Numerical Methods

Lecture 4

CS357

David Semeraro

Key Concepts

- Machine epsilon, ϵ, is the smallest machine number for which $1 + \epsilon \neq 1$.
- In single precision, $\epsilon = 2^{-23}$.
- The relative error in representing a normalized floating point number by a machine number using round to nearest is bounded by the *unit roundoff error u.*
- In single precision $u = 2^{-24}$.

Key Concepts

- Not all reals can be exactly represented as a machine floating point number. Then what?
- IEEE options:
- Round to next nearest FP (preferred), Round to 0, Round up, and Round down
- Let x+ and x- be the two floating point machine numbers closest to x
- round to nearest: round(x) = x- or x+, whichever is closest
- round toward 0: round(x) = x- or x+, whichever is between 0 and x
- round toward $-\infty$ (down): round(x) = x-
- round toward $+\infty$ (up): round(x) = x+

Errors in Representation

- 32 bit word example (single precision)
- $x = 2^{54897} \rightarrow$ overflow
  - Exponent is beyond the 8 bit range.
- $x = 2^{-45962} \rightarrow$ underflow
- Numbers that overflow or underflow have large relative errors when replaced by nearest machine numbers. They are said to be *out of range.*

Errors in Representation

- $x = q \times 2^m \;\left(\frac{1}{2} \le q < 1,\;\; -126 \le m \le 127\right)$
- Replace x with nearest machine number
  - Correct rounding
  - Roundoff error
- How large is the error in representing:

$$x = (0.1b_1b_2 \ldots b_{24}b_{25} \ldots)_2 \times 2^m$$

by the nearest machine number.

Errors in Representation

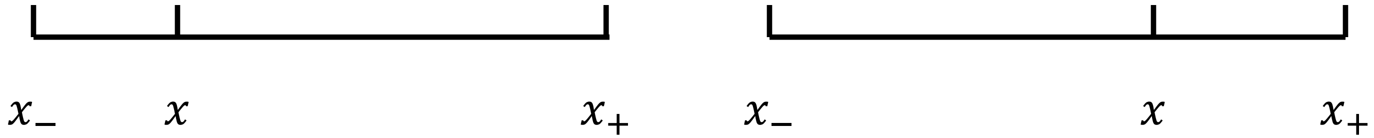
- 2 options
  - Round down (drop excess bits in mantissa)

    $$x_- = (0.1b_2b_3 \dots b_{24})_2 \times 2^m$$

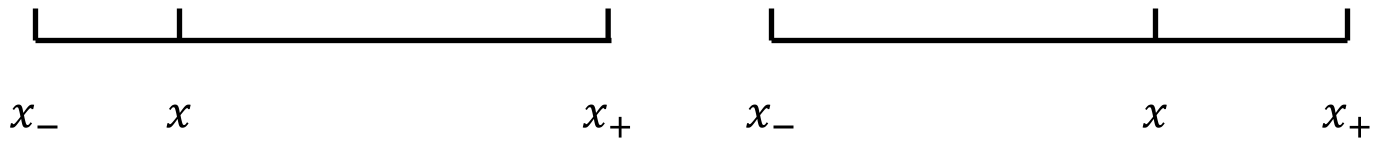
  - Round up (add 1 unit to b_{24} in x_-)

    $$x_+ = [(0.1b_2b_3 \dots b_{24})_2 + 2^{-24}] \times 2^m$$

- Closer number chosen to represent x.



Rounding Error



- Rounding down

$$|x - x_-| \leq \frac{1}{2}|x_+ - x_-|$$

- And

$$|x_+ - x_-| = [(0.1b_2b_3 \ldots b_{24})_2 + 2^{-24}] \times 2^m - (0.1b_2b_3 \ldots b_{24})_2 \times 2^m = 2^{-24+m}$$

- $|x - x_-| \leq 2^{-25+m}$

Unit roundoff error

- The relative error is

$$\left|\frac{x - x_-}{x}\right| \leq \frac{2^{-25+m}}{(0.1b_1b_2 \ldots b_{24}b_{25} \ldots)_2 \times 2^m} \leq \frac{2^{-25}}{\frac{1}{2}}$$

$$\frac{2^{-25}}{\frac{1}{2}} = 2^{-24} = u$$

$$\epsilon = 2^{-23} \rightarrow \epsilon = 2u$$

Unit roundoff

- $u = 2^{-k}$ where k is the number of binary digits in the mantissa including the hidden bit.
- $k = 24$ for single precision $k = 53$ in double precision.
- The same analysis holds for x closer to x_+. The relative error is still bounded by u.

Key concepts

- The set of representable machine numbers is finite.
- So not all math operations are well defined.
- Basic algebra breaks down in floating point arithmetic.

$$(a + b) + c \neq a + (b + c)$$

- How does roundoff impact computation errors ?

