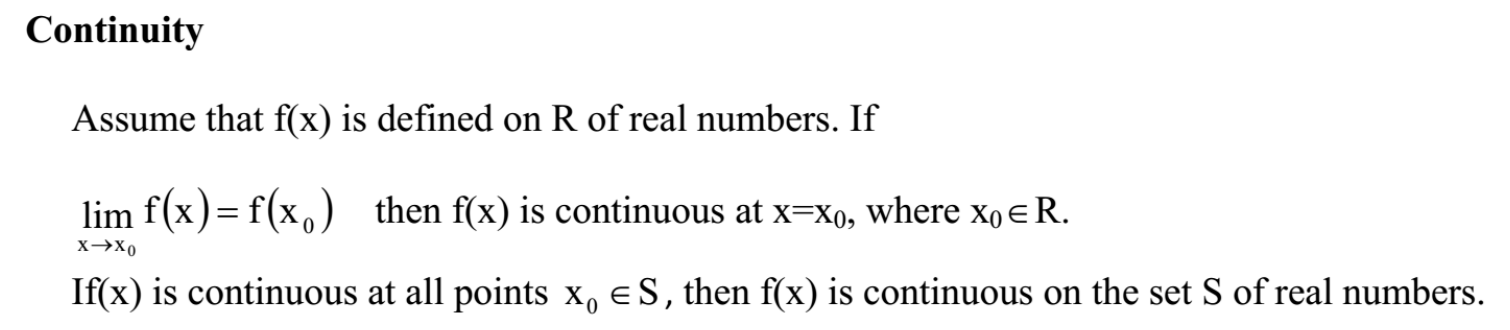
f(x) is a function of x and it is defined on the set R of real numbers. If

$$\lim_{x\to x_0} f(x) = L \qquad (1)$$

then f(x) is said to have limit L at $x = x_0$. If $x = x_0 + h$, the equation (1) can be written as

$$\lim_{h\to 0} f(x_0 + h) = L$$

where h is an increment and $x_0 \in R$.



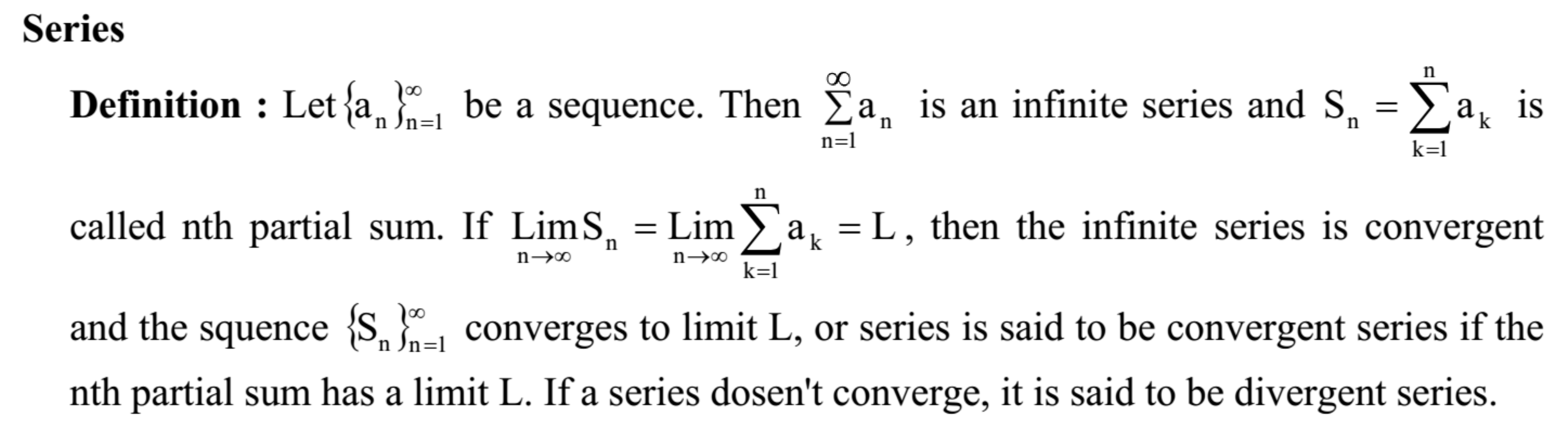
Continuity

Assume that f(x) is defined on R of real numbers. If

$\lim_{x\to x_0} f(x) = f(x_0)$ then f(x) is continuous at $x=x_0$, where $x_0 \in R$.

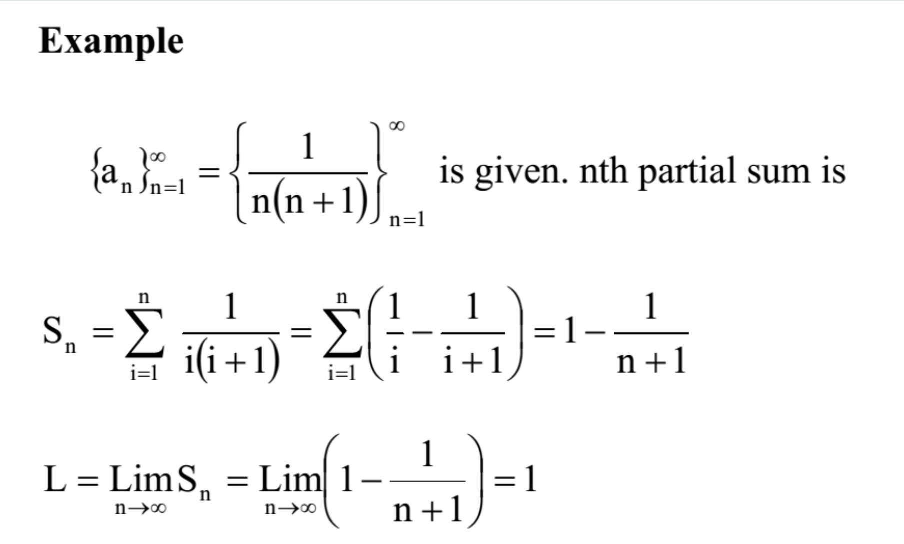
If(x) is continuous at all points $x_0 \in S$, then f(x) is continuous on the set S of real numbers.

Mathematical Preliminaries



Series

Definition : Let $\{a_n\}_{n=1}^{\infty}$ be a sequence. Then $\sum_{n=1}^{\infty} a_n$ is an infinite series and $S_n = \sum_{k=1}^{n} a_k$ is called nth partial sum. If $\lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{k=1}^{n} a_k = L$, then the infinite series is convergent and the squence $\{S_n\}_{n=1}^{\infty}$ converges to limit L, or series is said to be convergent series if the nth partial sum has a limit L. If a series dosen't converge, it is said to be divergent series.



Example

$\{a_n\}_{n=1}^{\infty} = \left\{ \frac{1}{n(n+1)} \right\}_{n=1}^{\infty}$ is given. nth partial sum is

$$S_n = \sum_{i=1}^{n} \frac{1}{i(i+1)} = \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i+1} \right) = 1 - \frac{1}{n+1}$$

$$L = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \left(1 - \frac{1}{n+1} \right) = 1$$

Mathematical Preliminaries

Convergent sequence

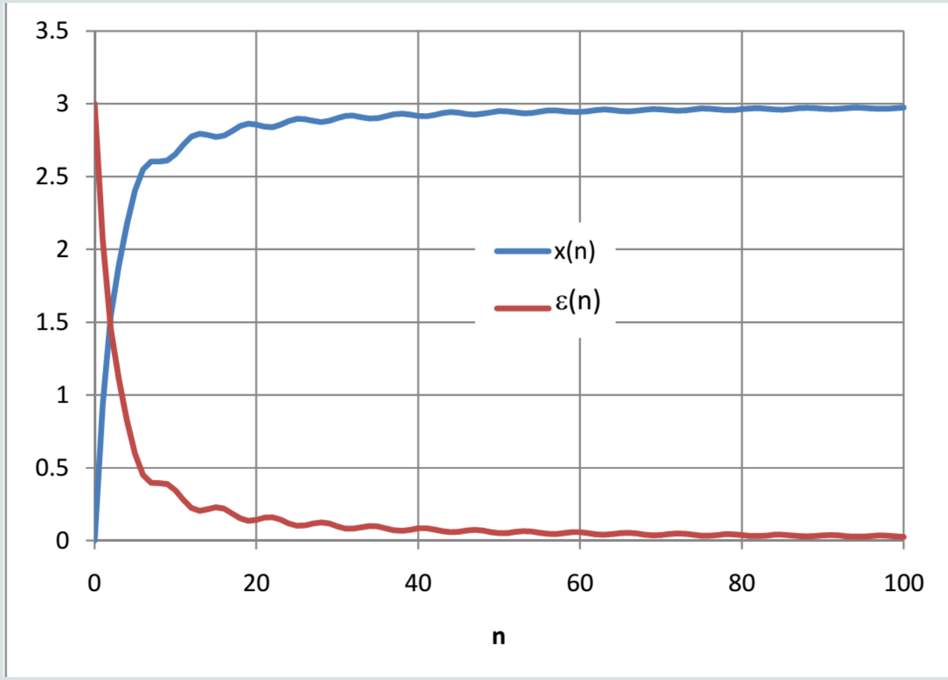
$\{x_n\}_{n=1}^{\infty}$ is an infinite sequence and x_n is a general term of the sequence. If x_n has a limit L, such that $\lim_{n\to\infty} x_n = L$ then sequence is a **convergent sequence**.

Example

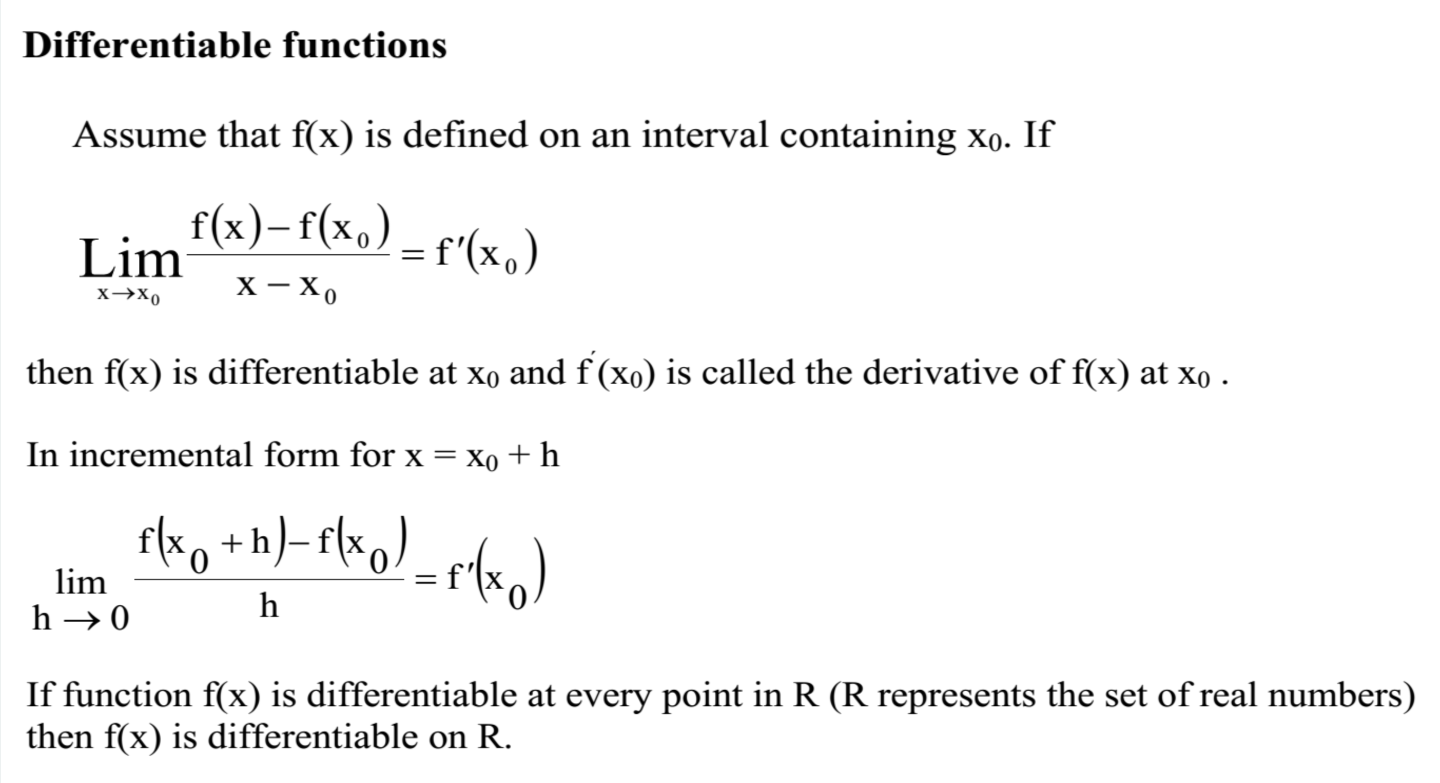
$x_n = \dfrac{6n^2 + n\cos(n)}{2n^2 + 2n + 3}$ is a squence.

