COMPSCI 514: Problem Set 3

Due: 11/17 by 11:59pm in Gradescope.

Instructions:

- You are allowed to work on this problem set in a group of up to three members.
- Each group should **submit a single solution set**: one member should upload a pdf to Gradescope, marking the other members as part of their group in Gradescope.
- You should separately submit the core competency problems from any challenge problems you choose to complete. These do not necessarily need to be submitted with the same groups.
- You may talk to members of other groups at a high level about the problems but **not work through the solutions in detail together**.
- You must show your work/derive any answers as part of the solutions to receive full credit.

Core Competency Problems

1. Linear Algebra Practice (8 points)

1. (2 points) Verify that for any two matrices $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times k}$, we have $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. **Hint**: Use the definition of matrix multiplication and the fact that, by definition, for any matrix $\mathbf{M}$, $\mathbf{M}_{ij} = (\mathbf{M}^T)_{ji}$.
2. (2 points) Use part (1) to conclude that for any set of z matrices, $\mathbf{A}_1 \in \mathbb{R}^{d_0 \times d_1}, \mathbf{A}_2 \in \mathbb{R}^{d_1 \times d_2}, \ldots, \mathbf{A}_z \in \mathbb{R}^{d_{z-1} \times d_z}$ that $(\mathbf{A}_1\mathbf{A}_2 \ldots \mathbf{A}_z)^T = \mathbf{A}_z^T\mathbf{A}_{z-1}^T \ldots \mathbf{A}_1^T$.
3. (2 points) Verify that for any $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^T\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}^T)$.
4. (2 points) Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD of $\mathbf{A}$. Prove that $\|\mathbf{A}\|_F^2 = \|\mathbf{U}\mathbf{\Sigma}\|_F^2 = \|\mathbf{V}\mathbf{\Sigma}\|_F^2 = \sum_{i=1}^{\text{rank}(\mathbf{A})} \sigma_i(\mathbf{A})^2$, where $\sigma_i(\mathbf{A})$ is the i^{th} singular value of $\mathbf{A}$. **Hint:** You might want to use part (3) here.

2. Projection Matrix Practice (8 points)

Throughout the following questions, let $\mathbf{V} \in \mathbb{R}^{d \times k}$ be an orthonormal matrix (i.e., its columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in \mathbb{R}^d$ all have unit norm and are orthogonal to each other). As discussed in class, $\mathbf{V}\mathbf{V}^T$ is the projection matrix onto the subspace $\mathcal{V} \subset \mathbb{R}^d$ spanned by $\mathbf{V}$'s columns.

1. (2 points) Prove that $\mathbf{V}\mathbf{V}^T$ projects vectors orthogonally to the subspace $\mathcal{V}$. Formally, show that for any $\mathbf{y} \in \mathbb{R}^d$ and any $\mathbf{x} \in \mathcal{V}$ we have $\langle \mathbf{x}, (\mathbf{y} - \mathbf{V}\mathbf{V}^T\mathbf{y}) \rangle = \mathbf{x}^T(\mathbf{y} - \mathbf{V}\mathbf{V}^T\mathbf{y}) = 0$. **Hint:** Start by using that we can write $\mathbf{x} = \mathbf{V}\mathbf{c}$ for some coefficient vector $\mathbf{c} \in \mathbb{R}^k$.

2. (2 points) Prove formally that $\mathbf{V}\mathbf{V}^T$ projects any point to the nearest point in the subspace $\mathcal{V}$. That is, prove that for any $\mathbf{y} \in \mathbb{R}^d$:

$$\mathbf{V}\mathbf{V}^T\mathbf{y} = \arg\min_{\mathbf{z}\in\mathcal{V}} \|\mathbf{y} - \mathbf{z}\|_2^2.$$

Hint: Use the Pythagorean theorem: For any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_2^2 = \|\mathbf{V}\mathbf{V}^T\mathbf{x}\|_2^2 + \|\mathbf{x} - \mathbf{V}\mathbf{V}^T\mathbf{x}\|_2^2$. Try a proof by contradiction. You may also want to use part (1) and perhaps draw a diagram to help your intuition.

3. (2 points) Use part (2) to prove that, when $k = d$, $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the $d \times d$ identity matrix. That is – a square matrix with orthonormal columns has orthonormal rows. Or in other words, is its own inverse.

4. (2 points) Use the Courant-Fischer Principal to prove that $\mathbf{V}\mathbf{V}^T$ has exactly k eigenvalues equal to 1 and exactly $d-k$ eigenvalues equal to 0. **Hint:** You might want to first show that $\|\mathbf{V}\mathbf{V}^T\mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_2^2$ for any vector $\mathbf{x} \in \mathbb{R}^d$.

