Numerical Analysis

Mathematics of Scientific Computing

主讲人　邱欣欣
幻灯片制作　邱欣欣

中国海洋大学　信息科学与工程学院

2013 年 9 月 13 日

Computer Arithmetic

1. Absolute and Relative Errors : Loss of Significance
    - Loss of Significance
    - Substraction of Nearly Equal Quantities
    - Loss of Precision
    - Evaluation of Functions
    - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
    - Numerical Instability
    - Conditioning

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Absolute and Relative Errors : Loss of Significance

- When a real number x is approximated by another number x^*, the error is $x – x^*$.
  - The absolute error is $|x - x^*|$
  - The relative error is $\left|\frac{x - x^*}{x}\right|$

In scientific measurements, it is almost always the relative error that is more significant.

Absolute and Relative Errors : Loss of Significance

- We have already considered relative error in our investigation of roundoff errors. The inequality

$$\left|\frac{x - fl(x)}{x}\right| \leqslant \varepsilon$$

is a statement about the relative error involved in representing a real number x by a nearby floating-pointing machine number.

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Loss of Significance

- Examples for large relative error can occur, subtract the two numbers

$$
\begin{aligned}
x &= 0.37214\,78693\\
y &= 0.37202\,30572\\
x{-}y &= 0.00012\,48121
\end{aligned}
$$

If this calculation were to be performed in a decimal computer having a five-digit mantissa, we would see

$$
\begin{aligned}
fl(x) &= 0.37215\\
fl(y) &= 0.37202\\
fl(x){-}fl(y) &= 0.00013
\end{aligned}
$$

Loss of Significance

- The relative error is then very large

$$\left|\frac{x-y-[fl(x)-fl(y)]}{x-y}\right| = \left|\frac{0.00012\,48121-0.00013}{0.00012\,48121}\right| \approx 4\%$$

- Whenever the computer must shift the digits in the mantissa to achieve a normalized floating-point number, additional 0's are supplied on the right. These 0's are spurious and do not represent additional accuracy. Thus, $fl(x) - fl(y)$ is represented in the computer as 0.13000×10^{-3}, but the 0's in the mantissa serve only as placeholders.

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Substraction of Nearly Equal Quantities

EXAMPLE The assignment statement

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

involves subtractive cancellation and loss of significance for small values of x. How can we avoid this trouble?

Solution Rewrite the function in this way

$$y = \left(\sqrt{x^2+1}-1\right)\left(\frac{\sqrt{x^2+1}+1}{\sqrt{x^2+1}+1}\right) = \frac{x^2}{\sqrt{x^2+1}+1}$$

Thus, the difficulty is avoided by reprogramming with a different assignment statement

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

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Loss of Precision

THEOREM 1 Theorem on Loss of Precision

If x and y are positive normalized floating-point binary machine numbers such that $x > y$ and

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

then at most q and at least p significant binary bits are lost in the subtraction $x - y$.

Loss of Precision

Proof The normalized binary floating-point forms for x and y are

$$x = r \times 2^n \qquad (\frac{1}{2} \leqslant r \leqslant 1)$$

$$y = s \times 2^m \qquad (\frac{1}{2} \leqslant s \leqslant 1)$$

We must write y as

$$y = (s \times 2^{m-n}) \times 2^n$$

and then we have

$$x - y = (r - s \times 2^{m-n}) \times 2^n$$

Then mantissa of this number satisfies

$$r - s \times 2^{m-n} = r(1 - \frac{s \times 2^m}{r \times 2^n}) = r(1 - \frac{y}{x}) < 2^{-p}$$

Loss of Precision

EXAMPLE Consider the assignment statement

$$y \leftarrow x - \sin x$$

Since $\sin x \approx x$ for small values of x, this calculation involves a loss of significance. How can this be avoided?

Solution Let us find an alternative form for the function $y = x - \sin x$. The Taylor series for $\sin x$ is helpful here. Thus, we have

$$\begin{aligned} y &= x - \sin x \\ &= x - (x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots) \\ &= \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \dots \end{aligned}$$

Loss of Precision

Solution If x is near 0, then a truncated series can be used, as in this assignment statement

$$y \leftarrow (x^3/6)(1 - (x^2/20)(1 - (x^2/42)(1 - x^2/72)))$$

If values of y are needed for a wide range of x-values in this function, it would be best to use both assignment statement, each in its proper range.