$\lim_{n\to\infty} x_n = 3 \Rightarrow L = 3$, therefore squence has a limit L=3 and it is convergent.

Error sequence is defined as $\varepsilon_n = 3 - x_n$. Limit of error sequence is $\lim_{n\to\infty} \varepsilon_n = 3 - x_n = 0$.



Mathematical Preliminaries



Differentiable functions

Assume that f(x) is defined on an interval containing x_0. If

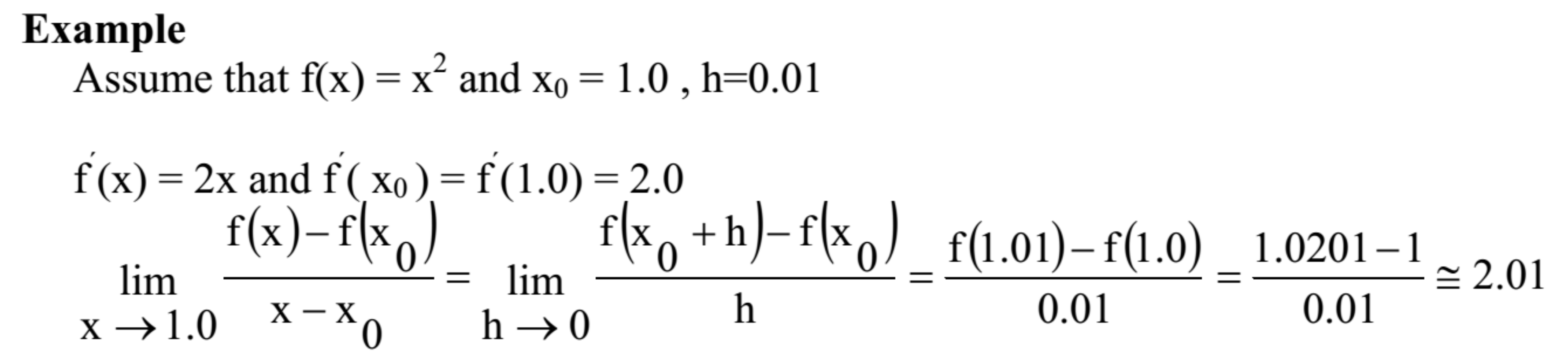
$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}=f'(x_0)$$

then f(x) is differentiable at x_0 and $f'(x_0)$ is called the derivative of f(x) at x_0 .

In incremental form for $x = x_0 + h$

$$\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}=f'(x_0)$$

If function f(x) is differentiable at every point in R (R represents the set of real numbers) then f(x) is differentiable on R.



Example

Assume that $f(x) = x^2$ and $x_0 = 1.0$, h=0.01

$f'(x) = 2x$ and $f'(x_0) = f'(1.0) = 2.0$

$$\lim_{x\to 1.0}\frac{f(x)-f(x_0)}{x-x_0}=\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}=\frac{f(1.01)-f(1.0)}{0.01}=\frac{1.0201-1}{0.01}\cong 2.01$$

Mathematical Preliminaries

Integrable functions

C[a,b] denotes the set of all continuous functions on the closed interval [a,b].

Assume that $f(x) \in C[a,b]$ If $\int_a^b f(x)dx = F(x)\Big|_{x=a}^{x=b} = F(b) - F(a)$ then it is said that f(x) is integrable and F(x) is integral of f(x) on the closed interval [a,b].

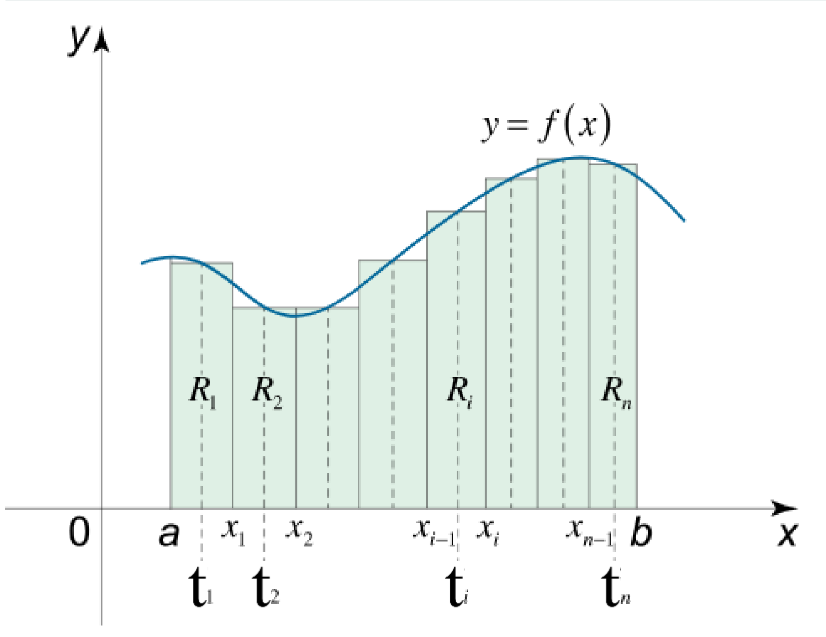
Summation Definition of Integral :

Assume that $f(x) \in C[a,b]$ and the interval [a,b] is subdivided into n intervals which are $[x_0,x_1]$,$[x_1,x_2]$,$[x_2,x_3]$, ...,$[x_{n-1},x_n]$, where x_0=a and x_n=b. Select an arbitrary point t_k in the interval $[x_{k-1},x_k]$, k=0,1,2,...,n and by introducing the difference $\Delta x_k = x_k - x_{k-1}$. Then

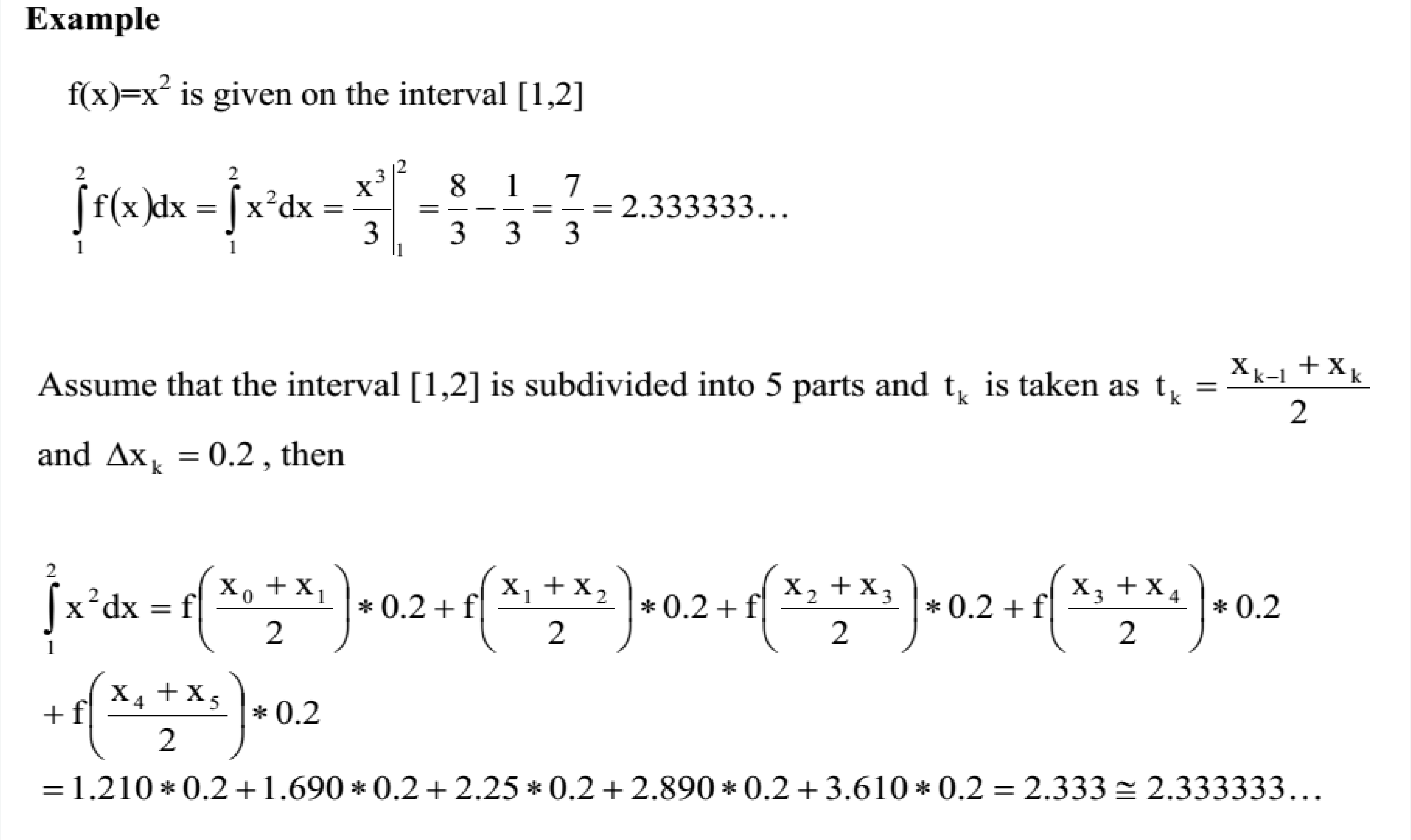
$\int_a^b f(x)dx = \lim_{n \to \infty} \sum_{k=1}^{n} f(t_k)\Delta x_k$ is called summation definition of integral and

$\sum_{k=1}^{n} f(t_k)\Delta x_k$ is called Riemann sum of f(x) on the interval [a,b]. Note that

$$\sum_{k=1}^{n} f(t_k)\Delta x_k \cong \int_a^b f(x)dx$$



Mathematical Preliminaries



Example

f(x)=x^2 is given on the interval [1,2]

$$\int_1^2 f(x)dx = \int_1^2 x^2 dx = \left.\frac{x^3}{3}\right|_1^2 = \frac{8}{3} - \frac{1}{3} = \frac{7}{3} = 2.333333\ldots$$

