MATH50003
Numerical Analysis

II.3 Floating Point Arithmetic

Dr Sheehan Olver

Part II

Representing Numbers

1. Reals via floating point
2. Floating point arithmetic and bounding errors
3. Interval arithmetic for rigorous computations

$$x = \pm 2^{q-\sigma}(1.b_1 b_2 \text{---} b_S b_{S+1} b_{S+2} \text{---})_2$$

Rounding

How does a computer round a real to a float?

Definition 7 (rounding). $\mathrm{fl}^{\mathrm{up}}_{\sigma,Q,S} : \mathbb{R} \to F_{\sigma,Q,S}$

$$\mathrm{fl}^{up}(x) := \min\{y \in F \text{ s.t. } y \geq x\}$$

Eg

$$\mathrm{fl}^{up}_{\underbrace{16}_{15,5,10}}(1/3) = \mathrm{fl}^{up}(2^{-2} \times (1.0101010101\,0101\text{---})_2)$$

$$= 2^{-2} \times (1.0101010110)_2$$

$\mathrm{fl}^{\mathrm{down}}_{\sigma,Q,S} : \mathbb{R} \to F_{\sigma,Q,S}$

$$\mathrm{fl}^{down}(x) := \max\{y \in F \text{ s.t. } y \leq x\}$$

Eg

$$\mathrm{fl}^{down}(1/3) = 2^{-2}(1.0101010101)_2$$

Note

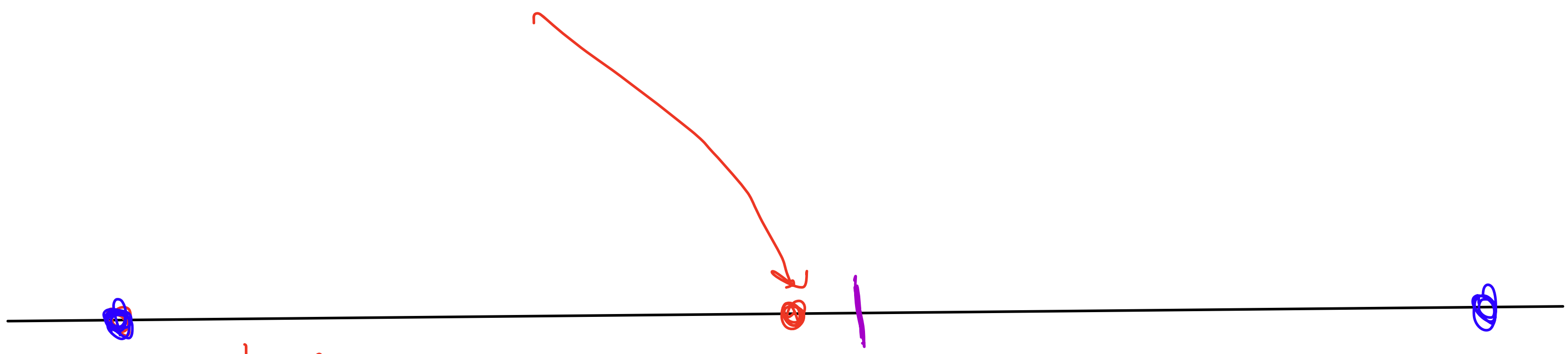
$$\mathrm{fl}^{up}(-x) = -\mathrm{fl}^{down}(x)$$

Default is round-to-nearest.

