2. Numerical Analysis Errors. Function Evaluation

NUMERICAL ANALYSIS. Prof. Y. Nishidate (323-B, nisidate@u-aizu.ac.jp)
http://web-int.u-aizu.ac.jp/∼nisidate/na/

ERROR ANALYSIS

In the practice of numerical analysis it is important to be aware that computed solutions are not exact mathematical solutions. The precision of a numerical solution can be diminished in several ways.
Definition. Suppose that $\tilde{p}$ is an approximation to p.The *error* is

$$\epsilon_p = \tilde{p} - p$$

and the *relative error* is

$$R_p = (\tilde{p} - p)\,/p$$

provided that $p \neq 0$.
Definition. The number $\tilde{p}$ is said to *approximate* p to d significant digits if d is the largest positive integer for which

$$|\tilde{p} - p|/|p| < 5 \cdot 10^{-d}$$

Truncation error

Truncation error is the error introduced when a more complicated mathematical expression is replaced with a more elementary formula. This terminology originates from the technique of replacing a function with a truncated Taylor series. For example, the infinite Taylor series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + ... + \frac{x^n}{n!} + ...$$

might be replaced with the first five terms in some specific computation thus introducing truncation error.

Round-off error

Computer representation of real numbers is limited to the fixed precision of the mantissa. True values are not stored exactly by computer representation in many cases. This is called *round-off error*. The real number $(0.1)_{10} = (0.0001100110011001100110 0...)_2$ is truncated when it is stored in a computer. The actual number that is stored in the computer may undergo *chopping* or *rounding* of the last digit. Since the computer works with limited number of digits in numbers, rounding errors are introduced and propagated in computations.

Loss of significance

Consider the two floating-point numbers $p = 3.1415926536$ and $q = 3.1415957341$, which are nearly equal and both carry 11 decimal digits of precision. Suppose that their difference is formed: $p - q = -0.0000030805$. Since the first six digits of p and q are the same, their difference $p - q$ contains only five decimal digits of precision. This phenomenon is called *loss of significance or subtractive cancellation.* This reduction in the precision of the final computed answer can creep in when it is not suspected.

Loss of trailing digits

Different type of information loss may occur when their scale is quite different. Suppose that two numbers $p = 2^{33}$ and $q = 2^1$ are added where both numbers are represented as 32-bit floating point numbers. Their addition $p + q$ is $2^{33}(1 + 2^{-32})$, and thus the mantissa should be able to express 2^{-32}. However, 32-bit floating point number consists of 23-bits mantissa, the amount q is ignored. This phenomenon is called *loss of trailing digits.*

Propagation of error

Let us investigate how error propagates in successive computations. Consider two numbers with true values p and q and two numbers with approximate values $\tilde{p}$ and $\tilde{q}$.

Addition and subtraction. Denoting errors for p and q as ϵ_p and ϵ_q the sum is

$$\tilde{p} \pm \tilde{q} = (p + \epsilon_p) \pm (q + \epsilon_q) = (p \pm q) + (\epsilon_p \pm \epsilon_q)$$

For addition and subtraction the error is the sum of errors of the operands:

$$\epsilon_{p \pm q} = \epsilon_p \pm \epsilon_q$$

Remark. Relative error of the difference of two positive numbers can be very high due to loss of significance.
Example. Consider two numbers $p = 1.137$ and $q = 1.093$ with errors $\epsilon_p = 0.011$ and $\epsilon_q = -0.011$. Relative error of their difference is

$$R_{p-q} = (\epsilon_p - \epsilon_q)/(p - q) = 0.5 = 50\%$$

As a result, we have no valid significant digits despite $R_p, R_q \approx 0.01 = 1\%$.

Multiplication. The propagation of error in multiplication is more complicated. The product of two numbers can be presented as:

$$\tilde{p}\tilde{q} = (p + \epsilon_p)(q + \epsilon_q) = pq + p\epsilon_q + q\epsilon_p + \epsilon_p\epsilon_q$$

Suppose that $p \neq 0$ and $q \neq 0$. Then we can obtain:

$$\frac{\tilde{p}\tilde{q} - pq}{pq} = \frac{\epsilon_q}{q} + \frac{\epsilon_p}{p} + \frac{\epsilon_p\epsilon_q}{pq}$$

Thus, we arrive at:

$$R_{pq} = R_p + R_q + R_pR_q$$

Taking into account that the third term is small in comparison to the first two terms, the relative error of the product is approximately the sum of the relative errors of the operands.

FUNCTION EVALUATION

Let us examine how numbers are combined when we evaluate functions. Mathematically equivalent formulas can produce very different roundoff errors during evaluation. If the way we evaluate a function gives a great deal of roundoff, then we cannot use it for practical purpose. We need to learn how to evaluate functions using computer and how to avoid roundoff in function computing.