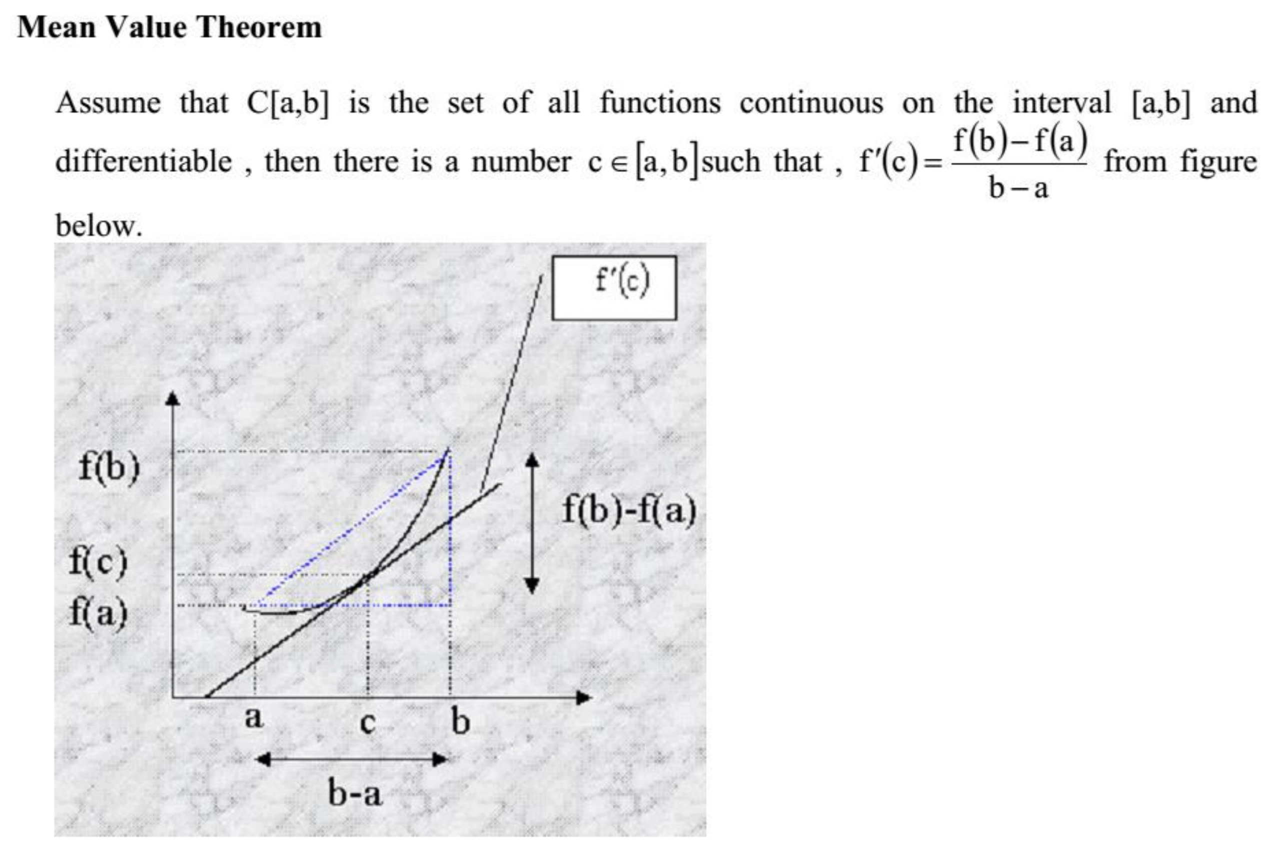
Assume that the interval [1,2] is subdivided into 5 parts and t_k is taken as $t_k = \dfrac{x_{k-1} + x_k}{2}$ and $\Delta x_k = 0.2$, then

$$\int_1^2 x^2 dx = f\left(\frac{x_0 + x_1}{2}\right) * 0.2 + f\left(\frac{x_1 + x_2}{2}\right) * 0.2 + f\left(\frac{x_2 + x_3}{2}\right) * 0.2 + f\left(\frac{x_3 + x_4}{2}\right) * 0.2$$

$$+ f\left(\frac{x_4 + x_5}{2}\right) * 0.2$$

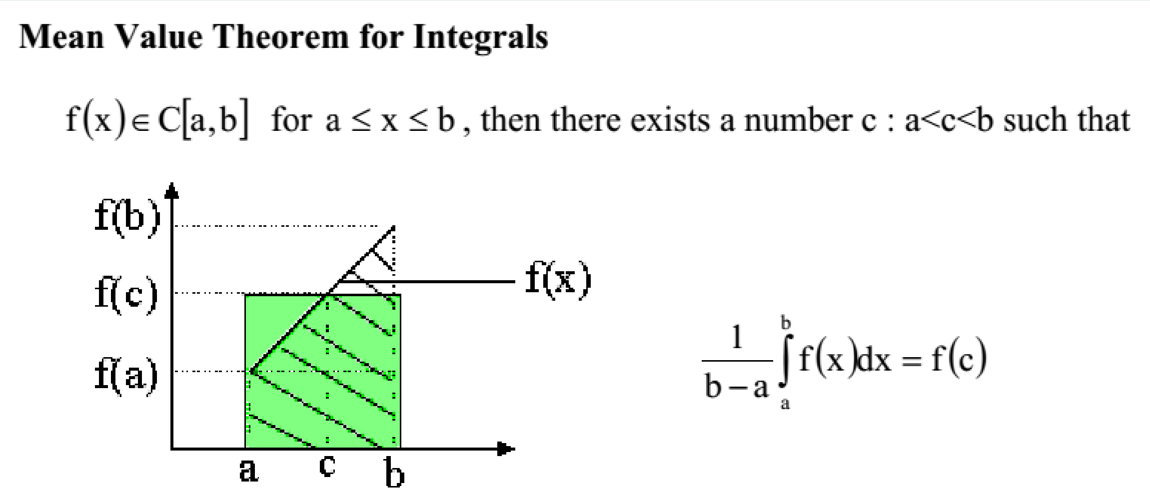
$$= 1.210 * 0.2 + 1.690 * 0.2 + 2.25 * 0.2 + 2.890 * 0.2 + 3.610 * 0.2 = 2.333 \cong 2.333333\ldots$$

Mathematical Preliminaries



Mean Value Theorem

Assume that C[a,b] is the set of all functions continuous on the interval [a,b] and differentiable , then there is a number $c \in [a,b]$ such that , $f'(c) = \frac{f(b) - f(a)}{b - a}$ from figure below.

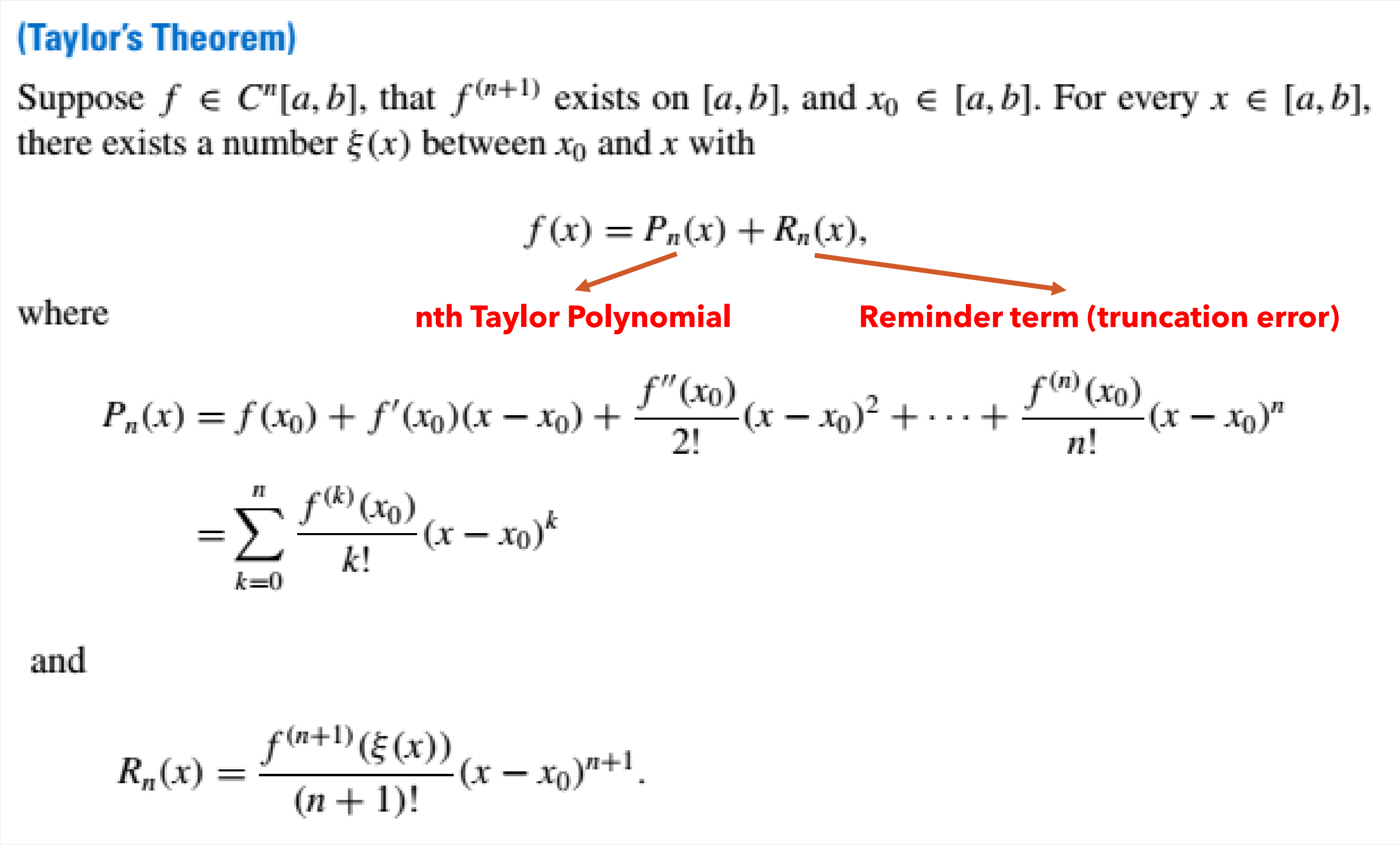


Mean Value Theorem for Integrals

$f(x) \in C[a,b]$ for $a \le x \le b$, then there exists a number c : a<c<b such that

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

Mathematical Preliminaries



(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

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

nth Taylor Polynomial

Reminder term (truncation error)

