Preliminaries

Minping Wan

Tel: 0755 8801 8278
Email: wanmp@sustc.edu.cn

Preliminaries

- Review of Calculus

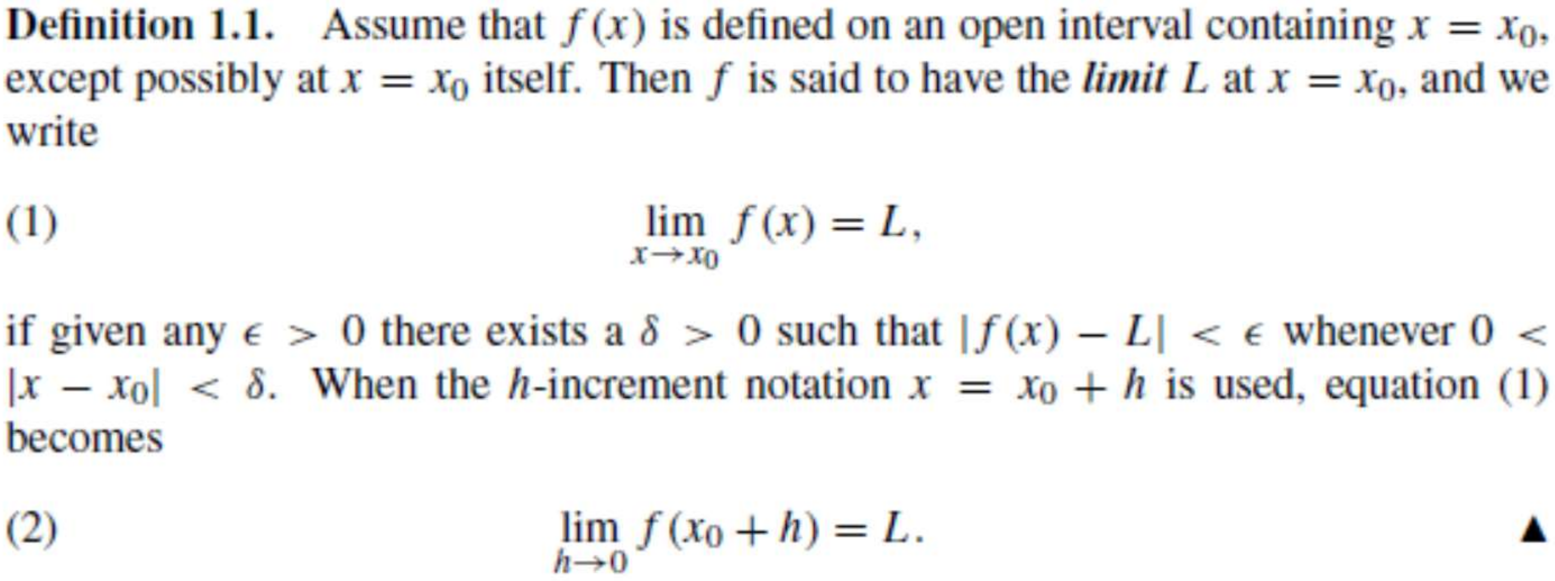
- Binary Numbers

- Error Analysis

- Vectors and Matrix

Review of Calculus

Limits and Continuity



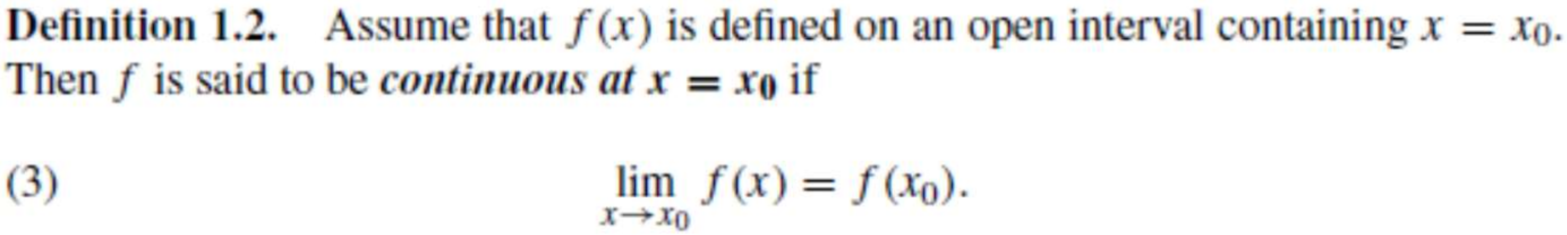
Definition 1.1. Assume that $f(x)$ is defined on an open interval containing $x = x_0$, except possibly at $x = x_0$ itself. Then f is said to have the ***limit*** L at $x = x_0$, and we write

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

if given any $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - x_0| < \delta$. When the h-increment notation $x = x_0 + h$ is used, equation (1) becomes

$$\lim_{h \to 0} f(x_0 + h) = L. \tag{2}$$

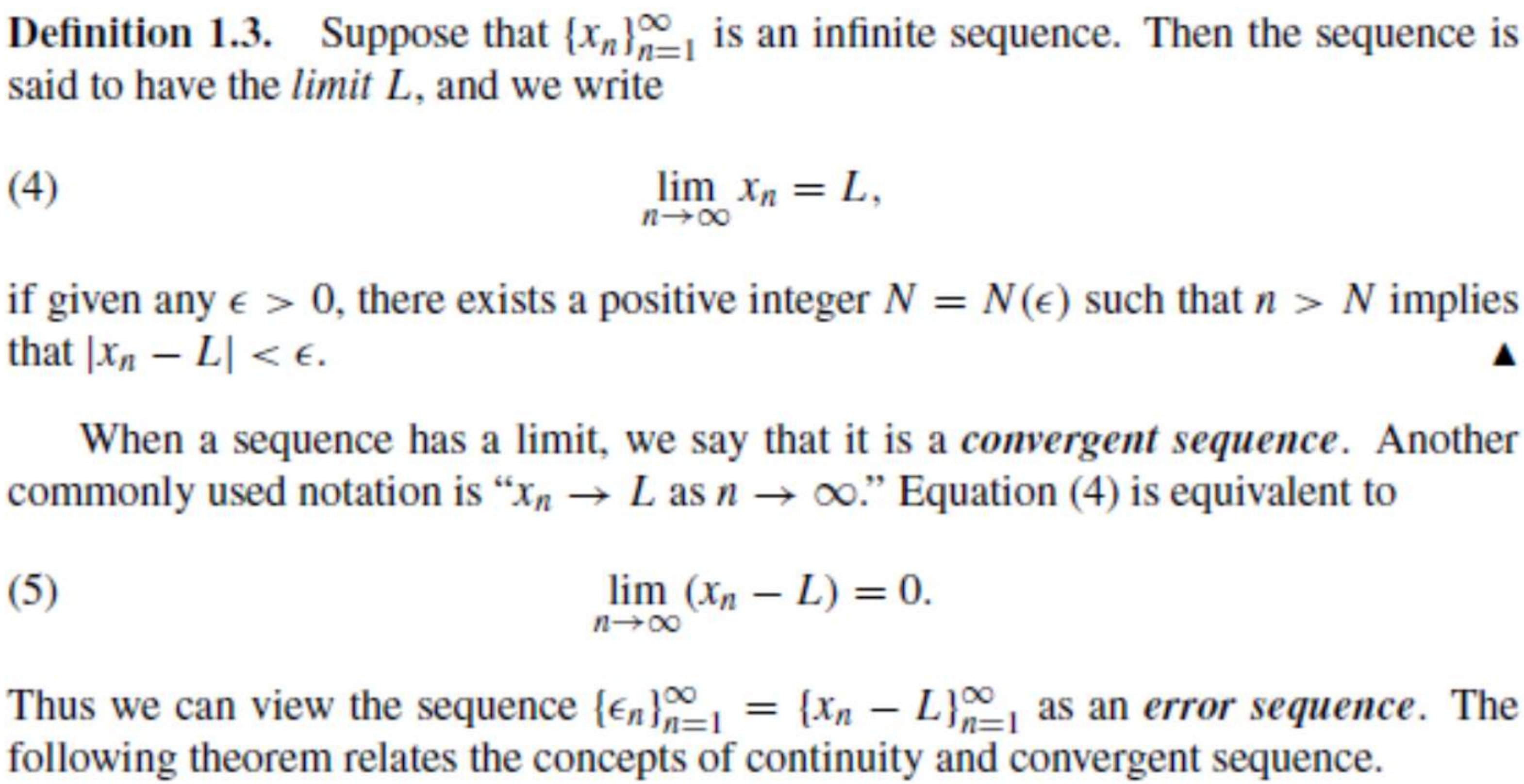
▲



Definition 1.2. Assume that $f(x)$ is defined on an open interval containing $x = x_0$. Then f is said to be ***continuous at*** $\boldsymbol{x = x_0}$ if

$$\lim_{x \to x_0} f(x) = f(x_0). \tag{3}$$

Limits and Continuity



Definition 1.3. Suppose that $\{x_n\}_{n=1}^{\infty}$ is an infinite sequence. Then the sequence is said to have the *limit* L, and we write

$$\lim_{n\to\infty} x_n = L, \tag{4}$$

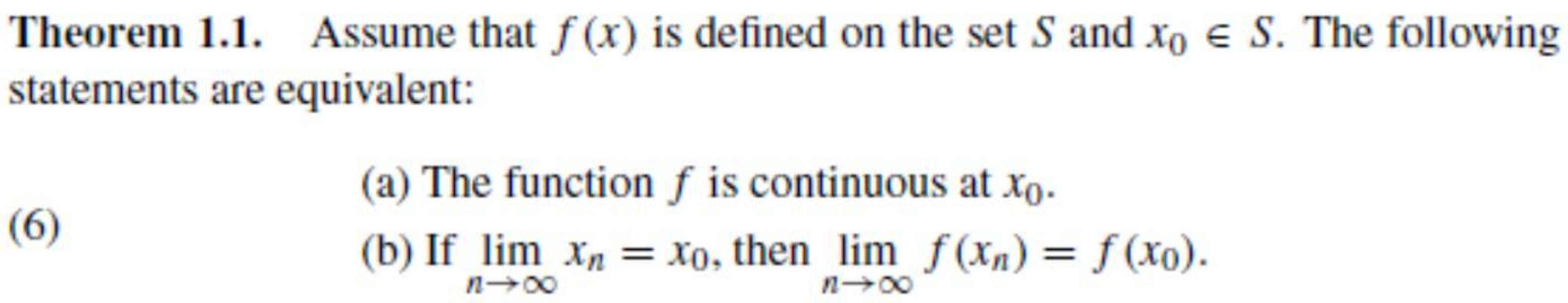
if given any $\epsilon > 0$, there exists a positive integer $N = N(\epsilon)$ such that $n > N$ implies that $|x_n - L| < \epsilon$. ▲

When a sequence has a limit, we say that it is a ***convergent sequence***. Another commonly used notation is "$x_n \to L$ as $n \to \infty$." Equation (4) is equivalent to

$$\lim_{n\to\infty} (x_n - L) = 0. \tag{5}$$

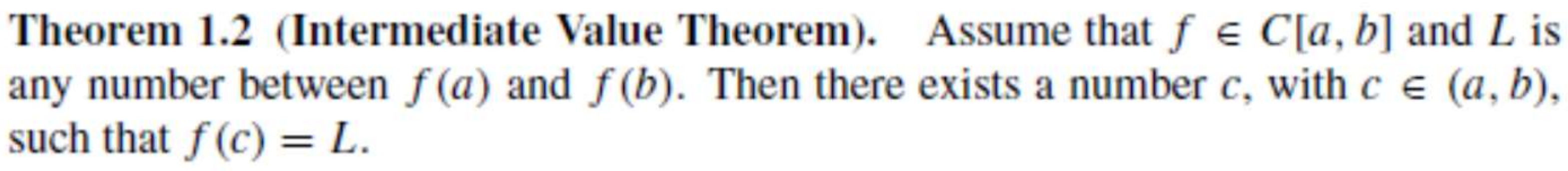
Thus we can view the sequence $\{\epsilon_n\}_{n=1}^{\infty} = \{x_n - L\}_{n=1}^{\infty}$ as an ***error sequence***. The following theorem relates the concepts of continuity and convergent sequence.

