2021 Spring MAS 365
Chapter 1: Mathematical Preliminaries and Error Analysis

Donghwan Kim

KAIST

Mar/2,4, 2021

Limits and Continuity

Definition 1

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

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

if, for any $\epsilon > 0$, there exists $\delta > 0$ such that

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

Limits and Continuity (cont'd)

Definition 2

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

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

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

Limits and Continuity (cont'd)

Definition 3

Let $\{x_n\}_{n=1}^{\infty}$ *be an infinite sequence of real numbers. This has the* **limit** x *(i.e.,* **converges to** x*), i.e.,*

$$\lim_{n\to\infty} x_n = x,$$

if, for any $\epsilon > 0$*, there exists a positive integer* $N(\epsilon)$ *such that* $|x_n - x| < \epsilon$ *whenever* $n > N(\epsilon)$*.*

Differentiability

Definition 4

Let f be a function defined an open interval containing x_0. The function f is **differentiable** *at x_0 if the* **derivative** *of f at $\boldsymbol{x_0}$:*

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

exists. A function that has a derivative at each number in a set X is **differentiable on** X.

Q. What should we do if computing a derivative (or integration) is expensive?

Differentiability (cont'd)

- Rolle's Theorem
- Mean Value Theorem
- Extreme Value Theorem
- Intermediate Value Theorem

Taylor Polynomials and Series

Theorem 1

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

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

where

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k \quad \text{and} \quad R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}.$$

- $P_n(x)$ is called the n**th Taylor polynomial** for f about x_0, and $R_n(x)$ is called the **remainder term** (or **truncation error**) associated with $P_n(x)$.

Two Objectives of Numerical Analysis

1. Find an approximation to the solution of a given problem.
2. Determine a bound for the accuracy of the approximation.

Taylor Polynomials and Series (cont'd)

Ex. Let $f(x) = \cos x$ and $x_0 = 0$.

- Determine the second Taylor Polynomial for f about x_0.
- Determine the third Taylor Polynomial for f about x_0.

Taylor Polynomials and Series (cont'd)

Ex. Let $f(x) = \cos x$ and $x_0 = 0$.

- Use the third Taylor polynomial to approximate $\int_0^{0.1} \cos x dx$.

Round-off Errors and Computer Arithmetic

- (Finite-digit) computer arithmetic

$$2 + 2 = 4 \quad \text{and} \quad (\sqrt{3})^2 = 3?$$

- Round-off error: the error that is produced when a calculator or computer is used to perform real number calculations

Binary Machine Numbers

- Double-precision floating-point format in IEEE 754-1985 (called binary64 in IEEE 754-2008) uses the following 64-bit (binary digit) representation for a real number.

$$s\ c_{10}\ \ldots\ c_0\ f_{51}\ \ldots\ f_0$$

Symbol	Bits	Description
s	1	sign (0 if positive, 1 if negative)
c	11	characteristic (exponent with base 2)
f	52	mantissa (fraction)

- This gives a floating-point number of the form

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

where $c = \sum_{k=0}^{10} c_k 2^k$ and $f = \sum_{k=0}^{51} \frac{f_k}{2^{52-k}}$.

Binary Machine Numbers (cont'd)

Ex. Consider the machine number

$$0\ 10000000010\ 1011000\cdots 0$$

Binary Machine Numbers (cont'd)

- Exponents range form -1022 to 1023, instead of -1023 to 1024, since -1023 and 1024 are reserved for special numbers (*e.g.*, NaN, infinity, zero).

- **Underflow**: set to zero when a magnitude is less than

$$2^{-1022}(1+0) \approx 0.22251 \times 10^{-307}.$$

- **Overflow**: typically causes the computations to stop when a magnitude is greater than

$$2^{1023}(2-2^{-52}) \approx 0.17977 \times 10^{309}.$$

Decimal Machine Numbers

- For simplicity, assume that machine numbers are represented in the normalized decimal floating-point form

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

for each $i = 2, \ldots, k$.

- Consider any positive real number in a 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; **chopping** and **rounding**.

Ex. $\pi = 0.3141592\ldots \times 10^1$

Decimal Machine Numbers (cont'd)

Definition 5

Suppose that $\hat{p}$ is an approximation to p. The **actual error** *is $\hat{p} - p$, the* **absolute error** *is $|\hat{p} - p|$, and the* **relative error** *is $\frac{|\hat{p}-p|}{|p|}$, provided that $p \neq 0$.*

- Relative error of the floating-point representation $fl(y)$

$$\frac{|fl(y) - y|}{|y|}$$

Ex. $p = 0.300 \times 10^1$ and $\hat{p} = 0.3100 \times 10^1$

Decimal Machine Numbers (cont'd)

Definition 6

The number $\hat{p}$ is said to approximate p to t **significant digits** *(or figures) if t is the largest nonnegative integer for which*

$$\frac{|\hat{p} - p|}{|p|} \leq 5 \times 10^{-t}$$

Ex. Determine $\max |\hat{p} - p|$ for $p = 0.1$ and 100, when $\hat{p}$ agrees with p to four signifant digits.

Finite-Digit Arithmetic

- The arithmetic performed in a computer is not exact. A simplified finite-digit arithmetic is given by

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \otimes y = fl(fl(x) \times fl(y))$$

Ex. Let $x = \frac{5}{7} = 0.\overline{714285}$ and $y = \frac{1}{3}$. Use five-digit chopping for $x + y$ and report the relative error.

