MATH 4513 Numerical Analysis

Chapter 1. Mathematical Preliminaries

Xu Zhang

Assistant Professor

Department of Mathematics

Oklahoma State University

Text Book: Numerical Analysis (10th edition)

R. L. Burden, D. J. Faires, A. M. Burden

Chapter 1. Mathematical Preliminaries

Content

1.1.1 Limit and Continuity

Definition 1 (Limit).

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

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

if, given any real number $\epsilon > 0$, there exists a real number $\delta > 0$ such that

$$|f(x) - L| < \epsilon, \quad \text{whenever} \quad x \in X \quad \text{and} \quad 0 < |x - x_0| < \delta.$$

Example 2.

Find $\lim\limits_{x \to 2} f(x)$ where

$$f(x) = \begin{cases} x^2 - 1, & \text{if } x \neq 2, \\ 4, & \text{if } x = 2. \end{cases}$$

Answer: 3

Definition 3 (Continuity).

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

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

Furthermore, f is **continuous on the set X** if it is continuous at every number in X.

Example 4.

Is the function $f(x)$ continuous on $(0, \infty)$?

$$f(x) = \begin{cases} x^2 - 1, & \text{if } x \neq 2, \\ 4, & \text{if } x = 2. \end{cases}$$

Answer: no, f is discontinuous at $x = 2$.

1.1.2 Differentiability

Definition 5 (Differentiability).

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

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

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

Theorem 6.

If a function f is differentiable at x_0, then f is continuous at x_0.

Example 7.

"If f is continuous, it is differentiable?" Is this statement true?

Answer: no, e.g. $f(x) = |x|$.

Theorem 8 (Intermediate Value Theorem).

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

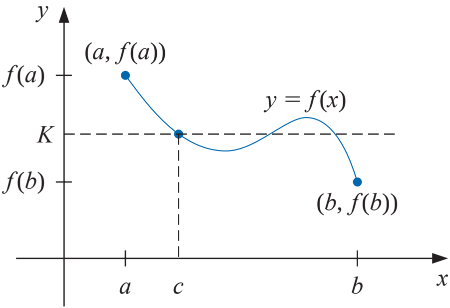

Figure: Intermediate Value Theorem

Example 9.

Show that $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a solution in the interval $[0, 1]$.

Proof.

- Let $f(x) = x^5 - 2x^3 + 3x^2 - 1$. It is clear that f is continuous on $[0, 1]$.
- Also note that

$$f(0) = -1 < 0 \quad \text{and} \quad f(1) = 1 > 0.$$

- By the IVT, there exists a number $c \in (0, 1)$ such that $f(c) = 0$.

Theorem 10 (Rolle's Theorem).

Suppose f is continuous on $[a, b]$ and is differentiable on (a, b). If $f(a) = f(b)$, then there exists a number $c \in (a, b)$ such that $f'(c) = 0$.

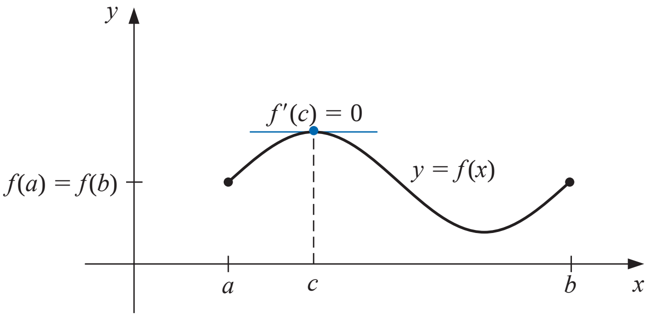

Figure: Rolle's Theorem

Theorem 11 (Mean Value Theorem).

Suppose f is continuous on $[a, b]$ and is differentiable on (a, b), then there exists a number $c \in (a, b)$ such that

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

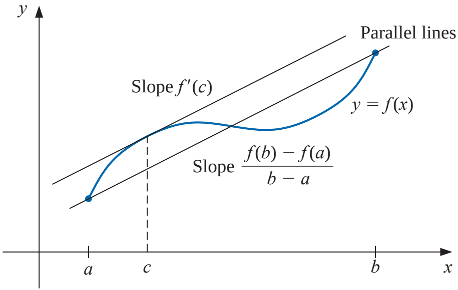

Figure: Mean Value Theorem

Theorem 12 (Extreme Value Theorem).

