MA 322: Scientific Computing

Department of Mathematics
Indian Institute of Technology Guwahati

January 11, 2023

CHAPTER 1: ERRORS

Floating-point representations

Table 1.1 Floating-point representations on various computers

Machine	S/D	R/C	β	t	L	U	δ	M
CDC CYBER 170	S	R	2	48	-976	1071	3.55E − 15	2.81E14
CDC CYBER 205	S	C	2	47	$-28,626$	28,718	1.42E − 14	1.41E14
CRAY-1	S	C	2	48	-8192	8191	7.11E − 15	2.81E14
DEC VAX	S	R	2	24	-127	127	5.96E − 8	1.68E7
DEC VAX	D	R	2	53	-1023	1023	1.11E − 16	9.01E15
HP-11C, 15C	S	R	10	10	-99	99	5.00E − 10	1.00E10
IBM 3033	S	C	16	6	-64	63	9.54E − 7	1.68E7
IBM 3033	D	C	16	14	-64	63	2.22E − 16	7.21E16
Intel 8087	S	R	2	24	-126	127	5.96E − 8	1.68E7
Intel 8087	D	R	2	53	-1022	1023	1.11E − 16	9.01E15
PRIME 850	S	R	2	23	-128	127	1.19E − 7	8.39E6
PRIME 850	S	C	2	23	-128	127	1.19E − 7	8.39E6
PRIME 850	D	C	2	47	$-32,896$	32,639	1.42E − 14	1.41E14

Sources of errors

- Mathematical modeling of a physical problem
- Blunders (arithmetic errors, programming errors)
- Uncertainty in physical data
- Machine errors
- Mathematical truncation error, e.g., computing $\sqrt{1+x}$ for small x.

Error propagation

- Let $*$ denotes the arithmetic operations $+, -, \times, \div$, and $\hat{*}$ be the computer version of the same operation.
- Let x_A and y_A be the numbers used for calculations, and suppose they are in errors, with true values $x_T = x_A + \epsilon, \qquad y_T = y_A + \eta$.
- $x_A * y_A$ is the number actually computed, and for its error,

$$x_T * y_T - x_A \hat{*} y_A = [x_T * y_T - x_A * y_A] + [x_A * y_A - x_A \hat{*} y_A]$$

- $[x_T * y_T - x_A * y_A]$ is called the propagated error and $[x_A * y_A - x_A \hat{*} y_A]$ is called the rounding or chopping error.
- $x_A \hat{*} y_A = \mathrm{fl}(x_A * y_A)$ – $x_A ast y_A$ is computed exactly and then rounded.
- $|x_A * y_A - x_A \hat{*} y_A| \leq \dfrac{\beta}{2} |x_A * y_A| \beta^{-t}$.

Error propagation

Multiplication:

- For the error in $x_A y_A$,

$$\begin{aligned} x_T y_T - x_A y_A &= x_T y_T - (x_T - \epsilon)(y_T - \eta) \\ &= x_T \eta + y_T \epsilon - \epsilon\eta \end{aligned}$$

- Relative error,

$$\begin{aligned} \mathrm{Rel}(x_A y_A) &\equiv \frac{x_T y_T - x_A y_A}{x_T y_T} = \frac{\eta}{y_T} + \frac{\epsilon}{x_T} - \frac{\epsilon}{x_T}\frac{\eta}{y_T} \\ &= \mathrm{Rel}(x_A) + \mathrm{Rel}(y_A) - \mathrm{Rel}(x_A)\mathrm{Rel}(y_A) \end{aligned}$$

- For $|\mathrm{Rel}(x_A)|$, $|\mathrm{Rel}(y_A)| \ll 1$,

$$\mathrm{Rel}(x_A y_A) \approx \mathrm{Rel}(x_A) + \mathrm{Rel}(y_A)$$

Error propagation

- Division:

$$\mathrm{Rel}\frac{x_A}{y_A} = \frac{\mathrm{Rel}(x_A) - \mathrm{Rel}(y_A)}{1 - \mathrm{Rel}(y_A)} \tag{1}$$

For $|\mathrm{Rel}(y_A)| \ll 1$,

$$\mathrm{Rel}\frac{x_A}{y_A} \approx \mathrm{Rel}(x_A) - \mathrm{Rel}(y_A) \tag{2}$$

- Addition and subtraction:

$$(x_T \pm y_T) - (x_A \pm y_A) = (x_T - x_A) \pm (y_T - y_A) = \epsilon \pm \eta$$

$$\mathrm{Err}(x_A \pm y_A) = \mathrm{Err}(x_A) \pm \mathrm{Err}(y_A)$$

Loss of significance

Example

Suppose we are supposed to compute

$$\sqrt{1+x^2}-1$$

and assign it to y.

- For x small, the accuracy can be jeopardized by the subtraction of nearly equal numbers.
- The difficulty is avoided by reprogramming with a different assignment statement as,

$$y \leftarrow \frac{x^2}{\sqrt{1+x^2}+1}$$

Loss of significance

Example

Suppose we are supposed to compute

$$\sqrt{1+x^2}-1$$

and assign it to y.

- For x small, the accuracy can be jeopardized by the subtraction of nearly equal numbers.
- The difficulty is avoided by reprogramming with a different assignment statement as,

$$y \leftarrow \frac{x^2}{\sqrt{1+x^2}+1}$$

Loss of significance

Example

Suppose we are supposed to compute

$$\sqrt{1+x^2}-1$$

and assign it to y.

