MATH50003 Numerical Analysis

II.2 Reals

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

Ariane 5 rocket explosion

Learn floating point, or else…

II.2.1 Real numbers in binary

We can represent any real number using binary digits

Definition 6 (real binary format)**.** For $b_1, b_2, \ldots \in \{0, 1\}$, Denote a non-negative real number in *binary format* by:

$$(B_p \ldots B_0.b_1 b_2 b_3 \ldots)_2 := (B_p \ldots B_0)_2 + \sum_{k=1}^{\infty} \frac{b_k}{2^k}.$$

Example 11 (rational in binary)

II.2.2 Floating-point numbers

How do we represent an uncountable set with only *p*-bits?

Bit Format: $s\ q_{Q-1} \ldots q_0\ b_1 \ldots b_S$

Definition 7 (floating-point numbers)**.** Given integers σ (the *exponential shift*), Q (the number of *exponent bits*) and S (the *precision*), define the set of *Floating-point numbers* by dividing into *normal*, *sub-normal*, and *special number* subsets:

$$F_{\sigma,Q,S} := F^{\text{normal}}_{\sigma,Q,S} \cup F^{\text{sub}}_{\sigma,Q,S} \cup F^{\text{special}}.$$

How do bits dictate whether its normal/sub/special?

Look at exponent. 3 examples:

0 10000 1010000000

1 00000 1100000000

1 11111 0000000000

II.2.3 IEEE float-point numbers

What exponent shift/number of bits/precision is used in practice?

$$F_{16} := F_{15,5,10}$$

$$F_{32} := F_{127,8,23}$$

$$F_{64} := F_{1023,11,52}$$

$$F^{\text{normal}}_{\sigma,Q,S} := \{\pm 2^{q-\sigma} \times (1.b_1 b_2 b_3 \ldots b_S)_2 : 1 \leq q < 2^Q - 1\}.$$

Half-precision
$F_{16} := F_{15,5,10}$

Example 12 (interpreting 16-bits as a float)**.** Consider the number with bits

$$0\ 10000\ 1010000000$$

Example 13 (rational to 16-bits). How is the number $1/3$ stored in F_{16}?

II.2.4 Sub-normal and special numbers

Sub-normal have exponent bits all 0, special have all 1

$$F^{\text{sub}}_{\sigma,Q,S} := \{\pm 2^{1-\sigma} \times (0.b_1 b_2 b_3 \ldots b_S)_2\}.$$

Example 14 (subnormal in 16-bits)**.** Consider the number with bits

$$1 \; 00000 \; 1100000000$$

$$F^{\text{special}} := \{\infty, -\infty, \text{NaN}\}$$

Example 15 (special in 16-bits)**.** The number with bits

$$\color{red}{1}\ \color{green}{11111}\ \color{blue}{0000000000}$$

On the other hand, the number with bits

$$\color{red}{1}\ \color{green}{11111}\ \color{blue}{0000000001}$$

Time for code.