1
Error Analysis

A real number x can either have one digit , finite number of digits or can have infinitely many digits. But a digital calculating device can hold only a finite number of digits and therefore, after a finite number of digits (depending on the capacity of the calculating device) the rest should be discarded in some sense. In this way, the representation of the real number x on a computing device is only approximate. Although the omitted part of x is very small in its value, this approximation can lead to considerably large error in the numerical computation.

Definition (Floating-point From): Let x be a non- zero real number. An n-digit floating- point number in base β has the form $f(x) = (-1)^s \times (.d_1 d_2 ... d_n)_\beta \times \beta^e$ where $(.d_1 d_2 ... d_n)_\beta = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + ... + \frac{d_n}{\beta^n}$

is a β-fraction called the mantissa or singificand, $s = 1$ or 0 is called the sign and e is an integer called the exponent. The number β is also called the radix and the point preceding d_1 is called the radix point.

Remark: (i) When β =2, the floating-point representation is called binary floating- point representation and when $\beta = 10$, it is called the decimal floating-point representation.

(ii) Note that there are only finite number of digits in the floating- point representation where as a real number can have infinite sequence of digits for instance 1/3 =0.33333...... Therefore, the floating-point representation is only an approximatation to a real number.

(iii) Afloating-point number is said to be normalized if either $d_1 \neq 0$ or $d_1 = d_2 = ... = d_n = 0$.

Ex. The following are examples of real number in the decimal floating point representing

I. The real number $x = 6.238$ can be represented as $6.238 = (-1)^0 \times 0.6238 \times 10^1$, in which case, we have $s = 0$, $\beta = 10$ $e= 1$, $d_1 = 2$, $d_3 = 3$ and $d_4 = 8$. Note that this representation is the normalized floating-point representation.

II. The real number $x = -0.0014$ can be represented in the decimal float-point representation as $-0.0014 = (-1)^1 \times 0.0014 \times 10^1$, which is not in the normalized from. The normalized representation is $x = (-1)^1 \times 0.14 \times 10^{-2}$.

Definition (Overflow and Underflow): The exponent e is limited to a range $m < e < M$ .During the calculation if some computed number has an exponent $e > M$ then we say, the memory over flow or if $e < m$, we say the memory underflow.

Remark: In the case of overflow, computer will usually produce meaningless results or simply prints the symbol NaN, which means, the quantity obtained due to such a calculation is 'not a number'. The symbol ∞ is also denoted as NaN on some computers. The underflow is less serious because in this case, a computer will simply consider the number as zero.

Chopping and Rounding a Number

Any real number x can be represented exactly as $x = (-1)^s \times (\cdot d_1 d_2 ... d_{n+1} .....)_\beta \times \beta^e$, With $d_1 \neq 0$ or $d_2 = d_3 = ... = 0$, $s = 0$ or 1, and $m < e < M$ for which the floating-point form is an approximate representation. Let us denote this approximation of *x* by fl (*x*). There are two ways to produce fl(*x*) from *x* as defined below.

(i) The chopped machine approximatation of *x* is given by $fl(x) = (-1)^s \times (.d_1 d_2 ... d_n)_\beta \times \beta^e$. In this mathod we cut the number upto first *n* decimal point irrespective of the values after *n* decimal point

Ex. Let $x = 0.2346513$....Then upto three decimal places we have $fl(x) = 0.234$ we simply chapped of all the digits.

(ii) The rounded machine opproximation of *x* is given by

$$fl(x) = \begin{cases} fl(x) = (-1)^s \times (.d_1 d_2 ... d_n)_\beta \times \beta^e & , 0 \le d_{n+1} < \dfrac{\beta}{2} \\ fl(x) = (-1)^s \times (.d_1 d_2 ... (d_n + 1))_\beta \times \beta^e & , \dfrac{\beta}{2} \le d_{n+1} < \beta \end{cases}$$

in this method if last decimal place is $> \dfrac{\beta}{2}$ we add 1 in last decimal palce otherwise it remain same

Ex. Let $x = 0.2346513....$. Note that at fourth decimal palce we have digit 6, which is greater then $\dfrac{\beta}{2} = \dfrac{10}{2} = 5$, hence upto three decimal places we have $fl(x) = 0.235$

Different Type of Errors

The approximate representation of a real number obviously differs from the actual number, whose difference is called an error.

(i) The error in a computed quantity is defined as **Error = True Value –Approximate Value**

(ii) The absolute error is the absolute value of the error defined above.

(iii) The relative error is a meansure of the error in relation to the size of the true value as given by

$$\text{Relative Error} = \frac{\text{Error}}{\text{True Value}}$$

(iv) The percentage error is defined as 100 times the relative error.

(v) The tern trunction error is used to denote error, which result from approximating a smooth function by truncting its Taylor series representation to a finite number of terms.

Remark: (i) Let x_A be the approximation of the real number *x*. Then

$$E(x_A) := \text{Error}(x_A) = x - x_A$$

$$E_a(x_A) := \text{absolute Error}(x_A) = |E(x_A)|$$

$$E_r(x_A) := \text{Relative Error}(x_A) = \frac{|E(x_A)|}{x}$$

(ii) If we denote the relative error in fl (*x*) as $\in > 0$, then we have $fl(x) = (1 - \in) x$, we *x* is a real number

Ex. A second degree polynomial approximation to $f(x) = \sqrt{x+1}, x \in [0,1]$

Using the Taylor series expansion about $x = 0$ is given by $f(x) \approx 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16\left(\sqrt{1+\xi}\right)^5}$

Therefore, the truncation error is given by $x^3 / \left(16\left(\sqrt{1+\xi}\right)^5\right)$

Loss of Significant Digits

If x_A is an approximation to x, then we say that x_A approximates x to r significant β-digits if

$$|x - x_A| \le \frac{1}{2}\beta^{s-r+1}$$

With s the largest integer such that $\beta^s \le |x|$

Ex. (a) For $x = 1/3$, the approximate number $x_A = 0.333$ has three significant digits, since $|x - x_A| \approx .00033 < 0.0005 = 0.5 \times 10^{-3}$. But $10^{-1} < 0.333... = x.$ Therefore, in this case $s = -1$ and hence r = 3.

