1 Format/Details

Held on Zoom 5/6 from 2pm-4pm. Will be aimed to have similar length to the midterm (1 hour, 15 minutes) but you can use the full 2 hours if needed. Will be open note. All students must have their videos on and will be able to ask questions over chat and in breakout rooms.

Format/difficulty will be similar to the midterm, with a mix of short answers with explanations and problem solving. Likely will have four main questions and a fifth bonus question.

2 Concepts to Study

Probability and Randomized Algorithms (First Half of Class)

- The exam will not specifically test this part of the class, but you should be able to apply foundational techniques. E.g., compute expectations, linearity of expectation, union bound, etc.

Low-Rank Approximation and PCA

- Understand and apply important linear algebraic manipulations used. E.g.:
  - \( y^T y = \|y\|_2^2 \) and using this to split \( \|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x^T y \).
  - \( \text{tr}(AA^T) = \text{tr}(A^TA) = \|A\|_F^2 = \sum_{i=1}^{\text{rank}(A)} \sigma_i(A)^2 \).
  - For \( V \in \mathbb{R}^{d \times k} \) with orthonormal columns, \( V^TV = I \) and \( VV^T \) is a projection matrix.
  - By Pythagorean theorem \( \|X - XVV^T\|_F^2 = \|X\|_F^2 - \|XVV^T\|_F^2 \).
  - Definition of eigenvectors and values.
  - Courant-Fischer theorem and connection to low-rank approximation.

- Low-rank approximation as projection onto a \( k \)-dimensional subspace. How this projection gives a compressed representation of a data matrix \( X \).

- Dual view of low-rank approximation as finding \( k \) vectors that approximately span the rows (data points) and the columns (features). High level understanding of why a data matrix may be nearly low-rank.

- Finding the best low-rank approximation (i.e., the best orthonormal span \( V \in \mathbb{R}^{d \times k} \) of \( X \) using the eigenvectors of \( X^TX \). Do not need to have full derivation memorized, but it is worth working through. Understand high level takeaways – eigenvectors (principal components) as directions of greatest variance, measuring the quality of the optimal low-rank approximation by plotting the eigenvalues (the spectrum).

- Singular value decomposition definition.
• Connection of SVD of \( X \) to eigendecompositions of \( X^T X \) and \( XX^T \). Connection of singular values to eigenvalues of \( X^T X \) and \( XX^T \).

• Computing PCA/optimal low-rank approximation from the SVD. Connection of left and right singular vectors to the dual view of low-rank approximation as row and column approximation.

• Application of the SVD to linear regression (as seen on Problem Set 3).

• Low-rank approximation of a similarity matrix and entity embeddings (high level idea, don’t need to know details).

• Iterative methods for SVD: power method and high level ideas of analysis. When does it converge fast, when does it converge slow.

Spectral Methods for Graphs

• Adjacency matrix \( A \) and Laplacian \( (L = D - A) \) definitions.

• Motivation behind using the second smallest eigenvector of the Laplacian to find a small but balanced cut. \( \vec{x}^T L \vec{x} \) as giving the size of a cut when \( \vec{x} \in \{-1, 1\}^n \) is a cut indicator vector.

• Graph clustering for non-linearly separable data and for community detection.

• Stochastic block model definition, expected adjacency matrix, Laplacian, and eigenvectors. Why spectral clustering works for stochastic block model.

• Understand the high level idea of stochastic block model proof.

• Connection of power method to random walks on a graph.

Optimization

• Definition of gradient and connection to directional derivative.

• Ability to compute gradient for basic functions like least square regression.

• Gradient descent.

• Convex function definition and corollary of what it implies about the gradient.

• Lipschitz function definition.

• Would not need to recreate the analysis of GD for convex Lipschitz functions and do not need to memorize the convergence theorem, but should understand the main ideas. Would be valuable to work through.

• Convex set definition, definition of projection, projected gradient descent for constrained optimization and why its analysis is essentially identical to that of gradient descent.
3 Practice Questions

I recommend trying to solve some problems first *without any resources or notes* first to better replicate the setting of a timed exam. Then, if you get stuck, go back to resources.