$\underbrace{\mathrm{fl}^{\text{nearest}}_{\sigma,Q,S}}_{=:\,\mathrm{fl}} : \mathbb{R} \to F_{\sigma,Q,S}$

Consider

$$x = 2^{q-\sigma}(1.b_1 b_2 \ldots b_S\, b_{S+1}\, b_{S+2} \ldots)_2 > 0$$



$$x_- = \mathrm{fl}^{\text{down}}(x) = 2^{q-\sigma}(1.b_1 \ldots b_S)_2$$

$$x_h = \frac{x_+ + x_-}{2} = x_- + 2^{q-\sigma-S-1} = 2^{q-\sigma}(1.b_1 \ldots b_S\, 1)_2$$

$$x_+ = \mathrm{fl}^{\text{up}}(x) = x_- + 2^{q-\sigma-S}$$

Different cases:

(down) $x_- \le x < x_h \Rightarrow \mathrm{fl}^{\text{near}}(x) = \mathrm{fl}^{\text{down}}(x) = x_-$

up $\quad x_b < x \le x_+ \Rightarrow fl^{near}(x) = fl^{up}(x) = x_+$

half $\quad x = x_b \Rightarrow fl^{near}(x) = \begin{cases} x_- & \text{if } b_s = 0 \\ x_+ & \text{otherwise} \end{cases}$

so that $fl^{near}(x)$ has last bit $\emptyset$.

$$\boxed{\begin{array}{l} \text{For } x < 0 \\ fl(x) = -fl(-x) \end{array}}$$

Arithmetic

Operations are exact up to rounding

$$x \oplus y := \mathrm{fl}(x + y)$$
$$x \ominus y := \mathrm{fl}(x - y)$$
$$x \otimes y := \mathrm{fl}(x * y)$$
$$x \oslash y := \mathrm{fl}(x / y)$$

"+" on a computer adds exactly, then rounds

Eg: in F_{16},

$$1 \oplus 2^{-11} = \mathrm{fl}(1 + 2^{-11})$$
$$= \mathrm{fl}((1.00000000001)_2)$$
$$= 1$$

$$x \,\hat{\circ}\, y := \mathrm{fl}(x^y)$$

Here, $x, y \in F$, not general reals!

Example 8 (decimal is not exact).

What's $1.1 + 0.1$ on a computer?

When we type "1.1" it makes $fl(1.1)$.

So this is

$$fl(1.1) \oplus fl(0.1)$$

We have

$b_{s+1} = 0 \Rightarrow$ round down

$$fl(1.1) = fl((1.0001100110\,01100\ldots)_2)$$

$$= (1.0001100110)_2$$

$b_{s+1} = 0 \Rightarrow$ round down

$$fl(0.1) = fl(2^{-4}(1.1001100110\,0110\ldots)_2)$$

$$= 2^{-4}(1.1001100110)_2$$

$\Rightarrow$

$$fl(1.1) \oplus fl(0.1) = fl(fl(1.1) + fl(0.1))$$

$$= fl((1.0011001100\,011)_2)$$

round down

$$= (1.0011001 1)_2$$

But

$$fl(1.2) = fl(1.0011001100\,11\,0\,0\,11\ldots)_2$$

$b_{s+1} = 1$

round up

$$= (1.0011001101)_2$$

II.2.1 Bounding errors
Analysis on rounding errors

Definition 8 (machine epsilon/smallest positive normal number/largest normal number)**.** *Machine epsilon* is denoted

$$\epsilon_{\mathrm{m},S} := 2^{-S}.$$

For F_{64}, $S = 52 \Rightarrow \epsilon_m = 2^{-52}$

$\approx 2.22 \times 10^{-16}$

Definition 9 (normalised range)**.** The *normalised range* $\mathcal{N}_{\sigma,Q,S} \subset \mathbb{R}$ is the subset of real numbers that lies between the smallest and largest normal floating-point number:

$$\mathcal{N}_{\sigma,Q,S} := \{x : \underbrace{\min |F^{\text{normal}}_{\sigma,Q,S}|}_{\text{smallest positive normal}} \leq |x| \leq \underbrace{\max F^{\text{normal}}_{\sigma,Q,S}}_{\text{largest normal}}\}$$

$$x \in N_{\sigma,Q,S} \Rightarrow fl^{\text{mode}}(x) \in F^{\text{normal}}$$

relative error

Proposition 2 (round bound). *If $x \in \mathcal{N}$ then*

$$\mathrm{fl}^{\mathrm{mode}}(x) = x(1 + \delta_x^{\mathrm{mode}}) = x + \underbrace{x\,\delta_x^{\mathrm{mode}}}$$

where the relative error *is bounded by:*

machine $\epsilon \approx 2.22 \times 10^{-16}$

$$|\delta_x^{\mathrm{nearest}}| \leq \frac{\epsilon_{\mathrm{m}}}{2}$$

$$|\delta_x^{\mathrm{up/down}}| < \epsilon_{\mathrm{m}}.$$

Proof Only show mode = near. Assume $x > 0$.

