Floating Point Number Systems

Simon Fraser University – Surrey Campus

MACM 316 – Spring 2005

Instructor: Ha Le

Overview

- Real number system
- Examples
- Absolute and relative errors
- Floating point numbers
- Roundoff error analysis
- Conditioning and stability
- A stability analysis
- Rate of convergence

Real number system

- The arithmetic of the mathematically defined real number system, denoted by $\mathbb{R}$, is used.
- $\mathbb{R}$ is infinite in

  (1) *extent*, i.e., there are numbers $x \in \mathbb{R}$ such that $|x|$ is arbitrarily large.

  (2) *density*, i.e., any interval $I = \{x \,|\, a \leq x \leq b\}$ of $\mathbb{R}$ is an infinite set.

- Computer systems can only represent finite sets of numbers, so all the actual implementations of algorithms must use *approximations* to $\mathbb{R}$ and *inexact arithmetic*.

Example 1: Evaluate $I_n = \int_0^1 \frac{x^n}{x+\alpha} dx$

$$\begin{cases} I_0 & = \int_0^1 \frac{1}{(x+\alpha)} dx & = \ln\left(\frac{\alpha+1}{\alpha}\right) \\ I_n + \alpha\, I_{n-1} & = \int_0^1 \frac{x^n + \alpha\, x^{n-1}}{x+\alpha} dx & = \frac{1}{n} \end{cases}$$

$$\Longrightarrow \quad I_n = \frac{1}{n} - \alpha\, I_{n-1}, \quad I_0 = \ln\left(\frac{\alpha+1}{\alpha}\right)$$

Using single precision *floating point* arithmetic:

$$\alpha = .5 \Rightarrow I_{100} = 6.64 \times 10^{-3}, \quad \alpha = 2.0 \Rightarrow I_{100} = 2.1 \times 10^{22}.$$

Note. If $\alpha > 1$, $(x+\alpha) > 1$ for $0 \le x \le 1$. Hence,

$$\int_0^1 \frac{x^n}{x+\alpha} dx \le \int_0^1 x^n dx = \frac{1}{n+1}.$$

Example 2: Evaluate $e^{-5.5}$

Recall. $e^y = \sum_{n=0}^{\infty} \frac{y^n}{n\,!} = 1 + y + \frac{y^2}{2\,!} + \frac{y^3}{3\,!} + \cdots$.

Using a calculator which carries five *significant figures*.

Method 1.

$$x_1 = e^{-5.5} = \sum_{n=0}^{20} \frac{(-5.5)^n}{n\,!} = .0026363$$

Method 2.

$$x_2 = e^{-5.5} = \frac{1}{e^{5.5}} = \frac{1}{\sum_{n=0}^{20} \frac{(5.5)^n}{n\,!}} = .0040865$$

Note. The correct answer, up to five significant digits, is

$$x_e = e^{-5.5} = .0040868$$

Absolute and Relative Error

Computed result: x, correct mathematical result: x_e.

$$Err_{abs} = |x_e - x|\,, \quad Err_{rel} = \frac{|x_e - x|}{|x_e|}$$

Definition. The *significant digits* in a number are the digits starting with the first, i.e., leftmost, nonzero digit (e.g., $.00\underbrace{40868}$).

- x is said to approximate x_e to about s significant digits if the relative error satisfies

$$0.5 \times 10^{-s} \le \frac{|x_e - x|}{|x_e|} < 5.0 \times 10^{-s}.$$

Example 3: Relative Error and Significant Digits

In Example 2,

$$x_e = .0040868, \quad x_1 = .0026363, \quad x_2 = .0040865.$$

Method 1.

$$0.5 \times 10^{-1} \leq Err_{rel} = \frac{|x_e - x_1|}{|x_e|} \approx 3.5 \times 10^{-1} < 5.0 \times 10^{-1}.$$

Hence, x_1 has approximately one significant digit correct (in this example, x_1 has zero correct digits).

Method 2.

$$0.5 \times 10^{-4} \leq Err_{rel} = \frac{|x_e - x_2|}{|x_e|} \approx 0.7 \times 10^{-4} < 5.0 \times 10^{-4}.$$

Hence, x_2 has approximately four significant digits correct (in this example, x_2 is indeed correct to four significant digits).