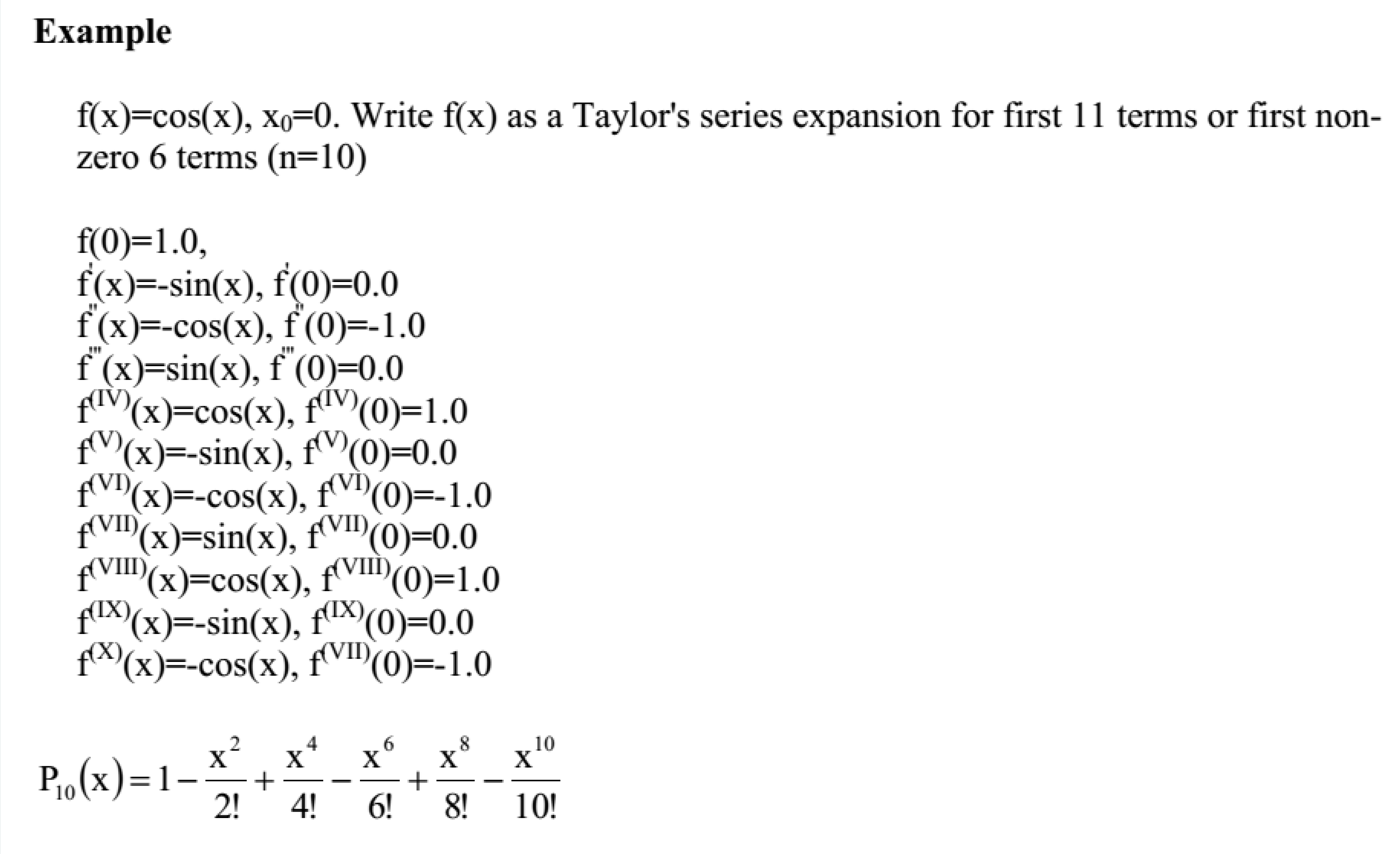
where

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

and

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

Mathematical Preliminaries



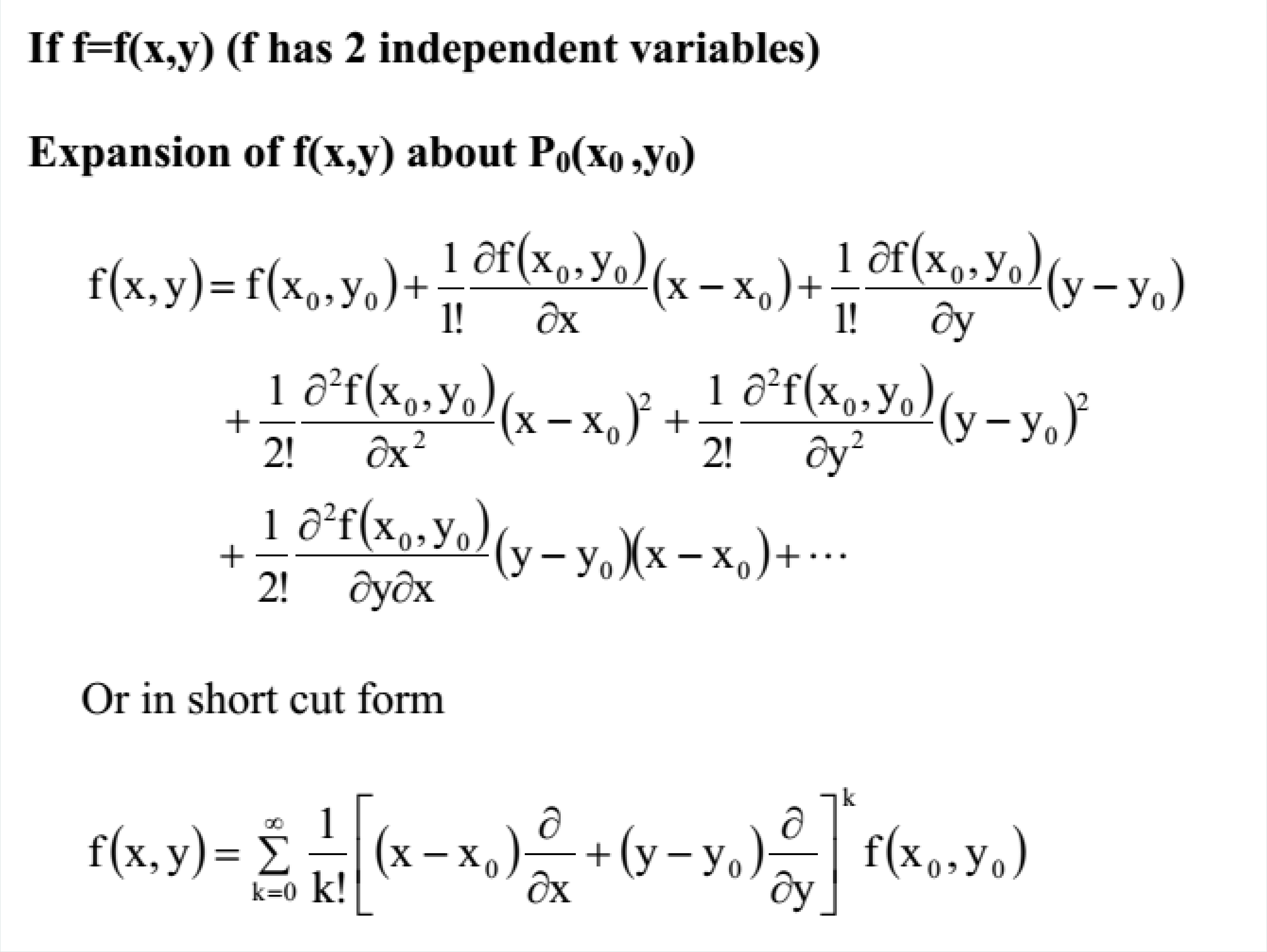
Example

$f(x)=\cos(x)$, $x_0=0$. Write $f(x)$ as a Taylor's series expansion for first 11 terms or first non-zero 6 terms ($n=10$)

$f(0)=1.0$,
$f'(x)=-\sin(x)$, $f'(0)=0.0$
$f''(x)=-\cos(x)$, $f''(0)=-1.0$
$f'''(x)=\sin(x)$, $f'''(0)=0.0$
$f^{(IV)}(x)=\cos(x)$, $f^{(IV)}(0)=1.0$
$f^{(V)}(x)=-\sin(x)$, $f^{(V)}(0)=0.0$
$f^{(VI)}(x)=-\cos(x)$, $f^{(VI)}(0)=-1.0$
$f^{(VII)}(x)=\sin(x)$, $f^{(VII)}(0)=0.0$
$f^{(VIII)}(x)=\cos(x)$, $f^{(VIII)}(0)=1.0$
$f^{(IX)}(x)=-\sin(x)$, $f^{(IX)}(0)=0.0$
$f^{(X)}(x)=-\cos(x)$, $f^{(VII)}(0)=-1.0$

$$P_{10}(x)=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\frac{x^8}{8!}-\frac{x^{10}}{10!}$$

Mathematical Preliminaries



If f=f(x,y) (f has 2 independent variables)

Expansion of f(x,y) about $P_0(x_0 ,y_0)$

$$f(x,y)=f(x_0,y_0)+\frac{1}{1!}\frac{\partial f(x_0,y_0)}{\partial x}(x-x_0)+\frac{1}{1!}\frac{\partial f(x_0,y_0)}{\partial y}(y-y_0)$$
$$+\frac{1}{2!}\frac{\partial^2 f(x_0,y_0)}{\partial x^2}(x-x_0)^2+\frac{1}{2!}\frac{\partial^2 f(x_0,y_0)}{\partial y^2}(y-y_0)^2$$
$$+\frac{1}{2!}\frac{\partial^2 f(x_0,y_0)}{\partial y\partial x}(y-y_0)(x-x_0)+\cdots$$

Or in short cut form

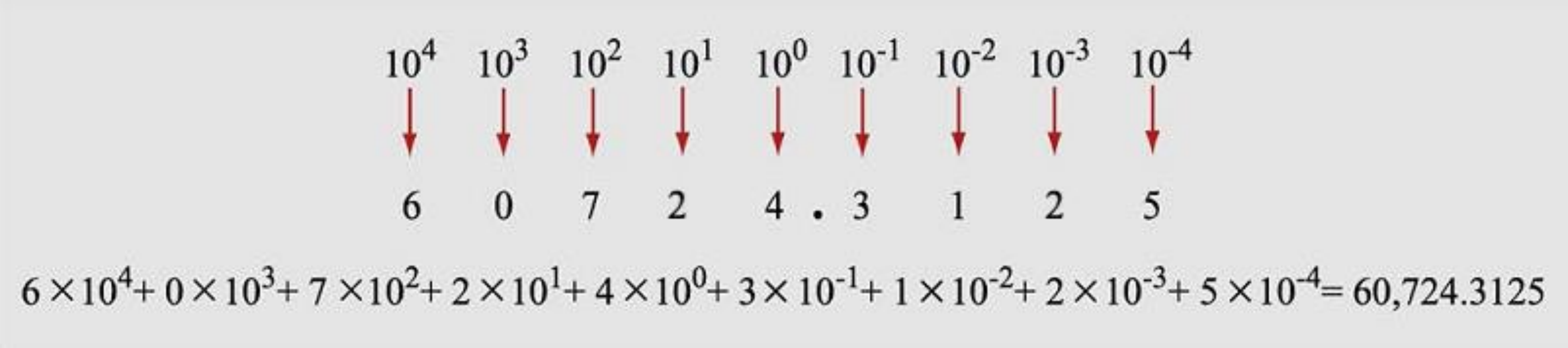
$$f(x,y)=\sum_{k=0}^{\infty}\frac{1}{k!}\left[(x-x_0)\frac{\partial}{\partial x}+(y-y_0)\frac{\partial}{\partial y}\right]^k f(x_0,y_0)$$

Representation of Numbers on a Computer

Decimal and Binary Representation

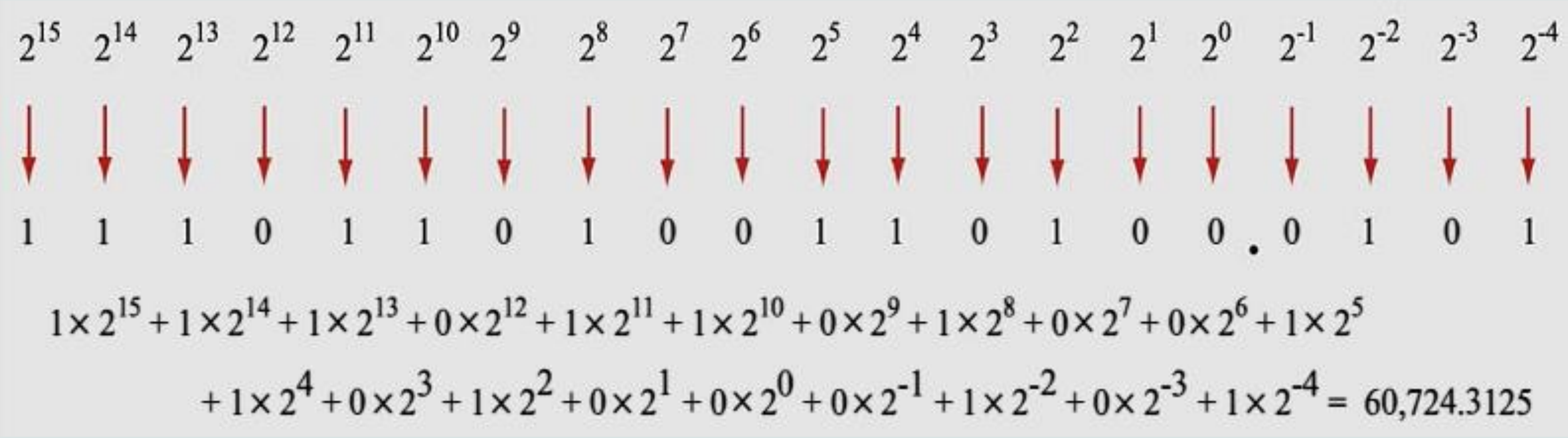
- Numbers can be represented in various forms. The familiar decimal system (base 10) uses ten digits 0, 1, ..., 9.
- A number is written by a sequence of digits that correspond to multiples of powers of 10.

Decimal



$$6 \times 10^4 + 0 \times 10^3 + 7 \times 10^2 + 2 \times 10^1 + 4 \times 10^0 + 3 \times 10^{-1} + 1 \times 10^{-2} + 2 \times 10^{-3} + 5 \times 10^{-4} = 60{,}724.3125$$

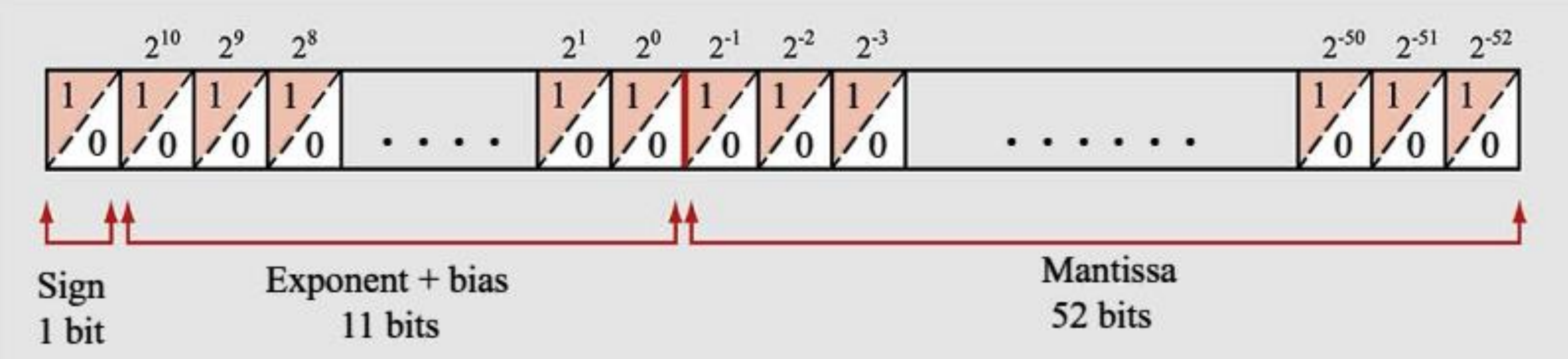
Binary



$$1 \times 2^{15} + 1 \times 2^{14} + 1 \times 2^{13} + 0 \times 2^{12} + 1 \times 2^{11} + 1 \times 2^{10} + 0 \times 2^9 + 1 \times 2^8 + 0 \times 2^7 + 0 \times 2^6 + 1 \times 2^5$$
$$+ 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} = 60{,}724.3125$$

Representation of Numbers on a Computer

- A 64-bit (binary digit) representation is used for a real number.
- The first bit is a **sign** indicator, denoted **s**.
- This is followed by an 11-bit exponent, **c**, called the **characteristic**, and a 52-bit binary fraction, **f**, called the **mantissa**.
- The base for the exponent is 2.