**Linear Algebra and Low-Rank Approximation**

1. Exercises 3.6, 3.7, 3.8, 3.10, 3.11 (here $|\vec{x}|$ denotes the Euclidean norm of $\vec{x}$), 3.12, 3.13, 3.15, 3.18, 3.20, 3.21 (how does $B$ here connect to Problem 2.1 of Problem Set 3?), 3.22, 3.26, 7.16, 12.31, 12.33 *Foundations of Data Science.*

2. Linear algebra practice (some off Piazza/Pset 3):

   (a) For any vector $y$ prove that $||y||_2^2 = \langle y, y \rangle = y^T y$.

   (b) If $X = AB$, $X$'s columns are spanned by the columns of $A$ and $X$'s rows are spanned by the rows of $B$. Check that you understand why. What about when $X = ABC$ for some matrices $A, B, C$. If rank($A$) = $k$, prove that rank($X$) $\leq k$.

   (c) For $V \in \mathbb{R}^{n \times k}$ with orthonormal columns and vector $x \in \mathbb{R}^n$ when is $\|V^T x\|_2 = \|x\|_2$? Always? Sometimes? Never?

   (d) Show that for any matrix $A$ with SVD $U \Sigma V^T$,

   $$\|A\|_F^2 = \text{tr}(A^T A) = \text{tr}(AA^T) = \|U \Sigma\|_F^2 = \|V \Sigma\|_F^2 = \sum_{i=1}^n \sigma_i(A)^2,$$

   where $\sigma_i(A)^2$ is the $i^{th}$ singular value of $A$ (the $i^{th}$ diagonal entry of $\Sigma$) squared.

   (e) Prove that if $V \in \mathbb{R}^{d \times k}$ has orthonormal columns, then for any matrix $X \in \mathbb{R}^{n \times d}$,

   $$\|X - XV V^T\|_F^2 = \|X\|_F^2 - \|XVV^T\|_F^2.$$  

   Hint: Use that $\|A\|_F^2 = \text{tr}(AA^T)$ for any $A$ and that trace is linear: $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$. The above is often called the ‘pythagorean theorem’. I used it in class when deriving PCA. Why intuitively is it like the pythagorean theorem you are used to?

   (f) For any $V \in \mathbb{R}^{d \times k}$ with orthonormal columns, $V V^T$ is the projection matrix onto the subspace spanned by the columns of $V$ (V’s column span). We used this fact many times when discussing low-rank approximation. Show that $VV^T = (VV^T)(VV^T)$. Why does this property make intuitive sense if $VV^T$ is a projection?

   (g) **More challenging:** Prove that for any $V \in \mathbb{R}^{d \times k}$ with orthonormal columns $VV^T$ is actually the projection onto the subspace spanned by the columns of $V$ (V’s column span). Hint: Formally to prove this, argue that for any vector $x \in \mathbb{R}^d$: $y = VV^T x$ satisfies: $y = \arg \min_{z:z \in \text{colspan}(V)} \|x - z\|_2^2$.

3. Let $X \in \mathbb{R}^{n \times d}$ be a matrix with singular values $\sigma_1, \ldots, \sigma_d$ and SVD $X = U \Sigma V^T$. What are the eigenvalues of $X^T X + \lambda I$. What are the corresponding eigenvectors?

4. Prove that for any matrix $X \in \mathbb{R}^{n \times d}$ and unit vector $y \in \mathbb{R}^d$, $\|X y\|_2 \leq \sigma_1(X)$.  

   **Hint:** Use Courant-Fischer: that the top eigenvector if a symmetric matrix $A$ is given by $v_1 = \arg \max_{\|v\|_2 = 1} v^T A v$.

5. You have a data matrix $X \in \mathbb{R}^{n \times d}$ where each row corresponds to a student and each column corresponds to their grade on an assignment. The final column is their cumulative grade. Give an upper bound on the rank of this matrix Do you think it is well approximated by an even lower rank matrix?
6. $X \in \mathbb{R}^{n \times 6}$ contains health information on $n$ patients. For each patient, the data includes their height in inches, height in cms, weight in pounds, weight in kgs, heart rate, and blood pressure. Give an upper bound on $\min_{B: \text{rank}(B)=4} \|X - B\|_F^2$.

7. What is one reason why you would want to compute a low-rank approximation of a matrix $X \in \mathbb{R}^{n \times d}$?

8. Letting $U_k \in \mathbb{R}^{n \times k}$ have columns equal to the top $k$ left singular vectors of $X$ and $V_k \in \mathbb{R}^{d \times k}$ have columns equal to the top $k$ right singular vectors of $X$, $U_k U_k^T X = X V_k V_k^T$. Always? Sometimes? Never?

