Welcome to Ma221! Lec 4, Aug 30

Floating Point + Error Analysis

Last time: evaluated $(x-2)^{13}$ using Horner's rule: $O(10^{-8})$ errors $>> |p(x)|$ near root

Floating point: how to represent real numbers

Long ago (<1985) most computers did arithmetic differently $\Rightarrow$ hard to write portable code $\Rightarrow$ IEEE 754 Standard Committee formed to create a standard, led by Prof. Kahan

First Standard 1985, then 2008, then 2019, forming committee for 2028 standard

Scientific Notation:

$$\pm d.dd\ldots d \cdot \text{radix}^e$$

Usually radix = 2 (or 10 for finance)

Store sign bit $\pm$
exponent (e)
mantissa d.d...d

p = # digits in mantissa

both p and # bits in e are limited to
16, 32, 64, 128 bits
Historically, only hardware for 32, 64
Lately 16 popular, for ML
lots of accelerators under construction
Also bfloat16 from Google
more exponent bits than IEEE16
smaller p
Latest committee: 8 bit arithmetic
e4p4 and e5p3
For starters, ignore limits on e,
so assume no overflow, no underflow

Normalization: use
$3.100 \cdot 10^0$ not $0.0031 \cdot 10^3$
$\Rightarrow$ unique representation
+ hardware simpler
+ in binary, leading $d = 1$
$\Rightarrow$ "free bit", don't store it
(decimal: store 3 digits in 10 bits
not 1 digit in 4 bits)

Def: $rnd(x)$ = nearest floating point number to x

(Default IEEE rounding: round to "nearest even", i.e. last bit is zero if there is a tie

i.e. $\frac{1}{2}$ time round up, $\frac{1}{2}$ time down

$\Rightarrow$ unbiased

Def: Relative Representation Error (RRE)

$$RRE(x) = \frac{|x - rnd(x)|}{|rnd(x)|}$$

Def: Maximum RRE $= \max_{x \neq 0} RRE(x)$

aka machine epsilon, macheps, ε

(note: matlab's eps $= 2\varepsilon$)

Max RRE = half distance from 1 to next larger floating point number

$= 1 + (radix)^{1-p}$

MaxRRE $= .5 \cdot radix^{1-p} = 2^{-p}$ in binary

Round off model (no over/underflow)

(*) $fl(a \ op \ b) = rnd(a \ op \ b)$

$= \text{true result rounded to nearest even}$

$= (a \; op \; b)(1+\delta) \quad |\delta| \le \varepsilon$

op can be $+, -, *, /$

754 also includes $\sqrt{\ }, \ldots$

(*) also true for complex arithmetic with larger ε

Existing IEEE Binary Formats

half (H), single (S), double (D), quad (Q)

S: 32 bits = 1 (sign)
+8 (exponent bits)
+23 (mantissa bits)

$p = 1 + 23 = 24 \Rightarrow \varepsilon = 2^{-24} \approx 6 \cdot 10^{-8}$

$-126 \le e \le 127 \Rightarrow$ OV = overflow thresh.
$\approx 2^{128} \approx 10^{38}$

$\Rightarrow$ UN = underflow thresh
$= 2^{-126} \approx 10^{-38}$

D: $64 = 1 + 11 + 52 \Rightarrow p = 53 \Rightarrow \varepsilon = 2^{-53} \approx 10^{-16}$

$-1022 \le e \le 1023$ OV $\approx 2^{1024} \approx 10^{308}$

UN $= 2^{-1022} \approx 10^{-308}$

Q: $128 = 1 + 15 + 112$

H: $16 = 1 + 5 + 10 \Rightarrow \varepsilon \approx 5 \cdot 10^{-4}$

$OV \sim 10^4$, $UN \sim 10^{-4}$

Bfloat16: $16 = 1 + \underset{\sim}{8} + 7 \Rightarrow \varepsilon \sim 4 \cdot 10^{-3}$

Same as single S

Use (*) $fl(a \; op \; b) = (a \; op \; b)(1+\delta)$

$|\delta| \le \varepsilon$

for error analysis

Horner's rule for $p(x) = \sum_{i=0}^{d} a_i \cdot x^i$

```
p = a_d
for i = d-1 : -1 : 0
    p = p * x + a_i
```

Label intermediate terms

```
p_d = a_d
for i = d-1 : -1 : 0
    p_i = p_{i+1} * x + a_i
```

Introduce roundoff

```
p_d = a_d
for i = d-1 : -1 : 0
    p_i = (x * p_{i+1} * (1 + δ_i) + a_i) * (1 + δ_i')
          |δ_i| ≤ ε   |δ_i'| ≤ ε
```

Simplify:

$$p_0 = \sum_{i=0}^{d-1} a_i x^i \left[(1+\delta_i') \prod_{j=0}^{i-1} (1+\delta_j)(1+\delta_j') \right]$$

$$+ a_d \cdot x^d \left[\prod_{j=0}^{d-1} (1+\delta_j)(1+\delta_i') \right]$$

How many factors $(1+\delta)$?

$$= \sum_{i=0}^{d-1} a_i' x^i \quad \text{where} \quad \leq 2d$$

$$a_i' = a_i \cdot (\text{factors } (1+\delta))$$

Horner is backwards stable:
exact value of slightly
different polynomial

Simplify:

$$\prod_{i=1}^{n} (1+\delta_i) \leq \prod_{i=1}^{n} (1+\varepsilon) = (1+\varepsilon)^n$$

$$= 1 + n\varepsilon + O(\varepsilon^2)$$

$$\leq 1 + \frac{n\varepsilon}{1-n\varepsilon} \quad \text{if } n\varepsilon < 1$$

$$\prod_{i=1}^{n} (1+\delta_i) \geq (1-\varepsilon)^n = 1 - n\varepsilon + O(\varepsilon^2)$$

$$\geq 1 - \frac{n\varepsilon}{1-n\varepsilon}, \quad n\varepsilon < 1$$

$$|\text{computed } p_d - p(x)| \le \sum_{i=0}^{d-1} (2i+1)\varepsilon |a_i x^i| + 2d\varepsilon |a_d x^d|$$

$$\text{relerr} = \frac{|\text{computed } p_d - p(x)|}{|p(x)|}$$

$$\le \underbrace{\frac{\sum_{i=0}^{d} |a_i x^i|}{|p(x)|}}_{\text{condition number}} \cdot \underbrace{2d\varepsilon}_{\text{backward error}}$$

k correct digits $\Leftrightarrow$ relative error bound $\le 10^{-k}$

$\Rightarrow -\log_{10}(\text{relative error}) \ge k$

Compiler absolute error bounds

```
p = a_d , ebnd = |a_d|
for i = d-1 : -1 : 0
    p = x·p + a_i , ebnd = |x|·ebnd + |a_i|
ebnd = ebnd · 2dε
```