Error Analysis

- $fl(x)$ is the floating point machine number closest to x.
- We have shown:

$$\frac{|x - fl(x)|}{|x|} < u$$

- For 32 bit word length $u = 2^{-24}$
- Assuming correct rounding is used (as in previous example)

Error Analysis

- Or written another way
$$fl(x) = x(1 + \delta),$$
- $|\delta| \leq 2^{-24}$ for single precision
- Consider some operation $* \in (\times, \div, +, -)$
- For machine numbers x, y combined arithmetically we get $fl(x * y)$ instead of $(x * y)$

Error Analysis

- Assume the operation is correctly formed, normalized, and rounded to form a machine number. Then,
- $fl(x * y) = (x * y)(1 + \delta)$
- $-2^{-24} \leq \delta \leq 2^{-24}$

Loss of Significance

- Subtraction can cause loss of significant digits when the two numbers are nearly equal.
- This error can be reduced by various techniques
  - Taylor series
  - Trigonometric identities
  - Logarithmic properties
  - Double precision
  - Range reduction

Loss of Significance

- Revisit significant digits.

$$x = 0.5823962 \times 10^5$$

- x has 7 significant digits
- 5 is the most significant
- 2 is the least significant

Example from the text

- Consider $y \leftarrow x - \sin x$
- Calculate for small x on 10 decimal digit computer.
- Use $x = {}^{1}/_{15}$
- Find machine number closest to x

$$x \leftarrow 0.6666666667 \times 10^{-1}$$

- Calculate $\sin x$

$$\sin(x) \leftarrow 0.6661729492 \times 10^{-1}$$

Example from Text

- Calculate $x - \sin(x)$

$$
\begin{array}{r l}
 & 0.6666666667 \times 10^{-1} \\
- & 0.6661729492 \times 10^{-1} \\
\hline
 & 0.0004937175 \times 10^{-1} \rightarrow 0.4937175000 \times 10^{-4}
\end{array}
$$

Normalized: 0.4937175 — Spurious digits: 000

Correct to 10 decimals $\approx 0.4937174327 \times 10^{-4}$

Loss of Precision Theorem

Let x and y be (normalized) floating point machine numbers with x > y > 0.

If $2^{-p} \leq 1 - \frac{y}{x} \leq 2^{-q}$ for positive integers p and q, the significant binary digits lost in calculating x - y is between q and p.

example

- Consider x = 37.593621 and y = 37.584216

$$0.000244 = 2^{-12} \leq 1 - \frac{y}{x} = 0.0002501754 \leq 2^{-11} = 0.000488$$

- <u>11 to 12 bits lost</u> in computing $x - y$

- What can we do to reduce loss of accuracy in subtraction?

Example from previous lecture

- Evaluate $y = \sqrt{x+\delta} - \sqrt{x}$
  - $x = 100$ and $\delta = 0.1$
  - using 2 decimals
- Solution

$$\sqrt{x+\delta} = \sqrt{100.1} = 10.0049987\ldots$$

$$\tilde{y} = 10.00 \; - \; \sqrt{100} = 0.00^{*}$$

$$\left|\frac{\tilde{y}-y}{y}\right| = 1 \;\; \text{(catastrophic cancellation)}$$

*The subtraction is carried out exactly.

Example from previous lecture

- Rewrite the formula

$$y = \left(\sqrt{x+\delta} - \sqrt{x}\right)\left(\frac{\sqrt{x+\delta}+\sqrt{x}}{\sqrt{x+\delta}+\sqrt{x}}\right)$$

$$= \frac{\delta}{\sqrt{x+\delta}+\sqrt{x}}$$

$$\tilde{y} = \frac{0.1}{10.0 + 10.0} = \frac{0.1}{20.0} = 0.005$$

$$\left|\frac{\tilde{y} - y}{y}\right| = 2.6 \times 10^{-4}$$

Taylor Series to the rescue.

- Revisit $f(x) = x - \sin(x)$, $x \to 0$
- Use Taylor series to approximate $\sin(x)$.

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$f(x) = x - (x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots)$$

Taylor series to the rescue

$$f(x) = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots$$

- For small x: $x \gg \frac{x^3}{3!}$ and so near zero cancelation occurs.
- By eliminating the large terms from $f(x)$ we eliminate the problem.

Taylor series to the rescue

How do we know for what values of x to use the expansion form over the original expression?

- From the loss of precision theorem, choosing x such that:

$$2^{-1} \leq 1 - \frac{\sin(x)}{x}$$

Ensures at most 1 lost bit of accuracy is lost in calculating $x - \sin(x)$.