3. Optimal Low-Rank Approximation From Scratch (10 points)

In class we used the Courant-Fischer theorem to prove that the best low-rank approximation to any matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is given by $\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T$ where $\mathbf{V}_k \in \mathbb{R}^{d\times k}$ contains the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$ (i.e., the top k singular vectors of $\mathbf{X}$). Here you will prove this from scratch, using just the basic properties of projection matrices and eigenvectors.

1. (2 points) Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be any matrix and $\mathbf{B} \in \mathbb{R}^{n\times d}$ be any rank-k matrix with SVD $\mathbf{B} = \mathbf{W}\mathbf{S}\mathbf{Z}^T$ for orthonormal $\mathbf{W} \in \mathbb{R}^{n\times k}$, $\mathbf{Z} \in \mathbb{R}^{d\times k}$, and diagonal $\mathbf{S} \in \mathbb{R}^{k\times k}$. Prove that $\|\mathbf{X} - \mathbf{B}\|_F^2 = \|\mathbf{X}\mathbf{Z}\mathbf{Z}^T - \mathbf{B}\|_F^2 + \|\mathbf{X} - \mathbf{X}\mathbf{Z}\mathbf{Z}^T\|_F^2$. **Hint:** Use the Pythagorean theorem.

2. (2 points) Use part (1) to show that if $\mathbf{B} = \arg\min_{\mathbf{M}:\text{rank}(\mathbf{M})=k} \|\mathbf{X} - \mathbf{M}\|_F^2$ then we have $\mathbf{X}\mathbf{Z}\mathbf{Z}^T = \mathbf{B}$.

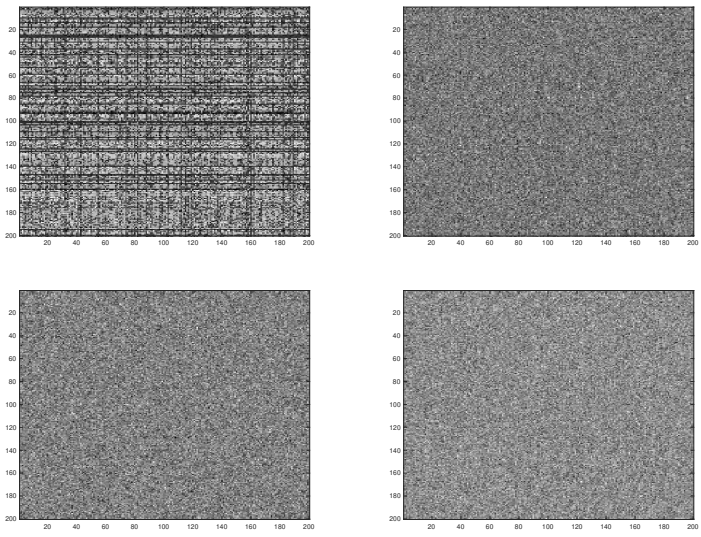
3. (2 points) Using a similar argument as above, one can show that if $\mathbf{B}$ is an optimal rank-k approximation of $\mathbf{X}$ then $\mathbf{W}\mathbf{W}^T\mathbf{X} = \mathbf{B}$. Use this and part (2) to show that: $\mathbf{X}\mathbf{Z} = \mathbf{W}\mathbf{S}$ and $\mathbf{W}^T\mathbf{X} = \mathbf{S}\mathbf{Z}^T$.

4. (2 points) Use part (3) to show that if $\mathbf{B}$ is an optimal rank-k approximation of $\mathbf{X}$ then $\mathbf{X}^T\mathbf{X}\mathbf{Z} = \mathbf{Z}\mathbf{S}^2$ and use this to argue that each column of $\mathbf{Z}$ is an eigenvector of $\mathbf{X}^T\mathbf{X}$.

5. (2 points) Complete the proof, showing that the best low-rank approximation of $\mathbf{X}$ is given by $\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T$ where $\mathbf{V}_k$ contains the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$.

4. Distinguishing Random Matrices (6 points)

Consider the four 200×200 random matrices shown below. They are represented as 200×200 images, where a pixel is lighter when an entry in the matrix is relatively large, and darker when it is relatively small. The raw matrices can be downloaded in the `four_matrices.mat` file from the assignment page.



