

1. Given  $A \in \mathbb{R}^{1000 \times 2}$ ,  $B \in \mathbb{R}^{2 \times 1000}$  and  $C \in \mathbb{R}^{1000 \times 1}$ , write down the number of real number multiplications and additions needed to compute the product  $ABC$  by the two procedures indicated below:

- $(AB)C$
- $A(BC)$

2. We build a computer, where the real numbers are represented using 5 digits as explained below:

S	A	B	C	E
---	---	---	---	---

where

- S is the sign bit; 0 for positive and 1 for negative
- A,B,C: First three significant digits in decimal expansion with the decimal point occurring between A and B
- E is the exponent in base 10 with a bias of 5
- All digits after the third significant digit are chopped off
- $+0$  is represented by setting  $S = 0$  and  $A = 0$  (B,C,E) can be anything
- $-0$  is represented by setting  $S = 1$  and  $A = 0$  (B,C,E) can be anything
- $+\infty$  is represented by  $S = 0$ ,  $A = B = C = E = 9$
- $-\infty$  is represented by  $S = 1$ ,  $A = B = C = E = 9$
- Not A Number is represented by setting  $S$  other than 0 and 1.

For example, the number  $\pi = 3.14159\dots$  is represented as follows. Chopping off after the third significant digit, we have  $\pi = +3.14 \times 10^0$ . Hence, the representation of  $\pi$  in our system is:

0	3	1	4	5
---	---	---	---	---

The number  $-0.001259$  is represented as follows. Chopping off after the third significant digit, we have the number as  $-1.25 \times 10^{-3}$ . Hence, the representation in our system is:

1	1	2	5	2
---	---	---	---	---

Answer the following questions:

- (a) How many non-zero floating point numbers (from now on abbreviated as FPN) can be represented by our machine (both positive and negative)?
- (b) How many FPNs are in the following intervals?
  - $(9, 10)$
  - $(10, 11)$
  - $(0, 1)$
- (c) Identify the smallest positive and largest positive FPN on this machine.
- (d) Identify the machine precision.
- (e) What is the smallest positive integer not representable exactly on this machine?
- (f) Consider the recurrence:

$$a_{n+1} = 5a_n - 4a_{n-1}$$

with  $a_1 = a_2 = 2.93$ . Note that  $a_1$ ,  $a_2$ , 5 and 4 are exactly represented on our machine. Compute  $a_n$  for  $n \in \{3, 4, \dots, 10\}$  in our machine (Work out the values by hand). Note that at each step in the recurrence  $5a_n$  and  $4a_{n-1}$  will be both chopped down to the first three significant digits before the subtraction is performed.

3. Consider the following differential equation:

$$\frac{d^2u}{dt^2} = 0 \text{ with } u(0) = 2.95 \text{ and } \frac{du}{dt}(t=0) = 0$$

- Solve the differential equation analytically.
- Using Taylor series show that

$$\frac{u(t+3\delta t) - 3u(t+\delta t) + 2u(t)}{3(\delta t)^2} = \frac{d^2u}{dt^2} + \mathcal{O}(\delta t)$$

- Discretize the differential equation by using the above finite difference for  $u_{tt}$ , i.e., setting  $u_n = u(n\delta t)$ , obtain

$$u_{n+3} = 3u_{n+1} - 2u_n$$

where  $n \geq 0$ .

- Take  $u_0 = 2.95$ ,  $u_1 = 2.95$  and  $u_2 = 2.95$  (Note that we have satisfied  $\frac{du}{dt}(t=0) = 0$  by taking the forward and central difference). Solve for  $u_n$  for all  $n$  manually (without using a computer). Does this match with the analytic solution?
- Solve the recurrence obtained above (use Octave/MATLAB) and display  $u_0, u_1, u_2, \dots, u_{69}$  till the first 16 digits after the decimal in successive lines. Does this match with the analytic solution or the solution of the discretized equation? Explain your observation.
- Rewrite the equations in matrix form, i.e.,

$$Au = b$$

- Plot the condition number of the matrix  $A$  as a function of  $n$ .
- Comment on how the condition number scales with  $n$ .
- Comment on the relationship of the condition number and accuracy of the solution  $u_n$  obtained.

4. Recall that  $\text{fl}(x \oplus y) = (x \oplus y)(1 + \delta)$ , where  $x, y \in \mathbb{R}$ ,  $\oplus \in \{+, -, \times, \div\}$ ,  $|\delta| \leq \mu$ ,  $\mu$  is the machine precision and  $\text{fl}(z)$  is the floating point representation of  $z$ . If  $a, b \in \mathbb{R}^{n \times 1}$  and  $\frac{|\text{fl}(a^T b) - a^T b|}{\|a\|_2 \|b\|_2} = \phi_\mu(n)$ , obtain an upper bound  $\phi_\mu(n)$  in terms of  $n$  and the machine precision  $\mu$ .