CS370 Notes

Minyang Jiang

January 9, 2017

1 Floating Point Number Systems

1.1 Introduction

$$F(\beta, t, L, u)$$

continas:

$$\begin{gathered} 0 \text{ or } \pm 0.\beta_1\beta_2 ... \beta_t * \beta^d \\ \text{where } \beta_1 \neq 0 \\ 0 \leq \beta_i \leq \beta \\ L \leq d \leq u \end{gathered}$$

Two common systems used today - Base 2

1. Single precision: $F(2, 24, -126, 127)$
2. Double precision: $F(2, 53, -1022, 1023)$

Important concepts
If x is any real number then set $fl(x) =$ floating point representation of x
if we write:

$$x = \pm 0.x_1x_2x_3...x_tx_{t+1}... * \beta^d$$

$$fl(x) = \pm 0.x_1x_2x_3...x_t * \beta^d$$

relative error

$$\delta_x = \frac{fl(x) - x}{x}$$

$$\|\delta_x\| \leq ?$$

$$\begin{aligned} \frac{\|fl(x) - x\|}{\|x\|} &= \frac{0.00 \ldots 0x_{t+1} \ldots * \beta^d}{0.x_1x_2 \ldots x_{t+1} \ldots * \beta^d} \\ &= \frac{0.x_{t+1}x_{t+2} \ldots * \beta^{-t}}{x_1.x_2 \ldots * \beta_{-1}} \\ \delta = \frac{fl(x) - x}{x} & \qquad \text{then } \|\delta\| \leq \epsilon = \begin{cases} \beta^{1-t} \\ \frac{\beta^{1-t}}{2} \end{cases} \\ fl(x) = x(1 + \delta) & \qquad |\delta| \leq \epsilon \end{aligned}$$

What about floating point arithmetic?
x, y real numbers, $x + y$ real
$x \oplus y$ =addition inside floating pt system $= fl(fl(x) + fl(y))$

1.2 Analysis some errors in computation

Example. Addition

$$
\begin{aligned}
\left\|\frac{(x+y)-(x\oplus y)}{x+y}\right\| &= \frac{\|(x+y)-fl(fl(x)+fl(y))\|}{x+y} = \frac{\|x+y-x(1+\delta)+y(1+\delta)\|}{x+y} \\
&= \frac{\|x+y-(x+y+\delta_1 x+\delta_2 y+x\delta_3+y\delta_3+\delta_1\delta_3 x+\delta_2\delta_3 y)\|}{x+y} \\
&\le \frac{\|\delta_1 x\|+\|\delta_2 y\|+\|\delta_3 x\|+\|\delta_3 y\|+\delta_1\delta_3 x+\|\delta 2\delta_3 y\|}{\|x+y\|} \\
&\le \frac{(\|x\|+\|y\|)(2\epsilon+\epsilon^2)}{\|x+y\|}
\end{aligned}
$$

$$\|\delta_1\| \le \epsilon, \qquad \|\delta_2\| \le \epsilon, \qquad \|\delta_3\| \le \epsilon$$

$\Rightarrow$ if x and y have same sign then relative error of addition

$$\left\|\frac{x\oplus y-(x+y)}{x+y}\right\| \le 2\epsilon+\epsilon^2$$

However if x and y have opposite sign, then you potentially have a problem particularly when $x+y \approx 0$, Situation is called **Catastrophic cancellation**

$$
\begin{aligned}
x &= 0.x_1x_2\ldots x_{t-1}x_tx_{t+1}\ldots * \beta^d \\
y &= -0.x_1x_2\ldots x_{t-1}x_tx_{t+1}\ldots * \beta^d \\
x+y &= 0.00\ldots 0??\ldots * \beta^d
\end{aligned}
$$

1.3 How about some algorithms?

1.3.1 Example

Given α, compute:

$$I_n = \int_0^1 \frac{x^n}{x+\alpha}dx \qquad n = 0,1,\ldots,100\ldots$$

Step 1

$$I_0 = \int_0^1 \frac{1}{x+\alpha}dx = Ln(x+\alpha) = Ln(1+\alpha) - Ln(\alpha) = Ln\left(\frac{1+\alpha}{\alpha}\right)$$

e.g.

$$
\begin{aligned}
\alpha = 0.5 \text{ then } I_0 = 1.098612288668\ldots \\
\alpha = 2.0 \text{ then } I_0 = 0.405465108108\ldots
\end{aligned}
$$

Step 2 Notice:

$$I_{n+1} = \int_0^1 \frac{x^{n+1}}{x+\alpha}dx = \int_0^1 \frac{x^n(x+\alpha-\alpha}{x+\alpha}dx = \int_0^1 x^n dx - \alpha\int_0^1 \frac{x^n}{x+\alpha}dx$$

$$I_{n+1} = \frac{1}{n+1} - \alpha I_n$$

$$
\begin{aligned}
&I_0 \\
&I_1 = 1 - \alpha I_0 \\
&I_2 = \frac{1}{2} - \alpha I_1 \\
&\vdots \\
&I_{100} = \frac{1}{100} - \alpha I_0 0
\end{aligned}
$$

If $\alpha = 0.5$ then $I_{100} = 0.00664$
if $\alpha = 2.0$ then $I_{100} = 2.1 * 10^{22}$

$$\|I_n\| \leq \frac{1}{1+\alpha}$$

Let's analyze what is happening
Math:

$$I_0^{ex}, I_{n+1}^{ex} = \frac{1}{n+1} - \alpha I_n^{ex}$$

CS:

$$I_0^{app}, I_{n+1}^{app} = \frac{1}{n+1} - \alpha I_n^{aop}$$

At every step, therer is some error:

$$
\begin{aligned}
e_0 &= I_0^{ex} - I_0^{app} \\
e_n &= I_0^{ex} - I_n^{app}
\end{aligned}
$$

Notice:

$$
\begin{aligned}
e_{n+1} &= I_{n+1}^{ex} - I_{n+1}^{app} \\
&= \left(\frac{1}{n+1} - \alpha I_n^{ex}\right) - \left(\frac{1}{n+1} - \alpha I_n^{app}\right) \\
&= -\alpha I_n^{ex} + \alpha I_n^{app} \\
&= -\alpha \left(I_n^{ex} - I_n^{app}\right) \\
&= -\alpha e_n \\
&= (-\alpha)^{n+1} e_0
\end{aligned}
$$

If $\|\alpha\| \leq 1$ then $\|e_n\| \to 0$ as $n \to \inf$
If $\|\alpha\| \geq 1$ then $\|e_n\| \to \inf$ as $n \to \inf$
What to do when $\|\alpha\| \geq 1$:

$$
\begin{aligned}
I_n &= \frac{1}{\alpha(n+1)} - \frac{1}{\alpha} I_{n+1} \\
e_n &= -\frac{1}{\alpha} e_{n+1}
\end{aligned}
$$

If $\|\alpha\| \geq 1$ then work backwards, e.g. I_{100}, Do $I_{200}, I_{199} \ldots$