Suppose $fl(x) = fl^{\text{down}}(x)$, i.e., $x_- \le x \le x_h$.

(Round Down)

$x = 2^{q-\sigma}(1.b_1 b_2 \ldots b_S b_{S+1} \ldots)_2$

$2^{q-\sigma}$

$x_- := \mathrm{fl}^{\text{down}}(x)$
$= 2^{q-\sigma}(1.b_1 b_2 \ldots b_S)_2$

$x_h := \dfrac{x_+ + x_-}{2}$
$= x_- + 2^{q-\sigma-S-1}$
$= 2^{q-\sigma}(1.b_1 b_2 \ldots b_S 1)_2$

$x_+ := \mathrm{fl}^{\text{up}}(x)$
$= x_- + 2^{q-\sigma-S}$

We have

$$fl(x) = x_- = x\left(1 + \underbrace{\frac{x_- - x}{x}}_{\delta_x}\right)$$

Where

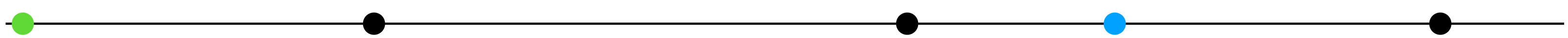
$$|\delta_x| \le \frac{x - x_-}{x} \underset{\text{since } x \le x_h}{\le} \frac{x_h - x_-}{x} \underset{x \ge x_-}{\le} \frac{x_h - x_-}{x_-}$$

$$\leq \frac{2^{q-\sigma-S-1}}{2^{q-\sigma}} = 2^{-S-1} = \frac{\epsilon_m}{2}$$

$\Uparrow$

(Round Up) $x_- \geq 2^{q-\sigma}$

$x = 2^{q-\sigma}(1.b_1 b_2 \ldots b_S b_{S+1} \ldots)_2$



$2^{q-\sigma}$

$x_- := \mathrm{fl}^{\mathrm{down}}(x)$

$= 2^{q-\sigma}(1.b_1 b_2 \ldots b_S)_2$

$x_{\mathrm{h}} := \dfrac{x_+ + x_-}{2}$

$= x_- + 2^{q-\sigma-S-1}$

$= 2^{q-\sigma}(1.b_1 b_2 \ldots b_S 1)_2$

$x_+ := \mathrm{fl}^{\mathrm{up}}(x)$

$= x_- + 2^{q-\sigma-S}$

Example 9 (bounding a simple computation).

$$(1.1 + 1.2) \times 1.3 = 2.99$$

What's a bound on error when done on computer?

$$\underbrace{[fl(1.1) \oplus fl(1.2)]}_{(*)} \otimes fl(1.3) = 2.99 + \underbrace{\delta}_{\text{absolute error}}$$

Show

$$|\delta| \le \underbrace{23\,\epsilon_m}_{\text{pecimistic bound}} \le 50 \times 10^{-16}$$

$$fl(\underbrace{1.1}_{\in N_{64,s}}) \overset{\text{round bound}}{=} 1.1\,(1 + \underbrace{\delta_1}_{|\delta_1| \le \frac{\epsilon_m}{2}})$$