Limits and Continuity

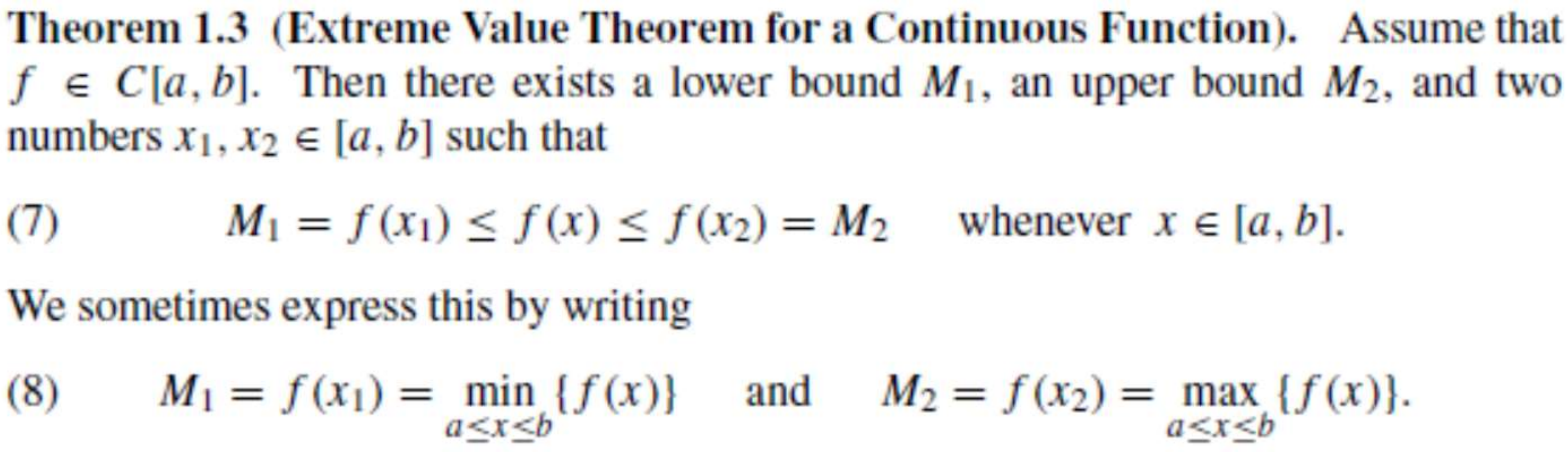


Theorem 1.1. Assume that $f(x)$ is defined on the set S and $x_0 \in S$. The following statements are equivalent:

$$(6)\qquad \begin{array}{l} \text{(a) The function } f \text{ is continuous at } x_0. \\ \text{(b) If } \lim_{n\to\infty} x_n = x_0, \text{ then } \lim_{n\to\infty} f(x_n) = f(x_0). \end{array}$$



Theorem 1.2 (Intermediate Value Theorem). Assume that $f \in C[a, b]$ and L is any number between $f(a)$ and $f(b)$. Then there exists a number c, with $c \in (a, b)$, such that $f(c) = L$.



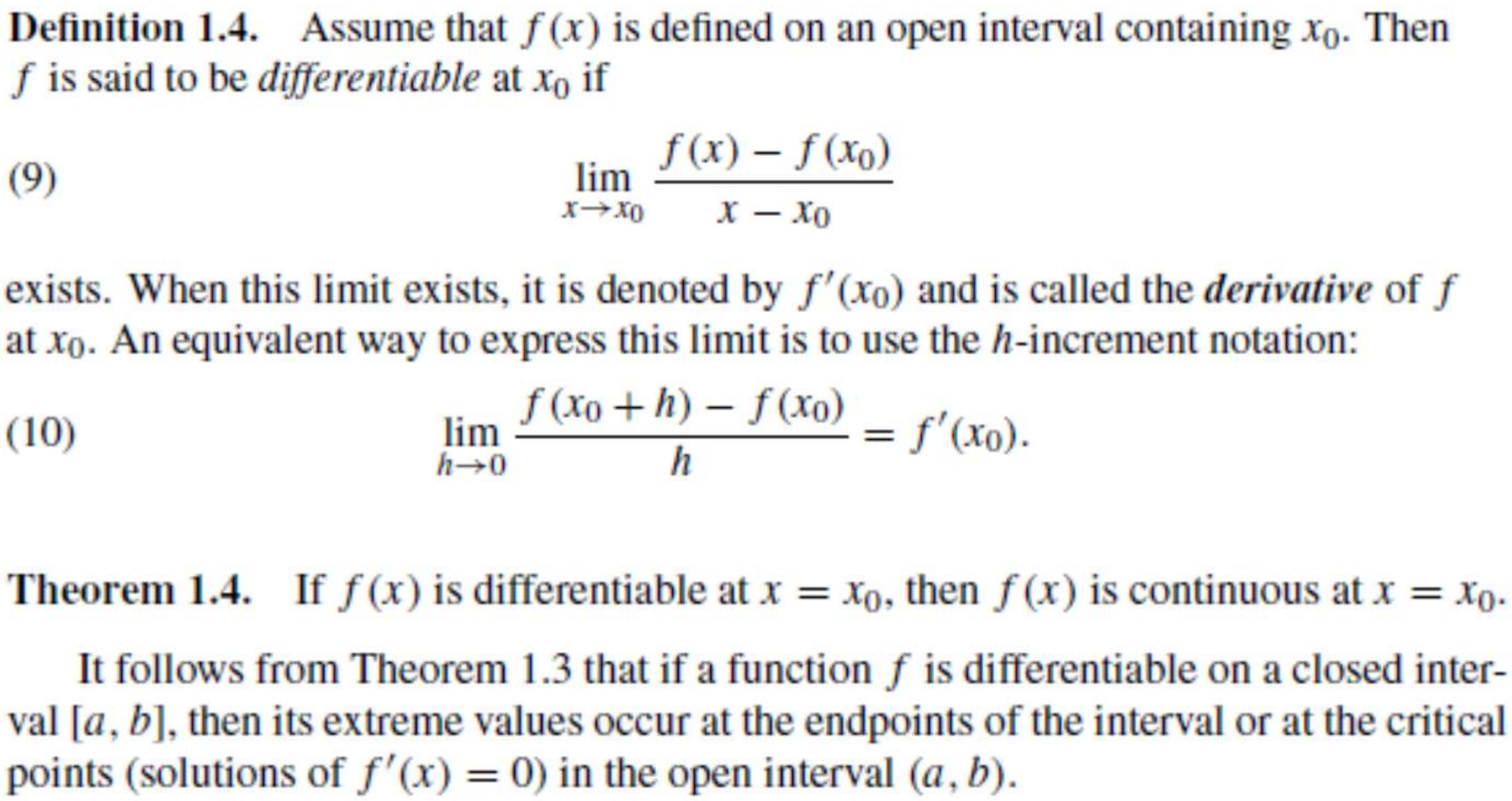
Theorem 1.3 (Extreme Value Theorem for a Continuous Function). Assume that $f \in C[a, b]$. Then there exists a lower bound M_1, an upper bound M_2, and two numbers $x_1, x_2 \in [a, b]$ such that

$$(7)\qquad M_1 = f(x_1) \le f(x) \le f(x_2) = M_2 \quad \text{whenever } x \in [a, b].$$

We sometimes express this by writing

$$(8)\qquad M_1 = f(x_1) = \min_{a\le x\le b}\{f(x)\} \quad \text{and} \quad M_2 = f(x_2) = \max_{a\le x\le b}\{f(x)\}.$$

Differentiable Functions



Definition 1.4. Assume that $f(x)$ is defined on an open interval containing x_0. Then f is said to be *differentiable* at x_0 if

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{9}$$

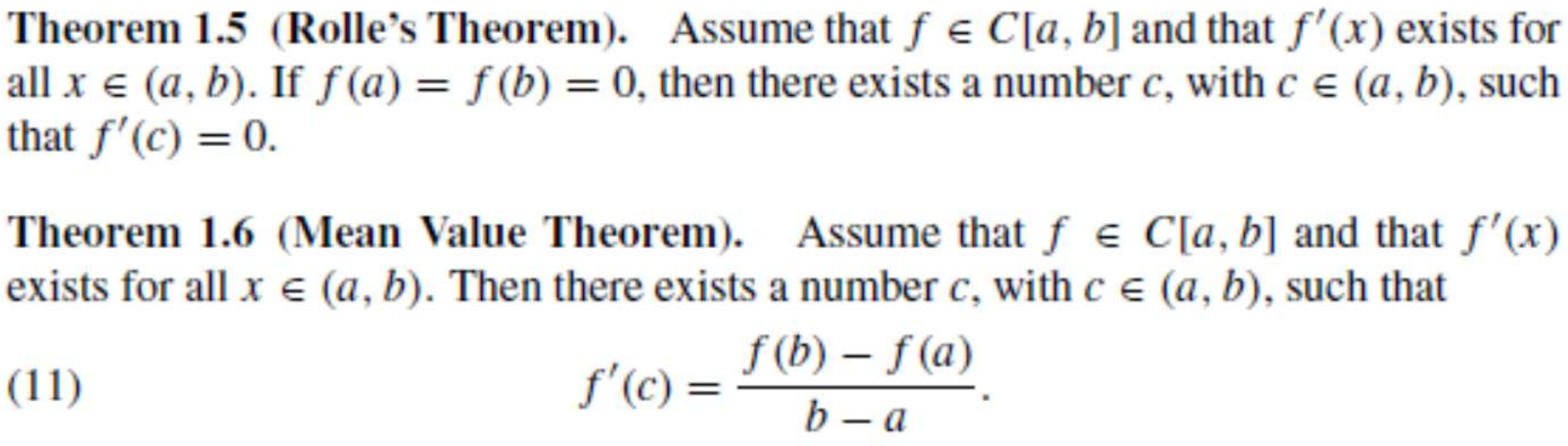
exists. When this limit exists, it is denoted by $f'(x_0)$ and is called the ***derivative*** of f at x_0. An equivalent way to express this limit is to use the h-increment notation:

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0). \tag{10}$$

Theorem 1.4. If $f(x)$ is differentiable at $x = x_0$, then $f(x)$ is continuous at $x = x_0$.

It follows from Theorem 1.3 that if a function f is differentiable on a closed interval $[a, b]$, then its extreme values occur at the endpoints of the interval or at the critical points (solutions of $f'(x) = 0$) in the open interval (a, b).

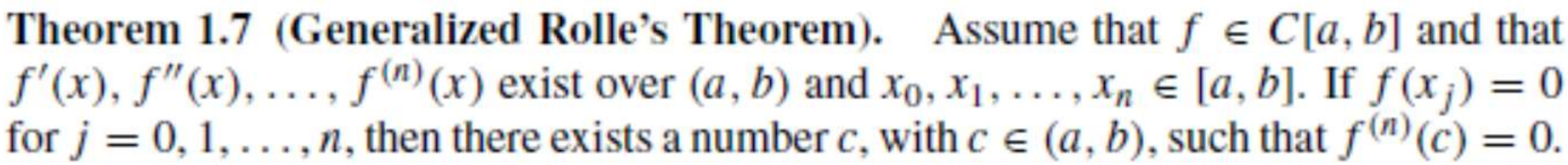
Differentiable Functions



Theorem 1.5 (Rolle's Theorem). Assume that $f \in C[a,b]$ and that $f'(x)$ exists for all $x \in (a,b)$. If $f(a) = f(b) = 0$, then there exists a number c, with $c \in (a,b)$, such that $f'(c) = 0$.

