Name: Suhas Kamath
SR No: 06-18-01-10-51-25-1-25945
Email ID: suhaskamath@iisc.ac.in
Date: September 1, 2025

Assignment No: Assignment 2
Course Code: DS284
Course Name: Numerical Linear Algebra
Term: AUG 2025

Question 1

Part (a)

A total of 24 numbers can be described using this system.

Part (b)

Normalised Binary Scientific Notation	Usual Binary Representation	Decimal Representation
$1.000 \times 2^{00-1}$	00000	0.5000
$1.001 \times 2^{00-1}$	00100	0.5625
$1.010 \times 2^{00-1}$	01000	0.6250
$1.011 \times 2^{00-1}$	01100	0.6875
$1.100 \times 2^{00-1}$	10000	0.7500
$1.101 \times 2^{00-1}$	10100	0.8125
$1.110 \times 2^{00-1}$	11000	0.8750
$1.111 \times 2^{00-1}$	11100	0.9375
$1.000 \times 2^{01-1}$	00001	1.000
$1.001 \times 2^{01-1}$	00101	1.125
$1.010 \times 2^{01-1}$	01001	1.250
$1.011 \times 2^{01-1}$	01101	1.375
$1.100 \times 2^{01-1}$	10001	1.500
$1.101 \times 2^{01-1}$	10101	1.625
$1.110 \times 2^{01-1}$	11001	1.750
$1.111 \times 2^{01-1}$	11101	1.875
$1.000 \times 2^{10-1}$	00010	2.00
$1.001 \times 2^{10-1}$	00110	2.25
$1.010 \times 2^{10-1}$	01010	2.50
$1.011 \times 2^{10-1}$	01110	2.75
$1.100 \times 2^{10-1}$	10010	3.00
$1.101 \times 2^{10-1}$	10110	3.25
$1.110 \times 2^{10-1}$	11010	3.50
$1.111 \times 2^{10-1}$	11110	3.75

Part (c)

The minimum (smallest) number that can be represented is 0.5. The maximum (largest) number that can be represented is 3.75.

Part (d)

The absolute gaps between two consecutive (or adjacent) numbers increases with the numbers. For example, during the start of the table, we can observe that the gap is $0.5625 - 0.5 = 0.0625$. Towards the end of the table, the gap has increased to $3.75 - 3.5 = 0.25$.

Part (e)

Machine epsilon. $\epsilon_{machine}$, is defined as follows:

Consider the discrete subset $\mathbb{F}$ of the read number system $\mathbb{R}$ to be our floating point number system, then for all $x \in \mathbb{R}$, there exists $x' \in \mathbb{F}$ such that $\frac{|x-x'|}{|x|} \leq \epsilon_{machine}$.

Let us take an example number to maximise the value of $\epsilon_{machine}$. Such a value will be exactly in the middle of two adjacent numbers. Let us take the number 1.0625 for an example. Since we cannot represent 1.0625 properly in our floating point representation, we shall instead round it down and store it as 1.000. In this case, the relative error is $\frac{1.0625-1}{1} = 0.0625$. Hence, $\epsilon_{machine} = 0.0625$.

Suhas Kamath
25945

Assignment 2

Quick Work
Page No.:
Date:
M T W T F S S

2) a) We have to prove that $|\underset{\sim}{x}^T \underset{\sim}{A} \underset{\sim}{x}| \leq \|\underset{\sim}{A}\|_2$, where $\underset{\sim}{x}$ is a unit vector and $\underset{\sim}{A}$ is a symmetric matrix.

Using Cauchy-Schwarz theorem, $|u^T v| \leq \|u\|_2 \cdot \|v\|_2$, we know that

$$|\underset{\sim}{x}^T \underset{\sim}{A} \underset{\sim}{x}| \leq \|\underset{\sim}{x}^T\|_2 \cdot \|\underset{\sim}{A}\underset{\sim}{x}\|_2$$

Because $\|x\|_2 = \|x^T\|_2 = 1$, we can simplify to : $|\underset{\sim}{x}^T \underset{\sim}{A} \underset{\sim}{x}| \leq \|\underset{\sim}{A}\underset{\sim}{x}\|_2$ (1)

From the definition of vector induced by norm, we know that $\|\underset{\sim}{A}\underset{\sim}{x}\|_2 \leq \|\underset{\sim}{A}\|_2$ (2)
matrix