$$fl(1.2) = 1.2(1+\delta_2)$$

$|\delta_2| \le \frac{\epsilon_m}{2}$

$$\Rightarrow$$

$$(*) = fl(\underbrace{1.1(1+\delta_1)}_{fl(1.1)} + 1.2(1+\delta_2))$$

from round bound

$$= (2.3 + 1.1\delta_1 + 1.2\delta_2)(1+\delta_3)$$

$|\delta_3| \le \frac{\epsilon_m}{2}$

$$= 2.3 + \underbrace{1.1\delta_1 + 1.2\delta_2 + 2.3\delta_3 + 1.1\delta_1\delta_3 + 1.2\delta_2\delta_3}_{=:\ \epsilon_1}$$

true

1.1 + 1.2

Where

$$|\epsilon_1| \le (\underbrace{1.1}_{\le 2} + \underbrace{1.2}_{\le 2} + \underbrace{2.3}_{\le 3})\frac{\epsilon_m}{2} + (\underbrace{1.1}_{\le 2} + \underbrace{1.2}_{\le 2})\underbrace{\frac{\epsilon_m^2}{4}}_{\le \frac{\epsilon_m}{4}2^{-5} \le \frac{\epsilon_m}{4}}$$

$$\le (2+2+3+1+1)\frac{\epsilon_m}{2} \le 5\epsilon_m.$$

$\Rightarrow$

error from $\otimes$

$$(*) \quad \otimes \quad \underbrace{fl(1.3)}_{1.3(1+\delta_4)} = (2.3+\epsilon_1)\,1.3(1+\delta_4)(1+\delta_5)$$

$$= 2.99 + \underbrace{1.3}_{\le 3/2}(\epsilon_1 + \underbrace{2.3}_{\le 3}\delta_4 + \underbrace{2.3}_{\le 3}\delta_5 + \underbrace{\epsilon_1\delta_4}_{\le \frac{5\epsilon_m^2}{2} \le \frac{5}{2}\epsilon_m} + \underbrace{\epsilon_1\delta_5}_{\le \frac{5\epsilon_m}{2}} + \underbrace{2.3}_{\le 3}\underbrace{\delta_4\delta_5}_{\le \frac{\epsilon_m}{4}} + \underbrace{\epsilon_1\delta_4\delta_5}_{\le \frac{5\epsilon_m}{4}})$$

δ

where $|\delta| \leq \cancel{2}\,3\,\epsilon_m$

II.2.2 Idealised floating point

A simplifed model for analysis

Definition 10 (idealised floating point)**.** An idealised mathematical model of floating point numbers for which the only subnormal number is zero can be defined as:

$$F_{\infty,S} := \{\pm 2^q \times (1.\underbrace{b_1 b_2 b_3 \ldots b_S}_{S \text{ digits}})_2 : q \in \mathbb{Z}\} \cup \{0\}$$

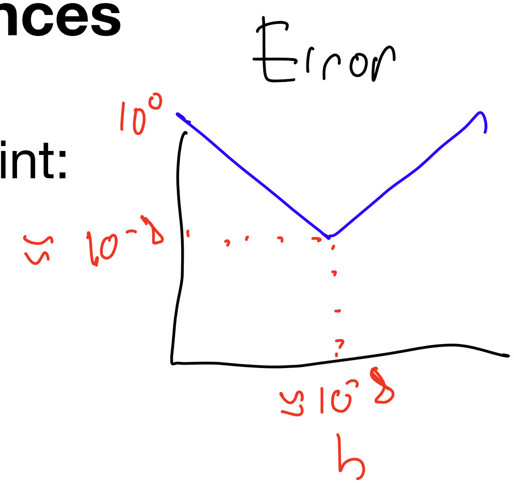
$Q=\infty$

anything

Round bound applies $\forall x \in F_{\infty,S}$

II.2.3 Divided differences floating point error bound

Explain the unexplained error in divided differences



General model of a function implemented in floating point:

