Indian Institute of Science, Bangalore
Department of Computational and Data Sciences (CDS)

DS284: Numerical Linear Algebra

Assignment 2 [Posted Aug 30, 2025]

Faculty Instructor: Dr. Phani Motamarri
TAs: Naman Pesricha, Harshit Rawat, Nikhil Kodali,
Sejal Maisheri, Ayush Kumar Pushp, Deepti Sahu

Submissions required: Problems 1 and 4. **Max Points: 50**

Notations: Vectors and matrices are denoted below by bold faced lower case and upper case alphabets respectively.

Problem 1

Solution to this problem needs to be submitted by 14 SEP and will be graded.

Recall in IEEE single precision binary floating point representation, we use 32 bits to represent numbers 1 bit is for sign, 8 bits for the exponent and 23 bits for the mantissa. Using normalized binary scientific notation a floating point number in IEEE single precision can be represented as

$$(-1)^s \times (1.f)_2 \times 2^{(exponent-127)}$$

Here $s = 0$ for positive numbers and $s = 1$ for negative numbers. f represents the bits in mantissa. Note the digit 1 in $1.f$ and is explicitly shown for clarity and all binary representations are normalized to take the form $1.f$. The subscript 2 in the above $1.f$ denotes that we are representing the digits in base 2.

Now, in this exercise we will construct a dummy floating point number system where we use 5 bits of precision to represent numbers. In this simplified floating point number system, let us assume that we are representing numbers such that the exponent field admits values -1, 0 and 1 only. Imagine only positive numbers are represented in this system. Then the normalized binary scientific notation in this toy system would be

$$(1.f)_2 \times 2^{(exponent-1)}$$

Note here we use a biased representation in the exponent field instead of using a separate sign bit for exponent. The value of this bias is 1. Recall our toy floating point system admits -1, 0, 1 in the exponent field. Thus a value of -1 in the exponent field means $exponent = 0$, value of 0 in exponent field means $exponent = 1$, value of 1 in exponent field means $exponent = 2$. In this normalized binary scientific notation, 3 bits are used to store f, 2 bits are used to store the $exponent$.

In this backdrop answer the following questions for the toy floating point system we constructed above:

(a) How many numbers can this toy system describe?

(b) Create a table with 3 columns. First column should contain normalized binary scientific notation of the form

$$(1.f)_2 \times 2^{(exponent-1)}$$

of all the above numbers. (Make sure the numbers you are representing here just use 5 bits to store them). Second column should contain the usual binary representation.

Third column should contain decimal representation (base 10 representation). Arrange the numbers in increasing order in the base 10 representation.

(c) From the table above, what is the minimum real number and maximum real number you can represent using our toy floating point number system.

(d) What can you say about absolute gaps between the numbers? Are they constant or do they change with the magnitude of the number you are representing?

(e) What can you say about machine epsilon for our toy floating point system?

(Hint: Pick $\mathbf{x} \in \mathbb{R}$, there exists $\mathbf{x}' \in \mathbb{F}$ such that $\frac{|\mathbf{x}-\mathbf{x}'|}{|\mathbf{x}|} \leqslant \epsilon_{machine}$)

Problem 2

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be invertible matrix and consider the solution of the problem $\mathbf{Ax} = \mathbf{b}$ for some given non-zero $\mathbf{b} \in \mathbb{R}^n$.

(a) Derive the relative condition number of the problem of computing $\mathbf{x}$ given $\mathbf{b}$ with respect to perturbations in $\mathbf{b}$.

(b) Find the value of the tight lower bound of the relative condition number obtained in (a).

(c) Suppose that $\mathbf{x}$ solves $\mathbf{Ax} = \mathbf{b}$ for some given non-zero $\mathbf{b}$. Consider perturbations $\Delta\mathbf{A} \in \mathbb{R}^{n\times n}$ of $\mathbf{A}$ satisfying the following in some given matrix norm induced by the vector norm $\|.\|$,

$$K(\mathbf{A})\|\Delta\mathbf{A}\| < \|\mathbf{A}\|$$

where $K(\mathbf{A})$ is the condition number of $\mathbf{A}$ in the given norm. Consider also some perturbation $\Delta\mathbf{b} \in \mathbb{R}^n$ of $\mathbf{b}$ and let $\mathbf{x} + \Delta\mathbf{x}$ solve

$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$$

Prove that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \left[\frac{K(\mathbf{A})}{1 - K(\mathbf{A})\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}}\left[\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right]\right]$$

Problem 3

Recall the following from class:

(i) For all $x \in \mathbb{R}$ there exists $|\epsilon| \leqslant \epsilon_{machine}$ such that $fl(x) = x(1 + \epsilon)$, where $fl(x)$ denotes floating point representation of x.

(ii) For all $x, y \in \mathbb{F}$ there exists $|\epsilon| \leqslant \epsilon_{machine}$ such that $x \circledast y = (x * y)(1 + \epsilon)$ where $*$ denotes one of the operators $+, -, \times, \div$ and let $\circledast$ be its floating point analogue. Note $\mathbb{F}$ is a discrete subset of $\mathbb{R}$ which denote floating point representation of the real numbers.