(b) For $x = 0.02138,$ the approximate number $x_A = .02144$ has the abolute error $|x - x_A| \approx .00006 < 0.0005 = 0.5 \times 10^{-3}$. But $10^{-2} < 0.02138 = x$. Therefore, in this case $s = -2$ and therefore, the number x_A has only two significant digits, but not three, with respect to x.

Remark: In a very simple way, the number of leading non-zero digits of x_A that are correct relative to the corresponding digits in the true value x is called the number of significant digits in x_A.

The role of significant digits in the numerical calculation is very important in the sense that the loss of significant digits many result in drastic amplification of the relative error.

The loss of significant digits in the process of calculation is refered to as **Loss of significance.**

Ex. Consider the function $f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$. On a six-digit decimal calculator, we have

$f(100000) = 100$ where as the true value is 158.113. This makes a drastic error in the calculation. This is the result of the loss of significant digits, which can be seen from the fact that as x increases, the terms $\sqrt{x+1}$ and $\sqrt{x}$ comes closer to each other and therefore loss of significant error in their computed value increasing.

Such loss can often be avoided by rewritting the given expression (whenever possible) in such a way that subtraction is avoided. For instance, the defination of $f(x)$ given in this example can be rewritten as

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

With this new definition, we see that on a six-dgit calculator, we have $f(100000) = 158.114000$

Propagation of Error

Let x_A and y_A denote the numbers used in the calculation, and let x_T and y_T be the corresponding true values. We will now see how error propagates with the four basic arithmeic operations.

Propagated Error in Addition and Subtraction

Let $x_T = x_A + \in$ and $y_T = y_A + \eta$ are positive numbers. The relative error $E_r(x_A \pm y_A)$ is given by

$$E_r\left(x_A \pm y_A\right) = \frac{\left(x_T + y_T\right) - \left(x_A \pm y_A\right)}{x_T \pm y_T} = \frac{\left(x_T \pm y_T\right) - \left(x_T - \in \pm\left(y_T - \eta\right)\right)}{x_T \pm y_T} = \frac{\in \pm \eta}{x_T \pm y_T}$$

This shows that relative error propagate solwly with addition, where as amplifies drastically with subtraction when $x_T \approx y_T$.

Popagated Error is Multiplication

The relative error $E_r\left(x_A \times y_A\right)$ is given by

$$E_r\left(x_A \times y_A\right) = \frac{\left(x_T \times y_T\right) - \left(x_A \times y_A\right)}{x_T \times y_T} = \frac{\left(x_T \times y^T\right) - \left(\left(x_T - \in\right) \times \left(y_T - \eta\right)\right)}{x_T \times y_T}$$

$$= \frac{\eta x_T + \in y_T - \in \eta}{x_T \times y_T} = \frac{\in}{x_T} + \frac{\eta}{y_T} - \left(\frac{\in}{x_T}\right)\left(\frac{\eta}{y_T}\right) = E_r\left(x_A\right) + E_r\left(y_A\right) - E_r\left(x_A\right)E_r\left(y_A\right)$$

This shows that relative error propagate slowly with multiplication.

Propagated Error in Division

The relative $E_r\left(x_A / y_A\right)$ is given by

$$E_r\left(x_A / y_A\right) = \frac{\left(x_T / y_T\right) - \left(x_A / y_A\right)}{x_T / y_T} = \frac{\left(x_T / y_T\right) - \left(\left(x_T - \in\right) / \left(y_T - \eta\right)\right)}{x_T / y_T}$$

$$= \frac{x_T\left(y_T - \eta\right) - y_T\left(x_T - \in\right)}{x_T\left(y_T - \eta\right)} = \frac{y_T \in - x_T \eta}{x_T\left(y_T - \eta\right)} = \frac{y_T}{y_T - \eta}\left(E_r\left(x_A\right) - E_r\left(y_A\right)\right)$$

$$= \frac{1}{1 - E_r\left(Y_A\right)}\left(E_r\left(x_A\right) - E_r\left(y_A\right)\right)$$

This show that relative error propagate slowly with division, unless $E_r\left(Y_A\right) \approx 1$. But this is very unlikely because we always expect the error to be very small, i.e. very close to zero in which case the right hand side is approximately equal to $E_r\left(x_A\right) - E_r\left(Y_A\right)$.

Propagated Error in Function Evaluation

Consider evaluating $f\left(x\right)$ at the approximate value x_A rather than at x. The consider how well does $f\left(x_A\right)$ approximate $f\left(x\right)$? Using the mean-value theorem, we get

$$f\left(x\right) - f\left(x_A\right) = f'(\xi)\left(x - x_A\right)$$

Where ξ is an unkonwn point between x and x_A.

The relative error of $f\left(x\right)$ with respect to $f\left(x_A\right)$ is given by

$$E_r\left(f\left(x\right)\right) = \frac{f'(\xi)}{f\left(x\right)}\left(x - x_A\right) = \frac{f'(\xi)}{f\left(x\right)} x E_r\left(x\right)$$