CS370 Lecture 1

Graham Cooper

January 4th, 2017

Five Topics in the Course (Jan 4th)

- Floating point numbers and Arithmetic
- Iterpolation, Splines, Parametric Curves
- Initial Value Problems - solve differencial equations
- Discrete Fourier Analysis
- Numerical Linear Algebra - solve equations - google pagerank

Topic 1: Floating Point Arithmetic

Examples where problems come whehn using approximation

eg1. $e^{-5.5} =$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... + \frac{x^n}{n!} + ...$$
$$e^{-x} = \frac{1}{x}$$
$$e^{-5.5} = \frac{1}{e^{5.5}} = \frac{1}{1 + 5.5 + \frac{5.5^2}{2} + ...}$$

Now do arithmetic keeping only 5 digits. In both cases infinite sums remain unchanged after 25 terms. There is no sense in going any further - we end up just truncating all of the terms after this as they are smaller than the 5th digit.

Method 1 gives $e^{-5.5} = 0.0026363$
Method 2 gives $e^{-5.5} = 0.0040868$

eg2. $ax^2 + bx + c = 0$
1

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

2

$$x_1 = \frac{2a}{-b + \sqrt{b^2 - 4ac}}$$

$$x_2 = \frac{2a}{-b - \sqrt{b^2 - 4ac}}$$

$$x^2 + 621x + 1 = 0$$

use 4 digit arithmetic

1: $x_1 = -0.200$ $x_2 = 62.1$
2: $x_1 = -0.0161$ $x_2 = -62.1$

Floating point number systems (Jan 6th)

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

Floating point number is either

$$0 + / - 0.x_1 x_2 ... x_t$$

Where x_n is a digit

- $x_1 \neq 0$
- $0 \leq x_i < \beta$
- $L \leq d \leq u$

eg $F(10, 6, L, u)$ for $\pi = 0.314519x10^1$

Single Precision: [+/-][8 bits][23 digits] F(2, 24, -126, 127)
Double Precision: [+/-][11 bits][52 digits] F(2, 53, -1022, 1023)

Given any real number:

$$+/ - 0.x_1 x_2 ... x_t x_{t+1} ... x\beta^d$$

you can truncate (rounding possible):

$$fl(x) = +/- 0.x_1 x_2 ... x_t$$

Question: How close is fl(x) to x? Relatively!

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

Claim: $|\delta| < \beta^{1-t}$ when truncating and $|\delta| < \frac{1}{2}\beta^{1-t}$ when rounding (Let this definition be ϵ)
Trucating:

$$|\delta| = \frac{0.00...0x_{t+1}...x\beta^d}{0.x_1 x_2 ... x\beta^d}$$

$$|\delta| = \frac{0.x_{t+1}x_{t+2}...x\beta^{-t}}{x_1.x_2 x_3...x\beta^{-1}}$$

$$|\delta| < \beta^{1-t}$$

In general $|\delta| < \epsilon$

$$fl(x) = x(1 + \delta)$$

$$|\delta| < \epsilon$$

What about arithmetic an errors?

x,y real
$x \oplus y =$
eg. $x = 0.1111...$ F(2,4,L,u)
$y = 0.1110...$

$fl(x) = 0.1111$
$fl(y) = 0.1110$
when adding them: 1.1101 then truncate to 0.1110×2^1

x,y real
$x \oplus y = fl(fl(x) + fl(y))$

$$|\frac{x \oplus y - (x+y)}{x+y}|$$

$$= |\frac{fl(fl(x) + fl(y)) - (x+y)}{x+y}|$$

since $|\delta_1| < \epsilon$ and $|\delta_2| < \epsilon$ and $|\delta_3| < \epsilon$

$$= |\frac{(x(1+\delta_1) + y(1+\delta_2))(1+\delta_3) - (x+y)}{x+y}|$$

$$|\frac{x + y + x\delta_1 + y\delta_2 + x\delta_3 + y\delta_3 + x\delta_1\delta_3 + y\delta_2\delta_3 - x - y}{x+y}|$$

$$|\frac{x\delta_1 + y\delta_2 + x\delta_3 + y\delta_3 + x\delta_1\delta_3 + y\delta_2\delta_3}{x+y}|$$

$$\leq \frac{|x||\delta_1| + |y||\delta_2| + |x||\delta_3| + |x||\delta_1||\delta_3| + |y||\delta_2\delta_3|}{|x+y|}$$

Since $|a+b| \leq |a| + |b|$ and $|a \times b| = |a||b|$

$$< \frac{(|x| + |y|)}{|x+y|}(2\epsilon + \epsilon^2)$$

$$|\frac{x \oplus y - (x+y)}{x+y}|$$

$$< \frac{|x| + |y|}{|x+y|}(2\epsilon + \epsilon^2)$$

If x and y have the same sign then $|x+y| = |x| + |y|$ and so addition has relative error bounded by $2\epsilon + \epsilon^2$

But if x and y are of opposite sign and perhaps of nearly the same size:

$$\frac{|x| + |y|}{|x+y|}$$

CATASTROPHIC CANCELLATION - Subtraction is deadly.