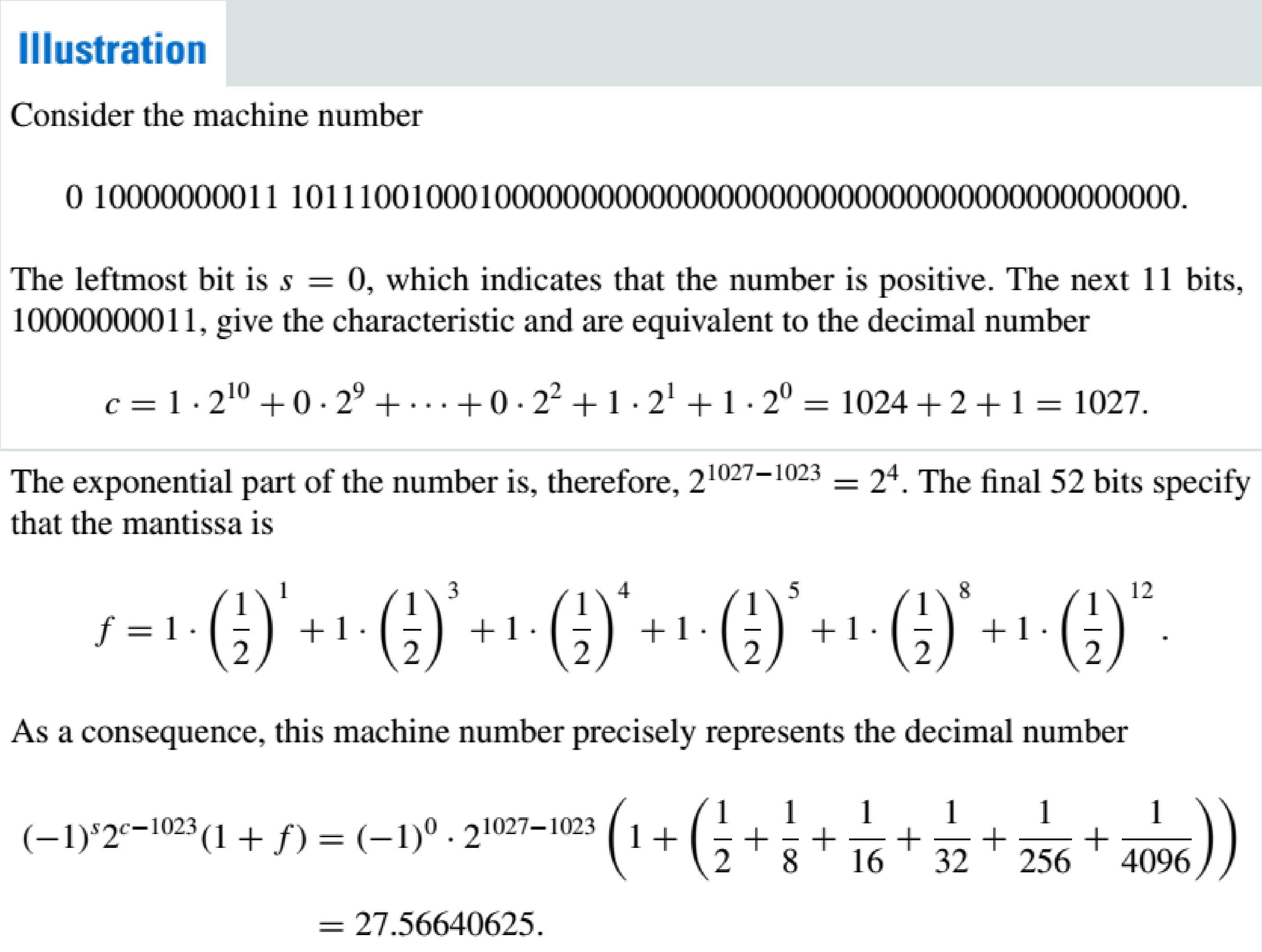


- Since 52 binary digits correspond to between 16 and 17 decimal digits, we can assume that a number represented in this system has at least 16 decimal digits of precision.
- The exponent of 11 binary digits gives a range of 0 to 2^11 – 1 = 2047.
- For a 32-bit representation, s=1, c=8, and f=23.

Approximations and Errors

- To save storage and provide a unique representation for each floating-point number, a normalization is imposed.
- IEEE Binary Floating Point Arithmetic Standard 754-2008.
- Using this system gives a floating-point number of the form

$$(-1)^s 2^{c-1023}(1+f).$$



Illustration

Consider the machine number

$$0\ 10000000011\ 1011100100010000000000000000000000000000000000000000.$$

The leftmost bit is $s = 0$, which indicates that the number is positive. The next 11 bits, 10000000011, give the characteristic and are equivalent to the decimal number

$$c = 1 \cdot 2^{10} + 0 \cdot 2^9 + \cdots + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 1024 + 2 + 1 = 1027.$$

The exponential part of the number is, therefore, $2^{1027-1023} = 2^4$. The final 52 bits specify that the mantissa is

$$f = 1 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^3 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^5 + 1 \cdot \left(\frac{1}{2}\right)^8 + 1 \cdot \left(\frac{1}{2}\right)^{12}.$$

As a consequence, this machine number precisely represents the decimal number

$$(-1)^s 2^{c-1023}(1+f) = (-1)^0 \cdot 2^{1027-1023}\left(1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096}\right)\right)$$

$$= 27.56640625.$$

Decimal Machine Numbers

- The machine numbers are represented in the normalized decimal floating-point form

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

for each i = 2, . . . , 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 common ways** of performing this termination.

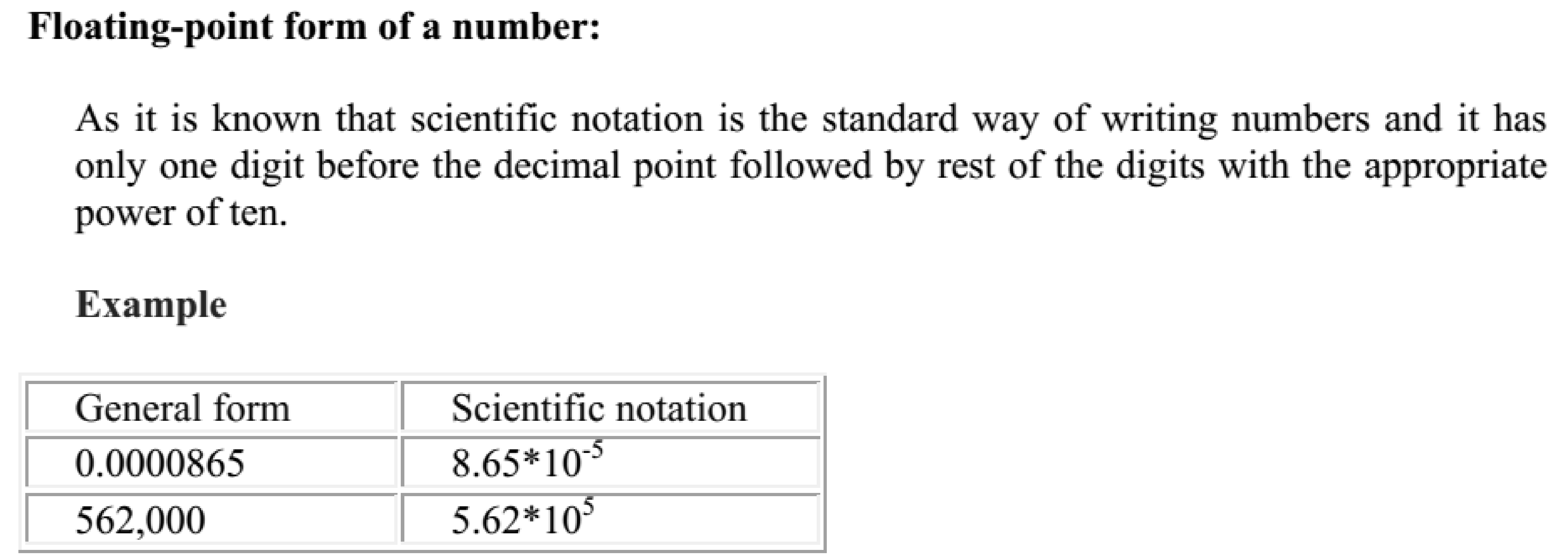
Chopping
It is to simply chop off the digits d(k+1)d(k+2) . . . . This produces the floating-point form

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

Rounding
It is added 5 × 10^[n−(k+1)] to y and then chops the result to obtain a number of the form

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

Floating-Point Form of a Number



Floating-point form of a number:

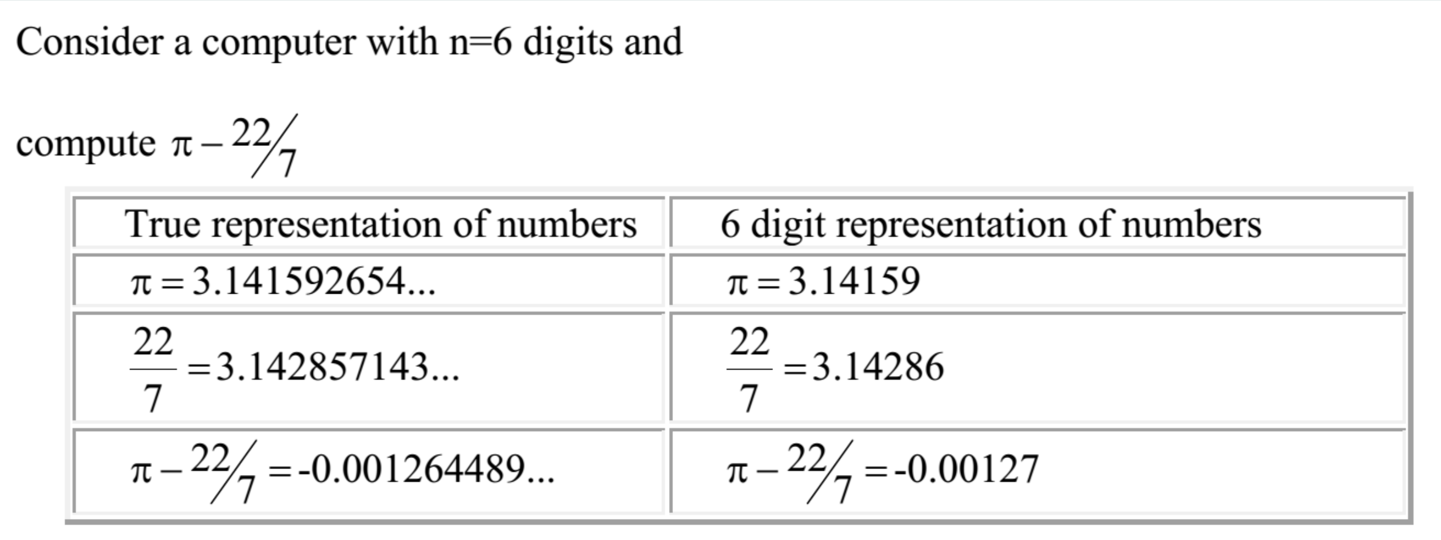
As it is known that scientific notation is the standard way of writing numbers and it has only one digit before the decimal point followed by rest of the digits with the appropriate power of ten.

Example

General form	Scientific notation
0.0000865	$8.65*10^{-5}$
562,000	$5.62*10^{5}$

Significant Figures (Digits)

- **Significant digits** of a number are those that can be used with confidence and designate the reliability of numerical value.
- All the following have 4 significant digits.
  - 0.00001987
  - 0.001987
  - 1.987
  - 1900
- Thus, zeros depending on their roles in the number may or may not be counted as significant digits.



