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

DS284: Numerical Linear Algebra

Final Exam – Aug 2024 Term

Faculty Instructor: Dr. Phani Motamarri

TAs: Srinibas, Nihar, Gourab, Rushikesh, Surya

Date and Duration: 29 Nov 2024, 2:00 PM to 5:00 PM Max Points: 100

Notations: (i) Vectors and matrices are denoted by bold faced lower case and upper case alphabets respectively. (ii) Set of all real numbers is denoted by $\mathbb{R}$ and all floating point numbers by $\mathbb{F}$ (iii) Set of all n dimensional column vectors is denoted by $\mathbb{R}^n$ and set of all $m \times n$ matrices is denoted by $\mathbb{R}^{m\times n}$ (iv) $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes inner product between the vectors.

Problem 1 [6x3.5=21 points]

Assert if the following statements are True or False/Yes or no. Give a detailed reasoning for your answer. Marks will be awarded only for your reasoning.

(a) Gourab conducted an experiment to collect m data points. Each data point is n-dimensional ($m < n$) i.e. $\mathbf{x}^{(i)} \in \mathbb{R}^n$ for all $i \in \{1, 2, \ldots, m\}$. Furthermore, the data points $\mathbf{x}^{(i)}$ are preprocessed and transformed into new vectors $\mathbf{q}^{(i)} \in \mathbb{R}^n$ such that $\|\mathbf{q}^{(i)}\|_2 = 1$, and $\langle \mathbf{q}^{(i)}, \mathbf{q}^{(j)} \rangle = 0$ for all $i \neq j$. In this way, the data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ is transformed into $\mathbf{Q} \in \mathbb{R}^{m\times n}$. Gourab claims that $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$, is this claim correct?

(b) Nihar and Rushikesh conducted an experiment together and constructed a data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ for further analysis. Further, they identified a reduced p-dimensional subspace ($p << m$) denoted as $\mathbb{V}$ to orthogonally project their matrix $\mathbf{X}$ onto $\mathbb{V}$ by constructing a projector $\mathbf{P} \in \mathbb{R}^{m\times m}$ corresponding to $\mathbb{V}$ so that projected data matrix is $\hat{\mathbf{X}} = \mathbf{P}\mathbf{X}$. Each of them uses a different basis set to compute the projector for the subspace $\mathbb{V}$. The columns of $\hat{\mathbf{Q}}_N \in \mathbb{R}^{m\times p}$ form the orthonormal basis vectors of $\mathbb{V}$ used by Nihar and the columns of $\hat{\mathbf{A}}_R \in \mathbb{R}^{m\times p}$ form the linearly independent basis vectors used by Rushikesh but are not orthonormal. Do you think the resulting projected data matrices $\hat{\mathbf{X}}_N$ and $\hat{\mathbf{X}}_R$ constructed by Nihar and Rushikesh using the projectors built with their choice of basis vectors would be identical?

(c) $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a full rank matrix ($m < n$), $\mathbf{A}\mathbf{x} = \mathbf{b}$ always has a solution for any $\mathbf{b} \in \mathbb{R}^m$

(d) Let $\{\phi_i(x)\}_{i=1\ldots m}$ denote m linearly independent basis functions (non-zero) defined over $[-1, 1]$ in an m-dimensional vector space. Srinibas constructed the following matrix $\mathbf{K}$

$$\mathbf{K} = \int_{-1}^{1} \frac{d\phi_i(x)}{dx} \frac{d\phi_j(x)}{dx} dx \;\text{ for }\; i, j = 1, 2...m, \tag{1}$$

and argues that $\mathbf{K}$ is a positive definite matrix. Do you agree with him?

(e) Consider the mathematical problem of computing the scalar $f(\mathbf{x}) = x_1 x_2$, where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. If the relative condition number of this problem is denoted by $\kappa(\mathbf{x})$ measured in $\|.\|_1$ norm, the problem becomes ill-conditioned when $|x_1| \gg |x_2|$ or $|x_2| \gg |x_1|$

(f) Let $\mathbf{A} \in \mathbb{C}^{m\times m}$ be a non-normal non-defective complex matrix with complex eigenvalues having non-zero imaginary part. Do you think "Power iteration" still converges to an eigenvector corresponding to a complex eigenvalue with the largest magnitude?

