1.2 Round-off Errors and Computer Arithmetic (binary numbers)

- In a computer model, a memory storage unit – **word** is used to store a number.
- A **word** has only a finite number of bits.
- These facts imply:
  1. Only a small set of real numbers (rational numbers) can be accurately represented on computers.
  2. (Rounding) errors are inevitable when computer memory is used to represent real, infinite precision numbers.
  3. Small rounding errors can be amplified with careless treatment.

So, do not be surprised that $(9.4)_{10} = (1001.\overline{0110})_2$ can not be represented exactly on computers.

IEEE floating point numbers

- Binary number: $(\ldots b_3 b_2 b_1 b_0 . b_{-1} b_{-2} b_{-3} \ldots)_2$
- Binary to decimal: $(\ldots b_3 b_2 b_1 b_0 . b_{-1} b_{-2} b_{-3} \ldots)_2 =$ $(\ldots b_3 2^3 + b^2\, 2^2 + b_1 2^1 + b_0 2^0 + b_{-1} 2^{-1} + b_{-2} 2^{-2} + b_{-3} 2^{-3} \ldots)_{10}$
- Double precision (long real) format
  - Example: “double” in C
- A 64-bit (binary digit) representation
  - 1 sign bit (s), 11 exponent bits – characteristic (c), 52 binary fraction bits – mantissa (f)

x	xxxxxxxxxxx	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
s	c	f

Represented number (Normalized IEEE floating point number):

$$(-1)^s 2^{c-1023}(1+f)$$

1023 is called exponent bias

$$0 \leq characteristic\ (c) \leq 2^{11} - 1 = 2047$$

- Smallest normalized positive number on machine has $s = 0, c = 1, f = 0$: $2^{-1022} \cdot (1 + 0) \approx 0.22251 \times 10^{-307}$
- Largest normalized positive number on machine has $s = 0, c = 2046, f = 1 - 2^{-52}$: $2^{1023} \cdot (1 + 1 - 2^{-52}) \approx 0.17977 \times 10^{309}$
- **Underflow**: $numbers\ < 2^{-1022} \cdot (1 + 0)$
- **Overflow**: $numbers > 2^{1023} \cdot (2 - 2^{-52})$
- **Machine epsilon** $(\epsilon_{mach}) = 2^{-52}$: this is the difference between 1 and the smallest machine floating point number greater than 1.

- Positive zero: $s = 0, c = 0, f = 0$.
- Negative zero: $s = 1, c = 0, f = 0$.
- Inf: $s = 0, c = 2047, f = 0$
- NaN: $s = 0, c = 2047, f \neq 0$