Finite-Digit Arithmetic (cont'd)

Ex. Let $x = \frac{5}{7} = 0.\overline{714285}$ and $u = 0.714251$. Determine the five-digit chopping value of $x \ominus u$ and report the relative error.

Common Error-producing Calculations

1. Subtraction of nearly equal numbers: consider two nearly equal numbers x and y such that $x > y$ and have the k-digit representations

$$fl(x) = 0.d_1 d_2 \ldots d_p \alpha_{p+1} \alpha_{p+2} \ldots \alpha_k \times 10^n$$
$$fl(y) = 0.d_1 d_2 \ldots d_p \beta_{p+1} \beta_{p+2} \ldots \beta_k \times 10^n$$

The floating-point form of $x - y$ is

$$fl(fl(x) - fl(y)) = ?$$

Common Error-producing Calculations (cont'd)

2. Division by a number with small magnitude: consider z and $\epsilon = 10^{-n}$ such that $fl(z) = z + \delta$

$$\frac{z}{\epsilon} \approx fl\left(\frac{fl(z)}{fl(\epsilon)}\right) = (z+\delta) \times 10^{n}$$

Common Error-producing Calculations (cont'd)

Ex. The roots of $ax^2 + bx + c = 0$, when $a \neq 0$, are

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

Nested Arithmetic

- Accuracy loss due to round-off error can be reduced by rearranging calculations, *e.g.*,

$$f(x) = x^3 - 6.1x^2 + 3.2x + 1.5 = ((x - 6.1)x + 3.2)x + 1.5$$

Algorithm and Pseudocode

- An **algorithm** is a procedure that describes, in an unambiguous manner, a finite sequence of steps to be performed in a specified order.
- We use a **pseudocode** to describe the algorithms.

For $i = 1, 2, \ldots, n$

Set $x_i = a + i * h$

While $i < N$ do Steps $3 - 6$

If $\ldots$ then

- The steps in the algorithms follow the rules of structured program construction.

Algorithm and Pseudocode (cont'd)

- INPUT: $N, x_1, x_2, \ldots, x_n$.
- OUTPUT: $SUM = \sum_{i=1}^{N} x_i$
- Step 1: Set $SUM = 0$.
- Step 2: For $i = 1, 2, \ldots, N$ do
  set $SUM = SUM + x_i$.
- Step 3: OUTPUT (SUM);
  STOP.

Characterizing Algorithms

- A variety of approximation problems will be studied throughout the course. We thus need a variety of conditions to categorize their accuracy.
- Stability: An algorithm is said to be **stable** if small changes in the initial data produce correspondingly small changes in the final results; otherwise it is said to be **unstable**. An algorithm is called **conditionally stable**, if it is stable only for certain choices of initial data.

Characterizing Algorithms (cont'd)

Definition 7

Suppose that E_0 denotes an error introduced at some stage in the calculations and E_n represents the magnitude of the error after n subsequent operations.

- *If $E_n \approx CnE_0$ for a constant C, then the growth of error is said to be* **linear**.
- *If $E_n \approx C^n E_0$ of some $C > 1$, then the growth of error is called* **exponential**.

- Linear growth of error is usually unavoidable, and such behavior is considered stable.

Characterizing Algorithms (cont'd)

Ex. For any constants c_1 and c_2,

$$p_n = c_1 \left(\frac{1}{3}\right)^n + c_2 3^n$$

is a solution to the recursive equation

$$p_n = \frac{10}{3} p_{n-1} - p_{n-2}, \quad \text{for } n = 2, 3, \ldots$$

- Consider five-digit rounding arithmetic.

Rates of Convergence

Definition 8

Suppose $\{\beta_n\}_{n=1}^{\infty}$ is a sequence known to converge to zero, and $\{\alpha_n\}_{n=1}^{\infty}$ converges to number α. If a positive constant K exists with

$$|\alpha_n - \alpha| \le K|\beta_n|, \quad \textit{for large } n,$$

then we say that $\{\alpha_n\}_{n=1}^{\infty}$ converges to α with **rate, or order, of convergence** *$O(\beta_n)$. (read "big oh of β_n".) It is indicated by writing $\alpha_n = \alpha + O(\beta_n)$.*

- In most of cases, we use

$$\beta_n = \frac{1}{n^p}$$

  for some number $p > 0$.
- We are generally interested in the largest value of p.
- $o(\beta_n)$ (read "small oh of β_n"), when "$<$" is used instead of "$\le$".

Rates of Convergence (cont'd)

Ex. Determine rate of convergence for the sequence $\{\alpha_n\}_{n=1}^{\infty}$, where $\alpha_n = \frac{3n^2+n+1}{n^2}$ and $\lim_{n\to\infty} \alpha_n = 3$.

Rates of Convergence (cont'd)

- $O(\cdot)$ notation for describing the rate at which functions converge.

Definition 9

Suppose that $\lim_{h\to 0} G(h) = 0$ *and* $\lim_{h\to 0} F(h) = L$. *If a positive constant* K *exists with*

$$|F(h) - L| \leq K|G(h)|, \qquad \textit{for sufficiently small } h,$$

then we write $F(h) = L + O(G(h))$.

- We usually consider $G(h) = h^p$, where $p > 0$.
- We are generally interested in the largest value of p.

Rates of Convergence (cont'd)

Ex. Use the third Taylor polynomial about 0 to show that $\cos h + \frac{1}{2}h^2 = 1 + O(h^4)$.