Problem 2 [4+3+4 =11 points]

Let us consider some data about a product sold in Reliance fresh. The data is related to normalised price and fiber content of various cornflakes brands available in the shop. There are 4 brands of cornflakes. Surya constructed a data matrix $\mathbf{X} \in \mathbb{R}^{4\times 2}$ with each of these feature vectors forming the two columns of the matrix $\mathbf{X}$. Upon computing the SVD of $\mathbf{X}$, he finds that the matrix $\mathbf{U}$ formed by the left singular vectors to be $\mathbf{U} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 \\ 0 & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 \\ 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}$ and the matrix $\mathbf{V}$ formed by the right singular vectors to be $\mathbf{V} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. Further the singular values are found to be $\sqrt{2}$ and 0.5

(a) Reconstruct the data matrix $\mathbf{X}$ from the singular values and singular vectors given above. Further compute the Frobenius norm of the data matrix $\mathbf{X}$. What will be the rank of this matrix $\mathbf{X}$? Justify.

(b) Construct the rank one approximation of $\mathbf{X}$ using the dominant singular vector and the corresponding singular value.

(c) Compute the matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. Verify that the given non-zero singular values of $\mathbf{X}$ are square roots of the eigenvalues of the matrix $\mathbf{C}$ by computing the eigenvalues of the matrix $\mathbf{C}$.

Problem 3 [2+3+5+8+3=21 points]

The notion of abstract vectors and linear operators is very useful in developing computational algorithms in linear algebra touching many scientific domains, including data science and quantum computing. These vectors and operators can be defined on both finite-dimensional and infinite-dimensional vector spaces. Let us restrict ourselves to m-dimensional finite-dimensional real vector space $\mathbb{V}^m$ in this exercise. Read the following points very carefully before you answer the questions below:

- Let $|v\rangle \in \mathbb{V}^m$ denote an abstract vector. The m components of this vector depend on the basis vectors you choose to represent $|v\rangle$ and these components can be conveniently represented as a column vector belonging to $\mathbb{R}^m$. To this end, let $\{|b_1\rangle, |b_2\rangle, \ldots, |b_m\rangle\}$ denote a set of m basis vectors that span $\mathbb{V}^m$ and are not necessarily orthogonal. We have the column vector $\mathbf{b}_i \in \mathbb{R}^m$ whose entries denote the components of the i^{th} basis vector $|b_i\rangle$ in the canonical(standard) basis. Hence, we can say that the set of column vectors $\{\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_m\}$ form a basis for $\mathbb{V}^m$.

- Let $\mathcal{O} : \mathbb{V}^m \to \mathbb{V}^m$ denote an abstract linear operator. The components of this operator will depend on the basis vectors you choose to represent $\mathcal{O}$ and can be conveniently represented as a matrix belonging to $\mathbb{R}^{m\times m}$.

- The matrix entries of the above linear operator $\mathcal{O}$ in the basis $\{|b_1\rangle, |b_2\rangle, \ldots, |b_m\rangle\}$ can be found using the following prescription: The i^{th} column entries of this matrix are nothing but the components of the vector $\mathcal{O}|b_i\rangle$ expressed in the basis $\{|b_1\rangle, |b_2\rangle, \ldots, |b_m\rangle\}$ spanning $\mathbb{V}^m$.

Now answer the following questions, keeping in mind that the basis set $\{\mathbf{b}_1, \mathbf{b}_2, \dots \mathbf{b}_m\}$ spanning $\mathbb{V}^m$ is not orthogonal and form the column vectors of the matrix $\mathbf{B} \in \mathbb{R}^{m \times m}$.

(a) Let the abstract vector $|v\rangle \in \mathbb{V}^m$ be expressed as a column vector $\mathbf{v} \in \mathbb{R}^m$ whose entries v_i are the components of $|v\rangle$ in the canonical(standard) basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$ spanning $\mathbb{V}^m$. Show that the expansion $\mathbf{v} = \sum_{i=1}^{m} \alpha_i \mathbf{b}_i$ is unique, i.e. argue that the scalars α_i are unique for a given $\mathbf{v}$ and the basis vectors $\mathbf{b}_i$.

(b) Let $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_m]^T$ be a $m \times 1$ vector, where α_i's are the coefficients defined in part (a). Derive a closed-form expression for $\boldsymbol{\alpha} \in \mathbb{R}^m$ in terms of the matrix $\mathbf{B}$. This exercise demonstrates that the same abstract vector $|v\rangle$ can be represented as $\mathbf{v}$ or $\boldsymbol{\alpha}$ depending on the choice of the basis.

