COMPSCI 514: Final Review

1 Format/Logistics

Date: Any 2 hour window contained in 12:00am 12/3 – 11:59pm 12/4.

General Info: 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.

Studying:

- I recommend focusing on practice problems – from this review sheet, the quizzes, the homeworks, and class. For quizzes/homeworks/in class questions – try to re-solve without looking at the answer key or a solution given in the next slide. Then check to see how you did.
- For all pratice questions, try to solve (and write down) a solution first without resources and somewhat quickly, as you would on the exam. Then go back and more slowly work through the problem, see if you solution is correct, etc.
- We encourage you to post on Piazza to check answers/discuss approaches.

Instructions for the Final:

- If you have a question, **post a private message in Piazza**. We will respond as quickly as possible. Generally, we will be actively answering questions from 8am-10pm on both 12/3 and 12/4. If you do not get a response, clearly state any assumptions you made about wording/intent of the question and move forward under those assumptions.
- You must **show your work/derive any answers** as part of the solutions to receive full credit (and partial credit if you make a mistake).
- Answer the questions, in a separate document, either handwritten or typed. After the exam, scan your work and submit the pdf under the Midterm Exam assignment on Gradescope. You will have a 15 minute buffer period to make the submission.
- The exam is open notes. If you refer to any other resources, you must cite them.
- You may use, but don't need a calculator on any questions. So if you find yourself needing one, maybe use this as a hint to change direction.
- **You may not discuss with any other students.** We take this very seriously, and any cheating on the exam will result in failing the class. Please don't do this! It is much easier to catch than you might think.

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, take the expectation of a random matrix or a random dot product, 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 $\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x^T y$.
  - $\text{tr}(\mathbf{A}\mathbf{A}^T) = \text{tr}(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_F^2 = \sum_{i=1}^{\text{rank}(\mathbf{A})} \sigma_i(\mathbf{A})^2$.
  - For $\mathbf{V} \in \mathbb{R}^{d \times k}$ with orthonormal columns, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T$ is a projection matrix.
  - By Pythagorean theorem $\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$.
  - Definition of eigenvectors and values.
  - Courant-Fischer theorem and connection to low-rank approxmiation.
- Low-rank approximation as projection onto a k-dimensional subspace. How this projection gives a compressed representation of a data matrix $\mathbf{X}$.
- 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 $\mathbf{V} \in \mathbb{R}^{d \times k}$) of $\mathbf{X}$ using the eigenvectors of $\mathbf{X}^T\mathbf{X}$. 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).
- Ability to recognize when a matrix will be low-rank or close to low-rank (e.g., given in image, be able to make an educated guess about what its spectrum looks like.)
- Singular value decomposition definition.
- Connection of SVD of $\mathbf{X}$ to eigendecompositions of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$. Connection of singular values to eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$.
- 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 4).
- 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 $\mathbf{A}$ and Laplacian ($\mathbf{L} = \mathbf{D} - \mathbf{A}$) definitions.
- Interpretation of the Laplacian as measuring how smooth a vector (a function) is over nodes of the graph.
- Motivation behind using the second smallest eigenvector of the Laplacian to find a small but balanced cut. $\vec{x}^T\mathbf{L}\vec{x}$ 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.

Optimization

- Definition of gradient and connection to directional derivative.
- Ability to compute the gradient for basic functions.
- 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.
- Online optimization set up and online gradient descent (OGD).
- Regret definition. Why regret can be negative.
- Don't need to recreate OGD analysis or memorize the regret bound, but should understand the main ideas and how it compares to regular GD analysis.

3 Practice Questions

1. 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 $\mathbf{B}$ here connect to Problem 1.2 of Problem Set 4?), 3.22, 3.26, 7.16, 12.31, 12.33 *Foundations of Data Science*.

2. Linear algebra practice (some off Problem Set 3):

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

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

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

   (d) Letting $\mathbf{U}_k \in \mathbb{R}^{n \times k}$ have columns equal to the top k left singular vectors of $\mathbf{X}$ and $\mathbf{V}_k \in \mathbb{R}^{d \times k}$ have columns equal to the top k right singular vectors of $\mathbf{X}$, $\mathbf{U}_k \mathbf{U}_k^T \mathbf{X} = \mathbf{X} \mathbf{V}_k \mathbf{V}_k^T$. Always? Sometimes? Never?

   (e) Show that for any matrix $\mathbf{A}$ with SVD $\mathbf{U\Sigma V}^T$,

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

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

   (f) Prove the matrix Pythagorean theorem: that if $\mathbf{V} \in \mathbb{R}^{d \times k}$ has orthonormal columns, then for any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\|\mathbf{X} - \mathbf{XVV}^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{XVV}^T\|_F^2$.

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

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

3. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have SVD $\mathbf{X} = \mathbf{U\Sigma V}^T$ with singular values $\sigma_1(\mathbf{X}), \ldots, \sigma_d(\mathbf{X})$.