Representation of Numbers in $\mathbb{R}$

Let $\beta \in \mathbb{N} \setminus \{0\}$ be *the base* for a number system, e.g.,

$$\beta = 10 \text{ (decimal) }, \quad \beta = 2 \text{ (binary)}, \quad \beta = 16 \text{ (hexadecimal)}.$$

Each $x \in \mathbb{R}$ can be represented by an *infinite* base β expansion in the *normalized* form

$$.d_0\, d_1\, d_2 \ldots d_{t-1}\, d_t \ldots \times \beta^p$$

where $p \in \mathbb{Z}$, d_k are digits in base β, i.e. $d_k \in \{0, 1, \ldots, \beta - 1\}$, and $d_0 \neq 0$.

Example.

$$732.5051 \Longrightarrow .7325051 \times 10^3, \quad -0.005612 \Longrightarrow -0.5612 \times 10^{-2}.$$

Floating Point Numbers

Recall. $\mathbb{R}$ is infinite in extent and density.

Floating point number systems limit

- the *infinite density* of $\mathbb{R}$ by representing only a *finite* number, t, of digits in the expansion;
- the *infinite extent* of $\mathbb{R}$ by representing only a finite number of integer values for the exponent p, i.e., $L \leq p \leq U$ for specified integers $L > 0$ and $U > 0$.

Therefore, each number in such a system is precisely of the form

$$.d_0\, d_1\, d_2 \ldots d_{t-1} \times \beta^p, \quad L \leq p \leq U, \quad d_0 \neq 0$$

or 0 (a very special floating point number).

Two Standardized Systems

A floating point number system is denoted by $F(\beta, t, L, U)$ or simply by F when the parameters are understood.

Two standardized systems for digital computers widely used in the design of software and hardware:

IEEE single precision: $\{\beta = 2; t = 24; L = -127; U = 128\}$,

IEEE double precision: $\{\beta = 2; t = 53; L = -1023; U = 1024\}$.

Note. An *exception* occurs if the exponent is out of range, which leads to a state called *overflow* if the exponent is too large, or *underflow* if the exponent is too small.

Truncation of a Real Number

Let

$$x = .d_0\, d_1 \ldots d_{n-1}\, d_n \ldots d_{t-1} \times \beta^p.$$

Using n digits:

- Rounding:

$$x = \begin{cases} .d_0\, d_1 \ldots d_{n-1} \times \beta^p & \text{if } 0 \le d_n \le 4, \\ .d_0\, d_1 \ldots (d_{n-1} + 1) \times \beta^p & \text{if } 5 \le d_n \le 9. \end{cases}$$

- Chopping:

$$x = .d_0\, d_1 \ldots d_{n-1} \times \beta^p.$$

Relationship between $x \in \mathbb{R}$ and $\mathrm{fl}(x) \in F$

For $x \in \mathbb{R}$, let $\mathrm{fl}(x) \in F(\beta, t, L, U)$ be its floating point approximation. Then

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \le \mathcal{E}. \tag{1}$$

$\mathcal{E}$: machine epsilon, or unit roundoff error.

$$\mathcal{E} = \begin{cases} \frac{1}{2}\beta^{1-t} & \text{for rounding,} \\ \beta^{1-t} & \text{for chopping.} \end{cases}$$

By (1), $\mathrm{fl}(x) - x = \delta\, x$, for some δ such that $|\delta| \le \mathcal{E}$. Hence, $\mathrm{fl}(x) = x(1+\delta)$, $-\mathcal{E} \le \delta \le \mathcal{E}$.

Example. Denote the addition operator in F by $\oplus$. For $w, z \in F$, $w \oplus z = \mathrm{fl}(w+z) = (w+z)(1+\delta)$.

Roundoff Error Analysis: an Exercise

How does $(a \oplus b) \oplus c$ differ from the true sum $a + b + c$?

$$
\begin{aligned}
(a \oplus b) \oplus c &= (a+b)(1+\delta_1) \oplus c = ((a+b)(1+\delta_1) + c)(1+\delta_2) \\
&= (a+b+c) + (a+b)\delta_1 + (a+b+c)\delta_2 + (a+b)\delta_1\,\delta_2.
\end{aligned}
$$

$$
\Longrightarrow \quad |(a+b+c) - ((a \oplus b) \oplus c)| \le (|a| + |b| + |c|)(|\delta_1| + |\delta_2| + |\delta_1||\delta_2|).
$$

If $(a + b + c) \neq 0$, then

$$
Err_{rel} = \frac{|(a+b+c) - ((a \oplus b) \oplus c)|}{|a+b+c|} \le \frac{|a| + |b| + |c|}{|a+b+c|}\,(2\mathcal{E} + \mathcal{E}^2).
$$

- If $|a+b+c| \approx |a| + |b| + |c|$ (e.g., $a, b, c \in \mathbb{R}^+$, or $a, b, c \in \mathbb{R}^-$, then Err_{rel} is bounded by $2\mathcal{E} + \mathcal{E}^2$ which is small;
- If $|a+b+c| << |a| + |b| + |c|$, then Err_{rel} can be quite large.

