# 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

$$1.\underbrace{0\cdots0}_{52\text{ bits}}\times 2^{-1024} \approx 10^{-307}$$

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)**.**

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

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

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

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 $\varepsilon$.

**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 $\delta$ is the relative error, $\varepsilon$ 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.**

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

Therefore,

$$a - b = 0.0061234$$

Approximating by ignoring the last digits,

$$\begin{aligned} \widetilde{a} &= 1.435 \\ \widetilde{b} &= 1.429 \end{aligned}$$

Therefore,

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

Therefore,

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

Therefore,

$$\begin{aligned} \delta &> 10^{-3} \\ \therefore \delta &> \varepsilon \end{aligned}$$

**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,

$$\begin{aligned} x_1 x_2 &= \frac{c}{a} \\ x_1 + x_2 &= -\frac{b}{a} \end{aligned}$$

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**([1,10ˆ8,1])
This gives the result

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

Therefore, the absolute error is

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

Therefore,

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

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.

# 4 Series of Approximations

## 4.1 Order of Convergence

**Definition 6.** Let $\{\alpha_n\}_{n=1}^{\infty}$ be a series. $\{\alpha_n\}$ is said to converge to $\alpha$, denoted as $\alpha_n \to \alpha$, if $\forall \varepsilon > 0$, $\varepsilon \in \mathbb{R}$, $\exists n_0(\varepsilon) \in \mathbb{N}$, such that $\forall n \in \mathbb{N}$, $n > n_0(\varepsilon)$, $|\alpha_n - \alpha| < \varepsilon$.

Usually, the series $\{\alpha_n\}$ is compared to a simpler series such as $\frac{1}{n}, \frac{1}{n^\beta}, \dots$.

**Definition 7.** $\alpha_n$ is said to be "big-O" of $\beta_n$, and is said to behave like $\beta_n$, if $\exists k \in \mathbb{R}$, $k > 0$, $\exists n_0 \in \mathbb{N}$, $n_0 > 0$, such that $\forall n > n_0$,

$$|\alpha_n| \leq k|\beta_n|$$

It is denoted as

$$\alpha_n = \mathrm{O}(\beta_n)$$

**Definition 8.** $\alpha_n$ is said to be "small-O" of $\beta_n$ if

$$\lim_{n \to \infty} \frac{\alpha_n}{\beta_n} = 0$$

It is denoted as

$$\alpha_n = \mathrm{o}(\beta_n)$$

**Exercise 5.**
Find the order of convergence of

$$\alpha_n = 2n^3 + 3n^2 + 4n + 5$$

**Solution 5.**

$$\begin{aligned} \alpha_n &= 2n^3 + 3n^2 + 4n + 5 \\ &\leq (2 + 3 + 4 + 5)n^3 \\ \therefore \alpha_n &\leq 14n^3 \end{aligned}$$

Therefore, comparing to the standard form,

$$\begin{aligned} k &= 14 \\ \beta_n &= n^3 \end{aligned}$$

Therefore, as $\forall n \geq 1$, $|a_n| \leq 14|\beta_n|$,

$$\alpha_n = \mathrm{O}(\beta_n)$$

Also,

$$\begin{aligned}\lim_{n\to\infty} \frac{\alpha_n}{\beta_n} &= \lim_{n\to\infty} \frac{2n^3 + 2n^2 + 4n + 5}{n^3} \\ &= 2\end{aligned}$$

Therefore, as the limits is not zero,

$$\alpha_n \neq \mathrm{o}(\beta_n)$$

However, $\forall \delta > 0$,

$$\alpha_n = \mathrm{o}\left(n^{3+\delta}\right)$$

## 4.2 Representation of Polynomials

### 4.2.1 Power series

**Definition 9** (Power series representation of polynomials)**.**

$$P_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

This representation may lead to loss of significant digits.

**Exercise 6.**

Let $P(x)$ represent a straight line.

$$\begin{aligned}P(6000) &= \frac{1}{3} \\ P(6001) &= -\frac{2}{3}\end{aligned}$$

If only 5 decimal digits are used, show that there is a loss of significant digits, if the power series representation of the polynomial is used.

**Solution 6.**
$P(x)$ represents a straight line. Therefore,

$$P(x) = ax + b$$

Therefore,

$$6000a + b = \frac{1}{3}$$
$$6001a + b = -\frac{2}{3}$$

Therefore,

$$\begin{pmatrix} 6000 & 1 \\ 6001 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$$
$$\therefore \begin{pmatrix} a \\ b \end{pmatrix}$$
$$= \frac{1}{|A|} \begin{pmatrix} 1 & -1 \\ -6001 & 6000 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$$
$$= - \begin{pmatrix} 1 \\ -6000.3 \end{pmatrix}$$
$$= \begin{pmatrix} -1 \\ 6000.3 \end{pmatrix}$$

Therefore,

$$a = -1$$
$$b = 6000.3$$

Therefore,

$$P(x) = -x + 6000.3$$

Substituting 6000 and 6001 in this expression,

$$P(6000) = 0.3$$
$$P(6001) = 0.7$$

However, the most accurate values of $P(6000)$ and $P(6001)$, using 5 decimal digits only, should be

$$P(6000) = 0.33333$$
$$P(6001) = -0.66666$$

Therefore, there is a loss of significant digits.

