Computer Arithmetic

Jamie M. foster
http://www.math.mcmaster.ca/∼jmfoster

Floating point numbers

Any real number, *i.e.* any number in $\mathbb{R}$, is represented on a computer by a **floating point** (FP) number. A particular floating point number system, $\hat{\mathbb{N}}$, is characterised by four integers:

1. β - Base or radix
2. p - Precision
3. $[l, u]$ - Exponent range

Any number $x \in \hat{\mathbb{N}}$ has the form

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{p-1}}{\beta^{p-1}} \right) \beta^m, \tag{1}$$

where d_i is an **non-zero** integer such that $0 \leq d_i \leq \beta - 1, \ i = 0, \cdots, p-1$. m is an integer such that $l \leq m \leq u$.

Properties of FP numbers:

1. FP representation is unique.
2. No digits wasted in leading zeros
3. If working in binary, *i.e.* with $\beta = 2$, leading digit, d_0 is always 1 and hence need not be stored.
4. FP numbers are finite and discrete.
5. There are, in total, $2(\beta - 1)\beta^{p-1}(U - L + 1) + 1$ in an FP system.
6. The smallest positive FP number is β^L (known as the under flow limit).
7. The largest positive FP number is $\beta^{U+1}(1 - \beta^{-p})$ (known as the over flow limit).
8. FP numbers are not uniformly distributed throughout their range.

Rounding

The two most common rules of rounding are **chop** and **round to nearest**.

Chop: Number is truncated after $p-1$ digits.

Round to nearest: x is represented by the $\hat{x} \in \hat{\mathbb{N}}$ that is the nearest to x. In the case of tie, round to the nearest even.

Table 1: Rounding and chopping in a floating point system.

Number	Chop	Round to nearest
1.649	1.6	1.6
1.650	1.6	1.6*
1.651	1.6	1.7
1.749	1.7	1.7
1.750	1.7	1.8*

* Round to even!

Machine precision

Characterises the accuracy of a computing system.
For rounding by **chopping**:- $\epsilon_{mach} = \beta^{1-p}$
For rounding to **nearest even**:- $\epsilon_{mach} = 0.5\beta^{1-p}$
For a general FP system

$$\left| \frac{\hat{x} - x}{x} \right| \leq \epsilon_{mach}. \tag{2}$$

Comment: In IEEE FP system $\epsilon_{mach} = 2^{-24} \approx 10^{-7}$ in single precision and $\epsilon_{mach} = 2^{-53} \approx 10^{-16}$ in double precision.

Subnormal FP numbers

If we relax the condition on the leading digit, d_0, and allow it be zero, then the extra numbers added to the FP system the *s*ubnormal (or *denormalised*) FP numbers.
Comment: No change in machine precision by denormalization. Subnormal numbers have lower digits of precision.

FP arithmetic

If the operation of 2 $p-$digit numbers contains more than p digits, then the excess digits are lost in in rounding.

1. For addition and subtraction, the exponents must match before their mantissas can be added or subtracted.

2. No such restriction for multiplication or division.

3. Overflow is more serious problem than underflow.