If f is continuous on $[a, b]$, then there exist $c_1, c_2 \in [a, b]$ such that

$$f(c_1) \leq f(x) \leq f(c_2), \quad \text{for all } x \in [a, b].$$

In addition, if f is differentiable on (a, b), then c_1 and c_2 occur either at the endpoints of $[a, b]$ or critical points (i.e., $f' = 0$) in (a, b).

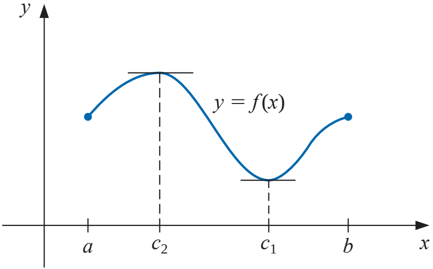

Figure: Extreme Value Theorem

Example 13.

Find the absolute minimum and absolute maximum values of

$$f(x) = \frac{x}{3 - x^2}$$

on the interval $[0, 1]$.

Solution.

- The derivative
$$f' = \frac{3 + x^2}{(3 - x^2)^2}$$
is continuous on $[0, 1]$.
- No critical points (verify) in this case. The absolute max/min value occur at endpoints ($x = 0, x = 1$).
- Note that $f(0) = 0$ and $f(1) = 1/2$. Hence,

$$\max_{x\in[0,1]} f(x) = \frac{1}{2}, \quad \min_{x\in[0,1]} f(x) = 0.$$

1.1.3 Integration

Definition 14 (Integral).

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

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

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

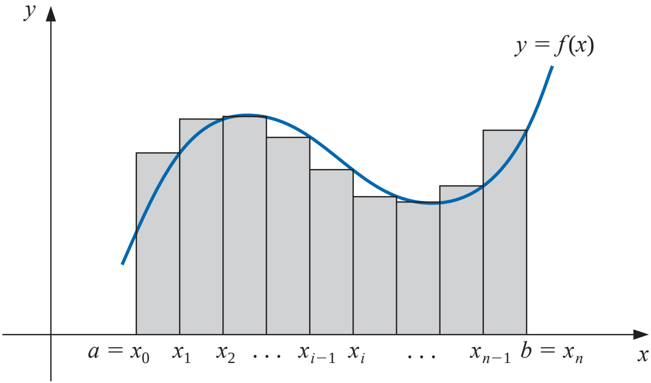

Figure: Definition of Integration

1.1.4 Taylor Polynomials

Theorem 15 (Taylor's Theorem).

Let $k \geq 1$ be an integer, and f be n times differentiable at the point a. Then

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

Here,

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

is called the n-th order Taylor polynomial of $f(x)$ at the point a, and

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

is called the remainder (or truncation error), where $\xi(x)$ is between a and x.

If x is close to a, then the remainder $R_n(x)$ is very small. In this case

$$f(x) \approx P_n(x).$$

Example 16.

Find the Taylor polynomials $P_2(x)$ and $P_3(x)$ for $f(x) = \sin(x)$ at the point 0.

Solution.

- Note that
$$f(x) = \sin(x),\ f'(x) = \cos(x),\ f''(x) = -\sin(x),\ f'''(x) = -\cos(x).$$
- Evaluating at $x_0 = 0$ yields
$$f(0) = 0,\ \ f'(0) = 1,\ \ f''(0) = 0,\ \ f'''(0) = -1.$$
- Thus,
$$P_2(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 = x.$$
$$P_3(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f'''(0)}{6}x^3 = x - \frac{x^3}{6}.$$

Section 1.2 Round-off Errors and Computer Arithmetic

Outline of this section

- The arithmetic performed by a computer is different from the arithmetic in calculus and algebra courses.
- For example, it is always true that

$$(\sqrt{2})^2 = 2.$$

In the traditional mathematics, the irrational number $\sqrt{2}$ is in fact a decimal with infinite number of digits

$$\sqrt{2} = 1.4142135623730950488016887242096980785696718753769...$$

- However, in the computational world, each number only has a fixed and finite number of digits such as

$$\sqrt{2} = 1.414213562373095$$

- Therefore, $(\sqrt{2})^2$ will not precisely equal 2 because of the finite precision of computers.
- The error that is produced when a computer is used to perform real-number calculations is called **round-off error**.

1.2.1 Binary Machine Numbers

Question: How do computers store numbers?

