Numerical Analysis

Dhyan Laad
2024ADPS0875G

1 Introduction to Computation

1.1 Floating Point Forms

A real number may potentially have an infinite decimal expansion, but computers are limited by hardware, and as such store numbers with a terminating approximation.

Definition 1.1. Given a real number x with digits $d_1, d_2, \ldots$, the *n-digit, base β floating point form*, or *n-β floating point form* is

$$(-1)^s \times (0.d_1 d_2 \ldots d_n)_\beta \times \beta^e$$

where $s \in \{0, 1\}$ is the *sign*, e is the *exponent*, and the β-fraction

$$(0.d_1 d_2 \ldots d_n)_\beta = \frac{d_1}{\beta^1} + \frac{d_2}{\beta^2} + \cdots + \frac{d_n}{\beta^n}$$

is called the *mantissa*. In the case that $d_1 \neq 0$, the representation is called the *normalized floating point form*.

For a fixed value of β and n as defined above, the notation $\mathrm{fl}(x)$ is used to denote the n-β floating point representation of x. Furthermore, for all computing systems, there are bounds on the values that the exponent e can take. This leads to the concepts of underflow and overflow.

Definition 1.2. Let a real number x have a floating point form with exponent e. For a computing system with exponential range (m, M) where m and M are integers,

(a) if $e > M$, then the system is said to *overflow*, and the result of the computation is denoted with a signed infinity: $\pm\infty$, and

(b) if $e < m$, then the system is said to *underflow*, and the result of the computation is simply 0.

There are two ways to determine the mantissa of the floating point representation of a real number with more than n digits: chopping and rounding. The chopped mantissa of the floating point representation of $x = 0.d_1 d_2 \ldots d_n d_{n+1} \ldots$ would simply be $(0.d_1 d_2 \ldots d_n)$, while the rounded mantissa would be

$$\begin{cases} (0.d_1 d_2 \ldots d_n) & d_{n+1} \in [0, \beta/2), \\ (0.d_1 d_2 \ldots (d_n + 1)) & d_{n+1} \in [\beta/2, \beta]. \end{cases}$$

1.2 Errors

The error of a floating point representation is a quantitification of how far removed it is from its true value.

Definition 1.3. Let $x \in \mathbb{R}$. The *absolute error* of its floating point representation is

$$x - \mathrm{fl}(x).$$

Note that since $\mathrm{fl}(x) \leq x$ for all $x \in \mathbb{R}$, the absolute error is always a positive quantity. Absolute error is the simplest quantitification but not the most useful, motivating a definition for relative error.

Definition 1.4. The ratio of the absolute error to the true value of a real number x is called its *relative error*. It is customarily denoted with ε:

$$\varepsilon = \frac{x - \mathrm{fl}(x)}{x}.$$

Another quantitification of how removed an approximation is from its true value is captured in the approximation's significant figures or significant digits.

Definition 1.5. Let x be a real number and x^* be an approximation of it. Then if

$$|x - x^*| \leq \frac{1}{2}\beta^{s-r+1}$$

where s is the largest integer such that $\beta^s \leq |x|$, then x^* is said to approximate x to r *significant figures* in β.

Theorem 1.6. *Let $\mathrm{fl}(x)$ be the n-β floating point representation for $x \in \mathbb{R}$, and set*

$$\varepsilon = \frac{x - \mathrm{fl}(x)}{x}.$$

Then,

(a) $\varepsilon \leq \beta^{-n+1}$ *for chopped systems, and*

(b) $\varepsilon \leq \frac{1}{2}\beta^{-n+1}$ *for rounded systems.*

Proof. Let x be a nonzero real number represented as

$$x = m \cdot \beta^e = (-1)^s \cdot (0.d_1 d_2 \dots d_n d_{n+1} \dots)\beta^e$$

where $d_1 \neq 0$. The smallest possible magnitude for the mantissa is $0.100\dots$ (in base β). Therefore, the bounds on m are

$$\frac{1}{\beta} \leq |m| < 1.$$

Since the floating point representation only stores n digits, the last digit stored is d_n, which is in the β^{-n} position relative to the decimal point.

Now, a chopped system truncates everything after d_n, and the absolute error would be given by

$$|x - \mathrm{fl}(x)| = 0.\underbrace{00\ldots0}_{n \text{ zeros}} d_n d_{n+1} \cdots \times \beta^e < \beta^{-n} \times \beta^e = \beta^{e-n}.$$

Now consider the relative error.

$$|\varepsilon| = \left|\frac{x - \mathrm{fl}(x)}{x}\right| < \frac{\beta^{e-n}}{|m \times \beta^e|} = \frac{\beta^{-n}}{|m|}.$$

To find the upper bound, we must minimize the denominator, whose minimum value we previously determined to be $1/\beta$, which yields

$$|\varepsilon| < \frac{\beta^{-n}}{1/\beta} \Rightarrow \varepsilon \leq \beta^{-n+1}. \tag{a}$$

In a rounded system, $\mathrm{fl}(x)$ is the number with n digits closest to x. The quantity analogous to a "least count" would be β^{e-n}. When rounding, the error cannot exceed half of this value

$$|x - \mathrm{fl}(x)| \leq \frac{1}{2}\beta^{-n} \times \beta^e = \frac{1}{2}\beta^{e-n}.$$

Dividing by x yields

$$|\varepsilon| = \frac{|x - \mathrm{fl}(x)|}{|x|} \leq \frac{1}{2} \cdot \frac{\beta^{-n}}{|m|}.$$

Once more, the error is maximized at $|m| = 1/\beta$. Therefore,

$$\varepsilon \leq \frac{1}{2}\beta^{-n+1}. \tag{b}$$

□

Propogation of Errors

When performing the arithmetic operations with approximate quantities, it is important to study the errors in the sum, difference, product, and quotient. Let $x = x^* + \varepsilon$ and $y = y^* + \eta$, where x and y are true real values, and x^* and y^* are their approximations. Let r_n denote the relative error in a quantity n. Then,

$$r_{xy} = \frac{xy - x^*y^*}{xy} = \frac{xy - (x+\varepsilon)(y+\eta)}{xy} = \frac{\varepsilon}{x} + \frac{\eta}{y} + \frac{\varepsilon\eta}{xy} \approx r_x + r_y,$$

$$r_{x/y} = \frac{x/y - x^*/y^*}{x/y} = \frac{\eta}{y+\eta} - \frac{y\varepsilon}{x(y+\eta)} \approx r_y - r_x.$$