Error continued.....

Department of Mathematics
IIT Guwahati

Nearby Machine number

We consider a $32-$ bit floating point system. Consider the real number written in binary form

$$x = q \times 2^m, \quad 1 \leq q < 2, \; -126 \leq m \leq 127, \; q = (1.a_1 a_2 a_3 \cdots a_{23} a_{24} \cdots)_2$$

what is the relative error in the closest machine number in this $32 - bit$ floating point system?

Chopping and rounding

To store $x = \pm(d_0.d_1d_2d_3\cdots d_{t-1}d_td_{t+1}\cdots)\times\beta^e$ using only t digits, it is possible to use one of a number of strategies. The two basic ones are

- *chopping:* ignore digits $d_t, d_{t+1}, d_{t+2}, d_{t+3}\ldots$, yielding $\tilde{d}_i = d_i$ and
$$\mathrm{fl}(x) = \pm\, d_0.d_1d_2d_3\cdots d_{t-1}\times\beta^e;$$
- *rounding:* consult d_t to determine the approximation
$$\mathrm{fl}(x) = \begin{cases} \pm\, d_0.d_1d_2d_3\cdots d_{t-1}\times\beta^e, & d_t < \beta/2, \\ \pm\left(d_0.d_1d_2d_3\cdots d_{t-1}+\beta^{1-t}\right)\times\beta^e, & d_t > \beta/2. \end{cases}$$
In case of a tie ($d_t = \beta/2$), round to the nearest even number.

Some results of chopping and rounding with $\beta = 10, t = 3$:

x	Chopped to 3 digits	Rounded to 3 digits
5.672	5.67	5.67
−5.672	−5.67	−5.67
5.677	5.67	5.68
−5.677	−5.67	−5.68
5.692	5.69	5.69
5.695	5.69	5.70

Roundoff error

Home work: Suppose that $fl(y)$ is a $k-$ digit rounding approximation to y. Show that

$$\left|\frac{y - fl(y)}{y}\right| \leq 0.5 \times 10^{1-k}.$$

Theorem: Floating Point Representation Error.
Let $x \mapsto \mathrm{fl}(x) = g \times \beta^e$, where $x \neq 0$ and g is the normalized, signed mantissa.

Then the absolute error committed in using the floating point representation of x is bounded by

$$|x - \mathrm{fl}(x)| \leq \begin{cases} \beta^{1-t} \cdot \beta^e & \text{for chopping,} \\ \frac{1}{2}\beta^{1-t} \cdot \beta^e & \text{for rounding,} \end{cases}$$

whereas the relative error satisfies

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq \begin{cases} \beta^{1-t} & \text{for chopping,} \\ \frac{1}{2}\beta^{1-t} & \text{for rounding.} \end{cases}$$

Machine epsilon

In any computer, it is desirable to know that the four arithmetic operations satisfy equations like:

- If x and y are machine numbers then

$$fl(x \odot y) = (x \odot y)(1 + \delta), \quad |\delta| < \eta = \frac{1}{2}\beta^{1-t}. \tag{1}$$

- If x and y are not necessarily machine number then

$$fl\Big(flx(x) \odot fl(y)\Big) = \Big(x(1 + \delta_1) \odot y(1 + \delta_2)\Big)(1 + \delta_3), \quad |\delta_i| \leq \eta.$$

Rude behavior of roundoff error!!

Cancellation error: if $x \simeq y$, then $x - y$ has a large relative error.

$$
\begin{aligned}
x &= .3721478693 \\
y &= .3720230572 \\
x - y &= .0001248121
\end{aligned}
$$

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

$$
\begin{aligned}
fl(x) &= .37215 \\
fl(y) &= .37202 \\
fl(x) - fl(y) &= .00013
\end{aligned}
$$

The relative error is then very large:

$$
\left| \frac{x - y - [fl(x) - fl(y)]}{x - y} \right| = \left| \frac{.0001248121 - .00013}{.0001248121} \right| \approx 4\%
$$

Does this contradict equation (1)?