These matrices were generated from the following four distributions:

- **A1:** Each entry of the matrix is i.i.d. $\mathcal{N}(0, 1)$.
- **A2:** The matrix is equal to $\mathbf{G}\mathbf{V}^T$ where $\mathbf{G} \in \mathbb{R}^{200\times k}$ has i.i.d. random Gaussian entries and $\mathbf{V} \in \mathbb{R}^{200\times k}$ is an orthonormal matrix for some $k < 200$.
- **A3:** The matrix is a mixture of the first two distributions. Specifically, it is equal to $0.2 \cdot \mathbf{B}_1 + 0.8 \cdot \mathbf{B}_2$ where $\mathbf{B}_1, \mathbf{B}_2$ are drawn from A_1 and A_2 respectively.
- **A4:** The matrix is generated by randomly permuting the rows and columns of the following 200×200 pixel image of the UMass Amherst campus:

1. (2 points) Let $\mathbf{M} \in \mathbb{R}^{n\times d}$ be an arbitrary matrix and let $\mathbf{P}_1 \in \mathbb{R}^{n\times n}$, $\mathbf{P}_2 \in \mathbb{R}^{d\times d}$ be permutation matrices. Prove that the singular values of $\mathbf{P}_1\mathbf{M}\mathbf{P}_2$ are equal to those of $\mathbf{M}$. I.e., if we change the order of the rows and columns of $\mathbf{M}$ this does not affect the spectrum of the matrix. **Hint:** Prove that a permutation matrix is an orthonormal matrix.
2. (2 points) Write code to compute the singular value spectrums of each of the four matrices. Show a plot of these spectrums and include a print out of your code.
3. (2 points) Use the spectrums computed above to match each matrix $\mathbf{M}_1, \ldots \mathbf{M}_4$ to the distribution in $A_1, \ldots, A_4$ that it was generated from. Explain why the spectrum is indicative of the distribution described. Identify the value of k used in distribution A_2.

Challenge Problems

C1. Johnson-Lindenstrauss for Clustering

As discussed in class, one of the most popular clustering objectives is *k-means* clustering. Given a set of n data points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^d$, the goal is to partition $[n]$ into k sets (clusters) $\mathcal{C} = \{C_1, \ldots C_k\}$ minimizing:

$$cost(\mathcal{C}, X) = \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2 \tag{1}$$

where $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{x}_i$ is the *centroid* of cluster C_j (i.e., the mean of the points in that cluster.)

1. Prove the fact discussed in class that $cost(\mathcal{C})$ can be equivalently written as:

$$cost(\mathcal{C}, X) = \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{i_1 \in C_j} \sum_{i_2 \in C_j} \|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|_2^2. \tag{2}$$

   Hint: Show that both (1) and (2) can be rewritten as $\sum_{j=1}^{k} \left[\left(\sum_{i \in C_j} \|\mathbf{x}_i\|_2^2 \right) - |C_j| \cdot \|\boldsymbol{\mu}_j\|_2^2 \right]$.

2. Suppose that $\boldsymbol{\Pi} \in \mathbb{R}^{m \times d}$ is a random projection matrix with each entry chosen independently as $\mathcal{N}(0, 1/m)$. For each $\mathbf{x}_i$ in the dataset, let $\tilde{\mathbf{x}}_i = \boldsymbol{\Pi}\mathbf{x}_i$ and let $\tilde{X} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n\}$ denote our set of sketched data points in $\mathbb{R}^m$.

   Conclude from part (1) that if $m = O\left(\frac{\log(n/\delta)}{\epsilon^2}\right)$, then with probability $\geq 1 - \delta$, for every possible clustering $\mathcal{C}$,

$$(1 - \epsilon) cost(\mathcal{C}, X) \leq cost(\mathcal{C}, \tilde{X}) \leq (1 + \epsilon) cost(C, X).$$

   Hint: To keep your calculations simple, you can use the version of the JL Lemma which says that all *squared norms* are preserved. I.e., that for all $\mathbf{x}_i, \mathbf{x}_j$ in our input set, $(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|\boldsymbol{\Pi}\mathbf{x}_i - \boldsymbol{\Pi}\mathbf{x}_j\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. This is what we actually proved in class and directly implies the version with unsquared norms by taking a square root.