Roundoff Error Analysis: a Generalization

- Addition of N numbers. If $\sum_{i=1}^{N} x_i \neq 0$, then

$$Err_{rel} = \frac{|\sum_{i=1}^{N} x_i - \mathrm{fl}(\sum_{i=1}^{N} x_i)|}{|\sum_{i=1}^{N} x_i|} \leq \frac{\sum_{i=1}^{N} |x_i|}{|\sum_{i=1}^{N} x_i} \, 1.01 \, N \, \mathcal{E}.$$

(The appearance of the factor 1.01 is an artificial technicality.)

- Product of N numbers. If $x_i \neq 0$, $1 \leq i \leq N$, then

$$Err_{rel} = \frac{|\prod_{i=1}^{N} x_i - \mathrm{fl}(\prod_{i=1}^{N} x_i)|}{|\prod_{i=1}^{N} x_i|} \leq 1.01 \, N \, \mathcal{E}.$$

Roundoff Error Analysis: an Example

In $F(10, 5, -10, 10)$, let

$$a = 10000., \quad b = 3.1416, \quad c = -10000.$$

Then $|a| + |b| + |c| = 20003.1416$ and $a + b + c = 3.1416$. Hence,

$$0.5 \times 10^0 \le Err_{rel} \le 6367.2(2\mathcal{E} + \mathcal{E}^2) \approx 0.6 < 5.0 \times 10^0.$$

This relative error implies that there may be no significant digits correct in the result. Indeed,

$$(a \oplus b) \oplus c = 10003. \oplus (-10000.) = 3.0000.$$

Therefore, the computed sum actually has one significant digit correct.

Conditioning

Consider a Problem $\mathcal{P}$ with input values $\mathcal{I}$ and output values $\mathcal{O}$. If a relative change of size $\Delta\mathcal{I}$ in one or more input values causes a relative change in the mathematically correct output values which is guaranteed to be small (i.e., not too much larger than $\Delta\mathcal{I}$), then $\mathcal{P}$ is said to be *well-conditioned*. Otherwise, $\mathcal{P}$ is said to be *ill-conditioned*.

Remark. The above definition is *independent* of any particular choice of algorithm and *independent* of any particular number system. *It is a statement about the mathematical problem.*

Condition Number

$$\mathcal{P}: \ \mathcal{I} = \{x\}, \ \mathcal{O} = \{f(x)\}.$$

From Taylor series expansion:

$$\begin{aligned} f(x+\Delta x) &= f(x) + f^{'}(x)\,\Delta x + \frac{1}{2} f^{''}(x)\Delta x^2 + O(\Delta x^3) \\ &\approx f(x) + f^{'}(x)\,\Delta x \end{aligned}$$

assuming that $|\Delta x|$ is small. Hence,

$$\frac{|f(x) - f(x+\Delta x)|}{|f(x)|} \approx \frac{|f^{'}(x)||\Delta x|}{|f(x)|} = \underbrace{\frac{|x||f^{'}(x)|}{|f(x)|}}_{\kappa(\mathcal{P})} \times \frac{|\Delta x|}{|x|}.$$

$\kappa(\mathcal{P})$: the *condition number* of $\mathcal{P}$.

Condition Number: an Example

In Example 2, $\mathcal{I} = \{x = -5.5\}$, $\mathcal{O} = \{f(x) = e^x\}$, and $\kappa(\mathcal{P}) = |x| = 5.5$. Hence, roundoff errors (in relative error) of size $\mathcal{E}$ can lead to relative errors in the output bounded by

$$\text{Err}_{rel} \approx \kappa(\mathcal{P})\mathcal{E}.$$

For example, if $\mathcal{E} \approx 10^{-5}$, then

$$0.5 \times 10^{-4} \leq \text{Err}_{rel} < 5.0 \times 10^{-4},$$

and we should expect to have about four significant digits correct.

Stability

Consider a Problem $\mathcal{P}$ with condition number $\kappa(\mathcal{P})$, and suppose that we apply Algorithm $\mathcal{A}$ to solve $\mathcal{P}$. If we can guarantee that the computed output values from $\mathcal{A}$ will have relative errors not too much larger than the errors due to the condition number $\kappa(\mathcal{P})$, then $\mathcal{A}$ is said to be *stable*. Otherwise, if the computed output values from $\mathcal{A}$ can have much larger relative errors, then $\mathcal{A}$ is said to be *unstable*.