- For x small, the accuracy can be jeopardized by the subtraction of nearly equal numbers.
- The difficulty is avoided by reprogramming with a different assignment statement as,

$$y \leftarrow \frac{x^2}{\sqrt{1+x^2}+1}$$

CHAPTER 2: ROOT FINDINGS

The Bisection Method

Steps:

- Fix a and b such that $f(a) \cdot f(b) < 0$. Set $a_0 = a$ and $b_0 = b$.
- Set $c_0 = \frac{1}{2}(a_0 + b_0)$. Check $f(a_0) \cdot f(c_0)$ and $f(c_0) \cdot f(b_0)$. Set

$$(a_1, b_1) = \begin{cases} (a_0, c_0), & \text{if } f(a_0) \cdot f(c_0) < 0, \\ (c_0, b_0), & \text{if } f(c_0) \cdot f(b_0) < 0. \end{cases}$$

  Observe that $f(a_1) \cdot f(b_1) < 0$, so that root $c \in [a_1, b_1]$.
- Set $c_1 = \frac{1}{2}(a_1 + b_1)$. If $f(c_1) = 0$, then stop.
- In general, for $n \geq 1$, we set (a_{n+1}, b_{n+1}) by

$$(a_{n+1}, b_{n+1}) = \begin{cases} (a_n, c_n), & \text{if } f(a_n) \cdot f(c_n) < 0, \\ (c_n, b_n), & \text{if } f(c_n) \cdot f(b_n) < 0. \end{cases}$$

The Bisection Method

Steps:

- Fix a and b such that $f(a) \cdot f(b) < 0$. Set $a_0 = a$ and $b_0 = b$.
- Set $c_0 = \frac{1}{2}(a_0 + b_0)$. Check $f(a_0) \cdot f(c_0)$ and $f(c_0) \cdot f(b_0)$. Set

$$(a_1, b_1) = \begin{cases} (a_0, c_0), & \text{if } f(a_0) \cdot f(c_0) < 0, \\ (c_0, b_0), & \text{if } f(c_0) \cdot f(b_0) < 0. \end{cases}$$

  Observe that $f(a_1) \cdot f(b_1) < 0$, so that root $c \in [a_1, b_1]$.
- Set $c_1 = \frac{1}{2}(a_1 + b_1)$. If $f(c_1) = 0$, then stop.
- In general, for $n \geq 1$, we set (a_{n+1}, b_{n+1}) by

$$(a_{n+1}, b_{n+1}) = \begin{cases} (a_n, c_n), & \text{if } f(a_n) \cdot f(c_n) < 0, \\ (c_n, b_n), & \text{if } f(c_n) \cdot f(b_n) < 0. \end{cases}$$

The Bisection Method

- Thus, we construct sequences $\{a_{n+1}\}$ and $\{b_{n+1}\}$ such that $f(a_{n+1}) \cdot f(b_{n+1}) < 0$, so that root $c \in [a_{n+1}, b_{n+1}]$. Set

$$c_{n+1} = \frac{1}{2}(a_{n+1} + b_{n+1}).$$

  If $f(c_{n+1}) = 0$, STOP, else REPEAT this step.

Error Analysis:

- Note that

$$a_0 \le a_1 \le a_2 \le \cdots \le b_0$$

$$b_0 \ge b_1 \ge b_2 \ge \cdots \ge a_0$$

$$b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n) \qquad (n \ge 0)$$

$$b_n - a_n = 2^{-n}(b_0 - a_0)$$

Theorem (Theorem on Bisection Method)

If $[a_0, b_0],\ [a_1, b_1],\ \cdots [a_n, b_n], \cdots$ denote the intervals in the bisection method, then the limits $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ exist, are equal, and represent a zero of f. If $c = \lim_{n\to\infty} c_n$ and $c_n = \frac{1}{2}(a_n + b_n)$, then

$$|c - c_n| \le 2^{-(n+1)}(b_0 - a_0).$$

Error Analysis:

- Note that

$$a_0 \leq a_1 \leq a_2 \leq \cdots \leq b_0$$

$$b_0 \geq b_1 \geq b_2 \geq \cdots \geq a_0$$

$$b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n) \qquad (n \geq 0)$$

$$b_n - a_n = 2^{-n}(b_0 - a_0)$$

Theorem (Theorem on Bisection Method)

If $[a_0, b_0],\ [a_1, b_1],\ \cdots [a_n, b_n], \cdots$ *denote the intervals in the bisection method, then the limits* $\lim_{n\to\infty} a_n$ *and* $\lim_{n\to\infty} b_n$ *exist, are equal, and represent a zero of* f*. If* $c = \lim_{n\to\infty} c_n$ *and* $c_n = \frac{1}{2}(a_n + b_n)$*, then*

$$|c - c_n| \leq 2^{-(n+1)}(b_0 - a_0).$$

The Bisection Method

Disadvantages:

- Converges very slowly when compared with the methods to be discussed in the upcoming lecture(s) – the bisection method converges linearly.

Advantages:

- If $f \in C[a, b]$, the method is guaranteed to converge.
- At each step, we get upper and lower bounds on the root.