(c) Consider a vector $\mathbf{w} \in \mathbb{R}^m$ lying in the p-dimensional subspace $\mathbf{V}^p \subset \mathbf{V}^m$ $(p < m)$. Let the linearly independent non-orthogonal column vectors $\{\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_p\}$ span this subspace $\mathbf{V}^p$ and hence one can write $\mathbf{w} = \sum_{i=1}^{p} \beta_i \mathbf{n}_i$. Derive a closed form expression for the vector $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]^T$ in terms of the vector $\mathbf{w}$ and the matrix $\mathbf{N} \in \mathbb{R}^{m \times p}$ where i^{th} column of this matrix $\mathbf{N}$ is given by $\mathbf{n}_i$.

(d) Let the linear operator $\mathcal{O} : \mathbb{V}^m \to \mathbb{V}^m$ be represented in the canonical(standard) basis by the matrix $\mathbf{O}^E \in \mathbb{R}^{m \times m}$. We now seek to find the matrix $\mathbf{O}^B \in \mathbb{R}^{m \times m}$ representing $\mathcal{O}$ in our non-orthogonal basis $\{\mathbf{b}_1, \mathbf{b}_2, \dots \mathbf{b}_m\}$ spanning $\mathbb{V}^m$. To this end, one can construct the entries of i^{th} column of the matrix $\mathbf{O}^B$ by representing the action of $\mathbf{O}^E$ on the basis vector $\mathbf{b}_i$ in the basis $\{\mathbf{b}_1, \mathbf{b}_2, \dots \mathbf{b}_m\}$. Using this information, derive a closed-form expression for the matrix $\mathbf{O}^B$ in terms of the matrices $\mathbf{O}^E$ and $\mathbf{B}$.

(e) Comment on the eigenvalues, geometric multiplicity, and algebraic multiplicity of the matrix $\mathbf{O}^B$ in relation to the matrix $\mathbf{O}^E$ examining the expression derived in part (d)?

Problem 4 [2+4+6+5+2+4+7+3+2=35 points]

Let $\mathbf{A} \in \mathbb{R}^{m \times m}$ not necessarily a symmetric matrix be a full rank matrix. Also let $\mathbf{A}_s = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$ be a symmetric positive definite matrix. You are now interested in solving the linear system of equations $\mathbf{Ax} = \mathbf{b}$ for some non-zero $\mathbf{b} \in \mathbb{R}^m$ using an iterative solver $\mathcal{S}$. A salient feature of this solver $\mathcal{S}$ is to seek an iterate $\mathbf{x}_n$ lying in a Krylov subspace $\mathbf{K}_n = \{\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}, ..., \mathbf{A}^{n-1}\mathbf{b}\}$ minimizing the norm of the residual vector corresponding $\mathbf{Ax} = \mathbf{b}$ at every n^{th} iteration. You are now going to deduce the convergence behavior of the solver $\mathcal{S}$ by answering the following questions.

(a) Let $\mathbf{K}_n = \{\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}, ..., \mathbf{A}^{n-1}\mathbf{b}\}$ be a Krylov subspace of dimension n. Let $\mathbf{x}_n \in \mathbf{K}_n$ be an iterate at n^{th} iteration of the iterative solver $\mathcal{S}$ employed to solve $\mathbf{Ax} = \mathbf{b}$. Show that the residual vector $\mathbf{r}_n = \mathbf{b} - \mathbf{Ax}_n$ can be written as $\mathbf{r}_n = p_n(\mathbf{A})\mathbf{b}$ where $p_n \in P_n$ with $P_n = \{\text{Polynomials } p \text{ of degree} \leqslant n \text{ with } p(0) = 1\}$ **(2 marks)**

(b) The solver $\mathcal{S}$ employed to solve $\mathbf{Ax} = \mathbf{b}$ seeks to find the iterate $\mathbf{x}_n \in \mathbf{K}_n$ such that $||\mathbf{r}_n||_2 = ||p_n(\mathbf{A})\mathbf{b}||_2$ is minimized. Deduce that for $n = 1$, $p_1(\mathbf{A}) = \mathbf{I} - c_0\mathbf{A}$ and show that minimal polynomial which minimizes $||p_1(\mathbf{A})\mathbf{b}||_2$ should have c_0 to be positive. **(4 marks)**

(c) Recall $\mathbf{A}_s = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$ is positive definite. (i) Show that $\mathbf{x}^T\mathbf{Ax} = \mathbf{x}^T\mathbf{A}_s\mathbf{x}$, $\forall \mathbf{x} \in \mathbb{R}^m$. (ii) If λ_{min} is the lowest eigenvalue of the matrix $\mathbf{A}_s$, show that $\mathbf{x}^T\mathbf{A}_s\mathbf{x} \geqslant \lambda_{min}(\mathbf{A}_s) > 0$, $\forall \mathbf{x} \in \mathbb{R}^m$ such that $||\mathbf{x}||_2 = 1$. (iii) If λ_{max} is the highest eigenvalue of $\mathbf{A}^T\mathbf{A}$, show that $(\mathbf{Ax})^T(\mathbf{Ax}) \leqslant \lambda_{max}(\mathbf{A}^T\mathbf{A})$, $\forall \mathbf{x} \in \mathbb{R}^m$ such that $||\mathbf{x}||_2 = 1$. **(6 marks)**