9. Consider two matrices $A = \begin{bmatrix} 1.01 & 0 \\ 0 & 1 \end{bmatrix}$ or $B = \begin{bmatrix} 1.1 & 0 \\ 0 & 1 \end{bmatrix}$.
   (a) What are their eigenvalues and eigenvectors?
   (b) On which matrix will power method converge more quickly?

10. Consider the matrices (each with 130 rows) and the singular value spectrums pictured below.

(a) Match each of the matrices to its corresponding singular value spectrum. Explain in a few sentences why you picked the matches you did.

(b) List the matrices in order of how well they would be approximated by a rank-20 approximation. Justify your answer.

**Spectral Methods for Graphs**

1. Consider a graph $G$ with Laplacian matrix $L$. Consider the problem: $x_* = \arg \min_{\|x\|=1} x^T L x$.
   (a) What is $x_*$. What value of $x_*^T L x_*$ does it achieve?
   (b) Is the above optimization problem a convex optimization problem? Is it over a convex constraint set?
2. Let $G$ be a $d$ regular graph (i.e., all vertices have $d$ neighbors).
   
   (a) What are the largest eigenvalue and and eigenvector of $G$’s adjacency matrix $A$?
   
   (b) What are the eigenvalues of $G$’s Laplacian $L$ in terms of the eigenvalues $\lambda_1, \ldots, \lambda_d$ of its adjacency matrix $A$?
   
   (c) What is the stationary distribution on $G$? I.e. with what probability will a random walk be at vertex $v_i$ as the number of steps in the walk goes to infinity?

3. Consider the stochastic block model.

   (a) Why is clustering with the second largest eigenvector of the expected adjacency matrix equivalent to clustering with the second smallest eigenvector of the expected Laplacian?
   
   (b) Are these two approaches identical when clustering using the actual rather than the expected matrices?
   
   (c) Describe a natural variant of the stochastic block model where these two algorithms would not be equivalent even on the expected matrices.

4. Consider the two graphs below:

   (a) Which of the graphs pictured below has the lowest second smallest Laplacian eigenvalue? Give a sentence or two justifying your answer.
   
   (b) Would partitioning Graph $B$ above using its minimum cut give a useful separation of the nodes into two communities? Give a sentence or two justifying your answer.

Optimization/Gradient Descent

1. Prove that the sum of two convex functions $f(x)$ and $g(x)$ (i.e., $[f + g](x)$) is also convex.

2. The difference of two convex functions $f(x)$ and $g(x)$ (i.e., $[f - g](x)$) is also convex. Always? Sometimes? Never?

3. The composition of two convex functions $f(x)$ and $g(x)$ (i.e., $[f \circ g](x)$) is also convex. Always? Sometimes? Never?
4. Prove that the intersection of two convex sets $A \cap B$ is also convex.

5. The sum of two $G$-Lipschitz functions is $2G$-Lipschitz. Always? Sometimes? Never?


7. Consider two vectors $x, y \in \mathbb{R}^d$ with $\|x\|_2 \geq \|y\|_2$ and let $\bar{x} = x \cdot \frac{\|y\|_2}{\|x\|_2}$. Show that $\|\bar{x} - y\|_2 \leq \|x - y\|_2$. Use an argument based around convex sets and projection.

8. Let $S$ be a convex set and let $f_S(\tilde{z}) = \begin{cases} 0 & \text{if } \tilde{z} \in S \\ 1 & \text{if } \tilde{z} \notin S \end{cases}$. Is $f_S$ a convex function? Either prove that it is, or give a counterexample.

9. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a $G$-Lipschitz function.
   
   (a) If $\theta^{(i+1)} = \theta^{(i)} - \eta \nabla f(\theta^{(i)})$, give an upper bound on $\theta^{(i+1)} - \theta^{(i)}$.
   
   (b) In our fixed step size gradient algorithm we set $t = \frac{R^2G^2}{\epsilon^2}$ and $\eta = \frac{R}{G\sqrt{t}}$. Under these settings, what is the worst case increase in function value from step $i$ to step $i + 1$.
   
   (c) Consider the case of projected gradient descent over a convex set $S$. So $\theta^{(i+1)} = P_S(\theta^{\text{out}})$ for $\theta^{\text{out}} = \theta^{(i)} - \eta \nabla f(\theta^{(i)})$. Show that the bound of (a) still holds.