MATH50003 Numerical Analysis

II.3 Floating Point Arithmetic

Dr Sheehan Olver

Part II
Representing Numbers

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

Rounding

How does a computer round a real to a float?

Definition 10 (rounding). $\mathrm{fl}^{\mathrm{up}}_{\sigma,Q,S} : \mathbb{R} \to F_{\sigma,Q,S}$

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

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

Arithmetic

Operations are exact up to rounding

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

Example 16 (decimal is not exact)

II.3.1 Bounding errors

Analysis on rounding errors

Definition 11 (machine epsilon/smallest positive normal number/largest normal number)**.**

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

Definition 12 (normalised range)

$$\mathcal{N}_{\sigma,Q,S} := \{x : \min |F^{\mathrm{normal}}_{\sigma,Q,S}| \leq |x| \leq \max F^{\mathrm{normal}}_{\sigma,Q,S}\}$$

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

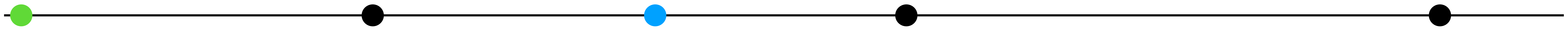
$$\mathrm{fl}^{\mathrm{mode}}(x) = x(1 + \delta_x^{\mathrm{mode}})$$

where the relative error *is bounded by:*

$$|\delta_x^{\mathrm{nearest}}| \leq \frac{\epsilon_{\mathrm{m}}}{2}$$
$$|\delta_x^{\mathrm{up/down}}| < \epsilon_{\mathrm{m}}.$$

(Round Down)

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



$$2^{q-\sigma}$$

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

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

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

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

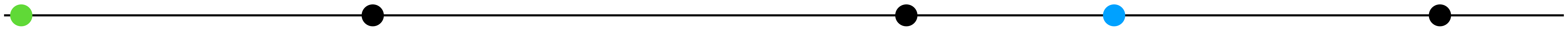
$$= 2^{q-\sigma}(1.b_1 b_2 \ldots b_S 1)_2$$

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

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

(Round Down)

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



$2^{q-\sigma}$

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

$$x_{\mathrm{h}} := \frac{x_+ + x_-}{2}$$
$$= x_- + 2^{q-\sigma-S-1}$$
$$= 2^{q-\sigma}(1.b_1 b_2 \dots b_S 1)_2$$

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

II.3.2 Idealised floating point

A simplifed model for analysis

Definition 13 (idealised floating point)**.** An idealised mathematical model of floating point numbers for which the only subnormal number is zero can be defined as:

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

Example 17 (bounding a simple computation)

II.3.3 Divided differences floating point error bound

Explain the unexplained error in divided differences

General model of a function implemented in floating point:

$$f(x) = f^{\mathrm{FP}}(x) + \delta_x^f$$

such that

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

Theorem 3 (divided difference error bound)**.**

$$\frac{f^{\mathrm{FP}}(x+h) \ominus f^{\mathrm{FP}}(x)}{h} = f'(x) + \delta_{x,h}^{\mathrm{FD}}$$

where

$$|\delta_{x,h}^{\mathrm{FD}}| \leq \frac{|f'(x)|}{2}\epsilon_{\mathrm{m}} + Mh + \frac{4c\epsilon_{\mathrm{m}}}{h}$$

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

Corollary 1 (divided differences in practice). *We have*

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

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* $x = \pm 2^j (1.b_1 \dots b_{S-1}0)_2$.

Heuristic (divided difference with floating-point step)

Now to Lab 4
To see rounding modes.