Combining the two inequalities (1) and (2), we get:

$$|\underset{\sim}{x}^T \underset{\sim}{A} \underset{\sim}{x}| \leq \|\underset{\sim}{A}\underset{\sim}{x}\|_2 \leq \|\underset{\sim}{A}\|_2$$

Hence, $|\underset{\sim}{x}^T \underset{\sim}{A} \underset{\sim}{x}| \leq \|\underset{\sim}{A}\|_2$ (proven)

b) Because we know that $\tilde{\lambda}$ and $\tilde{\underset{\sim}{u}}$ are the respective eigenvalue and eigenvector for $\tilde{\underset{\sim}{A}}$, we can say that:

$$\tilde{\underset{\sim}{A}} \cdot \tilde{\underset{\sim}{u}} = \tilde{\lambda}\,\tilde{\underset{\sim}{u}}$$

$$(\underset{\sim}{A} + \delta\underset{\sim}{A})(\underset{\sim}{u} + \delta\underset{\sim}{u}) = (\lambda + \delta\lambda)(\underset{\sim}{u} + \delta\underset{\sim}{u}) \quad \text{(Substitute the perturbations made)}$$

$$\underset{\sim}{A}\underset{\sim}{u} + \delta\underset{\sim}{A}\underset{\sim}{u} + \underset{\sim}{A}\delta\underset{\sim}{u} + \delta\underset{\sim}{A}\delta\underset{\sim}{u} = \lambda\underset{\sim}{u} + \delta\lambda\underset{\sim}{u} + \lambda\delta\underset{\sim}{u} + \delta\lambda\delta\underset{\sim}{u} \quad \text{(Expand)}$$

$$\delta\underset{\sim}{A}\underset{\sim}{u} + \underset{\sim}{A}\delta\underset{\sim}{u} + \delta\underset{\sim}{A}\delta\underset{\sim}{u} = \delta\lambda\underset{\sim}{u} + \lambda\delta\underset{\sim}{u} + \delta\lambda\delta\underset{\sim}{u} \quad (\text{Because } \underset{\sim}{A}\underset{\sim}{u} = \lambda\underset{\sim}{u})$$

$$\delta\underset{\sim}{A}\underset{\sim}{u} + \underset{\sim}{A}\delta\underset{\sim}{u} = \delta\lambda\underset{\sim}{u} + \lambda\delta\underset{\sim}{u} \quad (\text{Because } \delta\underset{\sim}{A}\delta\underset{\sim}{u} << \varepsilon_{machine} \text{ and } \delta\lambda\delta\underset{\sim}{u} << \varepsilon_{machine})$$

$$\underset{\sim}{u}^T\delta\underset{\sim}{A}\underset{\sim}{u} + \underset{\sim}{u}^T\underset{\sim}{A}\delta\underset{\sim}{u} = \underset{\sim}{u}^T\delta\lambda\underset{\sim}{u} + \underset{\sim}{u}^T\lambda\delta\underset{\sim}{u} \quad (\text{Pre-multiply with } \underset{\sim}{u}^T)$$

Quick Work

Page No.:

Date:

M T W T F S S

$\underset{\sim}{u}^T \delta A \underset{\sim}{u} + \lambda \underset{\sim}{u}^T \delta u = \delta\lambda \underset{\sim}{u}^T \underset{\sim}{u} + \lambda \underset{\sim}{u}^T \delta u$ (Because $\underset{\sim}{u}^T A \delta u = \delta\lambda \underset{\sim}{u}^T \delta u$, proven later)

$\underset{\sim}{u}^T \delta A \underset{\sim}{u} = \delta\lambda \underset{\sim}{u}^T \underset{\sim}{u}$ (Subtract $\lambda \underset{\sim}{u}^T \delta u$ from both sides)

$\underset{\sim}{u}^T \delta A \underset{\sim}{u} = \delta\lambda$ (Take u to be unit eigenvector.)

$|\delta\lambda| = |\underset{\sim}{u}^T \delta A \underset{\sim}{u}|$ (proven)

As proven in a), $|\underset{\sim}{u}^T \delta A \underset{\sim}{u}| \leq \|\delta A\|_2$

Because $|\delta\lambda| = |\underset{\sim}{u}^T \delta A \underset{\sim}{u}|$, $|\delta\lambda| \leq \|\delta A\|_2$

(proven)

Proof that $\underset{\sim}{u}^T A \delta u = \lambda \underset{\sim}{u}^T \delta u$

L.H.S. $= \underset{\sim}{u}^T A \delta u$

$= \underset{\sim}{u}^T A^T \delta u$ (Because A is symmetric, $A = A^T$)