Consider a computer with n=6 digits and

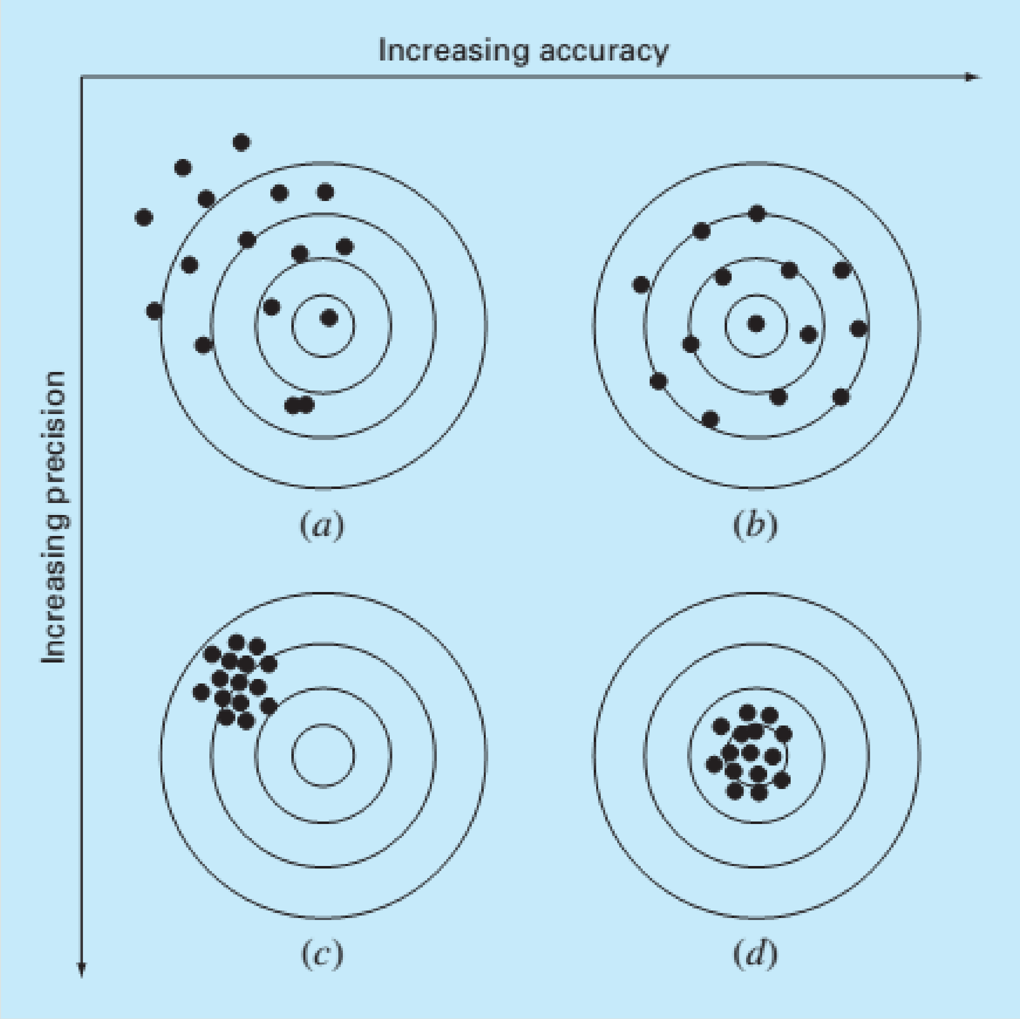
compute $\pi - {}^{22}\!/_{7}$

True representation of numbers	6 digit representation of numbers
$\pi = 3.141592654...$	$\pi = 3.14159$
$\frac{22}{7} = 3.142857143...$	$\frac{22}{7} = 3.14286$
$\pi - {}^{22}\!/_{7} = -0.001264489...$	$\pi - {}^{22}\!/_{7} = -0.00127$

Accuracy and Precision

- **Accuracy** refers to how closely a computed or measured value agrees with the true value.
  - **Inaccuracy** (also called bias) is defined as systematic deviation from the truth.

- **Precision** refers to how closely individual computed or measured values agree with each other.
  - **Imprecision** (also called uncertainty), on the other hand, refers to the magnitude of the scatter.



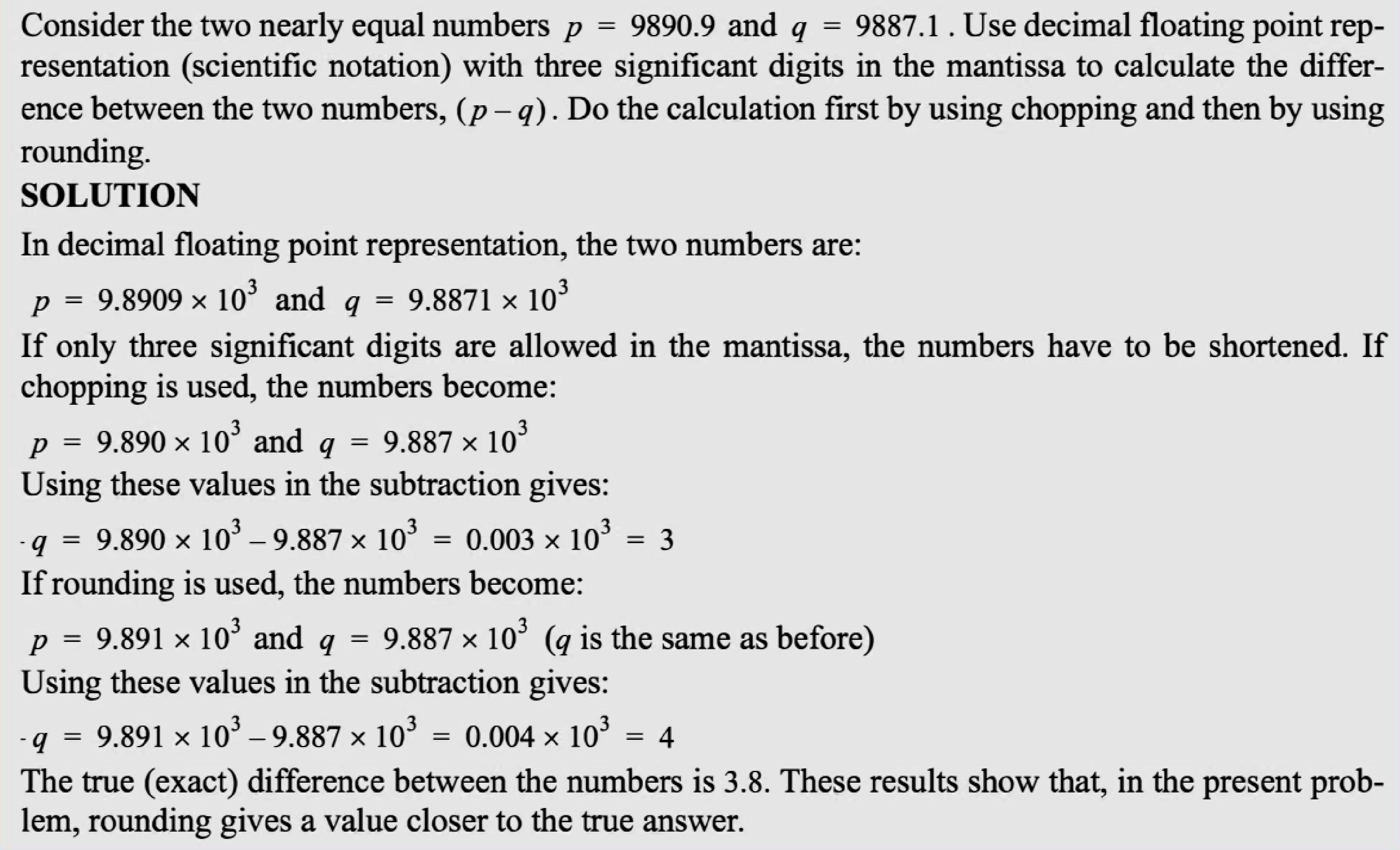
Sources of Errors

- Numerical solutions can be very accurate but in general are not exact.
- There are mainly **two kinds of errors** due to numerical methods.

  - **Round-Off Errors**
    - Round-off error is due to the fact that computers can represent only quantities with a finite number of digits.
  - **Truncation Errors**
    - Truncation error is the discrepancy introduced by the fact that numerical methods may employ approximations to represent exact mathematical operations and quantities.

- We briefly discuss errors not directly connected with the numerical methods themselves.
- The other sources of errors may be:
  - Mathematical modeling of a physical promlem
  - Uncertainity in physical data (measurement errors)
  - Programming errors (blunders)

Round-Off Errors



Consider the two nearly equal numbers $p = 9890.9$ and $q = 9887.1$. Use decimal floating point representation (scientific notation) with three significant digits in the mantissa to calculate the difference between the two numbers, $(p - q)$. Do the calculation first by using chopping and then by using rounding.

SOLUTION

In decimal floating point representation, the two numbers are:

$p = 9.8909 \times 10^3$ and $q = 9.8871 \times 10^3$

If only three significant digits are allowed in the mantissa, the numbers have to be shortened. If chopping is used, the numbers become:

$p = 9.890 \times 10^3$ and $q = 9.887 \times 10^3$

Using these values in the subtraction gives:

$- q = 9.890 \times 10^3 - 9.887 \times 10^3 = 0.003 \times 10^3 = 3$

If rounding is used, the numbers become:

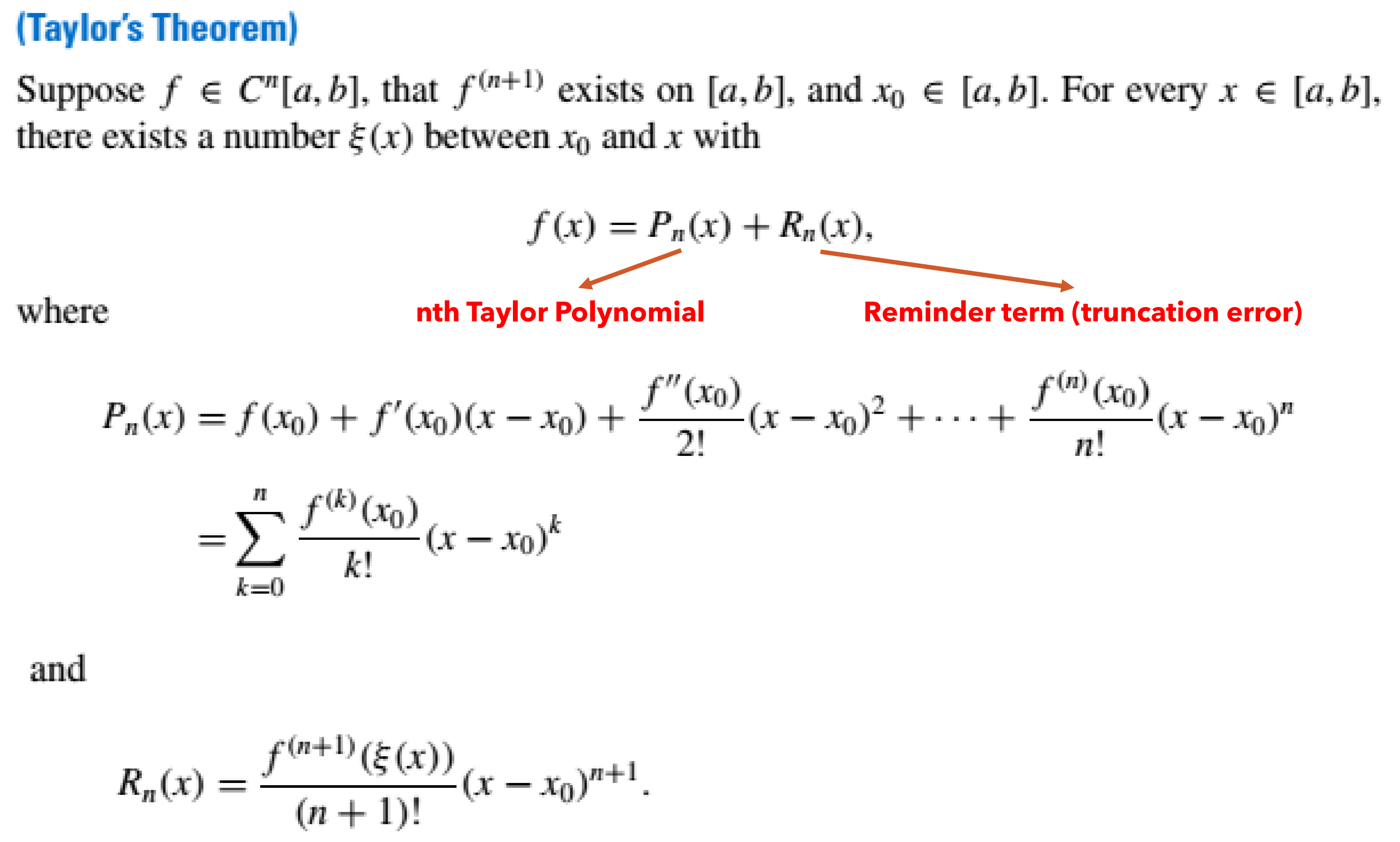
$p = 9.891 \times 10^3$ and $q = 9.887 \times 10^3$ (q is the same as before)

Using these values in the subtraction gives:

$- q = 9.891 \times 10^3 - 9.887 \times 10^3 = 0.004 \times 10^3 = 4$

The true (exact) difference between the numbers is 3.8. These results show that, in the present problem, rounding gives a value closer to the true answer.