### 4.2.2 Shifted Power Series

**Definition 10** (Shifted power series representation of polynomials)**.**

$$P_n(x) = a_0 + a_1(x-c) + \cdots + a_n(x-c)^n$$

This representation is a power series shifted by $c$. Hence, this representation does not lead to loss of significant digits.

**Exercise 7.**
Let $P(x)$ be a straight line.

$$P(6000) = \frac{1}{3}$$
$$P(6001) = -\frac{2}{3}$$

If only 5 decimal digits are used, show that there is no loss of significant digits, if the shifted power series representation of the polynomial is used, with $c = 6000$.

**Solution 7.**
$P(x)$ represents a straight line. Therefore,

$$P(x) = a(x-6000) + b$$

Therefore,

$$b = \frac{1}{3}$$
$$a + b = -0.66666$$
$$\therefore a = -0.99999$$

Therefore,

$$P(x) = -0.99999(x-6000) + 0.33333$$

Substituting 6000 and 6001 in this expression,

$$P(6000) = 0.33333$$
$$P(6001) = -0.66666$$

Therefore, there is no loss of significant digits, as the values of $P(6000)$ and $P(6001)$ are the most accurate values possible, using 5 decimal digits.

### 4.2.3 Newton's Form

**Definition 11** (Newton's form of representation of polynomials)**.**

$$P_n(x) = a_0 + a_1(x - c_1) + \cdots + a_n(x - c_1) \ldots (x - c_n)$$

The number of multiplications needed to calculate $P_n(x)$ is

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

The number of additions or subtractions needed to calculate $P_n(x)$ is

$$\sum_{i=1}^{n} i + n = \frac{n(n+1)}{2} + n$$

Therefore, the total number of operations needed to calculate $P_n(x)$ is $\mathrm{O}(n^2)$.

### 4.2.4 Nested Newton's Form

**Definition 12** (Nested Newton's form of representation of polynomials)**.**

$$P_n(x) = a_0 + (x - c_1)\Big(a_1 + (x - c_2)\left(a_2 + (x - c_3)\left(\ldots\right)\right)\Big)$$

The number of multiplications needed to calculate $P_n(x)$ is

$$\sum_{i=1}^{n} 1 = n$$

The number of additions or subtractions needed to calculate $P_n(x)$ is

$$\sum_{i=1}^{n} 2 = 2n$$

Therefore, the total number of operations needed to calculate $P_n(x)$ is big-O of $\mathrm{O}(n)$.

## 4.3 Properties of Polynomials

**Theorem 2.** *For a polynomial in shifted power series form,*

$$P_n(x) = P_n(c) + (x - c)q_{n-1}(x)$$

*Proof.*

$$\begin{aligned}
P_n(x) &= a_0 + a_1(x-c) + \cdots + a_n(x-c)^n \\
&= a_0 + (x-c)\left(a_1 + a_2(x-2) + \cdots + a_n(x-c)^{n-1}\right) \\
&= a_0 + (x-c)q_{n-1}(x) \\
&= P_n(c) + (x-c)q_{n-1}(x)
\end{aligned}$$

□

**Theorem 3.** *If $c$ is a root of $P_n(x)$, i.e., if*

$$P_n(c) = 0$$

*then*

$$P_n(x) = (x-c)q_{n-1}(x)$$

*If $c_1 \neq c_2$ are roots of $P_n(x)$, then*

$$P_n(x) = (x-c_1)(x-c_2)r_{n-2}(x)$$

*Similarly, if $P_n(x)$ has $n$ different roots, then*

$$P_n(x) = A(x-c_1)\dots(x-c_n)$$

*where $A \in \mathbb{R}$.*
*If $P_n(x)$ has $n+1$ different roots, then*

$$P_n(x) = A(x-c_1)\dots(x-c_n)(x-c_{n+1})$$

*where $A = 0$.*

**Theorem 4.** *If $p(x)$ and $q(x)$ are polynomials of degree at most $n$, that satisfy*

$$\begin{aligned}
p(x_i) &= f(x_i) \\
q(x_i) &= f(x_i)
\end{aligned}$$

*for $i \in \{0, \dots, n\}$, then*

$$p_n(x) \equiv q_n(x)$$

*This means that there exists a unique polynomial with degree $n$ which passes through $n+1$ points, i.e. $n+1$ points define a unique $n$ degree polynomial.*

*Proof.* Let

$$d_n(x) = p_n(x) - q_n(x)$$

Therefore, $d_n(x)$ is a polynomial of degree at most $n$, which has $n+1$ roots. Therefore,

$$d_n(x) \equiv 0$$

Therefore,

$$p_n(x) \equiv q_n(x)$$

□

## 4.4 Interpolation

**Theorem 5** (Weierstrass Approximation Theorem)**.** *Let $f(x) \in \mathrm{C}[a,b]$, i.e. it is continuous on $[a,b]$. Let $\varepsilon > 0$. Then there exists a polynomial $P(x)$ defined on $[a,b]$, such that $\forall x \in [a,b]$,*

$$|f(x) - P(x)| < \varepsilon$$