Example. For the Problem $\mathcal{P}$ in Example 2, $\mathcal{P}$ is well-conditioned. Method 1 is unstable, while Method 2 is stable.

A Stability Analysis

In Example 1, computing $I_n = \int_0^1 \frac{x^n}{x+\alpha} dx$ is reduced to solving

$$I_n = \frac{1}{n} - \alpha\, I_{n-1}, \quad I_0 = \ln\left(\frac{\alpha+1}{\alpha}\right).$$

We state *without proof* that this is a well-conditioned problem.

Suppose that the floating point representation of I_0 introduces some error ϵ_0. For simplicity, assume that no other errors are introduced at each stage of the computation after I_0 is computed. Let $(I_n)_A$ and $(I_n)_E$ be the approximate value and the exact value of I_n, respectively. Then

$$(I_n)_E = \frac{1}{n} - \alpha\,(I_{n-1})_E, \quad (I_n)_A = \frac{1}{n} - \alpha\,(I_{n-1})_A.$$

Set $\epsilon_n = (I_n)_A - (I_n)_E$. Then

$$\epsilon_n = (-\alpha)\,\epsilon_{n-1} = (-\alpha)^n\,\epsilon_0.$$

If $|\alpha| > 1$, then any initial error ϵ_0 is magnified by an unbounded amount as $n \to \infty$. On the other hand, if $|\alpha| < 1$, then any initial error is damped out. We conclude that the algorithm is stable if $|\alpha| < 1$, and unstable if $|\alpha| > 1$.

Taylor Series

The set of all functions that have n continuous derivatives on a set X is denoted $C^n(X)$, and the set of functions that have derivatives of all orders on X is denoted $C^\infty(X)$, where X consists of all numbers for which the functions are defined.

Taylor's theorem provides the most important tool for this course.

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

$$f(x) = \underbrace{P_n(x)}_{n\text{th Taylor polynomial}} + \underbrace{R_n(x)}_{\text{remainder term (or truncation error)}},$$

$$\begin{aligned} P_n(x) &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k\,!}(x-x_0)^k \\ &= f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n\,!}(x-x_0)^n, \\ R_n(x) &= \frac{f^{(n+1)}(\xi(x))}{(n+1)\,!}(x-x_0)^{n+1}. \end{aligned}$$

Taylor Series: an Example

- Find the third degree Taylor polynomial $P_3(x)$ for $f(x) = \sin x$ at the expansion point $x_0 = 0$.

Note that $f \in C^\infty(\mathbb{R})$. Hence, Taylor's theorem is applicable.

$$P_3(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f'''(0)}{3!}x^3.$$

$$\begin{aligned}
f'(x) &= \cos x &\Longrightarrow\quad f'(0) &= \cos 0 &= 1,\\
f''(x) &= -\sin x &\Longrightarrow\quad f''(0) &= -\sin 0 &= 0,\\
f'''(x) &= -\cos x &\Longrightarrow\quad f'''0) &= -\cos 0 &= -1.
\end{aligned}$$

Also, $f(0) = \sin 0 = 0$. Hence, $P_3(x) = x - \dfrac{x^3}{6}$.

- What about the truncation error ?

$$R_3(x) = \frac{f''''(\xi)}{4!}x^4 = \frac{\sin\xi}{24}\,x^4 \text{ where } \xi \in (0, x).$$

- How big could the error be at $x = \pi/2$?

$$R_3\left(\frac{\pi}{2}\right) = \frac{\sin\xi}{24}\,\left(\frac{\pi}{2}\right)^4 \text{ where } \xi \in \left(0, \frac{\pi}{2}\right).$$

Since $|\sin\xi| \leq 1$ for all $\xi \in \left(0, \frac{\pi}{2}\right)$,

$$\left|R_3\left(\frac{\pi}{2}\right)\right| \leq \frac{1}{24}\left(\frac{\pi}{2}\right)^4 \approx 2.025.$$

- Actual error ?

$$\left|f\left(\frac{\pi}{2}\right) - P_3\left(\frac{\pi}{2}\right)\right| = \left|\sin\frac{\pi}{2} - \left(\frac{\pi}{2} - \frac{(\pi/2)^3}{6}\right)\right| \approx 0.075.$$

In this example, the error bound is much larger than the actual error.

Rate of Convergence

Throughout this course, we will study numerical methods which solve a problem by constructing a sequence of (hopefully) better and better approximations which converge to the required solution. A technique is required to compare the convergence rates of different methods.