$= (A \underset{\sim}{u})^T \delta u$ (Reversal law: $A^T B^T = (BA)^T$)

$= \lambda \underset{\sim}{u}^T \delta u$ ($A \underset{\sim}{u} = \lambda \underset{\sim}{u}$)

(proven)

c) Let us first find out the inputs and outputs of the problem:

Input: δA & A Output: λ

We hence know that $\hat{\kappa}^R \leq \dfrac{\dfrac{\|\delta\lambda\|}{\|\lambda\|}}{\dfrac{\|\delta A\|_2}{\|A\|_2}}$

$\leq \dfrac{|\delta\lambda|}{|\lambda|} \times \dfrac{\|A\|_2}{\|\delta A\|_2}$ (Because $\delta\lambda$ and λ are scalar, the norm can be replaced with modulus)

(Because of result of b)) $\leq \dfrac{\|A\|_2}{|\lambda|}$

Hence, the condition number of this problem is $\dfrac{\|A\|_2}{\|\lambda\|}$.

d) From c), we know that $\hat{\kappa}^R = \dfrac{\|A\|_2}{\|\lambda\|}$

$\|A_2\| = a$ (Because $\|D\|_2 = \max_{1 \leq i \leq m} |d_{ii}|$ when D is an $m \times m$ diagonal matrix)

Hence, $\hat{\kappa}^R = \dfrac{\|A\|_2}{|\lambda|} = \dfrac{a}{a} = 1$

Because the relative condition number of this problem = 1, we know this problem is well-conditioned.

e) As discussed in class, if an algorithm is backward stable, the relative forward error is bounded as follows:

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq \hat{K}^R \cdot \varepsilon_{machine}$$

Because we calculated the value of $\hat{K}^R$ to be equal to 1 in d), we can state that:

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq \varepsilon_{machine}$$

Since the relative forward error is $O(\varepsilon_{machine})$, we can state that the relative forward error is minimal. In other words, we have achieved the "ambitious" error margin.

f) To calculate the eigenvalues, let us first obtain the characteristic polynomial. This can be obtained by:

$$\det(\underset{\sim}{M} - \lambda \underset{\sim}{I}) = \det\left(\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right)$$

$$= \det\left(\begin{pmatrix} a-\lambda & 0 \\ 0 & a-\lambda \end{pmatrix}\right)$$

$$= (a-\lambda)^2$$

$$= \lambda^2 - 2a\lambda + a^2$$

Quick Work

Page No.:

Date:

M T W T F S S

The steps followed by algorithm V are as follows:

1) Obtain the coefficients of the characteristic polynomial
2) Equate the characteristic polynomial to zero and calculate its roots.

In step 1 of the algorithm:

a) Coefficient of $\lambda^2 = 1$. Let us assume that 1 can be stored in the floating point representation exactly.

b) Coefficient of $\lambda = -2a$. Let us assume that (-2) can be represented in the floating point representation exactly. We already know that a can be represented exactly (without any error). However, errors may happen during multiplication.

$$(-2) \circledast a = fl(-2) * fl(a) * (1+\varepsilon_1)$$ (Floating point arithmetic axiom)

$$= -2 * a * (1+\varepsilon_1)$$ ($fl(-2) = -2$ and $fl(a) = a$)

Let us call this term as c_1. $c_1 = -2a(1+\varepsilon_1)$

c) Constant term = a^2. Just as before, there is no error in storing a, but error may occur during multiplication:

$$a \circledast a = fl(a) * fl(a) * (1+\varepsilon_2)$$ (Floating point arithmetic axiom)

$$= a^2 * (1+\varepsilon_2)$$ ($fl(a) = a$)

Let us call this term as c_2. $c_2 = a^2(1+\varepsilon_2)$

Now, we have a slightly perturbed polynomial to calculate the roots of,

$$\tilde{\lambda}^2 + c_1\tilde{\lambda} + c_2 = 0$$

Let us use the quadratic formula to calculate the roots:

$$\tilde{\lambda} = \frac{-c_1 \pm \sqrt{c_1^2 - 4c_2}}{2}$$

Let us first calculate the error margin in the discriminant, D. From the actual characteristic polynomial, we know that $D = 0$. However, the computer calculates the discriminant using the perturbed coefficients, not of the actual coefficients.

