

New York University Tandon School of Engineering  
Computer Science and Engineering  
CS-GY 6763: Final Exam Review.

## Logistics

- **Time/place:** 2:00pm - 3:30pm on Friday 5/9 in our usual classroom. It will be a 1.5 hour exam.
- **Material allowed:** One double sided sheet of paper with whatever you want on it is allowed.

## Concepts to Know

### Random variables and concentration.

- This was the first section of the course, but the tools we learned never stopped being important. You will not be tested on them in isolation, but don't be surprised if problems require applying techniques like computing expectations/variances or random variables, applying the basic concentration bounds, applying union bounds, etc.

### Optimization.

- Definition(s) of convex function.
- Definition of convex set.
- Gradient descent basic update rule.
- Definitions of  $G$ -Lipschitz,  $\beta$ -smooth,  $\alpha$ -strongly convex. Know the first order definition for high-dimensional functions. The second order definition you only need to know for low-dimensional functions. I.e. a twice differentiable function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is  $\beta$ -smooth,  $\alpha$ -strongly convex if for all  $x$ ,  $\alpha \leq f''(x) \leq \beta$ . I won't test on Hessians.
- How much time does it take to multiply an  $n \times d$  matrix by a  $d \times m$  matrix?
- Be able to compute gradients of basic functions from  $\mathbb{R}^d \rightarrow \mathbb{R}$ .
- Formulation of online convex optimization. Definition of online gradient descent.
- Definition of regret for online convex optimization and what bound online gradient descent achieves
- Update rule for stochastic gradient descent and to what problems it can be applied. Guarantee of SGD for sum of Lipschitz functions.
- Stationary point definition.
- What is the center of gravity method? How many iterations does it take to minimize a convex function over a convex set? Why isn't it used in practice?
- Grunbaum's Theorem, and why it is relevant to analyzing the center of gravity method.
- Be able to recognize a linear program (linear constraints, linear objective).
- Definition of a projection oracle.
- Definition of a separation oracle. Be able to describe a separation oracle for simple convex sets (e.g. for the unit ball or the  $\ell_1$  ball).
- Relax and round approach to solving combinatorial optimization problems.

## Singular Value Decomposition

- Basic properties of matrices with orthonormal columns and orthonormal rows. When does multiplying by such a matrix  $V$  preserve the norm of a vector?
- Singular value decomposition (SVD).
- Connection between SVD and eigendecomposition. Differences between SVD and eigendecomposition. For what matrices are they the same?
- Relation to norms:  $\|A\|_F^2 = \sum_{i=1}^d \sigma_i^2$ .  $\|A\|_2^2 = \max_{x: \|x\|_2=1} \|Ax\|_2^2 = \sigma_1^2$ .
- Optimal low-rank approximation.
- Power method and its analysis. What properties of the target matrix  $A$  control the convergence rate?
- Main idea behind Krylov methods.
- Polynomials of matrices and their eigenvalues.

## Spectral Graph Theory

- Matrix representations of graphs (adjacency matrix  $A$ , Laplacian matrix  $L$ , edge-vertex incidence matrix  $B$ ).
- Relationship between edge-vertex incidence matrix and Laplacian matrix.
- How to compute the value of cuts in a “linear algebraic way” using  $L$  or  $B$ .
- Intuitively what is the meaning of the smallest Laplacian eigenvector.
- Stochastic block model. Planted clique model. Random graphs in general.
- Basic strategy for recovering the partition in a stochastic block model using a spectral method.
- Matrix concentration inequalities (you don’t need to know the one stated in class, but know what it means to write down the expectation of a random matrix and bound its distance from that expectation).

## Randomized Numerical Linear Algebra

- What is a matrix sketch.
- Subspace embedding theorem.
- How subspace embeddings can be used to approximately solve least squares regression.
- $\epsilon$ -net arguments. Why was one needed for subspace embedding proof? Even if you can’t reproduce the proof yourself, know the main components.
- Size of  $\epsilon$ -net for the unit ball in  $\mathbb{R}^d$ .
- Definition of Randomized Hadamard matrices and how they are used to compute JL sketches faster.

## Compressed Sensing

- What is the compressed sensing problem.
- Restricted Isometry Property, why it allows for sparse recovery. Know the  $\ell_0$  minimization proof.
- Know how to prove that a random JL matrix satisfies RIP using the subspace embedding theorem.

## Practice Problems

- Problem 5 from <https://www.chrismusco.com/amlds2023/homeworks/hw3.pdf>.
- Problem 4 from <https://www.chrismusco.com/amlds2023/homeworks/hw4.pdf>
- Problem 4 from <https://www.chrismusco.com/amlds2022/homeworks/hw3.pdf>.