   (a) What are the eigenvalues of the matrix $(\mathbf{X}^T\mathbf{X})^2 + (\mathbf{X}^T\mathbf{X})^3$? What are its eigenvectors? How about the matrix $(\mathbf{XX}^T)^2 + (\mathbf{XX}^T)^3$?

   (b) What is the runtime required to compute $\left[(\mathbf{X}^T\mathbf{X})^2 + (\mathbf{X}^T\mathbf{X})^3\right] \vec{v}$ for any $\vec{v} \in \mathbb{R}^d$.

   (c) Name one method discussed in class which relies on efficiently applying a polynomial in $\mathbf{X}^T\mathbf{X}$ to a vector.

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

5. $\mathbf{X} \in \mathbb{R}^{500 \times 50}$ contains 500 well-clustered data points as its rows. In particular, there are ten cluster centers $\vec{y}_1, \ldots, \vec{y}_{10} \in \mathbb{R}^{50}$, such that each row $\vec{x}_i$ lies within Euclidean distance at most 1 of a center. Give an *upper bound* on $\min_{\mathbf{B}:\text{rank}(\mathbf{B})=10} \|\mathbf{X} - \mathbf{B}\|_F^2$.

6. Consider two matrices $\mathbf{A} = \begin{bmatrix} 1.01 & 0 \\ 0 & 1 \end{bmatrix}$ or $\mathbf{B} = \begin{bmatrix} 1.1 & 0 \\ 0 & 1 \end{bmatrix}$.

   (a) What are their eigenvalues and eigenvectors?

   (b) On which matrix will power method converge more quickly?

7. Let $\mathbf{X} \in \mathbb{R}^{n \times 900}$ have random entries drawn independently in $[0, 1]$. Let $\mathbf{Y} \in \mathbb{R}^{n \times 900}$ have rows corresponding to 30×30 pixel grayscale images of handwritten digits. Let $\mathbf{Z} \in \mathbb{R}^{n \times 900}$ have rows corresponding to 30×30 pixel grayscale images of handwritten letters from the English alphabet. All entries of $\mathbf{Y}$ and $\mathbf{Z}$ are in $[0, 1]$.

   (a) How do you expect $\sum_{i=11}^{900} \sigma_i(\mathbf{X})^2$, $\sum_{i=11}^{900} \sigma_i(\mathbf{Y})^2$, and $\sum_{i=11}^{900} \sigma_i(\mathbf{Z})^2$ to compare? Explain why in a few sentences.

   (b) Plot a guess at what the spectrums of these three matrices might look like. Do not worry about the scale of the y axis.

2. Spectral Methods for Graphs

1. Consider a graph G with Laplacian matrix $\mathbf{L}$. Consider the problem: $x_* = \arg\min_{x:\|x\|=1} x^T \mathbf{L} x$.

   (a) What is x_*? What value of $x_*^T \mathbf{L} x_*$ does it achieve?

   (b) Is the above optimization problem a convex optimization problem? Is it over a convex constraint set?

2. Consider the normalized adjacency matrix $\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ of a connected undirected graph G. The top eigenvalue of this matrix is $\lambda_1(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}) = 1$. Give an expression for the top eigenvector corresponding to this eigenvalue. **Hint:** It might be helpful to construct a small example and find the top eigenvector/value, then back-out a proof.

3. Let G be a d regular graph (i.e., all vertices have d neighbors).

   (a) What are the eigenvalues of G's Laplacian $\mathbf{L}$ in terms of the eigenvalues $\lambda_1(\mathbf{A}), \ldots, \lambda_d(\mathbf{A})$ of its adjacency matrix $\mathbf{A}$?

   (b) What are the largest eigenvalue and and eigenvector of G's adjacency matrix $\mathbf{A}$?

4. 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?

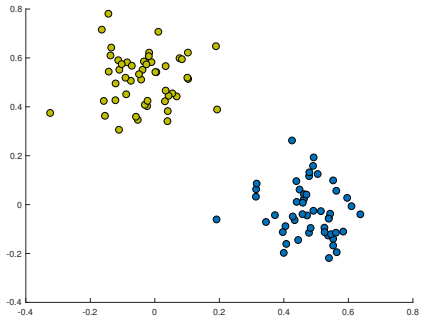
   (b) Are these two approaches identical when clustering using the actual rather than the expected matrices?

   (c) Describe a natural variant of the stochastic block model where these two algorithms would not be equivalent even on the expected matrices.