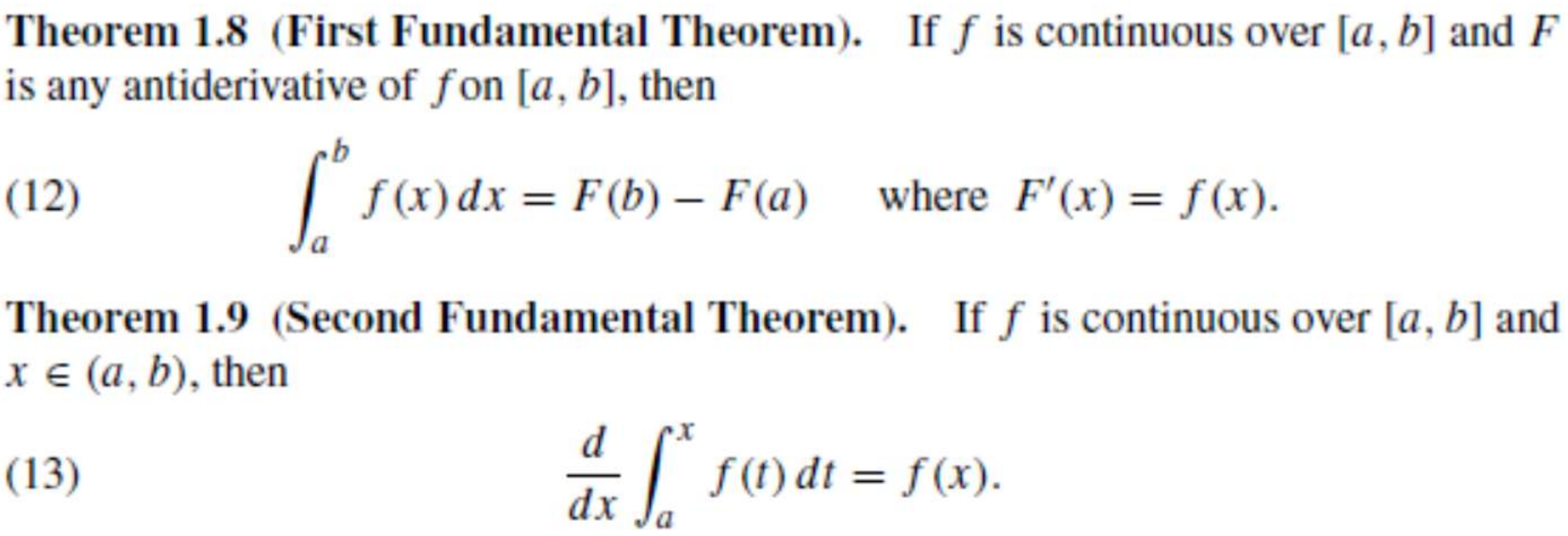
Theorem 1.6 (Mean Value Theorem). Assume that $f \in C[a,b]$ and that $f'(x)$ exists for all $x \in (a,b)$. Then there exists a number c, with $c \in (a,b)$, such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{11}$$



Theorem 1.7 (Generalized Rolle's Theorem). Assume that $f \in C[a,b]$ and that $f'(x), f''(x), \ldots, f^{(n)}(x)$ exist over (a,b) and $x_0, x_1, \ldots, x_n \in [a,b]$. If $f(x_j) = 0$ for $j = 0, 1, \ldots, n$, then there exists a number c, with $c \in (a,b)$, such that $f^{(n)}(c) = 0$.

Integrals



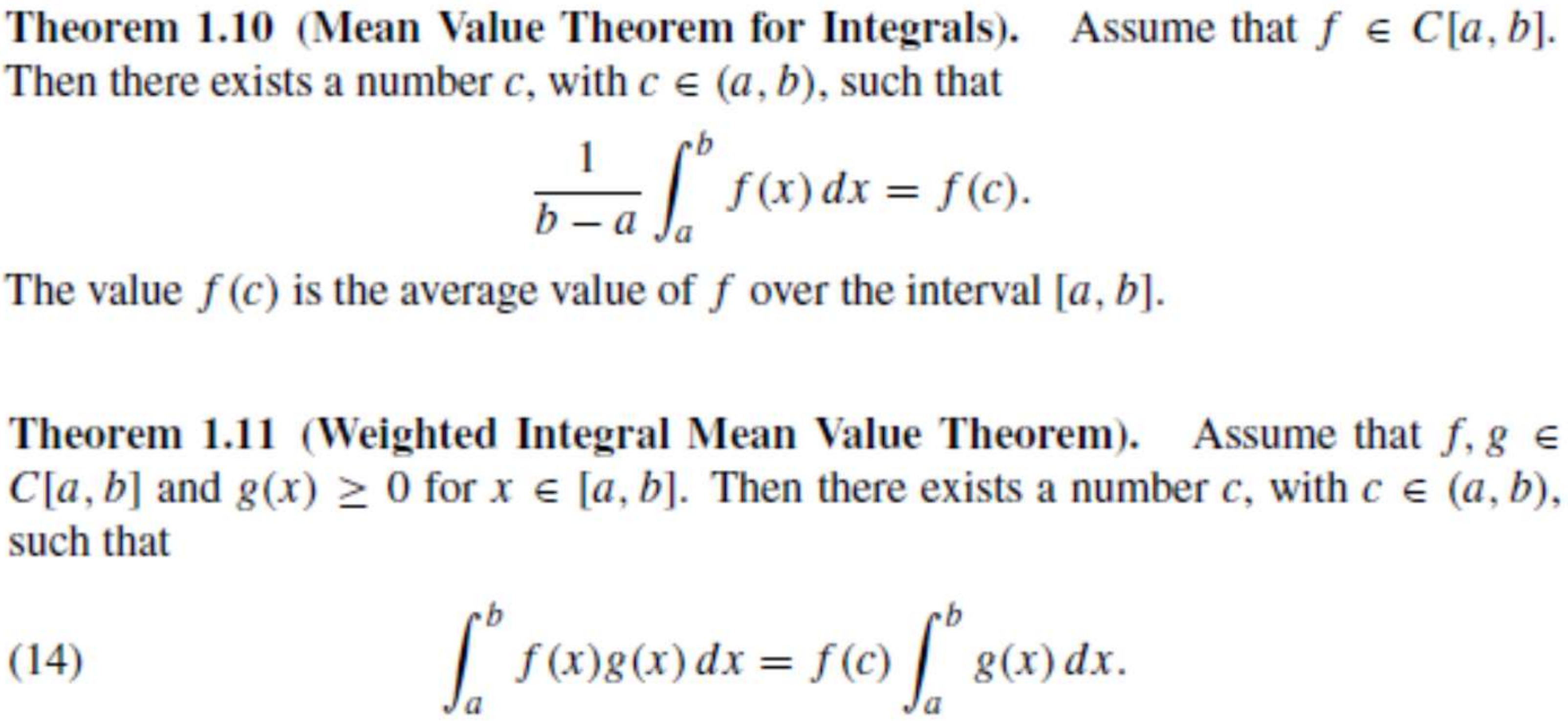
Theorem 1.8 (First Fundamental Theorem). If f is continuous over $[a, b]$ and F is any antiderivative of f on $[a, b]$, then

$$\int_a^b f(x)\,dx = F(b) - F(a) \qquad \text{where } F'(x) = f(x). \tag{12}$$

Theorem 1.9 (Second Fundamental Theorem). If f is continuous over $[a, b]$ and $x \in (a, b)$, then

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x). \tag{13}$$

Integrals



Theorem 1.10 (Mean Value Theorem for Integrals). Assume that $f \in C[a,b]$. Then there exists a number c, with $c \in (a,b)$, such that

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

The value $f(c)$ is the average value of f over the interval $[a,b]$.

Theorem 1.11 (Weighted Integral Mean Value Theorem). Assume that $f, g \in C[a,b]$ and $g(x) \geq 0$ for $x \in [a,b]$. Then there exists a number c, with $c \in (a,b)$, such that

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx. \tag{14}$$

Series

Definition 1.5. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence. Then $\sum_{n=1}^{\infty} a_n$ is an infinite series. The nth partial sum is $S_n = \sum_{k=1}^{n} a_k$. The infinite series ***converges*** if and only if the sequence $\{S_n\}_{n=1}^{\infty}$ converges to a limit S, that is,

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{k=1}^{n} a_k = S. \tag{15}$$

If a series does not converge, we say that it ***diverges***. ▲

Example 1.7. Consider the infinite sequence $\{a_n\}_{n=1}^{\infty} = \left\{\frac{1}{n(n+1)}\right\}_{n=1}^{\infty}$. Then the nth partial sum is

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

Therefore, the ***sum*** of the infinite series is

$$S = \lim_{n\to\infty} Sn = \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) = 1.$$

■

Series

Theorem 1.12 (Taylor's Theorem). Assume that $f \in C^{n+1}[a, b]$ and let $x_0 \in [a, b]$. Then, for every $x \in (a, b)$, there exists a number $c = c(x)$ (the value of c depends on the value of x) that lies between x_0 and x such that

$$f(x) = P_n(x) + R_n(x), \tag{16}$$

where

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{17}$$

and

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}. \tag{18}$$

Binary Numbers

- Decimal number system (base 10)
- Binary number system (base 2)
- Computer converts inputs to base 2 (or perhaps base 16), then performs base 2 arithmetic, and finally, translates the answer into base 10 before it displays a result.

$$\sum_{k=1}^{100,000} 0.1 = 9999.99447.$$

Base 2 Numbers

$$\begin{aligned}1563 = {} & (1 \times 2^{10}) + (1 \times 2^{9}) + (0 \times 2^{8}) + (0 \times 2^{7}) + (0 \times 2^{6}) \\ & + (0 \times 2^{5}) + (1 \times 2^{4}) + (1 \times 2^{3}) + (0 \times 2^{2}) + (1 \times 2^{1}) \\ & + (1 \times 2^{0}).\end{aligned}$$

$$1563 = 11000011011_{\text{two}}.$$

$$\begin{aligned}
1563 &= 2 \times 781 + 1, & b_0 &= 1 \\
781 &= 2 \times 390 + 1, & b_1 &= 1 \\
390 &= 2 \times 195 + 0, & b_2 &= 0 \\
195 &= 2 \times 97 + 1, & b_3 &= 1 \\
97 &= 2 \times 48 + 1, & b_4 &= 1 \\
48 &= 2 \times 24 + 0, & b_5 &= 0 \\
24 &= 2 \times 12 + 0, & b_6 &= 0 \\
12 &= 2 \times 6 + 0, & b_7 &= 0 \\
6 &= 2 \times 3 + 0, & b_8 &= 0 \\
3 &= 2 \times 1 + 1, & b_9 &= 1 \\
1 &= 2 \times 0 + 1, & b_{10} &= 1.
\end{aligned}$$

Binary Fractions

$$R = (d_1 \times 2^{-1}) + (d_2 \times 2^{-2}) + \cdots + (d_n \times 2^{-n}) + \cdots ,$$