Lots of self-contained exercises on starting on Page 113 on the *Convex Optimization* book ([https://web.stanford.edu/~boyd/cvxbook/bv\\_cvxbook.pdf](https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf)). Not exactly the flavor of the types of problems we typically see in this class, but good practice.

1. Let  $f_1(x), \dots, f_n(x)$  be  $\beta$ -smooth convex functions and let  $g(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$  be their average. Show that  $g$  is  $\beta$ -smooth.
2. Let  $f: \mathbb{R} \rightarrow \mathbb{R}$  be a  $\beta$ -smooth,  $\alpha$ -strongly convex function. Let  $g(x) = f(c \cdot x)$  for some constant  $0 < c < 1$ . How does  $g$ 's smoothness and strong convexity compare to that of  $f$ ? How about  $g$ 's condition number?
3. Let  $f(x) = x^4$ . Is  $f$   $G$ -Lipschitz for finite  $G$ ? Is  $f$   $\beta$ -smooth for finite  $B$ ?

From *Foundations of Data Science* book (<https://www.cs.cornell.edu/jeh/book.pdf>). Note that this book uses  $\|x\|$  to denote  $\|x\|_2$  for a vector  $x$ .

- **Exercises:** 3.6, 3.10, 3.11, 3.12, 3.13, 3.18 (we showed this in class, make sure you can prove), 3.20, 3.21, 3.22, 3.26, 12.36
  - **Exercises:** 12.33, 12.38 (try to figure out what the eigenvectors and eigenvalues of  $A$  are by hand)
  - **Exercises:** 6.16. Understand the second equation on page 194:  $\mathbb{E}[X] = AB$ .
1. For  $V \in \mathbb{R}^{n \times d}$  with orthonormal columns and vector  $x \in \mathbb{R}^n$ , when is  $\|V^T x\|_2 = \|x\|_2$ ? Always, sometimes, or never?
  2. Let  $f_1(x), \dots, f_n(x)$  be  $\beta$ -smooth convex functions and let  $g(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$  be their average. Show that  $g$  is  $\beta$ -smooth.
  3. For any  $V \in \mathbb{R}^{n \times d}$  with orthonormal columns,  $VV^T$  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?
  4. Let  $A_k$  be the optimal  $k$ -rank approximation for  $A$  obtained via an SVD. Prove that  $\|A - A_k\|_F^2 = \|A\|_F^2 - \|A_k\|_F^2$ . Is the same statement true if we switch these to spectral norms  $\|\cdot\|_2$ ?
  5. Let  $X \in \mathbb{R}^{n \times 900}$  have random entries drawn independently as  $\{0, 1\}$ , each with probability  $1/2$ . Let  $Y \in \mathbb{R}^{n \times 900}$  have rows corresponding to  $30 \times 30$  pixel black and white images of handwritten digits. All entries of  $Y$  are in  $\{0, 1\}$ . How do you expect  $\sum_{i=1}^{900} \sigma_i(X)^2$  and  $\sum_{i=1}^{900} \sigma_i(Y)^2$  to compare?
  6. Let  $X \in \mathbb{R}^{n \times d}$  have SVD  $U\Sigma V^T$  with singular values  $\sigma_1(X), \dots, \sigma_d(X)$ . What are the eigenvalues of  $(X^T X)^q$  What are its eigenvectors. How about  $(X X^T)^q$ . What is the runtime required to apply either of these two matrices to a vector?
  7. In the stochastic block model, 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? Are these two approaches identical when clustering using the actual rather than the expected matrices?
  8. Let  $\Pi \in \mathbb{R}^{m \times n}$  be a random Johnson-Lindenstrauss matrix, e.g. with random Gaussian or  $\pm 1$  entries. Let  $m = O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right)$ . Let  $\mathbf{v}_d$  be  $\mathbf{A}$ 's last right singular vector. Recall that  $\mathbf{v}_d = \arg \min_{\mathbf{v}: \|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$ . Let  $\tilde{\mathbf{v}}_d$  be the last right singular vector of  $\Pi\mathbf{A}$ . Show that with probability  $(1 - \delta)$ ,  $\|\mathbf{A}\tilde{\mathbf{v}}_d\|_2 \leq (1 + \epsilon) \min_{\mathbf{v}: \|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$ . In other words,  $\tilde{\mathbf{v}}_d$  is a good approximate last singular vector for  $\mathbf{A}$ .