Each of the following describes an algorithm implemented on a computer satisfying the properties (i) and (ii) described above. State with proper arguments whether the following algorithms are backward stable, stable but not backward stable, or unstable?

(a) Input data, $x \in \mathbb{R}$, computation of $2x$ as $x \oplus x$.

(b) Input data, $x \in \mathbb{R}$, computation of x^2 as $x \otimes x$.

(c) Input data, $x \in \mathbb{R} \setminus \{0\}$, computation of 1 as $x \oplus x$.

(d) Input data, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, computation of the inner product $\mathbf{x}^T\mathbf{y}$ as $(x_1 \otimes y_1) \oplus (x_2 \otimes y_2) \oplus (x_3 \otimes y_3) \oplus ...(x_m \otimes y_m)$.

(e) Input data $\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, computation of eigen-values of $\mathbf{A}$ by evaluating the roots of characteristic polynomial.

[Hint: You need to examine the stability by looking at how the eigen-values of perturbed matrix $\mathbf{A} + \delta\mathbf{A}$ can be computed by finding the roots of the corresponding characteristic polynomial]

Problem 4

Solution to this problem needs to be submitted by 14 SEP and will be graded.

Consider a symmetric matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$. Answer the following 8 questions:

(a) Show that $|\mathbf{x}^T\mathbf{A}\mathbf{x}| \leqslant \|\mathbf{A}\|_2$ for any non-zero unit vector $\mathbf{x} \in \mathbb{R}^m$.

(b) Let the vector $\mathbf{u} \in \mathbb{R}^m$ be an eigenvector of the above symmetric matrix $\mathbf{A}$ corresponding to an eigenvalue λ i.e. $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$. Further, let the matrix $\mathbf{A}$ undergo a symmetric matrix perturbation by $\delta\mathbf{A}$ such that $\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} = O(\epsilon_{mach})$. Let, $\tilde{\mathbf{u}} = \mathbf{u} + \delta\mathbf{u}$ and $\tilde{\lambda} = \lambda + \delta\lambda$ be the eigenvector-eigenvalue pair of the perturbed matrix $\tilde{\mathbf{A}} = \mathbf{A} + \delta\mathbf{A}$. Now, show that

$$|\delta\lambda| \leqslant \|\delta\mathbf{A}\|_2$$

(Hint:- Note that the perturbed matrix $\tilde{\mathbf{A}}$ is symmetric and start with the eigenvalue problem corresponding to $\tilde{\mathbf{A}}$ to first show that $|\delta\lambda| = |\mathbf{u}^T\,\delta\mathbf{A}\,\mathbf{u}|$. You may also assume that $\mathbf{A}$ is full rank and eigenvector-eigenvalue perturbations caused due to the symmetric perturbations in $\mathbf{A}$ are small and in the order of $||\delta\mathbf{A}||_2$.)

(c) Deduce the relative condition number for the problem of computing the eigenvalue λ of our symmetric matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ using the inequality derived in part (b).

(d) We now consider the problem of computing eigenvalues of the matrix $\mathbf{M} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$, where a is a non-zero real number. As you can see, the two eigenvalues of this matrix $\mathbf{M}$ are a, a. Find the relative condition number for the mathematical problem of computing the eigenvalues for the above matrix $\mathbf{M}$ using the result obtained in part(c).

(e) An Algorithm S is designed to compute the eigenvalues of the above matrix $\mathbf{M} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ on a computer using floating point arithmetic. Algorithm S is designed to be a backward stable algorithm. To this end, comment on the relative forward error incurred in computing an eigenvalue of $\mathbf{M}$ by employing this backward stable Algorithm S.

(f) Furthermore, another Algorithm U is designed to compute eigenvalues of the above matrix $\mathbf{M}$ by solving the roots of the characteristic polynomial of $\mathbf{M}$ i.e. $p_{\mathbf{M}}(z) = \det(\mathbf{M} - z\mathbf{I}) = 0$. Assume that in $\mathbf{M} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$, $a \in \mathbb{F}$, i.e. a does not have a floating point error when represented on a computer. List the steps of the Algorithm U required to compute eigenvalues of $\mathbf{M}$ on a computer. In doing so, deduce the floating-point approximation errors incurred both in evaluating the coefficients of $p_{\mathbf{M}}(z)$ and also, in the expressions employed to compute the roots of $p_{\mathbf{M}}(z)$ involving these coefficients. When doing this exercise, you may assume all the relative errors arising in floating point approximations to be the maximum possible relative error i.e. ϵ_M, the machine epsilon.

(g) Using the information in part (f), compute the forward relative error incurred in computing the eigenvalue a of $\mathbf{M}$ using the above Algorithm U on a computer and using this estimate, argue that the Algorithm U is unstable. (Hint: An Algorithm G is unstable if it is not both backward stable and stable. Also note that if the Algorithm G is backward stable or stable, then the relative forward error in the solution is $O(\kappa\epsilon_M)$ where κ is the condition number of the problem)