(d) Show that $||p_1(\mathbf{A})||_2^2 \leqslant 1 - 2c_0\lambda_{min}(\mathbf{A}_s) + c_0^2\lambda_{max}(\mathbf{A}^T\mathbf{A})$ for any $c_0 > 0$. (Hint:- Use the fact that $p_1(\mathbf{A}) = \mathbf{I} - c_0\mathbf{A}$ and the results derived in part(c)) **(5 marks)**

(e) For a general n^{th} iteration, let $\mathbf{x}_n^* \in \mathbf{K}_n$ minimizes $||\mathbf{r}_n||_2 = ||\mathbf{b} - \mathbf{A}\mathbf{x}_n||_2$ and let the corresponding residual vector $(\mathbf{b} - \mathbf{A}\mathbf{x}_n^*)$ be denoted as $\mathbf{r}_n^*$. Show that $\frac{||\mathbf{r}_n^*||_2}{||\mathbf{b}||_2} \leqslant \min_{p_n \in P_n} ||p_n(\mathbf{A})||_2$ where P_n is defined in (a). **(2 marks)**

(f) Let $p_n^* \in P_n$ minimize $||p_n(\mathbf{A})||_2$. Hence argue that $||p_n^*(\mathbf{A})||_2 \leqslant ||(p_1(\mathbf{A}))^n||_2$. Now show that for the matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$, $||\mathbf{A}^k||_2 \leqslant ||\mathbf{A}||_2^k$ and hence argue that $||p_1(\mathbf{A})^n||_2 \leqslant ||p_1(\mathbf{A})||_2^n$. **(4 marks)**

(g) Show that $||p_1(\mathbf{A})||_2 \leqslant \sqrt{\left[1 - \frac{\lambda_{min}^2(\mathbf{A}_s)}{\lambda_{max}(\mathbf{A}^T\mathbf{A})}\right]}$. (Hint:- Use the result from (d) for a particular choice of c_0 which minimizes $1 - 2c_0\lambda_{min}(\mathbf{A}_s) + c_0^2\lambda_{max}(\mathbf{A}^T\mathbf{A})$)
(7 marks)

(h) Use all the above results to show $\frac{||\mathbf{r}_n^*||_2}{||\mathbf{b}||_2} \leqslant \left[1 - \frac{\lambda_{min}^2(\mathbf{A}_s)}{\lambda_{max}(\mathbf{A}^T\mathbf{A})}\right]^{n/2}$. **(3 marks)**

(i) If $\mathbf{A}$ is symmetric and positive definite, show that $\frac{||\mathbf{r}_n^*||_2}{||\mathbf{b}||_2} \leqslant \left[\frac{(\kappa(\mathbf{A}))^2 - 1}{(\kappa(\mathbf{A}))^2}\right]^{n/2}$ where $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$ in 2-norm. This result establishes the fact that the convergence behaviour of the solver $\mathcal{S}$ depends on condition number of matrix $\mathbf{A}$.
(2 marks)

Problem 5 [3+9 =12 points]

Let us analyze the two different algorithms designed for summing up entries of an array. These algorithms are related to iterative summation vs pairwise summation. Consider a floating-point array of n (power of 2) elements, $\text{arr}[i] \in \mathbb{F} \subset \mathbb{R}\ \forall\ i = 1...n$.

Algorithm 1 Iterative summation

```
sum ← 0
for i = 1 to n do
    sum ← sum + arr[i]
end for
return sum
```

Algorithm 2 Pairwise summation

```
for i = 1 to log n do
    for j = 1 to n/2^i do
        arr[j] ← arr[j] + arr[j + n/2^i]
    end for
end for
return sum = arr[1]
```

The algorithms shown above are implemented on a computer conforming to Floating-Point Arithmetic (FPA) axioms, with a machine epsilon ϵ_M. Based on the algorithms given above, answer the following questions:

(a) Compare the number of floating point arithmetic operations carried out for both the algorithms.

(b) Which of the given algorithms do you think is more accurate? In other words, estimate the worst-case floating-point approximation error with respect to the ground-truth (summation done in infinite precision) in each of the above algorithms (**Hint:** Use

induction and also use the fact that the relative error in floating point approximation is bounded by ϵ_M)