A 64-bit (binary digit) representation is used for a real number.

- The first bit, denoted as s, is a **sign indicator**.
- The next 11-bit exponent, denoted as c, is called the **characteristic**.
- The last 52-bit fraction, denoted as f, called the **mantissa**.

Remark

- The 52 binary digits correspond to between 16 and 17 decimal digits of precision ($2^{52} \approx 0.45 \times 10^{16}$). We can assume that a number represented in this system has at least 16 decimal digits of precision.
- The 11 binary digits gives a range of 0 to $2^{11} - 1 = 2047$. However, to ensure numbers with small magnitude are well representable, the range is adjusted to -1023 to 1024.
- Using this system gives a floating-point number of the form

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

Example 17.

Consider the 64-bit machine number

$0\ 10000000011\ 1011100100010000000000000000000000000000000000000000$

What is the decimal floating number?

Solution.

- Since $s = 0$, then $(-1)^s = 1$. The number is positive.
- The characteristic 10000000011 is 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 final 52-bit specify the mantissa

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

- Thus the decimal floating number is

$$\begin{aligned}(-1)^s 2^{c-1023}(1+f) &= (-1)^0 \cdot 2^{1027-1023}\left(1 + \frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096}\right) \\ &= 27.56640625\end{aligned}$$

1.2.2 Decimal Machine Numbers

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

$$\pm 0.d_1 d_2 \cdots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad 0 \le d_i \le 9 \text{ for } i = 2, \cdots, k.$$

This form is called **k-digit decimal machine numbers**.

- Any positive real number

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

can be converted into a floating-point form, denoted as $fl(y)$, by terminating the mantissa of y at k decimal digits.

There are two ways to do it.

- **Chopping**: simply chop off the digits $d_{k+1}d_{k+2}\dots$.

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

- **Rounding**: add $5 \times 10^{n-(k+1)}$ to y and then chop the result.
  - when $d_{k+1} \geq 5$, add 1 to d_k, that is **round up**.
  - when $d_{k+1} < 5$, chop off the digits $d_{k+1}d_{k+2}\dots$, that is **round down**.

Example 18.

Determine the five-digit (a) chopping and (b) rounding values of the irrational number $\pi = 3.1415926\ldots$.

Solution.

Write π in normalized decimal form

$$\pi = 0.31415926\ldots \times 10^1$$

- The floating-point form of π using **five-digit chopping** is
$$fl(\pi) = 0.31415\ldots \times 10^1 = 3.1415$$
- The floating-point form of π using **five-digit rounding** is
$$fl(\pi) = 0.31416\ldots \times 10^1 = 3.1416$$

Definition 19 (Approximation Error).

Suppose p^* is an approximation to p. The following errors are often used.

- the **actual error**: $p - p^*$.
- the **absolute error**: $|p - p^*|$.
- the **relative error**: $\dfrac{|p - p^*|}{|p|}$, if $p \neq 0$.

Example 20.

Find the actual, absolute, and relative errors when approximating p by p^*

1. $p = 0.3000 \times 10^1$, and $p^* = 0.31 \times 10^1$.
2. $p = 0.3000 \times 10^{-3}$, and $p^* = 0.31 \times 10^{-3}$.
3. $p = 0.3000 \times 10^4$, and $p^* = 0.31 \times 10^4$.

Solution.

No.	$p - p^*$	$\|p - p^*\|$	$\dfrac{\|p - p^*\|}{\|p\|}$
1	-0.1	0.1	$0.0333\bar{3}$
2	-0.00001	0.00001	$0.0333\bar{3}$
3	-100	100	$0.0333\bar{3}$

Remark

- The example shows that the relative errors are the same, although the absolute errors are widely varying.
- As a measure of accuracy, the absolute error can be misleading, and the relative error is more meaningful because the relative error takes into account the size of the value.

1.2.3 Finite-Digit Arithmetic

- In addition to inaccurate representation of numbers, the arithmetic performed in a computer is not exact.
- The symbols $\oplus$, $\ominus$, $\otimes$, and $\oslash$ are used to represent machine addition, subtraction, multiplication, and division, respectively.
- A finite-digit arithmetic is given by

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \ominus y = fl(fl(x) - fl(y)),$$
$$x \otimes y = fl(fl(x) \times fl(y)), \quad x \oslash y = fl(fl(x) \div fl(y)).$$

Example 21.

