Numerical Analysis

Aakash Jog

2015-16

Contents

1 Lecturer Information

Prof. Nir Sochen

Office: Schreiber 201
Telephone: +972 3-640-8044
E-mail: sochen@post.tau.ac.il
Office Hours: Sundays, 10:00–12:00

2 Required Reading

1. S. D. Conte and C. de Boor, Elementary Numerical Analysis, 1972

3 Floating Point Representation

Exercise 1.
Represent 9.75 in base 2.

Solution 1.

$$\begin{aligned}
9.75 &= 8 + 1 + \frac{1}{2} + \frac{1}{4} \\
&= 2^3 + 2^0 + 2^{-1} + 2^{-2} \\
&= 2^3 \left(2^0 + 2^{-3} + 2^{-4} + 2^{-5}\right) \\
&= \left(2^{11} \left(1 + 0.001 + 0.0001 + 0.00001\right)\right)_2 \\
&= \left(2^{11} \left(1.00111\right)\right)_2
\end{aligned}$$

Definition 1 (Double precision floating point representation)**.** A floating point representation which uses 64 bits for representation of a number is called a double precision floating point representation.
The standard form of double precision representation is

$$a = \underbrace{\pm}_{1 \text{ bit}} \underbrace{1}_{1 \text{ bit}} . \underbrace{\cdots}_{52 \text{ bits}} \times w^{\underbrace{\pm}_{1 \text{ bit}} \underbrace{\cdots}_{10 \text{ bits}}}$$

Theorem 1 (Range of double precision floating point representation)**.** *The largest number which can be represented with double precision floating point representation is approximately* 10^{307} *and the smallest number which can be represented is approximately* 10^{-307}.

Proof. As the exponent has 10 bits for representation,

$$-\left(10^{10} - 1\right) \leq \text{exponent} \leq \left(10^{10} - 1\right)$$

Therefore,

$$-1023 \leq \text{exponent} \leq 1023$$

Therefore, the smallest number, in terms of absolute value, which can be represented, is

$$1.\underbrace{0 \cdots 0}_{52 \text{ bits}} \times 2^{-1024} \approx 10^{-307}$$

Therefore, the smallest number which can be represented is approximately 10^{-307}, and the largest number which can be represented is approximately 10^{307}. $\square$

Definition 2 (Overflow)**.** If a result is larger than the largest number which can be represented, it is called overflow.

Definition 3 (Underflow)**.** If a result is smaller than the smallest number which can be represented, it is called underflow.

Definition 4 (Least significant digit)**.**

$$1 = 1.\underbrace{0 \cdots 0}_{\text{52 zeros}} \times 2^0$$

Let 1_ε be the smallest number larger than 1, which can be represented in double precision floating point representation.
Therefore,

$$\begin{aligned} 1 &= 1.\underbrace{0 \cdots 0}_{\text{51 zeros}} 1 \times 2^0 \\ &= 1 + 2^{-52} \\ &\approx 1 + 2 \times 10^{-16} \end{aligned}$$

Therefore,

$$\begin{aligned} 1 - 1_\varepsilon &= 2^{-52} \\ &\approx 2 \times 10^{-16} \end{aligned}$$

This number is called the least significant digit, or the machine precision. It is the maximum possible error in representation. It is represented by ε.

Definition 5 (Error)**.** Let the DPFP representation of a number x be $\widetilde{x}$. The absolute error in representation is defined as

$$\begin{aligned} \text{absolute error} &= |x - \widetilde{x}| \\ &= 0.0 \cdots 01 \times 2^{\text{exponent}} \end{aligned}$$

The relative error in representation is defined as

$$\begin{aligned} \delta &= \frac{|x - \widetilde{x}|}{x} \\ &= 0.0 \cdots 01 \\ &< \varepsilon \end{aligned}$$

The maximum error, $2^{-52} \approx 2 \times 10^{-16}$, is called the machine precision.
In general,

$$\widetilde{x} \widetilde{\star} \widetilde{y} = (x \star y)(1 + \delta)$$

where δ is the relative error, ε is the machine precision, $\delta < \varepsilon$, and $\star$ is an operator.

3.1 Loss of Significant Digits in Addition and Subtraction

Exercise 2.
Represent $\pi + \frac{1}{30}$ in base 10 with 4 digits.

Solution 2.

$$\pi \approx 3.14159$$

Approximating by ignoring the last digits,

$$\tilde{\pi} = 3.141$$

Similarly,

$$\widetilde{\frac{1}{30}} = 3.333 \times 10^{-2}$$

Therefore, adding,

$$\begin{aligned} \tilde{\pi} + \widetilde{\frac{1}{30}} &= 3.141 + 0.03333 \\ &= 3.174 \end{aligned}$$

Therefore,

$$\begin{aligned} \delta &= \left| \frac{\left(\tilde{\pi} + \widetilde{\frac{1}{30}}\right) - \left(\pi + \frac{1}{30}\right)}{\pi + \frac{1}{30}} \right| \\ &= 0.0003 \end{aligned}$$

Therefore, $\delta < \varepsilon = 0.001$

Exercise 3.
Given

$$\begin{aligned} a &= 1.435234 \\ b &= 1.429111 \end{aligned}$$

Find the relative error.

Solution 3.

$$a = 1.435234$$
$$b = 1.429111$$

Therefore,

$$a - b = 0.0061234$$

Approximating by ignoring the last digits,

$$\widetilde{a} = 1.435$$
$$\widetilde{b} = 1.429$$

Therefore,

$$\widetilde{a} - \widetilde{b} = 0.006$$

Therefore,

$$\delta = \left| \frac{(a-b) - \left(\widetilde{a} - \widetilde{b}\right)}{a-b} \right|$$

Therefore,

$$\delta > 10^{-3}$$
$$\therefore \delta > \varepsilon$$

Exercise 4.
Solve

$$x^2 + 10^8 x + 1 = 0$$

Solution 4.

$$x = \frac{-10^8 \pm \sqrt{10^{16} - 4}}{2}$$

Therefore,

$$x_- \approx -10^8$$

Therefore, by Vietta Rules,

$$x_1 x_2 = \frac{c}{a}$$
$$x_1 + x_2 = -\frac{b}{a}$$

Therefore,

$$\begin{aligned} x_+ x_- &= 1 \\ \therefore x_+ &= \frac{1}{x_-} \\ &\approx -10^{-8} \end{aligned}$$

In MATLAB, this can be executed as x = **roots**([1,10^8,1])
This gives the result

$$x_+ = -7.45 \times 10^{-9}$$

Therefore, the absolute error is

$$\begin{aligned} |\tilde{x} - x| &= \left| -7.45 \times 10^{-9} - \left(-10^{-8}\right) \right| \\ &= 2.55 \times 10^{-9} \end{aligned}$$

Therefore,

$$\begin{aligned} \delta &= \left| \frac{\tilde{x} - x}{x} \right| \\ &= \left| \frac{2.55 \times 10^{-9}}{10^{-8}} \right| \\ &= 0.255 \\ &= 25\% \end{aligned}$$

The algorithm used by MATLAB is

if $b \geq 0$ **then**
  $x_1 = \frac{-b-\sqrt{b^2-4ac}}{2a}$
  $x_2 = \frac{x}{ax_1}$
else
  $x_2 = \frac{-b+\sqrt{b^2-4ac}}{2a}$
  $x_1 = \frac{c}{ax_2}$
end if

This is done to avoid subtraction of numbers close to each other, and hence avoid the possible error.