

Errors in Representing Numbers

non-machine numbers are represented by a nearest machine number in computer

correctly rounding and **roundoff error**

for a 32-bit single-precision machine with 23 bits for mantissa, the relative error in correct rounding is $\frac{1}{2} \times 2^{-23}$

the unit round error is

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

Notation $\mathbf{fl}(x)$

most computers use double-length arithmetic operations. Numbers are extended to double length, arithmetic operations are performed, and the result is rounded to a single length number

use $\mathbf{fl}(x)$ to denote the floating-point machine number corresponding to a real number x

for the 32-bit machine, we have

$$\mathbf{fl}(x) = x(1 + \delta), \quad \left(|\delta| \leq \frac{1}{2}\epsilon \right)$$

it is easy to see that if $\epsilon < 2^{-23}$, then

$$\mathbf{fl}(1 + \epsilon) = 1$$

Inverse-Error Analysis

more generally known as **backward error analysis**

denote \odot as one of the basic arithmetic operations, then

$$\mathbf{fl}(x \odot y) = (x \odot y)(1 + \delta), \quad (|\delta| \leq 2^{-24})$$

two interpretations

$$\mathbf{fl}(x + y) = (x + y)(1 + \delta)$$

$$\mathbf{fl}(x + y) = x(1 + \delta) + y(1 + \delta)$$

direct-error analysis and reverse-error analysis

forward-error analysis and backward-error analysis

Loss of Significance

subtraction of two nearly equal numbers may result in loss of significant digits on a finite precision machine

cure: reprogram or use higher precision arithmetic

avoid using unnecessary higher precision arithmetic

Loss of Precision Theorem

Let x and y be normalized floating-point machine numbers with $x > y > 0$. If $2^{-p} \leq 1 - y/x \leq 2^{-q}$ for some positive integers p and q , then at most p and at least q significant digits are lost in the subtraction $x - y$

An Example from Book

$$x = 37.593621 \text{ and } y = 37.584216$$

$$1 - \frac{y}{x} = 0.0002501754,$$

which is between $2^{-12} = 0.000244$ and $2^{-11} = 0.000488$. At least 11 and at most 12 binary digits are lost when computing $x - y$

Exactly how many digits lost depends on a computer

$$x = 0.37593612 \times 10^2 \text{ and } y = 0.37584216 \times 10^2$$

suppose a machine has 5 decimal digits of accuracy

we have $\tilde{x} = 0.37594 \times 10^2$ and $\tilde{y} = 0.37584 \times 10^2$, a machine computes

$$\tilde{x} - \tilde{y} = 0.00010 = 0.10000 \times 10^{-3}$$

An Example Cont.

exact computation

$$x - y = 0.00009396 = 0.9396 \times 10^{-4}$$

the relative error is

$$\frac{|(x - y) - (\tilde{x} - \tilde{y})|}{|x - y|} = \frac{0.604 \times 10^{-5}}{0.9396 \times 10^{-4}} = 0.06428$$

this relative error is considered large since the machine has 5 decimal digits of accuracy

if we use a machine with at least 8 decimal digits of accuracy, we can have the exact value as $0.93960000 \times 10^{-4}$

Avoiding Loss of Significance

analyze possible loss of significance, reschedule computation to avoid subtraction of two nearly equal numbers, modify algorithm

example. Evaluating

$$f(x) = \sqrt{x^2 + 1} - 1 \quad \text{at} \quad x \approx 0$$

using 5-decimal-digit arithmetic for $x = 10^{-3}$, we have $f(x) = 0$

rationalizing the numerator, we have

$$f(x) = \frac{x^2}{\sqrt{x^2 + 1} + 1}$$

computing $\sqrt{(10^{-3})^2 + 1} + 1 = 0.2 \times 10^1$, and $f(x) = 0.5 \times 10^{-6}$.

More Examples

evaluating $f(x) = x - \sin x$ at $x \approx 0$

using Taylor series for $\sin x$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

then

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

compute $x = 0.1$ with four-decimal-digit arithmetic, $\sin(0.1) = 0.9983 \times 10^{-1}$. So $x - \sin x = 0.17 \times 10^{-3}$. But $x^3/3! = 0.1667 \times 10^{-3}$. This strategy is not good for large x . For $x = \pi$, $x - \sin \pi = \pi$, but $x^3/3! = 5.1677$.

Range reduction for periodic functions

$$\sin(12532.14) \approx \sin(3.47 + 1994 \times 2\pi) = \sin(3.47)$$