⬇

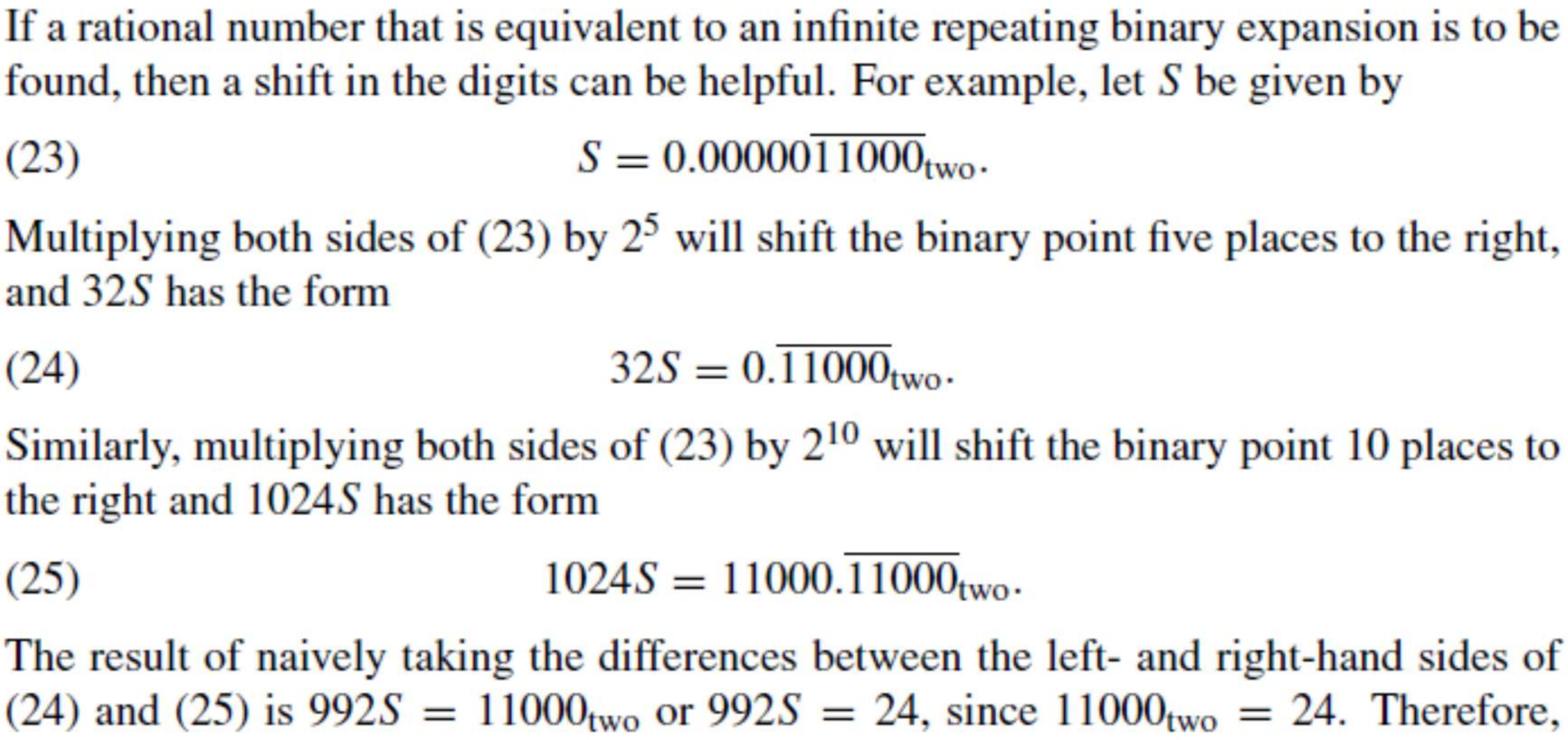
$$R = 0.d_1 d_2 \cdots d_n \cdots_{\text{two}}.$$

Example:

$$\frac{7}{10} = 0.1\overline{0110}_{\text{two}}.$$

$2R = 1.4$	$d_1 = \text{int}(1.4) = 1$	$F_1 = \text{frac}(1.4) = 0.4$
$2F_1 = 0.8$	$d_2 = \text{int}(0.8) = 0$	$F_2 = \text{frac}(0.8) = 0.8$
$2F_2 = 1.6$	$d_3 = \text{int}(1.6) = 1$	$F_3 = \text{frac}(1.6) = 0.6$
$2F_3 = 1.2$	$d_4 = \text{int}(1.2) = 1$	$F_4 = \text{frac}(1.2) = 0.2$
$2F_4 = 0.4$	$d_5 = \text{int}(0.4) = 0$	$F_5 = \text{frac}(0.4) = 0.4$
$2F_5 = 0.8$	$d_6 = \text{int}(0.8) = 0$	$F_6 = \text{frac}(0.8) = 0.8$
$2F_6 = 1.6$	$d_7 = \text{int}(1.6) = 1$	$F_7 = \text{frac}(1.6) = 0.6$

Binary Shifting



If a rational number that is equivalent to an infinite repeating binary expansion is to be found, then a shift in the digits can be helpful. For example, let S be given by

$$S = 0.00000\overline{11000}_{\text{two}}. \tag{23}$$

Multiplying both sides of (23) by 2^5 will shift the binary point five places to the right, and $32S$ has the form

$$32S = 0.\overline{11000}_{\text{two}}. \tag{24}$$

Similarly, multiplying both sides of (23) by 2^{10} will shift the binary point 10 places to the right and $1024S$ has the form

$$1024S = 11000.\overline{11000}_{\text{two}}. \tag{25}$$

The result of naively taking the differences between the left- and right-hand sides of (24) and (25) is $992S = 11000_{\text{two}}$ or $992S = 24$, since $11000_{\text{two}} = 24$. Therefore,

S=3/124

Machine Numbers

- computer stores a binary approximation to x

$$x \approx \pm q \times 2^n.$$

Where q is the ***mantissa***, n is the ***exponent.***

- Only a small subset of the real number system is used
- An example: $0.d_1 d_2 d_3 d_{4\,\text{two}} \times 2^n,$

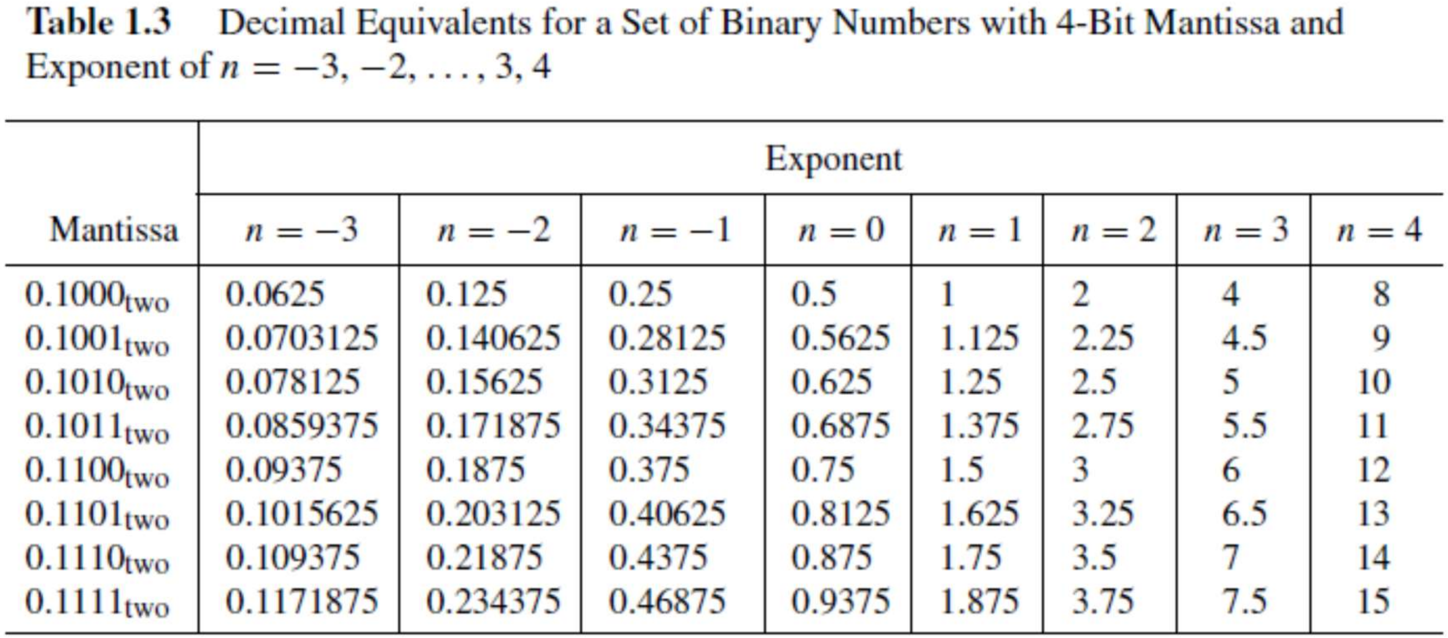


Table 1.3 Decimal Equivalents for a Set of Binary Numbers with 4-Bit Mantissa and Exponent of $n = -3, -2, \ldots, 3, 4$

	Exponent							
Mantissa	$n = -3$	$n = -2$	$n = -1$	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$
0.1000_{two}	0.0625	0.125	0.25	0.5	1	2	4	8
0.1001_{two}	0.0703125	0.140625	0.28125	0.5625	1.125	2.25	4.5	9
0.1010_{two}	0.078125	0.15625	0.3125	0.625	1.25	2.5	5	10
0.1011_{two}	0.0859375	0.171875	0.34375	0.6875	1.375	2.75	5.5	11
0.1100_{two}	0.09375	0.1875	0.375	0.75	1.5	3	6	12
0.1101_{two}	0.1015625	0.203125	0.40625	0.8125	1.625	3.25	6.5	13
0.1110_{two}	0.109375	0.21875	0.4375	0.875	1.75	3.5	7	14
0.1111_{two}	0.1171875	0.234375	0.46875	0.9375	1.875	3.75	7.5	15

Machine Numbers

- What would happen if a computer had only a 4-bit mantissa and was restricted to perform the computation $\left(\frac{1}{10}+\frac{1}{5}\right)+\frac{1}{6}$?

$$\begin{array}{ccccc}
\frac{1}{10} & \approx & 0.1101_{\text{two}} \times 2^{-3} & = & 0.01101_{\text{two}} \times 2^{-2} \\
\frac{1}{5} & \approx & 0.1101_{\text{two}} \times 2^{-2} & = & 0.1101_{\text{two}} \quad \times 2^{-2} \\
\hline
\frac{3}{10} & & & & 1.00111_{\text{two}} \times 2^{-2}.
\end{array}$$

$$\begin{array}{ccccc}
\frac{3}{10} & \approx & 0.1010_{\text{two}} \times 2^{-1} & = & 0.1010_{\text{two}} \quad \times 2^{-1} \\
\frac{1}{6} & \approx & 0.1011_{\text{two}} \times 2^{-2} & = & 0.01011_{\text{two}} \times 2^{-1} \\
\hline
\frac{7}{15} & & & & 0.11111_{\text{two}} \times 2^{-1}.
\end{array}$$

$$\frac{7}{15} \approx 0.10000_{\text{two}} \times 2^{0}.$$

- Error

$$\frac{7}{15} - 0.10000_{\text{two}} \approx 0.466667 - 0.500000 \approx 0.033333.$$

Computer Accuracy

- To store numbers accurately, computers must have floating-point binary numbers with at least 24 binary bits used for the mantissa (seven decimal places);
- A 32-bit mantissa can result in numbers with nine decimal places.

- Suppose that the mantissa contains 32 binary bits, an example

$$\frac{1}{10} \approx 0.11001100110011001100110011001100_{\text{two}} \times 2^{-3}.$$

- Compared with 1/10, the error is

$$0.\overline{1100}_{\text{two}} \times 2^{-35} \approx 2.328306437 \times 10^{-11}.$$

Computer Floating-Point Numbers

- Computers have both an ***integer mode*** and a ***floating-point mode*** for representing numbers.