Truncation Errors



(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

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

nth Taylor Polynomial

Reminder term (truncation error)

where

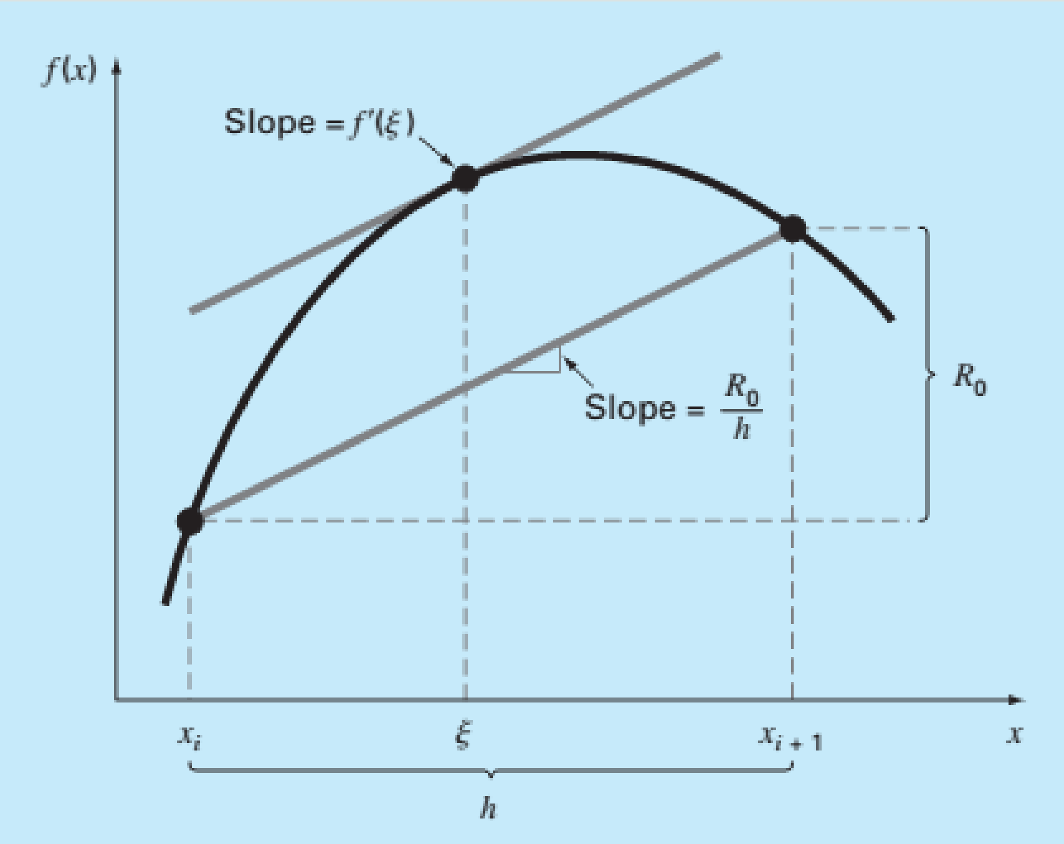
$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

$$= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

and

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

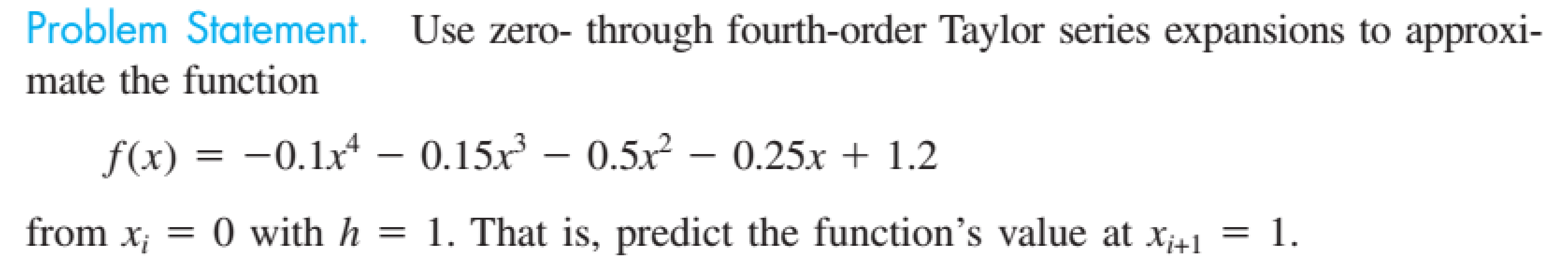
Truncation Errors



$$\frac{R_1}{t_{i+1} - t_i} = O(t_{i+1} - t_i)$$

- The error of our derivative approximation should be proportional to the step size.
- If we halve the step size, we would expect to halve the error of the derivative.

Approximations and Errors



Problem Statement. Use zero- through fourth-order Taylor series expansions to approximate the function

$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$

from $x_i = 0$ with $h = 1$. That is, predict the function's value at $x_{i+1} = 1$.

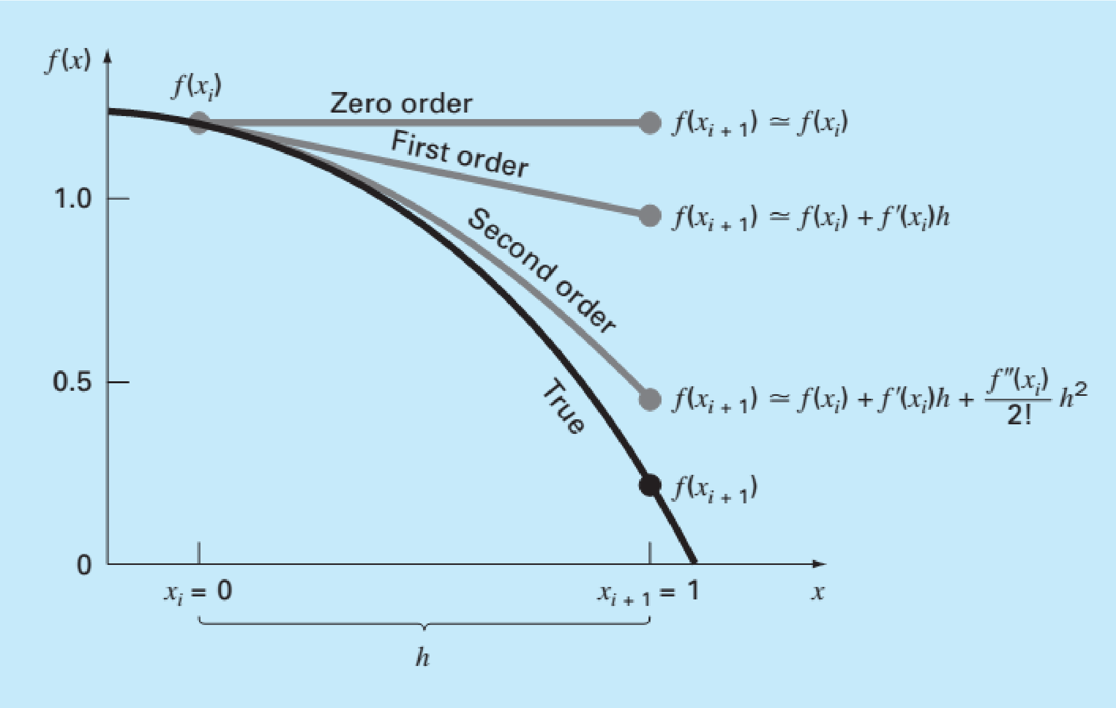


FIGURE 4.1
The approximation of $f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$ at $x = 1$ by zero-order, first-order, and second-order Taylor series expansions.

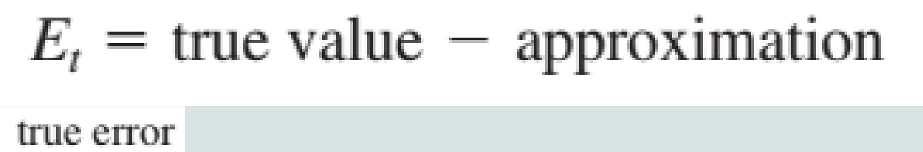
True Error

- **Numerical errors** arise from the use of approximations to represent exact mathematical operations and quantities.

- These include **truncation errors**, which result when approximations are used to represent exact mathematical procedures, and **round-off errors**, which result when numbers having limited significant figures are used to represent exact numbers.

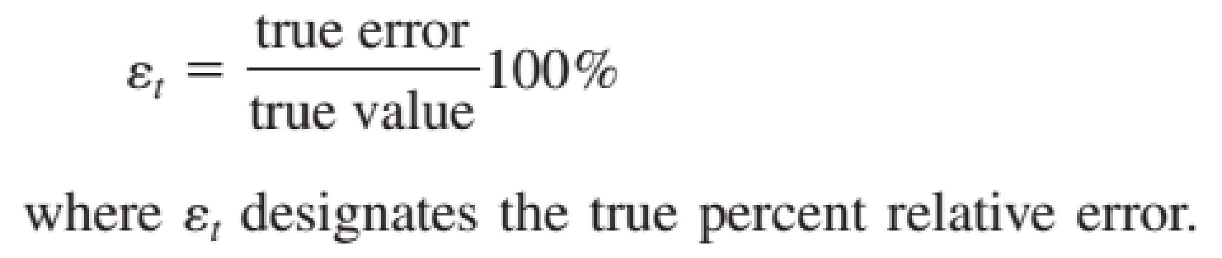
- For both types, the relationship between the exact, or true, result and the approximation can be formulated as

$$\text{True value} = \text{approximation} + \text{error}$$



$$E_t = \text{true value} - \text{approximation}$$

true error



$$\varepsilon_t = \frac{\text{true error}}{\text{true value}} 100\%$$

where ε_t designates the true percent relative error.

Approximate Error

- In actual situations such information is rarely available.
- For numerical methods, the true value will be known only when we deal with functions that can be solved analytically.
- In real-world applications, we will obviously not know the true answer a priori. For these situations, an alternative is to normalize the error using the best available estimate of the true value, that is, to the approximation itself, as in

$$\varepsilon_a = \frac{\text{approximate error}}{\text{approximation}} 100\%$$

- One of the challenges of numerical methods is to determine error estimates in the absence of knowledge regarding the true value.
- For example, certain numerical methods use an iterative approach to compute answers.
- In such an approach, a present approximation is made on the basis of a previous approximation.

