Math 381 - Fall 2020

Jay Newby

University of Alberta

Week 2

Today

1. ~~Discretization error~~
2. Floating point number system
3. Roundoff error

Review from last time

A computer can store a finite number of bits ($\{0, 1\}$ values) to represent a single floating point number

$$\mathrm{fl}(x) = \mathrm{sign}(x) \times (1.\tilde{d}_1 \tilde{d}_2 \tilde{d}_3 \cdots \tilde{d}_{t-1} \tilde{d}_t) \times 2^e,$$

where e is an integer exponent and $\tilde{d}_n \in \{0, 1\}$.

Review from last time

For IEEE standard

$$\mathcal{E}_{\text{rel}} = \frac{|\text{fl}(x) - x|}{|x|} \leq 2^{-t-1}$$

float32 : $\mathcal{E} \approx 10^{-8}$

float64 : $\mathcal{E} \approx 10^{-16}$

Let $fl(x) = \hat{x}$, $fl(y) = \hat{y}$

Floating point arithmetic error

Let ϵ be machine epsilon. for some constant $|c| \leq 1$, we have

$$\mathrm{fl}(x) \oplus \mathrm{fl}(y) = (\mathrm{fl}(x) + \mathrm{fl}(y))(1 + c\epsilon).$$

Likewise, for some constant $|c| \leq 1$,

$$\mathrm{fl}(x) \otimes \mathrm{fl}(y) = (\mathrm{fl}(x) \cdot \mathrm{fl}(y))(1 + c\epsilon).$$

$\frac{\hat{x} \oplus \hat{y}}{(\hat{x}+\hat{y})} = \frac{(\hat{x}+\hat{y})(1+c\epsilon)}{(\hat{x}+\hat{y})} = 1 + c\epsilon$

Relative error is bounded by ϵ

$|c| \leq 1$

$\Rightarrow \frac{\hat{x} \oplus \hat{y}}{(\hat{x}+\hat{y})} - 1 = \frac{\hat{x} \oplus \hat{y} - (\hat{x}+\hat{y})}{\hat{x}+\hat{y}} = c\epsilon \Rightarrow \frac{|\hat{x} \oplus \hat{y} - (\hat{x}+\hat{y})|}{|\hat{x}+\hat{y}|} \leq 1 \cdot \epsilon$

Common sources of roundoff error

float32 has 8 digits

exact $x+y = 1,000,000,010$

1 2 3 4 5 6 7 8 9 10

Ex: float32 $x = 10^9$, $y = 10$

approx $fl(x+y) = 1.00000000 \times 10^9$

1 2 3 4 5 6 7 8

of ①

Absolute error $|x+y - fl(x+y)| = 10$

1. If x and y differ greatly in magnitude then $x + y$ has a large *absolute* error
2. If y is small in absolute value then x/y may have large *relative* and *absolute* error
3. If y is large in absolute value then $x \cdot y$ may have large *relative* and *absolute* error
4. If $x \approx y$ then $x - y$ has large *relative* error

example continued

Relative error $$\frac{|x+y - fl(x+y)|}{|x+y|} = \frac{10}{10^9+10} \approx 10^{-8} \approx \varepsilon$$

small relative error

Overflow and Underflow

$\hat{x} = fl(x)$, $\hat{y} = fl(y)$

$|z|$ is small

$$z = x - y \Rightarrow \frac{|z - (\hat{x} - \hat{y})|}{|z|} = \frac{|x - y - (\hat{x} - \hat{y})|}{|x - y|}$$

$$= \frac{|x - \hat{x} - (y - \hat{y})|}{|x - y|} \quad \leftarrow \text{can be as small as we want}$$

$\approx 2\varepsilon$ ← relative error could be large

triangle ineq.

$$\leq \frac{|x - \hat{x}| + |y - \hat{y}|}{|x - y|}$$

[Week 2 Jupyter notebook examples]