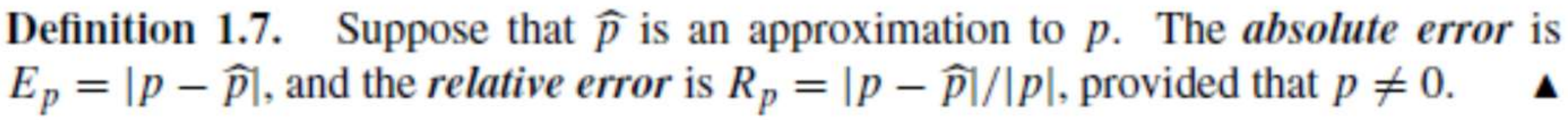
- Computers that use 32 bits to represent single-precision real numbers use 8 bits for the exponent and 24 bits for the mantissa. Represent real numbers with magnitudes in the range 2.938736E-39 to 1.701412E+38, with six decimal digits of numerical precision.

- Computers that use 48 bits to represent single-precision real numbers might use 8 bits for the exponent and 40 bits for the mantissa. Represent real numbers from 2.9387358771E-39 to 1.7014118346E+38, with 11 decimal digits of precision.

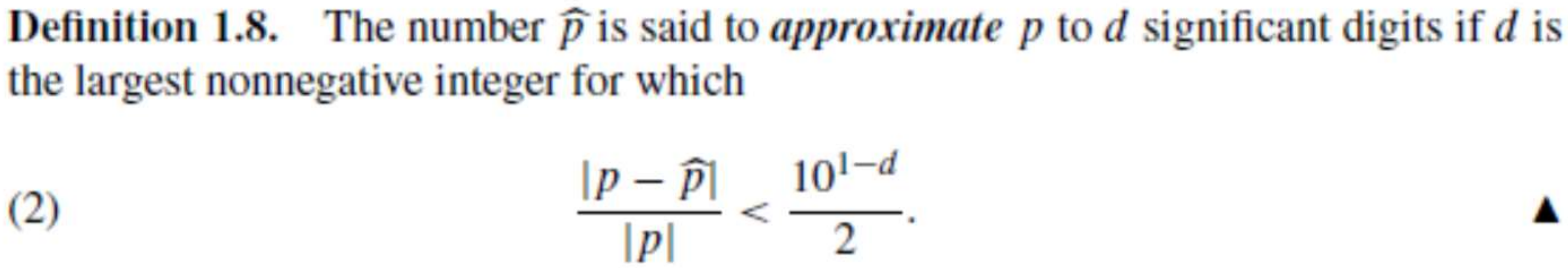
- For 64-bit double-precision real numbers, it might use 11 bits for the exponent and 53 bits for the mantissa, represents number from 5.562684646268003E-309 to 8.988465674311580E+307, with 16 decimal digits of precision.

Error Analysis

Absolute error and relative error



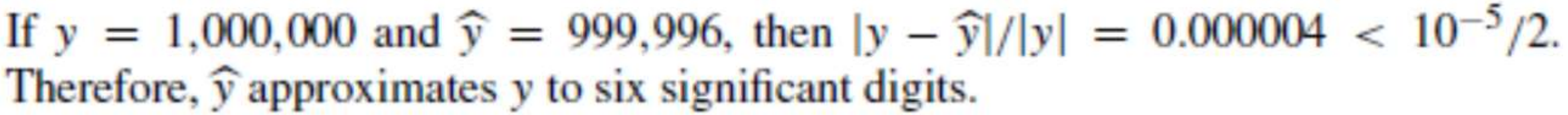
Definition 1.7. Suppose that $\widehat{p}$ is an approximation to p. The ***absolute error*** is $E_p = |p - \widehat{p}|$, and the ***relative error*** is $R_p = |p - \widehat{p}|/|p|$, provided that $p \neq 0$. ▲



Definition 1.8. The number $\widehat{p}$ is said to ***approximate*** p to d significant digits if d is the largest nonnegative integer for which

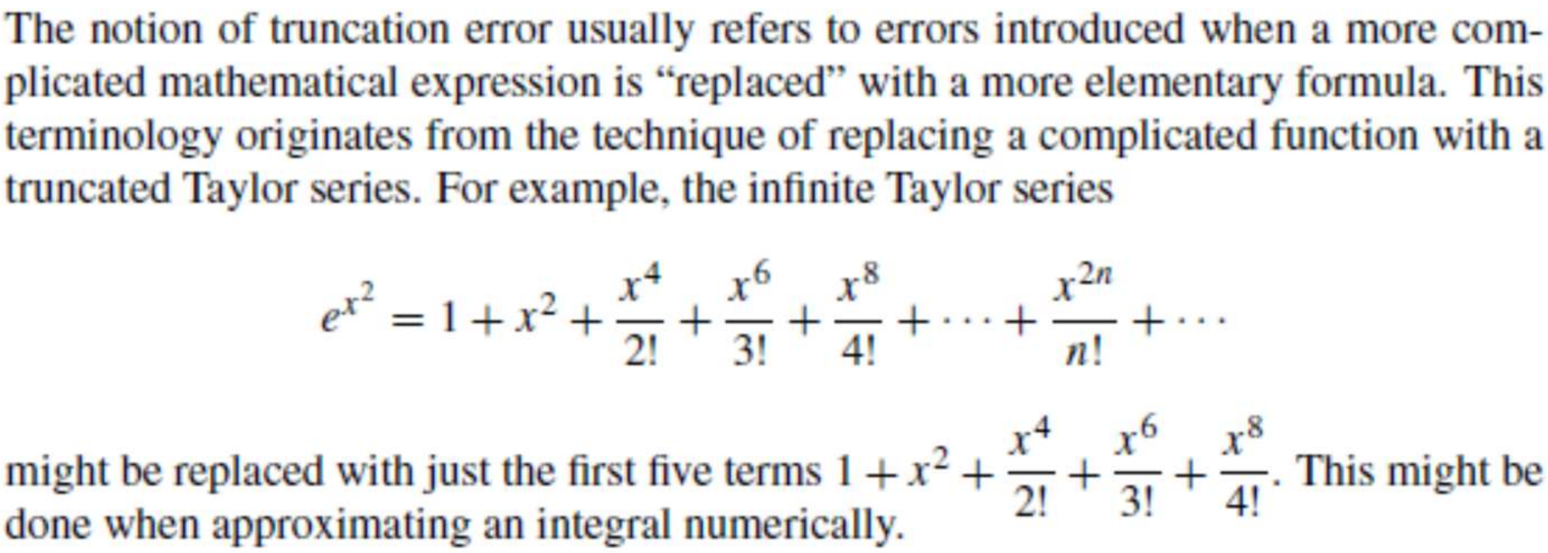
$$\frac{|p - \widehat{p}|}{|p|} < \frac{10^{1-d}}{2}. \qquad (2)$$

▲



If $y = 1{,}000{,}000$ and $\widehat{y} = 999{,}996$, then $|y - \widehat{y}|/|y| = 0.000004 < 10^{-5}/2$. Therefore, $\widehat{y}$ approximates y to six significant digits.

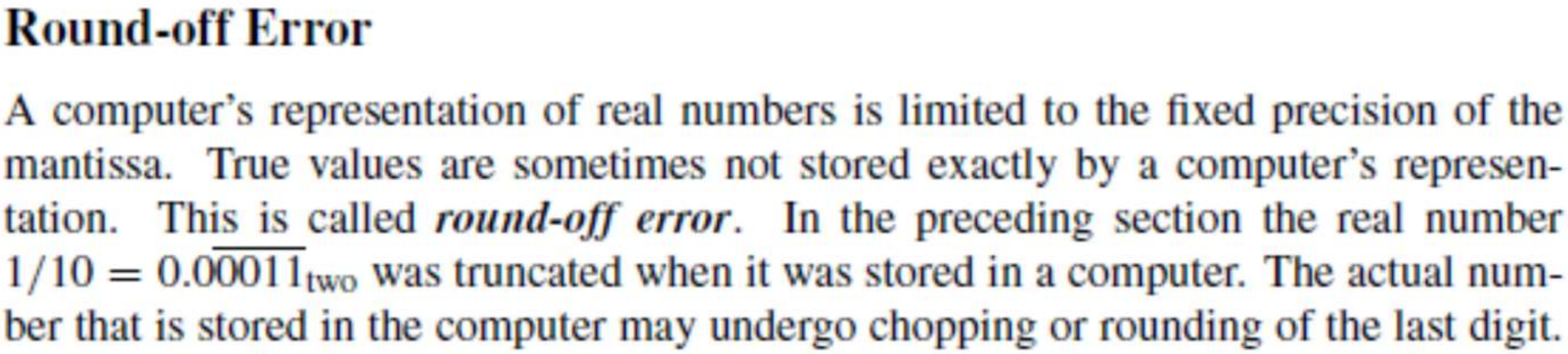
Truncation error and round-off error



The notion of truncation error usually refers to errors introduced when a more complicated mathematical expression is "replaced" with a more elementary formula. This terminology originates from the technique of replacing a complicated function with a truncated Taylor series. For example, the infinite Taylor series

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \cdots + \frac{x^{2n}}{n!} + \cdots$$

might be replaced with just the first five terms $1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$. This might be done when approximating an integral numerically.



Round-off Error

A computer's representation of real numbers is limited to the fixed precision of the mantissa. True values are sometimes not stored exactly by a computer's representation. This is called ***round-off error***. In the preceding section the real number $1/10 = 0.0\overline{0011}_{\text{two}}$ was truncated when it was stored in a computer. The actual number that is stored in the computer may undergo chopping or rounding of the last digit.

Loss of significance

- Consider the two numbers p = 3.1415926536 and q = 3.1415957341, which are nearly equal and both carry 11 decimal digits of precision. Suppose that their difference is formed: p − q = −0.0000030805. Since the first six digits of p and q are the same, their difference p − q contains only five decimal digits of precision. This phenomenon is called loss of significance or subtractive cancellation. This reduction in the precision of the final computed answer can creep in when it is not suspected.

- Example $f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$ and $g(x) = \dfrac{x}{\sqrt{x+1} + \sqrt{x}}$.

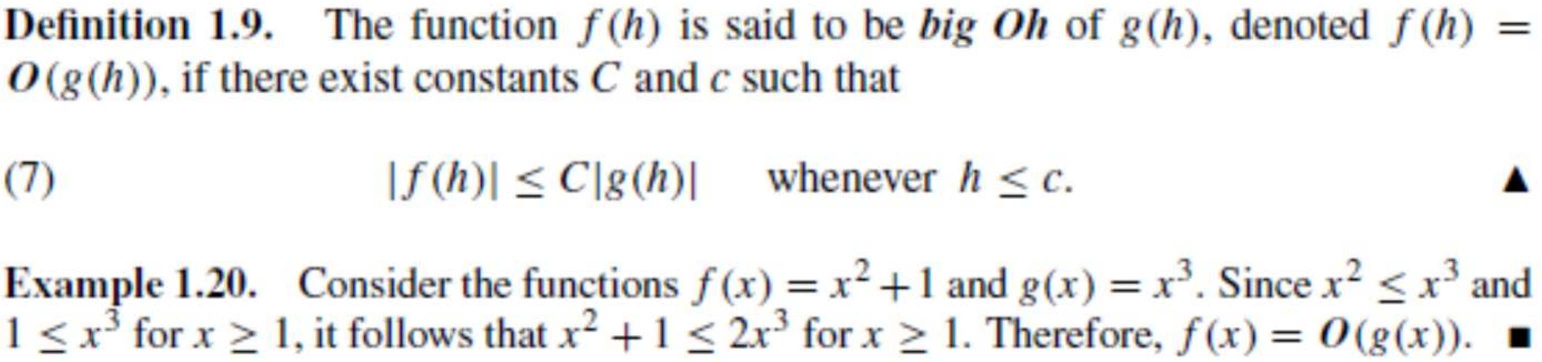
$$\begin{aligned} f(500) &= 500\left(\sqrt{501} - \sqrt{500}\right) \\ &= 500(22.3830 - 22.3607) = 500(0.0223) = 11.1500. \end{aligned}$$

True: 11.174755300747198..