Loss of Precision

Solution We see that the loss of bits in the subtraction of the first assignment statement can be limited to at most one bit by restricting x so that

$$1 - \frac{\sin x}{x} \geqslant \frac{1}{2}$$

It is easy to determine that x must be at least 1.9. Thus for $|x| \geqslant 1.9$, we should use the first assignment statement involving $x - \sin x$, and for $|x| < 1.9$ we should use a truncated series. We can verify that for the worst case $(x = 1.9)$, seven terms in the series give y with an error at most 10^{-9}.

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Evaluation of Functions

- There is another situation in which a drastic loss of significant digits occurs. This is in the evaluation of certain functions for very large arguments. Let us illustrate with the cosine function

  $$\cos(x + 2n\pi) = \cos x \qquad (n \textit{ is an integer})$$

  By the use of this property, the evaluation of $\cos x$ for any argument can be effected by evaluating at a reduced argument in the interval $[0,2\pi]$.
- For example, the evaluation of $\cos x$ at $x = 33278.21$ proceeds by finding the reduced argument

  $$y = 33278.21 - 5296 \times 2\pi = 2.46$$

Contents

1 Absolute and Relative Errors : Loss of Significance
- Loss of Significance
- Substraction of Nearly Equal Quantities
- Loss of Precision
- Evaluation of Functions
- Interval Arithmetic

2 Stable and Unstable Computations : Conditioning
- Numerical Instability
- Conditioning

Interval Arithmetic

- A method of controlling computations to know the extent of roundoff error is interval arithmetic. In this manner of computing, each calculated number is accompanied by an interval that is guaranteed to contain the correct value. Ideally, of course, these intervals are very small, and final answers can be given with only small uncertainties. However, the cost of carrying intervals(instead of simple machine numbers) throughout a lenthy computation may make the procedure cumbersome. Consequently, it is used only when great reliance must be placed on the computations. Also, it may be difficult to keep the intervals from growing much larger than is realistic.

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Contents

1 Absolute and Relative Errors : Loss of Significance
- Loss of Significance
- Substraction of Nearly Equal Quantities
- Loss of Precision
- Evaluation of Functions
- Interval Arithmetic

2 Stable and Unstable Computations : Conditioning
- Numerical Instability
- Conditioning

Numerical Instability

- We say that a numerical process is unstable if small errors made at one stage of the process are magnified in subsequent stages and seriously degrade the accuracy of the overall calculation.

  An example:

$$\begin{cases} x_0 = 1 \qquad x_1 = \frac{1}{3} \\ x_{n+1} = \frac{13}{3}x_n - \frac{4}{3}x_{n-1} \quad (\text{n}\leqslant 1) \end{cases}$$

  It is easily seen that this recurrence relation generates the sequence

$$x_n = (\frac{1}{3})^n$$

  The equation is true for $n = 0$ and $n = 1$. If its validity is granted for $n \leqslant m$, then its validity for $n = m + 1$ follows from

$$x_{m+1} = \frac{13}{3}x_m - \frac{4}{3}x_{m-1} = \frac{13}{3}(\frac{1}{3})^m - \frac{4}{3}(\frac{1}{3})^{m-1} = (\frac{1}{3})^{m-1}[\frac{13}{9} - \frac{4}{3}] = (\frac{1}{3})^{m+1}$$

Numerical Instability

- Here are some of the computed terms, calculated on a 32-bit computer similar to the Marc-32:

$x0 = 1.00000\,00$

$x1 = 0.33333\,33$ (7 correctly rounded significant digits)

$x2 = 0.11111\,12$ (6 correctly rounded significant digits)

$x3 = 0.03703\,73$ (5 correctly rounded significant digits)

$x4 = 0.01234\,66$ (4 correctly rounded significant digits)

$x5 = 0.00411\,87$ (3 correctly rounded significant digits)

$x6 = 0.00138\,57$ (2 correctly rounded significant digits)