$$D = c_1^2 - 4c_2$$

$$\tilde{D} = fl(c_1^2) \ominus fl(4c_2)$$

First, let us calculate $fl(c_1^2)$ and $fl(4c_2)$

$$fl(c_1^2) = fl(c_1) * fl(c_1) * (1+\varepsilon_3)$$

$$= (fl(c_1))^2 (1+\varepsilon_3)$$

$$= (-2a(1+\varepsilon_1))^2 (1+\varepsilon_3)$$

$$= 4a^2 (1+\varepsilon_1)^2 (1+\varepsilon_3)$$

$$= 4a^2 (1+2\varepsilon_1+\varepsilon_1^2)(1+\varepsilon_3)$$

$$= 4a^2 (1+2\varepsilon_1)(1+\varepsilon_3) \qquad (\varepsilon_1^2 << \varepsilon_{machine})$$

Quick Work

Page No.:

Date:

M T W T F S S

$$fl(c_1^2) = 4a^2(1+2\varepsilon_1+\varepsilon_3+2\varepsilon_1\varepsilon_3)$$

$$= 4a^2(1+2\varepsilon_1+\varepsilon_3) \qquad (2\varepsilon_1\varepsilon_3 << \varepsilon_{machine})$$

$$fl(4c_2) = fl(4) * fl(c_2) * (1+\varepsilon_4)$$

$$= 4 * a^2(1+\varepsilon_2)(1+\varepsilon_4) \qquad (\text{Assume } fl(4) = 4)$$

$$= 4a^2(1+\varepsilon_2+\varepsilon_4+\varepsilon_2\varepsilon_4)$$

$$\simeq 4a^2(1+\varepsilon_2+\varepsilon_4) \qquad (\varepsilon_2\varepsilon_4 << \varepsilon_{machine})$$

Hence, $\tilde{D} = fl(c_1^2) \ominus fl(4c_2)$

$$= [fl(c_1^2) - fl(4c_2)](1+\varepsilon_5)$$

$$= [4a^2(1+2\varepsilon_1+\varepsilon_3) - 4a^2(1+\varepsilon_2+\varepsilon_4)](1+\varepsilon_5)$$

$$= 4a^2(1+2\varepsilon_1+\varepsilon_3-1-\varepsilon_2-\varepsilon_4)(1+\varepsilon_5)$$

$$= 4a^2(\varepsilon_6)(1+\varepsilon_5) \qquad (\text{Assume } (2\varepsilon_1+\varepsilon_3-\varepsilon_2-\varepsilon_4) = \varepsilon_6)$$

$$= 4a^2(\varepsilon_6+\varepsilon_6\varepsilon_5)$$

$$= 4a^2\varepsilon_6 \qquad (\varepsilon_5 << \varepsilon_{much})$$

Hence, the discriminant is a small positive number, rather than 0. This causes a drastic change in the roots. If the discriminant is 0, it would result in repeated roots. In this case, because the discriminant is positive, it would result in two distinct roots

Here, let us calculate the final roots:

$$\hat{\lambda} = \frac{-c_1 \pm \sqrt{\delta}}{2} = \frac{-(-2a(1+\varepsilon_1)) \pm \sqrt{4a^2\varepsilon_6}}{2}$$

$$= \frac{2a(1+\varepsilon_1) \pm \sqrt{4a^2\varepsilon_6}}{2}$$

$$= \frac{2a(1+\varepsilon_1) \pm 2a\sqrt{\varepsilon_6}}{2}$$

$$= a(1+\varepsilon_1) \pm a\sqrt{\varepsilon_6}$$

$$\approx a \pm a\sqrt{\varepsilon_6} \quad (\text{Because } \sqrt{\varepsilon_6} \gg \varepsilon_1)$$

g) Let us now calculate forward relative error:

$$\text{Forward relative error} = \frac{|\hat{\lambda} - \lambda|}{|\lambda|} = \frac{|a \pm |a|\sqrt{\varepsilon_6} - a|}{|a|}$$

$$= \frac{a\sqrt{\varepsilon_6}}{a}$$

$$= \sqrt{\varepsilon_6}$$

Because $\sqrt{\varepsilon_6} \gg \varepsilon_{machine}$, the algorithm is not backward stable.

For an algorithm to be backward stable or stable, the forward relative error should be in order of $\varepsilon_{machine}$, i.e $O(\varepsilon_{machine})$

To further prove that the algorithm is not stable, let us substitute numbers. Assuming a IEEE 754 double precision floating point representation, which has a $\varepsilon_{machine}$ in the order of 10^{-16}.

The relative forward error = $\sqrt{10^{-16}} = 10^{-8}$, which is most definetly not $O(10^{-16})$.