3. Assuming that the guarantee of part (2) holds for some $\epsilon < 1/2$, prove that if $\tilde{\mathcal{C}}$ is an *optimal* clustering for the compressed dataset $\tilde{X}$, i.e., $cost(\tilde{\mathcal{C}}, \tilde{X}) = \min_{\mathcal{C}} cost(\mathcal{C}, \tilde{X})$, then:

$$cost(\tilde{\mathcal{C}}, X) \leq (1 + 4\epsilon) \min_{\mathcal{C}} cost(\mathcal{C}, X).$$

   That is, $\tilde{\mathcal{C}}$ is a *near optimal* clustering for the original dataset X. **Hint:** You will have to apply the guarantee of part (2) to two different clusterings in the course of the proof. You may want to recall from Problem Set 1 that for any $x \in (0, 1/2)$, $\frac{1}{1-x} \leq 1 + 2x$.

4. Download the provided MNIST dataset available in the `mnist.mat` file on the assignment page. We will work just with the test set (`testX`) here, which contains 10,000 images. Run k-means clustering (you can use the default implementation from any library you like – I used Matlab's implementation) and report on its performance: what is $cost(\mathcal{C}, X)$ for the clustering $\mathcal{C}$ that is returned? Beyond this numerical value, does the clustering look good? E.g., you may want to check if the cluster centroids look like actual handwritten digits, or how well each cluster aligns with a ground truth digit class (which are given in `testY`).

5. Now, apply Johnson-Lindenstrauss random compression and then run k-means clustering on the compressed dataset. Try this for various choices of embedding dimension m, ranging from very small ($m = 10$) to fairly large ($m = 400$). How does the performance change? In particular, report $\frac{cost(\tilde{\mathcal{C}},X)}{cost(\mathcal{C},X)}$ where $\tilde{\mathcal{C}}$ is returned by clustering the compressed dataset and $\mathcal{C}$ is returned by clustering the full dataset. Also compare the clusterings in terms of some of the other quality metrics suggested in part (4). **Hint:** You should see very good performance at least for large m, say $m = 400$. If you do not, there might be a bug in your implementation.

6. Based on the above, report the smallest choice of m that gives a reasonably good clustering of the original dataset. There is not a single correct answer here. Justify the answer that you do give.

In answering parts (4)-(6), include plots to support your answers. E.g. you may want to plot the error $\frac{cost(\tilde{\mathcal{C}},X)}{cost(\mathcal{C},X)}$ as a function of m, plot the cluster centroids (reshaped to be 28×28 images), etc. Attach all code used in solving the problem.

C2. Subspace Embedding

In this problem we will prove that if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is a random matrix with i.i.d. entries drawn from $\mathcal{N}(0, 1/m)$, then for any k-dimensional subspace $\mathcal{V}$, if we set $m = O\left(\frac{k^2 \log(k/\delta)}{\epsilon^2}\right)$, with probability at least $1 - \delta$, $\mathbf{\Pi}$ satisfies:

$$\forall \mathbf{x} \in \mathcal{V}, \quad (1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2. \tag{3}$$

That is, $\mathbf{\Pi}$ preserves the length of *all* vectors in the subspace. Note that this means also that $\mathbf{\Pi}$ preserves the distances between *all pairs* of vectors in the subspace since if $\mathbf{z}, \mathbf{y} \in \mathcal{V}$, letting $\mathbf{x} = \mathbf{z} - \mathbf{y}$, we have $\mathbf{x} \in \mathcal{V}$. This goes beyond the standard JL Lemma, which only applies to sets of n vectors for finite n, due to having a $\log n$ dependence in the embedding dimension. $\mathbf{\Pi}$ satisfying this property is known as a *subspace embedding* for $\mathcal{V}$.[1]

1. Let $\mathbf{V} \in \mathbb{R}^{d \times k}$ be an orthonormal basis for $\mathcal{V}$. Consider the $k \times k$ matrix $\mathbf{M} = \mathbf{I} - \mathbf{V}^T\mathbf{\Pi}^T\mathbf{\Pi}\mathbf{V}$. Argue that if $\|\mathbf{M}\|_2 = \max_{\mathbf{x} \in \mathbb{R}^k} \frac{|\mathbf{x}^T\mathbf{M}\mathbf{x}|}{\|\mathbf{x}\|_2^2} \leq \epsilon$ then (3) holds. **Hint:** Rewrite $\mathbf{x} = \mathbf{V}\mathbf{c}$ for some coefficient vector $\mathbf{c}$ and expand out the norms in (3) as inner products.

2. Prove that for $m = O\left(\frac{k^2 \log(k/\delta)}{\epsilon^2}\right)$, with probability at least $1 - \delta$, for every pair of columns $\mathbf{v}_i, \mathbf{v}_j$ of $\mathbf{V}$, we have $2 - \frac{\epsilon}{k} \leq \|\mathbf{\Pi}\mathbf{v}_i - \mathbf{\Pi}\mathbf{v}_j\|_2^2 \leq 2 + \frac{\epsilon}{k}$, and further for every $\mathbf{v}_i$, $1 - \frac{\epsilon}{2k} \leq \|\mathbf{\Pi}\mathbf{v}_i\|_2^2 \leq 1 + \frac{\epsilon}{2k}$.