Use 3-digit chopping arithmetic to compute $\frac{1}{3} \oplus \frac{7}{6}$.

$$\frac{1}{3} \oplus \frac{7}{6} = fl\left(fl(\frac{1}{3}) + fl(\frac{7}{6})\right) = fl(0.333 + 1.16) = fl(1.493) = 1.49$$

The relative error is $\dfrac{|1.49 - 1.5|}{1.5} = 0.667\%$

Example 22.

Let $p = 0.54617$, $q = 0.54601$. Use four-digit arithmetic to approximate $p + q$ and $p - q$ and determine the absolute error and relative error using (i) chopping and (ii) rounding.

Solution (1/2).

The true sum and difference are $s = p + q = 1.09218$, and $d = p - q = 0.00016$.

- Using four-digit chopping arithmetic,

$$\begin{aligned} s^* = p \oplus q &= fl(fl(p) + fl(q)) = fl(0.5461 + 0.5460) \\ &= fl(0.10921 \times 10^1) = 0.1092 \times 10^1 \end{aligned}$$

Abs. Error $|s - s^*| = 0.00018$, Rel. Error $\dfrac{|s - s^*|}{|s|} = 0.0001648.$

$$\begin{aligned} d^* = p \ominus q &= fl(fl(p) - fl(q)) = fl(0.5461 - 0.5460) \\ &= fl(0.1 \times 10^{-3}) = 0.1 \times 10^{-3} \end{aligned}$$

Abs. Error $|d - d^*| = 0.00006$, Rel. Error $\dfrac{|d - d^*|}{|d|} = 0.375.$

Solution (2/2).

- Using four-digit rounding arithmetic,

$$\begin{aligned} s^* = p \oplus q &= fl(fl(p) + fl(q)) = fl(0.5462 + 0.5460) \\ &= fl(0.10922 \times 10^1) = 0.1092 \times 10^1 \end{aligned}$$

Abs. Error $|s - s^*| = 0.00018$, Rel. Error $\dfrac{|s - s^*|}{|s|} = 0.0001648.$

$$\begin{aligned} d^* = p \ominus q &= fl(fl(p) - fl(q)) = fl(0.5462 - 0.5460) \\ &= fl(0.2 \times 10^{-3}) = 0.2 \times 10^{-3} \end{aligned}$$

Abs. Error $|d - d^*| = 0.00004$, Rel. Error $\dfrac{|d - d^*|}{|d|} = 0.25.$

Remark

As shown in the example above, one of the most common error-producing calculations involves Cancelation of significant digits due to the subtraction of nearly equal numbers.

Example 23.

The quadratic formula of the roots of $ax^2 + bx + c = 0$ is

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

We apply these formulas for solving the equation $x^2 + 62.10x + 1 = 0$, whose roots are approximately

$$x_1 = -0.01610723, \quad x_2 = -62.08390.$$

We will use four-digit rounding arithmetic in the calculation.

Solution (1/2)

$$\sqrt{b^2-4ac} = \sqrt{(62.10)^2-(4.000)(1.000)(1.000)}$$
$$= \sqrt{3856-4.000} = \sqrt{3852} = 62.06$$

Therefore

$$fl(x_1) = \frac{-62.10+62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000$$

The relative error is

$$\frac{|x_1 - fl(x_1)|}{|x_1|} \approx 0.24$$

Using the same approach, we found

$$\frac{|x_2 - fl(x_2)|}{|x_2|} \approx 0.00032.$$

The large relative error for x_1 is again because of subtracting two nearly equal number $\sqrt{b^2-4ac}$ and b in computing x_1.

Solution (2/2)

A remedy for this is to modify the root formula as follows

$$\begin{aligned} x_1 &= \frac{-b+\sqrt{b^2-4ac}}{2a}\cdot\left(\frac{-b-\sqrt{b^2-4ac}}{-b-\sqrt{b^2-4ac}}\right) \\ &= \frac{b^2-(b^2-4ac)}{2a(-b-\sqrt{b^2-4ac})} = \frac{-2c}{b+\sqrt{b^2-4ac}}. \end{aligned}$$

Using this formula, we have

$$fl(x_1) = \frac{-2.000}{62.10+62.06} = -0.01610.$$

Now the relative error of x_1 is

$$\frac{|x_1 - fl(x_1)|}{|x_1|} \approx 0.00062.$$

This is much more accurate than the previous approximation of x_1.