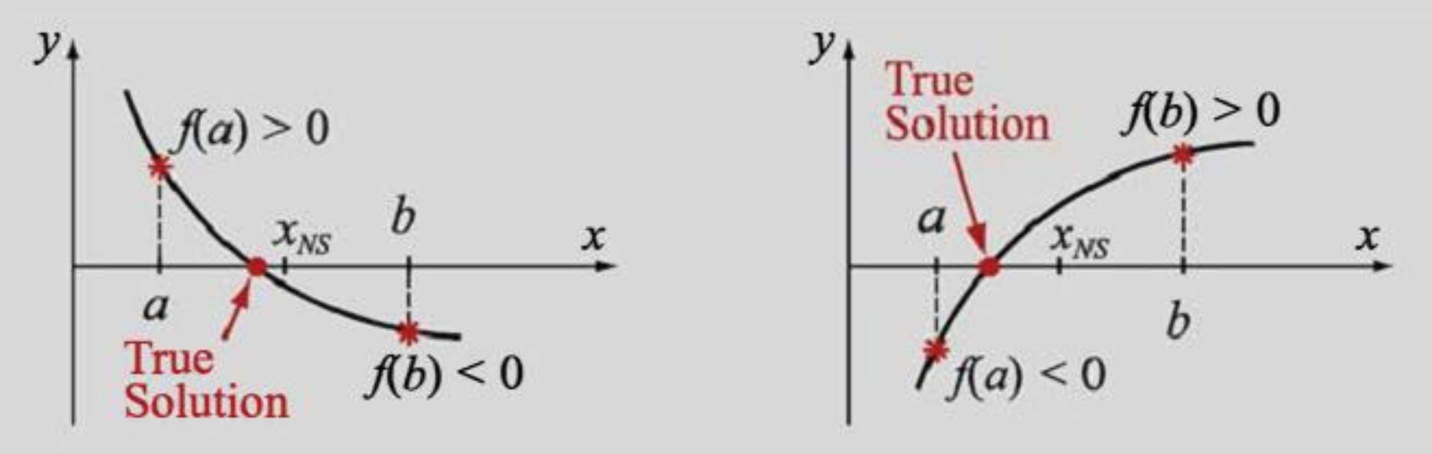
$$\varepsilon_a = \frac{\text{current approximation} - \text{previous approximation}}{\text{current approximation}} 100\%$$

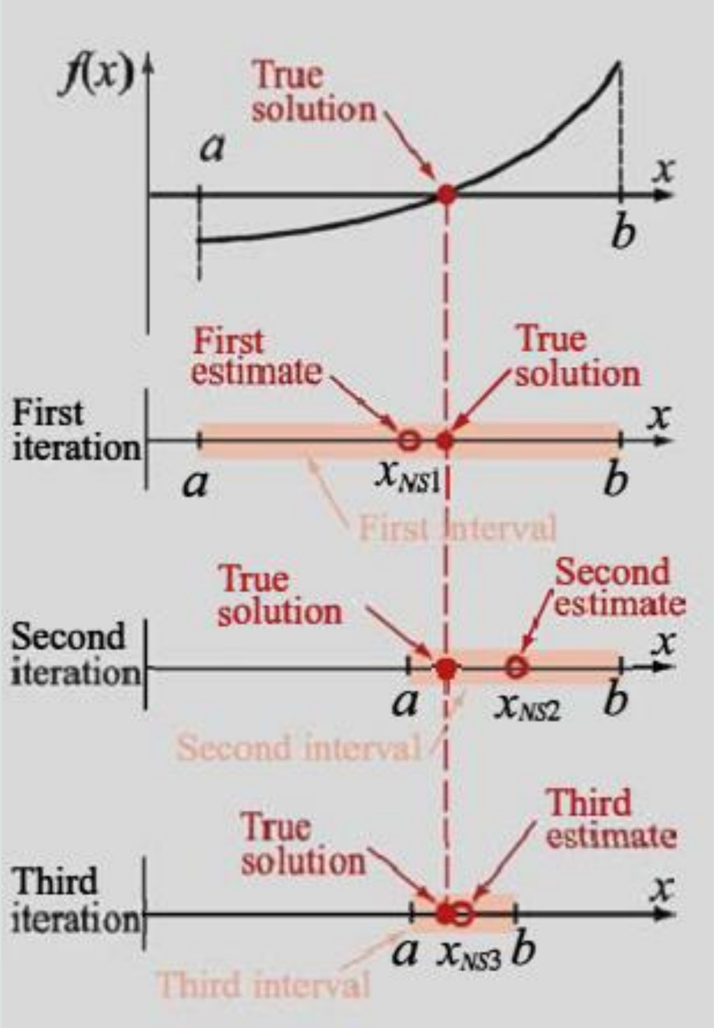
ES 361
COMPUTING METHODS IN ENGINEERING

Chapter 2
Solution of Nonlinear Equations of Single Variable

Bisection Method

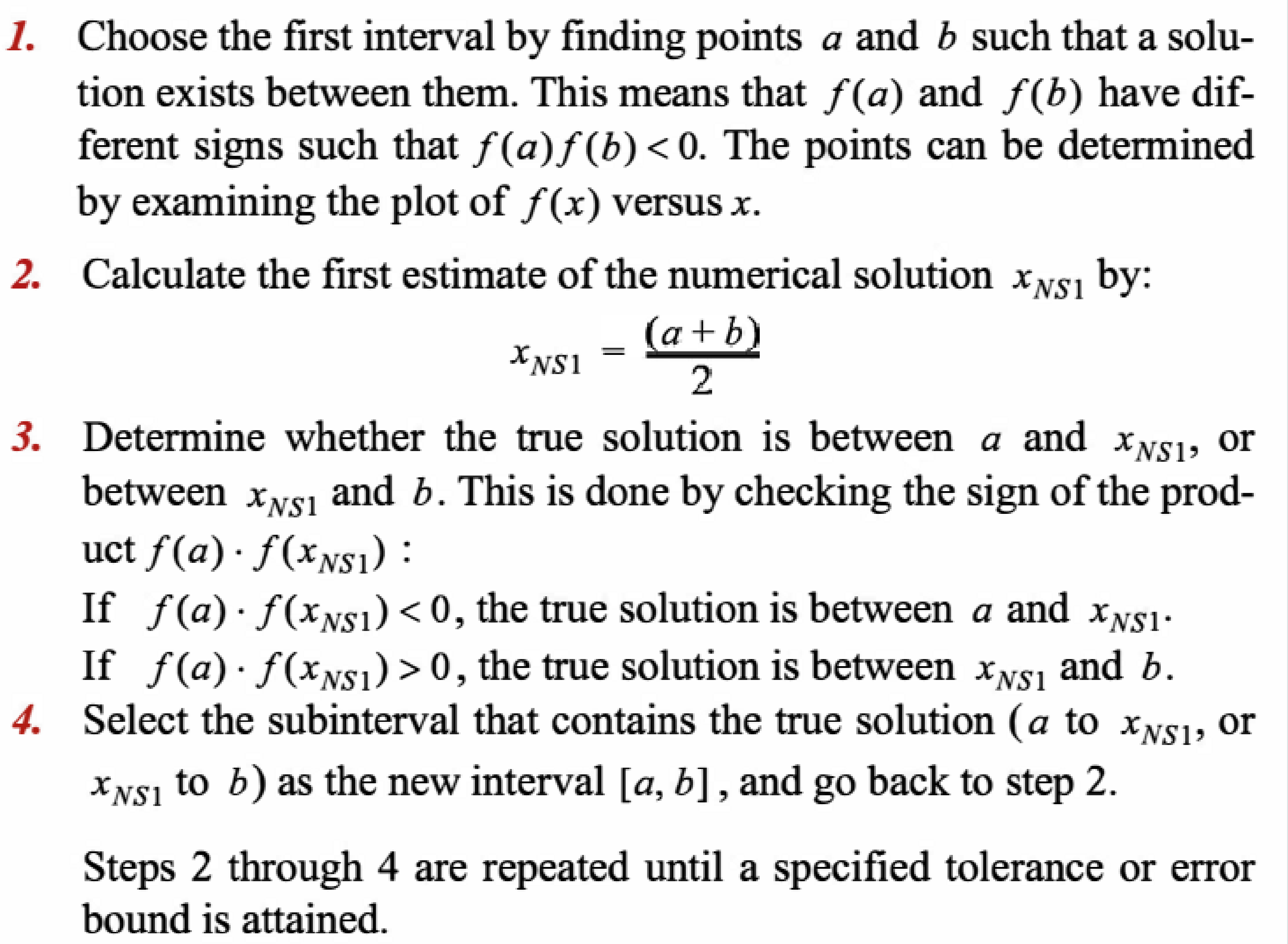
- The bisection method is a bracketing method of finding a numerical **solution of an equation of the form f(x) =0** when it is known that within a **given interval [a, b].**
- f(x) is continuous and the equation has a solution.





Bisection Method

Algorithm for the Bisection Method



1. Choose the first interval by finding points a and b such that a solution exists between them. This means that $f(a)$ and $f(b)$ have different signs such that $f(a)f(b) < 0$. The points can be determined by examining the plot of $f(x)$ versus x.

2. Calculate the first estimate of the numerical solution x_{NS1} by:

$$x_{NS1} = \frac{(a+b)}{2}$$

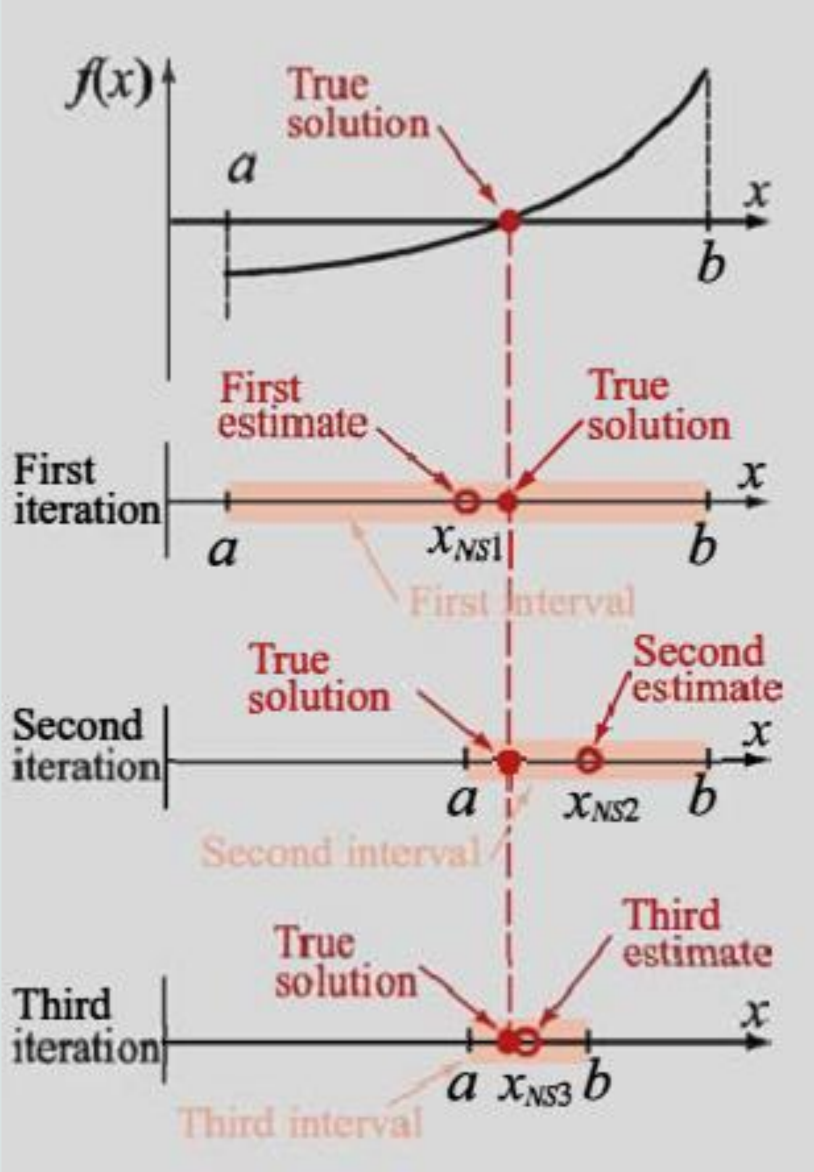
3. Determine whether the true solution is between a and x_{NS1}, or between x_{NS1} and b. This is done by checking the sign of the product $f(a) \cdot f(x_{NS1})$:

If $f(a) \cdot f(x_{NS1}) < 0$, the true solution is between a and x_{NS1}.

If $f(a) \cdot f(x_{NS1}) > 0$, the true solution is between x_{NS1} and b.

4. Select the subinterval that contains the true solution (a to x_{NS1}, or x_{NS1} to b) as the new interval $[a, b]$, and go back to step 2.

Steps 2 through 4 are repeated until a specified tolerance or error bound is attained.



Bisection Method

Additional Notes on the Biseciton Method

- The method always converges to an answer, provided a root was trapped in the interval [a,b] to begin with.

- The method may fail when the function is tangent to the axis and does not cross the x-axis at f(x)=0.

- The method converges slowly relative to other methods.

QUESTIONS?

E-Mail: beratcan@metu.edu.tr

Room: MM Building, 9th Floor, No:906