   Hint: Apply the JL Lemma on an appropriate set of vectors and with an appropriate error parameter. To keep your calculations simple, you can use the version of the JL Lemma which says that all *squared norms* are preserved. I.e., that for all $\mathbf{x}_i, \mathbf{x}_j$ in our input set, $(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|\mathbf{\Pi}\mathbf{x}_i - \mathbf{\Pi}\mathbf{x}_j\|_2^2 \leq (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. This is what we actually proved in class and directly implies the version with unsquared norms by taking a square root.

3. Prove that if the bounds for part (2) hold, then for all pairs $\mathbf{v}_i, \mathbf{v}_j$ with $i \neq j$, we have $|\mathbf{v}_i^T\mathbf{\Pi}^T\mathbf{\Pi}\mathbf{v}_j| \leq \frac{\epsilon}{k}$. **Hint:** Expand out $\|\mathbf{\Pi}\mathbf{v}_i - \mathbf{\Pi}\mathbf{v}_j\|_2^2$ as an inner product.

[1] Via a more advanced proof technique, we can actually show that $m = O\left(\frac{k + \log(1/\delta)}{\epsilon^2}\right)$ suffices for subspace embedding. This will be covered in CS 614 next semester.

4. Use part (3) to prove that if the bounds from part (2) hold, then $\|\mathbf{M}\|_F \leq \epsilon$ and in turn that $\|\mathbf{M}\|_2 \leq \epsilon$, completing the proof via part (1). **Hint:** You may want to use Core Problem 1.4 here, to conclude that $\|\mathbf{M}\|_2 \stackrel{\text{def}}{=} \sigma_1(\mathbf{M}) \leq \|\mathbf{M}\|_F$.

C3. Faster Random Projection 🌶🌶🌶🌶

A Johnson-Lindenstrass random projection matrix $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ with independent Gaussian entries has md non-zero entries and thus takes $O(md)$ time to multiply by a single vector $\mathbf{y} \in \mathbb{R}^d$. Here we will show that a much sparser (and thus faster to multiply by) distribution of matrices can be used instead. We will prove a variant of the distributional JL lemma for this class of matrices.

We will generate a random projection matrix $\mathbf{S} \in \mathbb{R}^{m \times d}$ as follows: In each column of $\mathbf{S}$, we will independently pick a single entry uniformly at random and set it to 1 or -1, each with probability $1/2$. All other entries will be set to 0.

1. What is the runtime required to multiply $\mathbf{S}$ by a vector $\mathbf{y} \in \mathbb{R}^d$? How does this compare to the Gaussian random projections studied in class?

2. Consider a vector $\mathbf{y} \in \mathbb{R}^d$. Prove that $\mathbb{E}[\|\mathbf{S}\mathbf{y}\|_2^2] = \|\mathbf{y}\|_2^2$. **Hint:** First compute the expectation of the j^{th} entry squared, $\mathbb{E}[[\mathbf{S}\mathbf{y}](j)^2]$, by writing $[\mathbf{S}\mathbf{y}](j)$ as a sum of random variables and then squaring the sum.

3. Prove that $\text{Var}[\|\mathbf{S}\mathbf{y}\|_2^2] \leq \frac{c\|\mathbf{y}\|_2^4}{m}$ for some constant c. My analysis gives $c = 3$ but any constant is fine. **Note:** This problem is quite challenging! Letting $\mathbf{Z} = \|\mathbf{S}\mathbf{y}\|_2^2$, write $\text{Var}[\mathbf{Z}] = \mathbb{E}[\mathbf{Z}^2] - \mathbb{E}[\mathbf{Z}]^2$. Carefully expand out $\mathbb{E}[\mathbf{Z}^2]$ as a quadruple sum.

4. Conclude that for $m = O\left(\frac{1}{\epsilon^2 \delta}\right)$, with probability at least $1 - \delta$, we have

$$(1 - \epsilon)\|\mathbf{y}\|_2^2 \leq \|\mathbf{S}\mathbf{y}\|_2^2 \leq (1 + \epsilon)\|\mathbf{y}\|_2^2.$$

5. How does the above result compare to the distributional JL lemma proven for random Gaussian matrices in class?