Example: Quadratic equation The formula for finding zeros of the quadratic equation

$$ax^2 + bx + c = 0$$

is given in textbooks as

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

To show possible difficulties with evaluating the zeros, consider the particular case

$$a = 1, \quad b = -8000, \quad c = 1: \qquad x^2 - 8000x + 1 = 0$$

After evaluation of the above formula for zeros with single precision we have:

$$x_1 = 8000, \quad x_2 = 2.44 \cdot 10^{-4}$$

The result is the cancellation of one root. The value x_2 is unsatisfactory (it should be equal to $1.25 \cdot 10^{-4}$). How can we avoid this cancellation? The simple way to compute x_2 in the case $|4ac| << b^2$ is to omit the term ax^2 and to make the following estimation:

$$x_2 = -c/b = 0.000125$$

Rearrangement of Formulas

Typical methods of function expression rearrangements are shown below on several examples. Similar approaches are used in the calculus course to rearrange the expressions that arose in the delta process of formally taking the derivative of a function.

Example. Evaluate the function $f(x) = \sqrt{x+1} - \sqrt{x}$ for large x. A simple approach is to rationalize the numerator by

$$f(x) = \left(\sqrt{x+1} - \sqrt{x}\right)\left(\frac{\sqrt{x+1}+\sqrt{x}}{\sqrt{x+1}+\sqrt{x}}\right) = \frac{1}{\sqrt{x+1}+\sqrt{x}}$$

After rearrangement we have no cancellation.

Example. Evaluate $f(x) = \sin(x+\varepsilon) - \sin x$ for small ε. It is possible to use the trigonometric identity

$$\sin a - \sin b = 2\cos\frac{a+b}{2}\sin\frac{a-b}{2}$$

Then we have

$$f(x) = 2\cos\left(x + \frac{\varepsilon}{2}\right)\sin\frac{\varepsilon}{2}$$

which is a suitable form for the computation.

Series Expansions

In some cases rearrangements for functions cannot be found to remove cancellation, and some other approach must be used. One effective approach is the expansion of the functions into a Taylor series around some suitable point.

If function $f(x)$ is analytic about $x = a$, then in the neighborhood of $x = a$ it can be represented by the Taylor series, which is a power series given by

$$f(x) = f(a) + (x-a)\frac{f'(a)}{1!} + (x-a)^2\frac{f''(a)}{2!} + \\ ... + (x-a)^n\frac{f^{(n)}(a)}{n!} + O((x-a)^{n+1})$$

where $O(h^{n+1})$ represents the error caused by truncating the terms of order (n+1) and higher.

For example, let $f(x) = \sin x$. Then $f'(x) = \cos x$, $f''(x) = -\sin x$, $f'''(x) = -\cos x$ *etc.* The numerical values $f(0) = 0$, $f'(0) = 1$, $f''(0) = 0$, $f'''(0) = -1$ ... must be substituted in the formula for the Taylor expansion. After that we obtain the following polynomial of degree $n = 9$ which approximates $\sin x$ about $x = 0$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}$$

Example. Compute $\frac{\sin x}{x}$ for small x. If we represent $\sin x$ as a series

$$\sin x = x - \frac{x^3}{6} + ...$$

then

$$\frac{\sin x}{x} = 1 - \frac{x^2}{6} + \frac{x^4}{120} - ...$$

which is accurate and simple expression for our purpose.

Example. Evaluate the function $f(x) = e^x - 1$ for small x. Expanding e^x about $x = 0$ with Taylor series, we get

$$e^x - 1 = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + ... - 1 = x\left(1 + \frac{x}{2} + \frac{x^2}{6} + ...\right)$$

This expression can be computed without difficulty.

Evaluation of Polynomials

Polynomials play a central role in computing. They occur naturally in various computational algorithms. Polynomials are also widely used in approximation and interpolation.

The polynomial

$$P(x) = a_0 + a_1x + a_2x^2 + ... + a_{n-1}x^{n-1} + a_nx^n$$

can be written in the "chain" (nested) form

$$P(x) = a_0 + x(a_1 + x(a_2 + ... + x(a_{n-1} + xa_n)...)$$

This expression allows to evaluate the polynomial value using n additions and n multiplications. An algorithm of the polynomial evaluation by nested rule can be expressed by the following pseudocode

```
$f = a_n$
for $i = n - 1$ to $0$
    $f = f * x + a_i$
end for
```

Use of Approximations

In some cases functions are defined in such a way that it is not easy to evaluate them according to the definition. In particular this is true for functions defined as integrals with infinite or semi-infinite intervals. If good approximation of such function is known then it may be more efficient to use this approximation than to evaluate the function value by numerical integration.

Example: Approximation of the error function. The error function is the integral of the normal distribution. The error function is used in statistics to evaluate the probability of finding a measurement lower than a given value when the measurements are distributed according to a normal distribution with a mean of zero and a standard deviation of 1. The error function is expressed by the following integral:

$$erf(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt$$

The result of the error function lies between 0 and 1.

One could perform numerical integration but good approximations exist. The following formula is given by Abramovitz and Stegan:

$$erf(x) = 1 - \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\sum_{i=1}^{5} a_i r^i(x) \quad \text{for} \quad x \geq 0$$

where

$$\begin{aligned} &r(x) = 1/(1 + 0.2316419x) \\ &a_1 = 0.31938153 \\ &a_2 = -0.356563782 \\ &a_3 = 1.781477937 \\ &a_4 = -1.821255978 \\ &a_5 = 1.330274429 \end{aligned}$$

The error of this approximation is better that $7.5 \cdot 10^{-8}$ for positive x. To compute the value for negative x, it is possible to use the fact that $erf(x) = 1 - erf(-x)$.