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

$$|\alpha_n - \alpha| \leq K|\beta_n| \quad \text{for large } n,$$

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

In nearly every situation, we use

$$\beta_n = \frac{1}{n^p} \quad \text{for some number } p > 0.$$

Usually we compare how fast $\{\alpha_n\}_{n=1}^{\infty} \to \alpha$ with how fast $\beta_n = 1/n^p \to 0$.

Example. $\{\alpha_n\}_{n=1}^{\infty} \to \alpha$ like $1/n$ or $1/n^2$.

$$\frac{1}{n^2} \to 0 \quad \text{faster than} \quad \frac{1}{n},$$
$$\frac{1}{n^3} \to 0 \quad \text{faster than} \quad \frac{1}{n^2}.$$

We are most interested in the largest value of p with $\alpha_n = \alpha + O(1/n^p)$.

To find the rate of convergence, we can use the definition: find the largest p such that

$$|\alpha_n - \alpha| \leq K\,|\beta_n| = K\,\frac{1}{n^p} \quad \text{for } n \text{ large},$$

or equivalently, find the largest p so that

$$\lim_{n\to\infty} \frac{|\alpha_n - \alpha|}{|\beta_n|} = \lim_{n\to\infty} \frac{|\alpha_n - \alpha|}{1/n^p} = K.$$

Note. K must be a constant, and cannot be "∞".

Rate of Convergence: an Example

For $\alpha_n = (n+1)/n^2$, $\hat{\alpha}_n = (n+3)/n^3$,
$\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \hat{\alpha}_n = 0 = \alpha$.

$$\lim_{n\to\infty} \frac{|\alpha_n - \alpha|}{1/n^p} = \lim_{n\to\infty} \frac{n+1}{n^2} n^p = \begin{cases} 1 & \text{if } p = 1, \\ \infty & \text{if } p \geq 2; \end{cases}$$

$$\lim_{n\to\infty} \frac{|\hat{\alpha}_n - \alpha|}{1/n^p} = \lim_{n\to\infty} \frac{n+3}{n^3} n^p = \begin{cases} 0 & \text{if } p = 1, \\ 1 & \text{if } p = 2, \\ \infty & \text{if } p \geq 3. \end{cases}$$

Hence, $\alpha_n = 0 + O(1/n)$, and $\hat{\alpha}_n = 0 + O(1/n^2)$.

Rate of Convergence: another Example

For $\alpha_n = \sin(1/n)$, we have $\lim_{n\to\infty} \alpha_n = 0$. For all $n \in \mathbb{N} \setminus \{0\}$, $\sin(1/n) > 0$. Hence, to find the rate of convergence of α_n, we need to find the largest p so that

$$\lim_{n\to\infty} \frac{|\alpha_n - \alpha|}{|\beta_n|} = \lim_{n\to\infty} \frac{|sin(1/n) - 0|}{1/n^p} = \lim_{n\to\infty} \frac{sin(1/n)}{1/n^p} = K.$$

Use change of variable $h = 1/n$: $\lim_{n\to\infty} \frac{\sin(1/n)}{1/n^p} \equiv \lim_{h\to 0} \frac{\sin h}{h^p}$.

Apply Taylor series expansion to $\sin h$ at $h = 0$:

$$\lim_{h\to 0} \frac{\sin h}{h^p} = \lim_{h\to 0} \frac{h - h^3/6 + \cdots}{h^p} = \begin{cases} 1 & \text{if } p = 1, \\ \infty & \text{if } p \in \mathbb{N} \setminus \{0, 1\}. \end{cases}$$

Hence, the rate of convergence is $O(h)$ or $O(1/n)$.

Rate of Convergence for Functions

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

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

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

In general, $G(h) = h^p$, where $p > 0$, and we are interested in finding the largest value of p for which $F(h) = L + O(h^p)$.

Rate of Convergence for Functions: an Example

Let $F(h) = \cos h + \frac{1}{2}h^2$. Then $L = \lim_{h \to 0} F(h) = 1$. We have

$$\begin{aligned}
\lim_{h \to 0} \frac{|F(h) - L|}{h^p} &= \lim_{h \to 0} \frac{\cos h + 1/2h^2 - 1}{h^p} \\
&= \lim_{h \to 0} \frac{\left(1 - 1/2h^2 + 1/24h^4 - \cdots\right) + 1/2h^2 - 1}{h^p} \\
&= \lim_{h \to 0} \frac{1/24h^4 - \cdots}{h^p} \\
&= \begin{cases} 1/24 & \text{if } p = 4, \\ \infty & \text{if } p > 4. \end{cases}
\end{aligned}$$

Hence, $F(h) = 1 + O(h^4)$.