$\forall x,$

$$f(x) = f^{\rm FP}(x) + \underbrace{\delta_x^f}_{\text{absolute error}}$$

such that

$$|\delta_x^f| \leq c\epsilon_{\rm m}$$

some reasonable c

Theorem 4 (divided difference error bound). For $x \in F_{\infty,S}$,

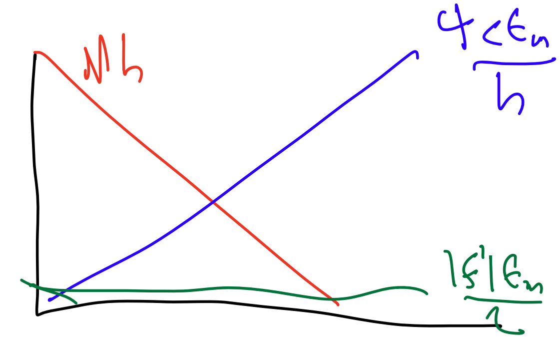
$$\frac{f^{\rm FP}(x+h) \ominus f^{\rm FP}(x)}{h} = f'(x) + \delta^{\rm FD}_{x,h}$$

abs error

where

$$|\delta^{\rm FD}_{x,h}| \leq \underbrace{\frac{|f'(x)|}{2}\epsilon_{\rm m}}_{\text{small } \approx 10^{-16}} + \underbrace{Mh}_{\to 0,\ h\to 0} + \underbrace{\frac{4c\epsilon_{\rm m}}{h}}_{\to \infty,\ h \to 0}$$

for $M = \sup_{x \leq t \leq x+h} |f''(t)|$.



Of course, bound is just upper bound so haven't shown it will blow up.

Proof

$$\frac{f^{FP}(x+h) \ominus f^{FP}(x)}{h} = \frac{\left(f(x+h) - \delta^f_{x+h} - f(x) + \delta^f_x\right)}{h}(1+\underbrace{\delta_1}_{|\delta_1| \leq \frac{\epsilon_m}{2}})$$

$$= \underbrace{\frac{f(x+h) - f(x)}{h}}_{\text{Prop 1}} (1+\delta_1) + \frac{\delta^f_x - \delta^f_{x+h}}{h} (1+\delta_1)$$

$$= f'(x) - \delta^{dd}$$

where

$$|\delta^{dd}| \leq \frac{Mh}{2}$$

$$= f'(x) + \underbrace{f'(x)\delta_1}_{|\cdot| \leq \frac{|f'(x)|}{2}\epsilon_m} - \underbrace{\delta^{dd}(1+\delta_1)}_{\leq |\delta^{dd}|(1+|\delta_1|) \leq Mh}$$

$$+ \underbrace{\frac{\delta^f_x - \delta^f_{x+h}}{h} (1+\delta_1)}$$

$$|\cdot| \leq \frac{|\delta_x^f| + |\delta_{x+h}^f|}{h}(1+|\delta_1|) \leq \frac{4 c \epsilon_m}{h}$$

Corollary 2 (divided differences in practice). *We have* (Non-examinable)

$$(f^{\mathrm{FP}}(x \oplus h) \ominus f^{\mathrm{FP}}(x)) \oslash h = \frac{f^{\mathrm{FP}}(x+h) \ominus f^{\mathrm{FP}}(x)}{h}$$

whenever $h = 2^{j-n}$ *for* $0 \leq n \leq S$ *and the last binary place of* $x \in F_{\infty,S}$ *is zero, that is* $x = \pm 2^j(1.b_1 \ldots b_{S-1}0)_2$.

No error in $\oslash$ since $\frac{\pm 2^q (1.b_1 \ldots b_S)_2}{2^{j-n}}$

$= \pm 2^{q+j-n}(1.b_1 \ldots b_S)_2 \in F_{\infty,S}$.

Also, since $b_S = 0$ we have

$x \oplus h = x + h$.

Heuristic (divided difference with floating-point step)

What's the best choice of h?

Balance

$$Mh \simeq \frac{4c\,\epsilon_m}{h} \Rightarrow$$

$$h^2 \simeq \frac{4c}{M}\epsilon_m$$

might not know

$$\Rightarrow \quad h = \text{const.}\ \sqrt{\epsilon_m}$$

$\hookrightarrow 1.49 \times 10^{-8}$

Now to Lab 3
To see rounding modes.