$$\begin{aligned} g(500) &= \frac{500}{\sqrt{501} + \sqrt{500}} \\ &= \frac{500}{22.3830 + 22.3607} = \frac{500}{44.7437} = 11.1748. \end{aligned}$$

$O(h^n)$ Order of Approximation

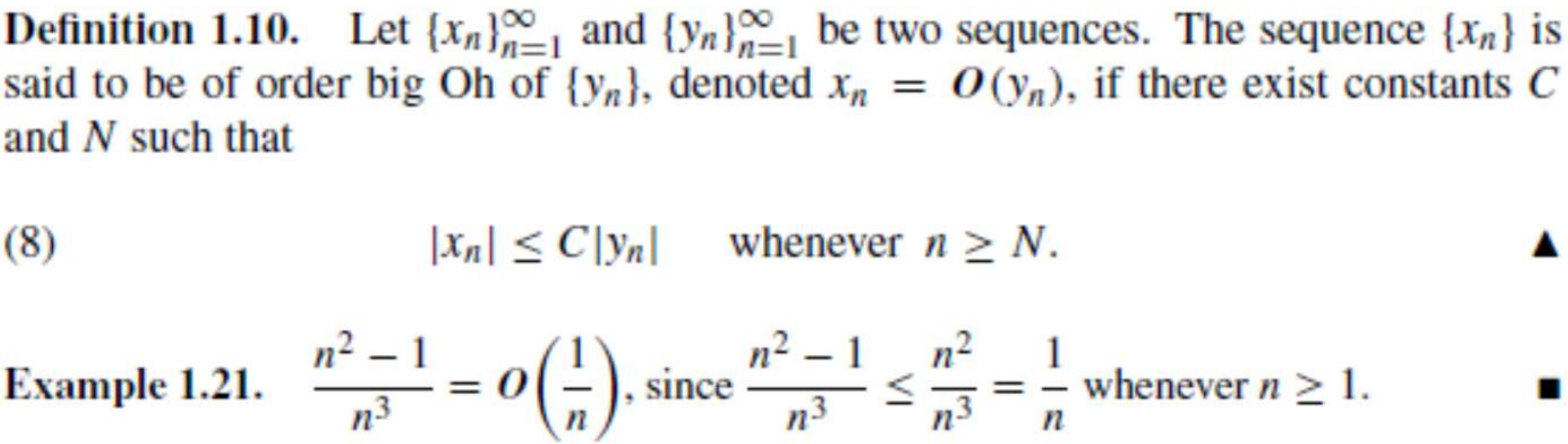
- Sequences $\left\{\frac{1}{n^2}\right\}_{n=1}^{\infty}$ and $\left\{\frac{1}{n}\right\}_{n=1}^{\infty}$ are both converging to zero; the first sequence is converging to zero more rapidly than the second one.



Definition 1.9. The function $f(h)$ is said to be ***big Oh*** of $g(h)$, denoted $f(h) = \boldsymbol{O}(g(h))$, if there exist constants C and c such that

$$(7) \qquad |f(h)| \le C|g(h)| \quad \text{whenever } h \le c. \qquad \blacktriangle$$

Example 1.20. Consider the functions $f(x) = x^2 + 1$ and $g(x) = x^3$. Since $x^2 \le x^3$ and $1 \le x^3$ for $x \ge 1$, it follows that $x^2 + 1 \le 2x^3$ for $x \ge 1$. Therefore, $f(x) = \boldsymbol{O}(g(x))$. ■

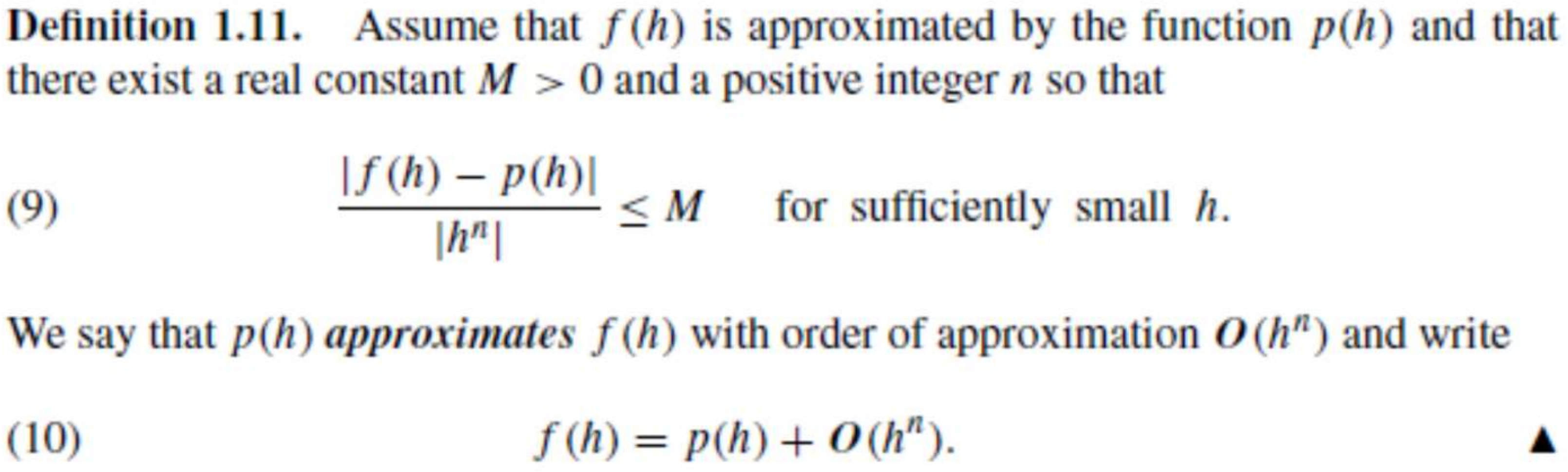


Definition 1.10. Let $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ be two sequences. The sequence $\{x_n\}$ is said to be of order big Oh of $\{y_n\}$, denoted $x_n = \boldsymbol{O}(y_n)$, if there exist constants C and N such that

$$(8) \qquad |x_n| \le C|y_n| \quad \text{whenever } n \ge N. \qquad \blacktriangle$$

Example 1.21. $\dfrac{n^2-1}{n^3} = \boldsymbol{O}\left(\dfrac{1}{n}\right)$, since $\dfrac{n^2-1}{n^3} \le \dfrac{n^2}{n^3} = \dfrac{1}{n}$ whenever $n \ge 1$. ■

$O(h^n)$ Order of Approximation



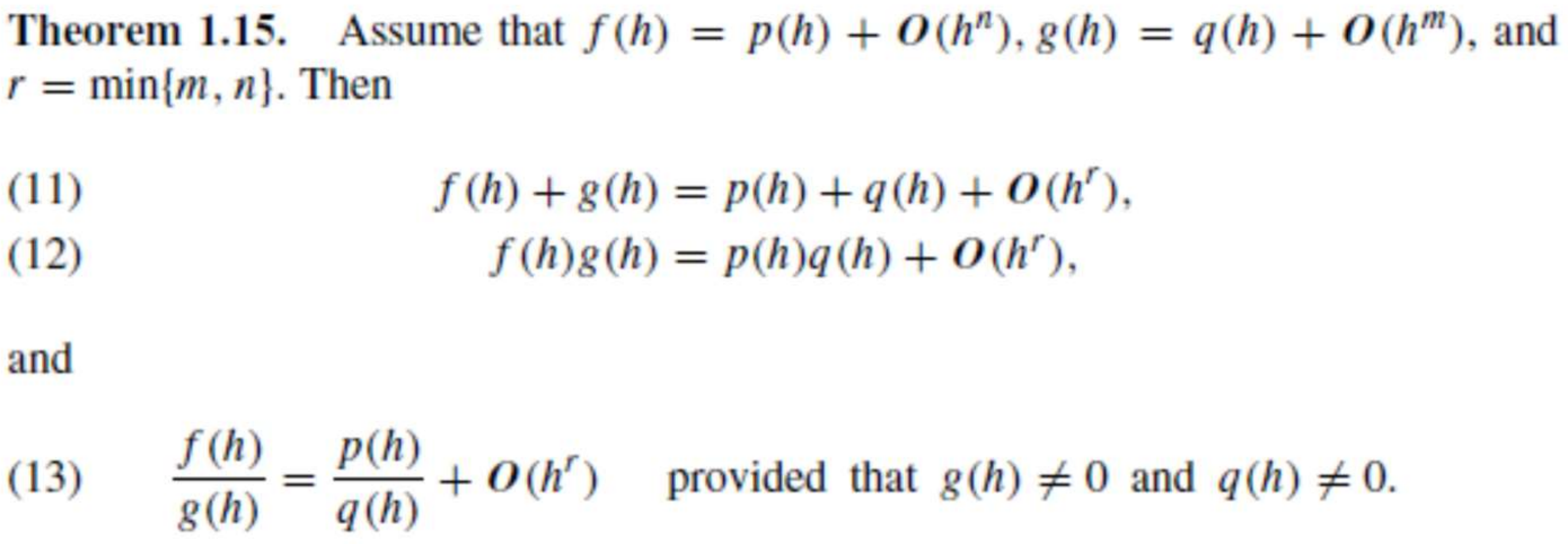
Definition 1.11. Assume that $f(h)$ is approximated by the function $p(h)$ and that there exist a real constant $M > 0$ and a positive integer n so that

$$\frac{|f(h) - p(h)|}{|h^n|} \le M \quad \text{for sufficiently small } h. \tag{9}$$

We say that $p(h)$ ***approximates*** $f(h)$ with order of approximation $\boldsymbol{O}(h^n)$ and write

$$f(h) = p(h) + \boldsymbol{O}(h^n). \tag{10}$$

▲



Theorem 1.15. Assume that $f(h) = p(h) + \boldsymbol{O}(h^n)$, $g(h) = q(h) + \boldsymbol{O}(h^m)$, and $r = \min\{m, n\}$. Then

$$f(h) + g(h) = p(h) + q(h) + \boldsymbol{O}(h^r), \tag{11}$$

$$f(h)g(h) = p(h)q(h) + \boldsymbol{O}(h^r), \tag{12}$$

and

$$\frac{f(h)}{g(h)} = \frac{p(h)}{q(h)} + \boldsymbol{O}(h^r) \quad \text{provided that } g(h) \ne 0 \text{ and } q(h) \ne 0. \tag{13}$$

$O(h^n)$ Order of Approximation

Theorem 1.16 (Taylor's Theorem). Assume that $f \in C^{n+1}[a, b]$. If both x_0 and $x = x_0 + h$ lie in $[a, b]$, then

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} h^k + O(h^{n+1}). \tag{15}$$

