

CS370 Notes

Minyang Jiang

January 11, 2017

1 Floating Point Number Systems

1.1 Introduction

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

continuas:

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

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:

$$\begin{aligned} x &= \pm 0.x_1x_2x_3\dots x_tx_{t+1}\dots * \beta^d \\ fl(x) &= \pm 0.x_1x_2x_3\dots x_t * \beta^d \end{aligned}$$

relative error

$$\begin{aligned} \delta_x &= \frac{fl(x) - x}{x} \\ \|\delta_x\| &\leq ? \end{aligned}$$

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

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

$$\begin{aligned}
 \left\| \frac{(x+y) - (x \oplus y)}{x+y} \right\| &= \frac{\|(x+y) - fl(fl(x) + fl(y))\|}{x+y} = \frac{\|x+y - x(1+\delta) + y(1+\delta)\|}{x+y} \\
 &= \frac{\|x+y - (x+y + \delta_1x + \delta_2y + x\delta_3 + y\delta_3 + \delta_1\delta_3x + \delta_2\delta_3y)\|}{x+y} \\
 &\leq \frac{\|\delta_1x\| + \|\delta_2y\| + \|\delta_3x\| + \|\delta_3y\| + \|\delta_1\delta_3x\| + \|\delta_2\delta_3y\|}{\|x+y\|} \\
 &\leq \frac{(\|x\| + \|y\|)(2\epsilon + \epsilon^2)}{\|x+y\|} \\
 \|\delta_1\| &\leq \epsilon, \quad \|\delta_2\| \leq \epsilon, \quad \|\delta_3\| \leq \epsilon
 \end{aligned}$$

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

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

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**

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

1.3 How about some algorithms?

1.3.1 Example

Given α , compute:

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

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.

$$\alpha = 0.5 \text{ then } I_0 = 1.098612288668 \dots$$

$$\alpha = 2.0 \text{ then } I_0 = 0.405465108108 \dots$$

Step 2

Notice:

$$\begin{aligned}
 I_{n+1} &= \int_0^1 \frac{x^{n+1}}{x+\alpha} dx = \int_0^1 \frac{x^n(x+\alpha-\alpha)}{x+\alpha} dx = \int_0^1 x^n dx - \alpha \int_0^1 \frac{x^n}{x+\alpha} dx \\
 I_{n+1} &= \frac{1}{n+1} - \alpha I_n
 \end{aligned}$$

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

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:

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

At every step, there is some error:

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

Notice:

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

If $\|\alpha\| \leq 1$ then $\|e_n\| \rightarrow 0$ as $n \rightarrow \infty$
 If $\|\alpha\| \geq 1$ then $\|e_n\| \rightarrow \infty$ as $n \rightarrow \infty$
 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} ...

2 Interpolation

Given n points $(x_1, y_1), \dots, (x_N, y_N)$ x_i distinct $x_1 < x_2 < \dots < x_N$