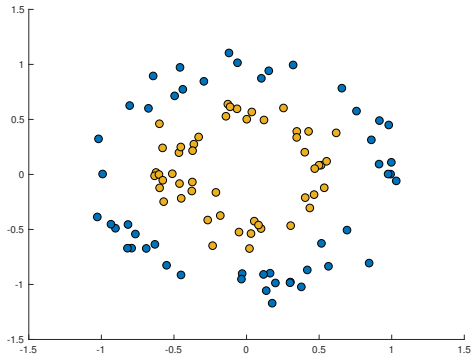
5. Describe in a sentence or two the difference between finding a minimum cut and partitioning a graph with the second smallest Laplacian eigenvector.

6. Consider the datasets below. You must run standard k-means clustering on one and spectral clustering on the other. Which would you apply each method to? Why?

A



B

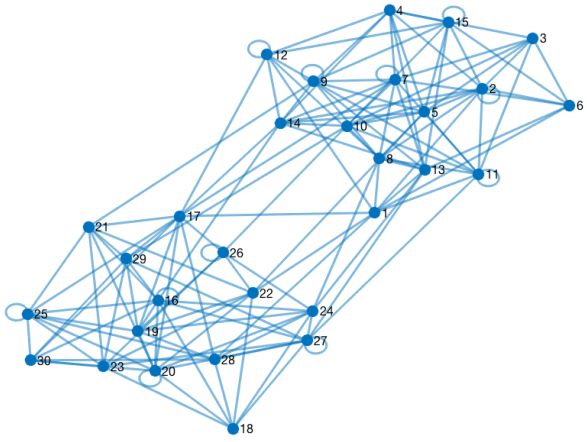


7. Consider the two stochastic block model graphs shown below.

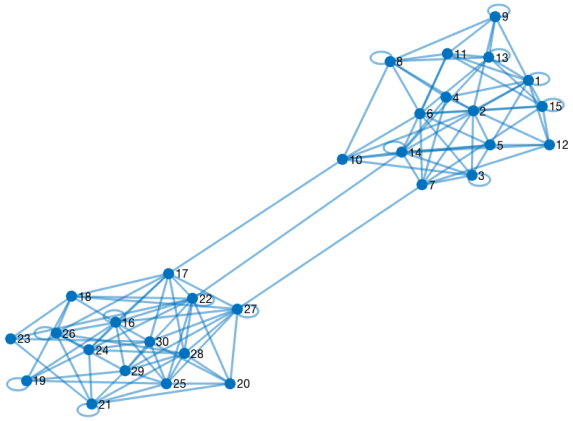
   (a) Which do you expect to have the largest spectral gap $\sigma_1(\mathbf{A}) - \sigma_2(\mathbf{A})$? Why?

   (b) Which do you expect to have the lowest second-smallest Laplacian eigenvalue?

A



B



3. Optimization/Gradient Descent

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

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

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

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

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

6. Which of the following loss functions would our analysis of gradient descent for convex Lipschitz functions apply to? For each, explain why or why not.

$$f(\theta) = \frac{1}{\theta} + \theta \qquad g(\theta) = \sin(\theta) + \theta \qquad h(\theta) = 3 \cdot |\theta - 4|.$$

7. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\vec{y} \in \mathbb{R}^n$ be fixed. Let $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2$.

   (a) What is $\vec{\nabla} f(\vec{\theta})$?

   (b) What method other than gradient descent have we learned in class to minimize $f(\vec{\theta})$? What is the optimal solution $\vec{\theta}^*$?

8. In our gradient descent analysis we showed that for large enough t, $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}^{(i)}) \leq f(\vec{\theta}^*) + \epsilon$ which implies that $f(\hat{\theta}) \leq f(\vec{\theta}^*) + \epsilon$ for the best iterate $\hat{\theta} = \arg\min_{\vec{\theta}^{(1)},\ldots,\vec{\theta}^{(t)}} f(\vec{\theta}^{(i)})$. Prove that if we instead set $\hat{\theta} = \frac{1}{t}\sum_{i=1}^{t} \vec{\theta}^{(i)}$ (i.e., set $\hat{\theta}$ to the average iterate) then we also have $f(\hat{\theta}) \leq f(\vec{\theta}^*) + \epsilon$. **Hint:** Use that f is convex.

9. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz function.

   (a) If $\theta^{(i+1)} = \theta^{(i)} - \eta \nabla f(\theta^{(i)})$, give an upper bound on $\|\theta^{(i+1)} - \theta^{(i)}\|_2$.

   (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 $\mathcal{S}$. So $\theta^{(i+1)} = P_{\mathcal{S}}(\theta^{out})$ for $\theta^{out} = \theta^{(i)} - \eta \nabla f(\theta^{(i)})$. Show that the bound of (a) still holds.

10. Consider optimizing $f_1(x) = x^2$, $f_2(x) = (x-1)^2$ and $f_3(x) = (x+1)^2$ in an online fashion. What is θ^{ol}. What is the regret for the sequence of solutions $\theta^{(1)} = 0$, $\theta^{(2)} = .5$, $\theta^{(3)} = -.5$.