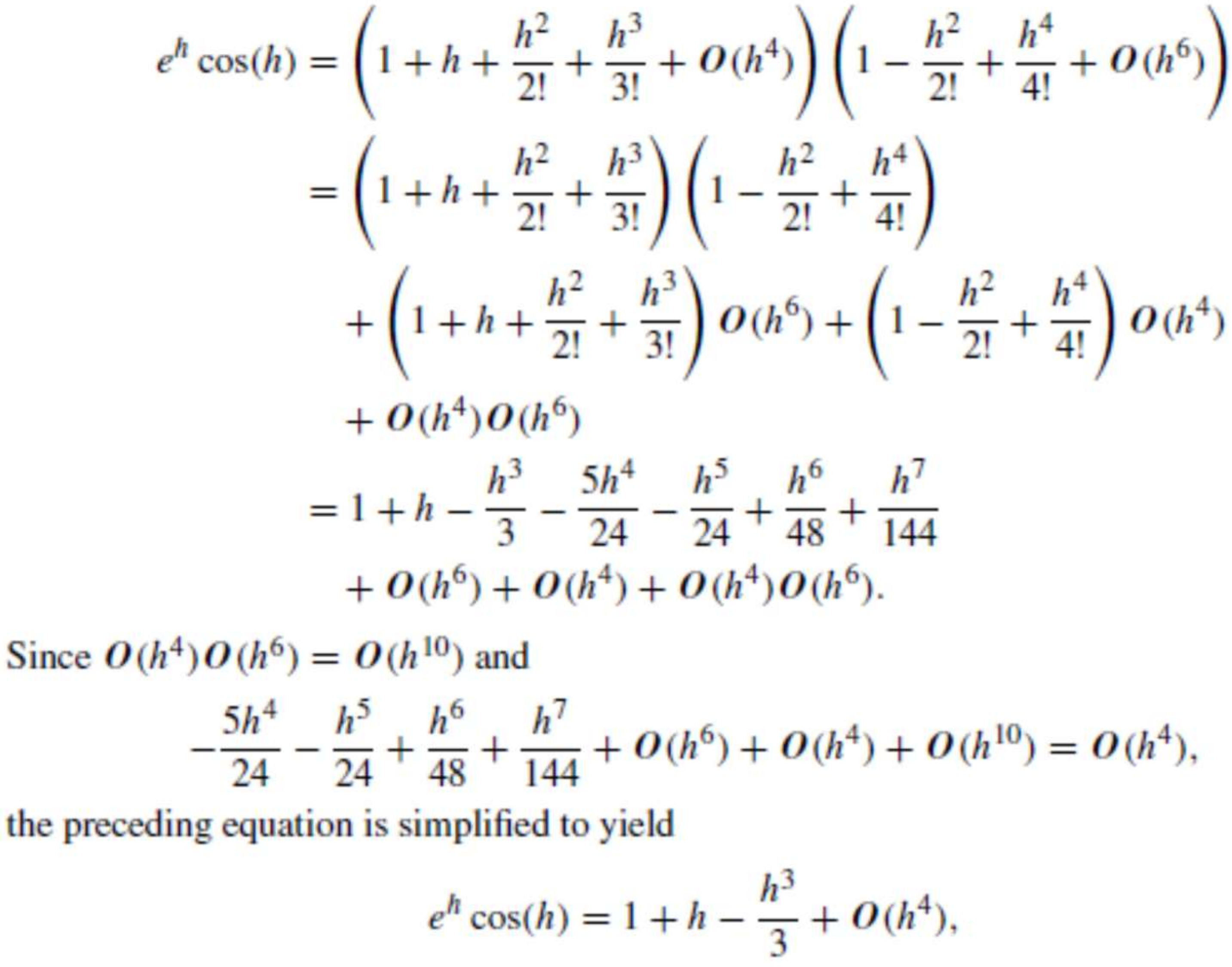
$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4) \quad \text{and} \quad \cos(h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6).$$

$$\begin{aligned} e^h + \cos(h) &= 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4) + 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6) \\ &= 2 + h + \frac{h^3}{3!} + O(h^4) + \frac{h^4}{4!} + O(h^6). \end{aligned}$$

Since $O(h^4) + \frac{h^4}{4!} = O(h^4)$ and $O(h^4) + O(h^6) = O(h^4)$, this reduces to

$$e^h + \cos(h) = 2 + h + \frac{h^3}{3!} + O(h^4),$$

$O(h^n)$ Order of Approximation



$$\begin{aligned}
e^h\cos(h) &= \left(1+h+\frac{h^2}{2!}+\frac{h^3}{3!}+O(h^4)\right)\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}+O(h^6)\right)\\
&= \left(1+h+\frac{h^2}{2!}+\frac{h^3}{3!}\right)\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}\right)\\
&\quad+\left(1+h+\frac{h^2}{2!}+\frac{h^3}{3!}\right)O(h^6)+\left(1-\frac{h^2}{2!}+\frac{h^4}{4!}\right)O(h^4)\\
&\quad+O(h^4)O(h^6)\\
&= 1+h-\frac{h^3}{3}-\frac{5h^4}{24}-\frac{h^5}{24}+\frac{h^6}{48}+\frac{h^7}{144}\\
&\quad+O(h^6)+O(h^4)+O(h^4)O(h^6).
\end{aligned}$$

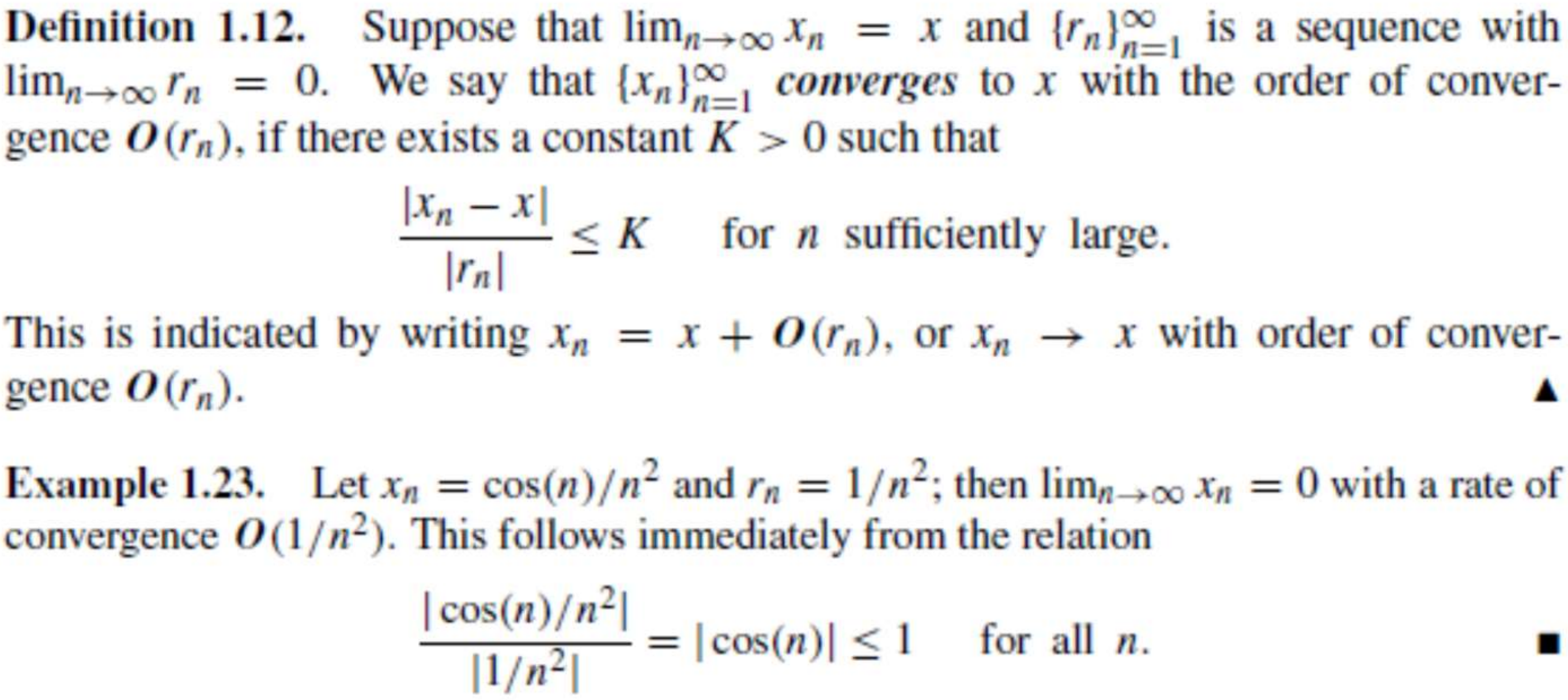
Since $O(h^4)O(h^6)=O(h^{10})$ and

$$-\frac{5h^4}{24}-\frac{h^5}{24}+\frac{h^6}{48}+\frac{h^7}{144}+O(h^6)+O(h^4)+O(h^{10})=O(h^4),$$

the preceding equation is simplified to yield

$$e^h\cos(h)=1+h-\frac{h^3}{3}+O(h^4),$$

Order of Convergence of a Sequence



Definition 1.12. Suppose that $\lim_{n\to\infty} x_n = x$ and $\{r_n\}_{n=1}^{\infty}$ is a sequence with $\lim_{n\to\infty} r_n = 0$. We say that $\{x_n\}_{n=1}^{\infty}$ ***converges*** to x with the order of convergence $\boldsymbol{O}(r_n)$, if there exists a constant $K > 0$ such that

$$\frac{|x_n - x|}{|r_n|} \le K \qquad \text{for } n \text{ sufficiently large.}$$

This is indicated by writing $x_n = x + \boldsymbol{O}(r_n)$, or $x_n \to x$ with order of convergence $\boldsymbol{O}(r_n)$. ▲

Example 1.23. Let $x_n = \cos(n)/n^2$ and $r_n = 1/n^2$; then $\lim_{n\to\infty} x_n = 0$ with a rate of convergence $\boldsymbol{O}(1/n^2)$. This follows immediately from the relation

$$\frac{|\cos(n)/n^2|}{|1/n^2|} = |\cos(n)| \le 1 \qquad \text{for all } n.$$

■

Propagation of Error

- Consider $p = \widehat{p} + \epsilon_p$ and $q = \widehat{q} + \epsilon_q,$
- The sum

$$p + q = (\widehat{p} + \epsilon_p) + (\widehat{q} + \epsilon_q) = (\widehat{p} + \widehat{q}) + (\epsilon_p + \epsilon_q).$$

- The multiplication

$$pq = (\widehat{p} + \epsilon_p)(\widehat{p} + \epsilon_q) = \widehat{p}\widehat{q} + \widehat{p}\epsilon_q + \widehat{q}\epsilon_p + \epsilon_p\epsilon_q.$$

$$R_{pq} = \frac{pq - \widehat{p}\widehat{q}}{pq} = \frac{\widehat{p}\epsilon_q + \widehat{q}\epsilon_p + \epsilon_p\epsilon_q}{pq} = \frac{\widehat{p}\epsilon_q}{pq} + \frac{\widehat{q}\epsilon_p}{pq} + \frac{\epsilon_p\epsilon_q}{pq}.$$

$$\Downarrow$$

$$R_{pq} = \frac{pq - \widehat{p}\widehat{q}}{pq} \approx \frac{\epsilon_q}{q} + \frac{\epsilon_p}{p} + 0 = R_q + R_p.$$