1.2.4 Nested Arithmetic

Accuracy loss due to round-off error can also be reduced by rearranging calculations, as shown in the next example.

Example 24.

Evaluate $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ at $x = 4.71$ using three-digit arithmetic.

Solution (1/3)

- Let us first find x^2 using 3-digit rounding arithmetic.

$$x \odot x = fl(4.71 \times 4.71) = fl(22.1841) = 22.2$$

- Then x^3 and $6.1x^2$ can be found by

$$x \odot (x \odot x) = fl(4.71 \times 22.2) = fl(104.562) = 105$$
$$6.1 \odot (x \odot x) = fl(6.1 \times 22.2) = fl(135.42) = 135$$

Solution (2/3)

- Using the three-digit arithmetic, we have the following results

	x	x^2	x^3	$6.1x^2$	$3.2x$
rounding	4.71	22.2	105.	135.	15.1
chopping	4.71	22.1	104.	134.	15.0
Exact	4.71	22.1841	104.487111	135.32301	15.072

- The exact value of $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ is

$$f(4.71) = 104.487111 - 135.32301 + 15.072 + 1.5 = -14.263899.$$

- Using the 3-digit chopping arithmetic, we have

$$f(4.71) = fl(fl(fl(104. - 134.) + 15.0) + 1.5) = -13.5, \quad \text{(verify)}$$

- Using the 3-digit rounding arithmetic we have

$$f(4.71) = fl(fl(fl(105. - 135.) + 15.1) + 1.5) = -13.4. \quad \text{(verify)}$$

Solution (3/3)

The relative error for these three-digit methods are

- Chopping:
$$\left|\frac{-14.263899+13.5}{-14.263899}\right| \approx 0.05$$
- Rounding:
$$\left|\frac{-14.263899+13.4}{-14.263899}\right| \approx 0.06$$

Remark

You should carefully verify these steps to make sure your notion of finite-digit arithmetic is correct.

- If we write the polynomial in the **nested** manner as

$$\begin{aligned} f(x) &= x^3 - 6.1x^2 + 3.2x + 1.5 \\ &= ((x - 6.1)x + 3.2)x + 1.5. \end{aligned}$$

- Then, by 3-digit chopping arithmetic we have

$$\begin{aligned} f(4.71) &= ((4.71 \ominus 6.1) \otimes 4.71 \oplus 3.2) \otimes 4.71 \oplus 1.5 \\ &= (-1.39 \otimes 4.71 \oplus 3.2) \otimes 4.71 \oplus 1.5 \\ &= (-6.54 \oplus 3.2) \otimes 4.71 \oplus 1.5 \\ &= -3.34 \otimes 4.71 \oplus 1.5 \\ &= -15.7 \oplus 1.5 \\ &= -14.2 \end{aligned}$$

- The relative error is reduced to

$$\frac{|-14.2639 + 14.2|}{14.2639} \approx 0.0045 \quad \text{(error is } 0.05 \text{ w/o nesting).}$$

- Similarly, using 3-digit rounding we have

$$
\begin{aligned}
f(4.71) &= ((4.71 \ominus 6.1) \otimes 4.71 \oplus 3.2) \otimes 4.71 \oplus 1.5 \\
&= (-1.39 \otimes 4.71 \oplus 3.2) \otimes 4.71 \oplus 1.5 \\
&= (-6.55 \oplus 3.2) \otimes 4.71 \oplus 1.5 \\
&= -3.35 \otimes 4.71 \oplus 1.5 \\
&= -15.8 \oplus 1.5 \\
&= -14.3
\end{aligned}
$$

- The relative error is $\dfrac{|-14.2639 + 14.3|}{14.2639} \approx 0.0025$. (The error is 0.06 w/o nesting).

Remark

In this example, we need 4 multiplications and 3 additions for standard evaluation, but only 2 multiplications and 3 additions in nested format.

- In general a n-th order polynomial

$$p_n(x) = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$$

can be re-written in nested format

$$p_n(x) = (\cdots((a_n x + a_{n-1})x + a_{n-2})x \cdots + a_1)x + a_0$$

- The former requires n additions and $2n-1$ multiplication. The latter requires n additions and n multiplications

Remark

Polynomials should always be expressed in nested form before performing an evaluation because this form minimizes the number of arithmetic calculations.