Taylor series to the rescue

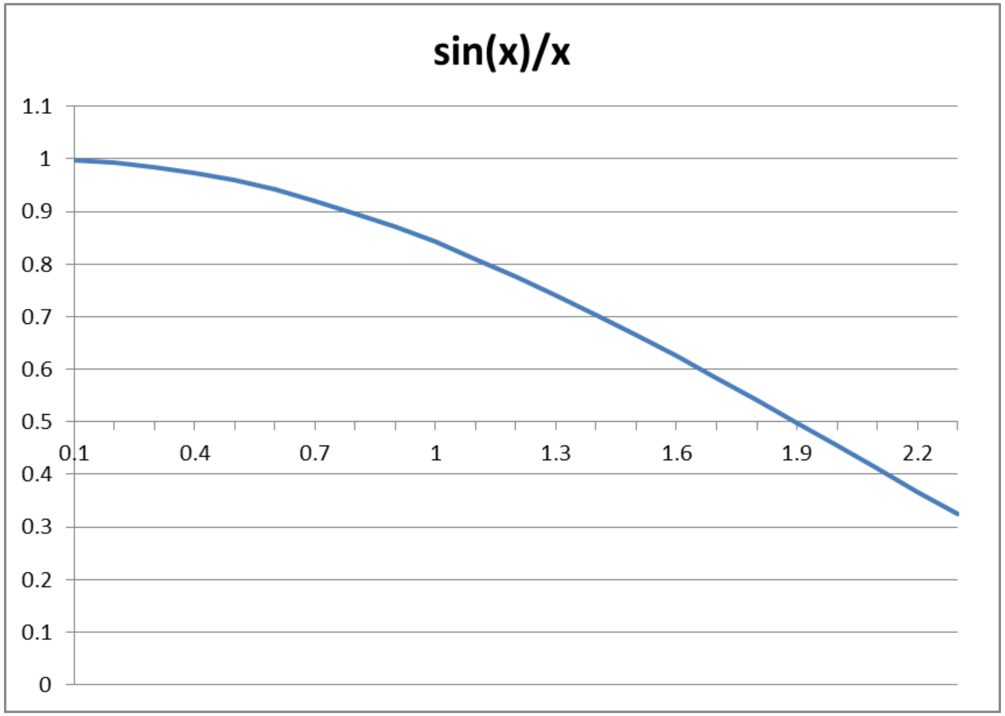
$$2^{-1} + \frac{\sin(x)}{x} \leq 1 - \frac{\sin(x)}{x} + \frac{\sin(x)}{x}$$

$$2^{-1} + \frac{\sin(x)}{x} \leq 1$$

$$2^{-1} - 2^{-1} + \frac{\sin(x)}{x} \leq 1 - 2^{-1}$$

$$\frac{\sin(x)}{x} \leq \frac{1}{2}$$

Taylor series to the rescue



- For $|x| \geq 1.9$ use $x - \sin(x)$
- For $|x| < 1.9$ use 10 term Taylor form.

Taylor series to the rescue

- So for $|x| \geq 1.9$ we ensure less than 1 bit of lost accuracy in the calculation of $f(x)$ by subtraction.
- What about the accuracy in the region where we use the Taylor series?
- The 11th term is: $\frac{x^{23}}{23!}$ which for $x = 1.9$ (the largest value of x in the interval for which we use the series) is $\approx 10^{-16}$

Taylor series to the rescue

$$f(x) = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots$$

- This is an alternating series.
- For a 10 term approximation, by the alternating series theorem, the error does not exceed the 11th term.
- The 11th term is: $\frac{x^{23}}{23!}$ which for $x = 1.9$ (the largest value of x in the interval for which we use the series) is $\approx 10^{-16}$

Using Trigonometry.

$$y \leftarrow cos^2(x) - sin^2(x)$$

- This subtraction loses significant digits when $x \rightarrow {}^{\pi}/_{4}$ because $cos^2({}^{\pi}/_{4}) = sin^2({}^{\pi}/_{4})$.
- Avoid the cancelation by using the identity:

$$\cos(2x) = cos^2(x) - sin^2(x)$$

$$y \leftarrow \cos(2x)$$

Logarithmic Properties

$$y \leftarrow \ln(x) - 1$$

- Cancelation occurs as $x \rightarrow e$

$$y = \ln(x) - 1 = \ln(x) - \ln(e)$$

- Eliminate the subtraction with:

$$\ln(x) - \ln(e) = \ln\left(\frac{x}{e}\right)$$

$$y \leftarrow \ln\left(\frac{x}{e}\right)$$

Range Reduction

$$\sin(x) = \sin(x + 2n\pi)$$

- Only require values for $0 < x \leq 2\pi$.
- Evaluation of $\sin(12532.14)$ is equivalent to evaluation of $\sin(3.47)$.

$$\frac{12532.14}{2\pi} \approx 1994.55$$

$$12532.14 - (2\pi \times 1994) \approx 3.47$$

Retaining 2 decimal digits of accuracy.

Range Reduction

- The computer uses this range reduction to evaluate trigonometric functions.
- The subtraction has reduced the number of significant digits in the argument from seven to three.
- The computed value of $\sin(12532.14)$ will have no more than 3 significant figures.

Summary

- Loss of significance may be avoided by reformulating the expression or other techniques such as series expansion.
- If x and y are positive normalized floating point machine numbers and

$$2^{-p} \leq 1 - \frac{y}{x} \leq 2^{-q}$$

Then at most p and at least q significant binary bits are lost in computing $x - y$.