Propagation of Error

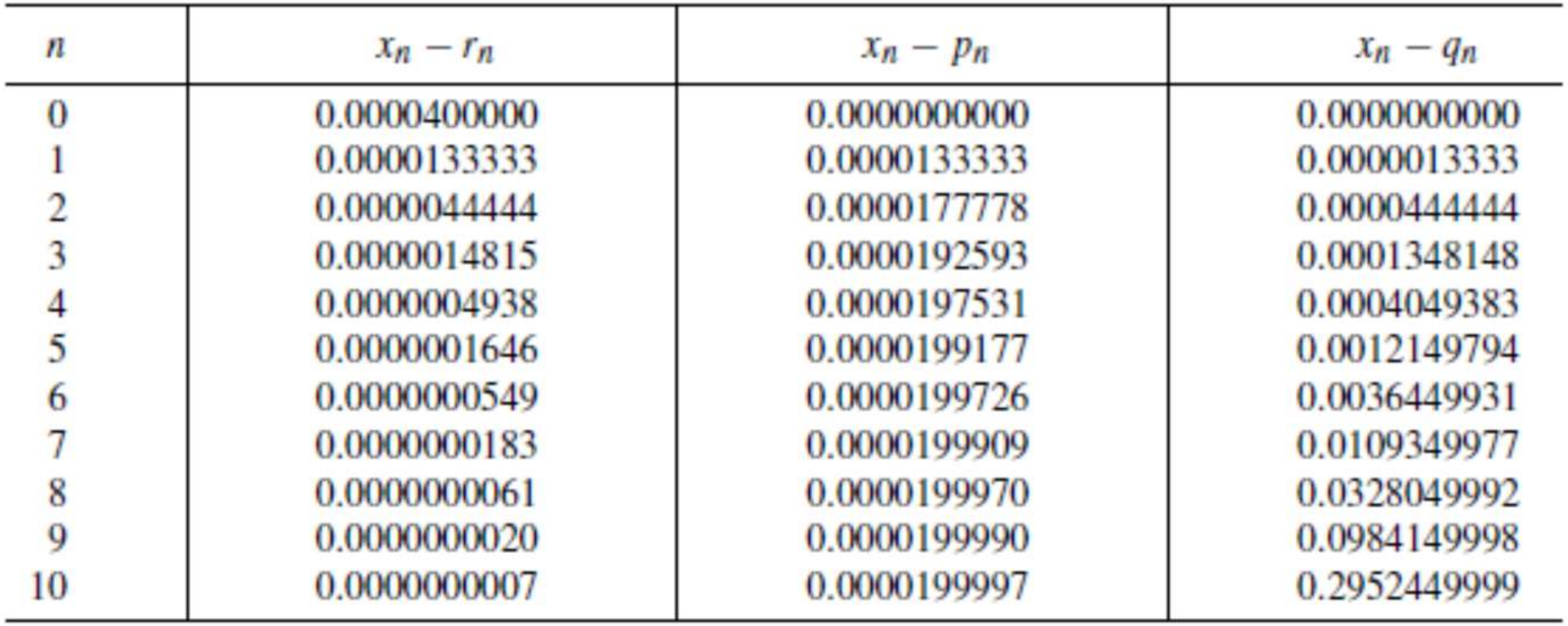
- Stable and unstable

- An example: $r_0 = 1 \quad \text{and} \quad r_n = \frac{1}{3} r_{n-1} \quad \text{for } n = 1, 2, \ldots,$

$$p_0 = 1,\ p_1 = \frac{1}{3}, \quad \text{and} \quad p_n = \frac{4}{3} p_{n-1} - \frac{1}{3} p_{n-2} \quad \text{for } n = 2, 3, \ldots,$$

$$q_0 = 1,\ q_1 = \frac{1}{3}, \quad \text{and} \quad q_n = \frac{10}{3} q_{n-1} - q_{n-2} \quad \text{for } n = 2, 3, \ldots.$$

- Consider r_0=0.99996, p_1=0.33332, q_1=0.33332



n	$x_n - r_n$	$x_n - p_n$	$x_n - q_n$
0	0.0000400000	0.0000000000	0.0000000000
1	0.0000133333	0.0000133333	0.0000013333
2	0.0000044444	0.0000177778	0.0000444444
3	0.0000014815	0.0000192593	0.0001348148
4	0.0000004938	0.0000197531	0.0004049383
5	0.0000001646	0.0000199177	0.0012149794
6	0.0000000549	0.0000199726	0.0036449931
7	0.0000000183	0.0000199909	0.0109349977
8	0.0000000061	0.0000199970	0.0328049992
9	0.0000000020	0.0000199990	0.0984149998
10	0.0000000007	0.0000199997	0.2952449999

Uncertainty in Data

- Data from real-world problems contain uncertainty or error. This type of error is referred to as ***noise***

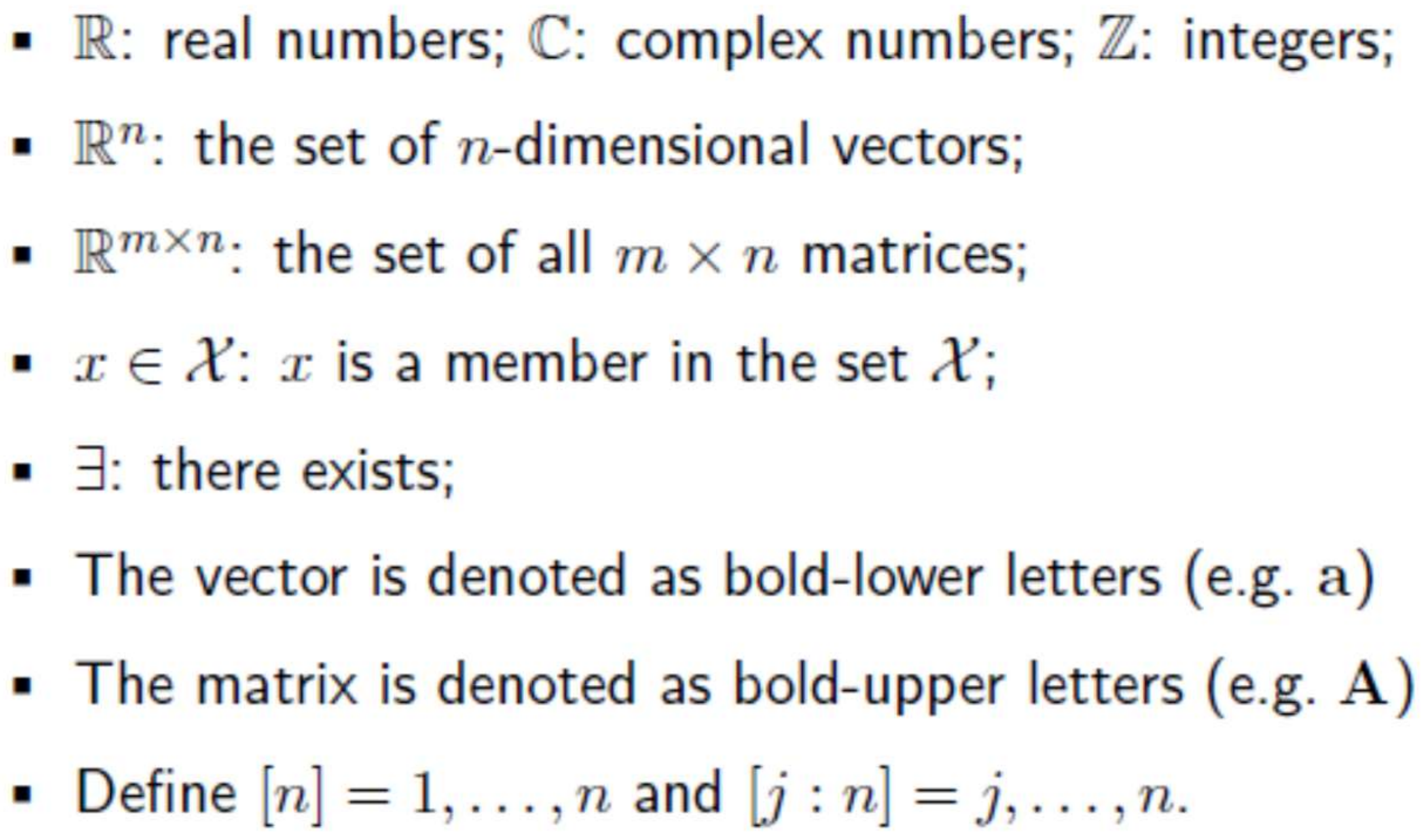
- An improvement of precision is not accomplished by performing successive computations using noisy data.

- If start with data with *d* significant digits of accuracy, then the result of a computation should be reported in *d* significant digits of accuracy.

- Example: for data p_1=4.152 and p_2=0.07931, then p_1+p_2=4.231, instead of p_1+p_2=4.23131.

Vectors and Matrix

Notations



- $\mathbb{R}$: real numbers; $\mathbb{C}$: complex numbers; $\mathbb{Z}$: integers;
- $\mathbb{R}^n$: the set of n-dimensional vectors;
- $\mathbb{R}^{m \times n}$: the set of all $m \times n$ matrices;
- $x \in \mathcal{X}$: x is a member in the set $\mathcal{X}$;
- $\exists$: there exists;
- The vector is denoted as bold-lower letters (e.g. $\mathbf{a}$)
- The matrix is denoted as bold-upper letters (e.g. $\mathbf{A}$)
- Define $[n] = 1, \ldots, n$ and $[j : n] = j, \ldots, n$.

Notations

- Vector: $\mathbf{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ and a_i is the i-th entry of $\mathbf{a}$.
- Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$,

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

- Matrix-vector multiplication. $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{A}\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n \in \mathbb{R}^m.$$

- Matrix-matrix multiplication. $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$.

$$\mathbf{A}\mathbf{B} = \mathbf{A}(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p) = (\mathbf{A}\mathbf{b}_1, \mathbf{A}\mathbf{b}_2, \ldots, \mathbf{A}\mathbf{b}_p) \in \mathbb{R}^{m \times p}.$$

Special Matrices

- Symmetric matrices $\mathbb{S}^n$:

$$\mathbf{A} \in \mathbb{S}^n \Leftrightarrow \mathbf{A} = \mathbf{A}^\top \Leftrightarrow a_{ij} = a_{ji}, \forall i, j = [n].$$

- Lower triangular matrices $\mathcal{L}$

$$\mathbf{A} \in \mathcal{L} \quad \Leftrightarrow \quad a_{ij} = 0 \text{ if } i < j.$$

- Upper triangular matrices $\mathcal{U}$

$$\mathbf{A} \in \mathcal{U} \Leftrightarrow a_{ij} = 0, \text{ if } i > j \Leftrightarrow \mathbf{A}^\top \in \mathcal{L}$$

- Positive semi-definite (definite) matrices $\mathbb{S}^n_+ (\mathbb{S}^n_{++})$.

$$\mathbf{A} \in \mathbb{S}^n_+ (\mathbb{S}^n_{++}) \quad \Leftrightarrow \quad \mathbf{x}^\top \mathbf{A} \mathbf{x} \geq (>) 0, \ \forall \mathbf{x} \neq \mathbf{0}.$$

- Orthogonal matrices $\mathcal{O}^n$: $\mathbf{A} \in \mathcal{O}^n \Leftrightarrow \mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$.

Vector norms

Let $\mathbf{x} \in \mathbb{R}^n$. There are several vector norms:

1. ℓ_2-norm

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = (|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2)^{1/2}$$

2. ℓ_1-norm

$$\|\mathbf{x}\|_1 = (|x_1| + |x_2| + \ldots + |x_n|)$$

3. ℓ_∞ norm

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Proposition (Norm equivalence)

For all $\mathbf{x} \in \mathbb{R}^n$ *have:*

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty$$

Basic concepts for vectors

- Inner product of $\mathbf{x}$ and $\mathbf{y}$:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(\theta)$$

  where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$.

- $\mathbf{x}$ is orthogonal to $\mathbf{y}$:

$$\mathbf{x}^\top \mathbf{y} = 0.$$

- For all $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ have:

$$\mathbf{x}^\top \mathbf{A} \mathbf{y} = \mathbf{y}^\top \mathbf{A}^\top \mathbf{x}.$$

Matrix norms

- Trace: $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\mathrm{tr}(\mathbf{A}) = a_{11} + a_{22} + \ldots + a_{nn}.$$

- Frobenius norm: $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2}.$$

- Matrix norm: $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{A}\mathbf{x}\| / \|\mathbf{x}\|$$

Matrix properties

- $\mathrm{diag}(\mathbf{A}) = (a_{11}, a_{22}, \ldots, a_{nn})$ and $\mathrm{Diag}(\mathbf{a})$ is

$$\mathrm{Diag}(\mathbf{a}) = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix}$$

- If $\mathbf{A} \in \mathbb{S}^n$, there exist $\mathbf{P} \in \mathcal{O}^n$ and $\mathbf{D} = \mathrm{Diag}(\mathbf{d})$ such that $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$.
- If $\mathbf{A} \in \mathbb{S}^n$, then $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$, where $\lambda_{\max}(\mathbf{A})$ denotes the maximal eigenvalue of $\mathbf{A}$.
- If $\mathbf{A} \in \mathbb{S}^n$, then $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^\top\mathbf{A}) = \sum_i d_i^2$, where d_is are eigenvalues of $\mathbf{A}$.