$x7 = 0.00051\,31$ (1 correctly rounded significant digits)

$x8 = 0.00037\,57$ (0 correctly rounded significant digits)

$x9 = 0.00094\,37$

$\ldots$

$x14 = 0.91437\,35$

$x15 = 3.65749\,3$ (incorrect with relative error of 10^8)

Numerical Instability

- Whether a process is numerically stable or unstable should be decided on the basis of relative errors. Thus, if there are large errors in a computation, that situation may be quite acceptable if the answers are large.
- Let us start with initial values $x_0 = 1$ and $x_1 = 4$. Now the correct solution is now $x_n = 4^n$, and the results of computation are correct to seven significant figures. Here are three of them:

$$\begin{aligned} x_1 &= 4.000006 \\ x_{10} &= 1.0485776 \times 10^6 \\ x_{20} &= 1.099512 \times 10^{12} \end{aligned}$$

In this case, the correct values are large enough to overwhelm the errors.

Contents

1. Absolute and Relative Errors : Loss of Significance
   - Loss of Significance
   - Substraction of Nearly Equal Quantities
   - Loss of Precision
   - Evaluation of Functions
   - Interval Arithmetic
2. Stable and Unstable Computations : Conditioning
   - Numerical Instability
   - Conditioning

Conditioning

- Condition or conditioning:
    - Used to indicated how sensitive the solution of a problem to the small relative changes of the input data.
    - Small changes of the input can produce large changes of the output: ill conditioned.
    - A conditon number can be defined. If the number is large, it indicates an ill-conditioned problem.

Conditioning

- Suppose our problem is simply to evaluate a function f at a point x. If x is perturbed slightly, what is the effect on $f(x)$? If this question refers to absolute errors, we can invork the Mean-Value Theorem and write

$$f(x+h) - f(x) = f'(\xi)h \approx hf'(x)$$

Thus, if $f'(x)$ is not too large, the effect of the perturbation on $f(x)$ is small.

Conditioning

- In perturbing x by the amount h, we have h/x as the relative size of the perturbation. Likewise, when $f(x)$ is perturbed to $f(x+h)$, the relative size of that perturbation is

$$\frac{f(x+h)-f(x)}{f(x)} \approx \frac{hf'(x)}{f(x)} = \left[\frac{xf'(x)}{f(x)}\right]\left(\frac{h}{x}\right)$$

Thus, the factor $xf'(x)/f(x)$ serves as a condition number for this problem.

Conditioning

EXAMPLE 1 What is the condition number for the evaluation of the inverse sine function?

Solution Let $f(x) = \arcsin x$. Then

$$\frac{xf'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\arcsin x}$$

Conditioning

- Let f and g be two functions that belong to class C^2 in a neighborhood of r, where r is a root of f. We assume that r is a simple root, so that $f'(r) \neq 0$. If we perturb the function f to $F \equiv f + \varepsilon g$, where is the new root? Suppose the new root is $r + h$, The perturbuation h satisfies the equation
$$f(r+h) + \varepsilon g(r+h) = 0$$
Since f and g belong to C^2, we can use Taylor's Theorem to express $F(r+h)$:
$$\left[f(r) + hf'(r) + \frac{1}{2}h^2 f''(\xi)\right] + \varepsilon\left[g(r) + hg'(r) + \frac{1}{2}h^2 g''(\eta)\right] = 0$$
Discarding terms in h^2 and using the fact that $f(r) = 0$, we obtain
$$h \approx -\varepsilon \frac{g(r)}{f'(r) + \varepsilon g'(r)} \approx -\varepsilon \frac{g(r)}{f'(r)}$$

Conditioning

EXAMPLE 2 We consider a classic example given by Wilkinson. Let

$$f(x) = \prod_{k=1}^{20}(x-k) = (x-1)(x-2)\dots(x-20)$$
$$g(x) = x^{20}$$

The root of f are obviously the integers 1, 2, ..., 20. How is the root $r = 20$ affected by perturbing f to $f + \varepsilon g$?

Solution The answer is

$$h \approx -\varepsilon \frac{g(20)}{f'(20)} = -\varepsilon \frac{20^{20}}{19!} \approx -\varepsilon 10^9$$