COMPSCI 514: Problem Set 3

Due: 11/8 by 11:59pm in Gradescope.

Instructions:

- You are allowed to, and highly encouraged 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 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.

1. Projection Matrix Applied to Vectors (6 points)

1. (3 points) Let $\mathbf{V} \in \mathbb{R}^{d \times k}$ be an orthonormal matrix (i.e., its columns $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k$ all have unit norm and are orthogonal to each other). $\mathbf{V}\mathbf{V}^T \in \mathbb{R}^{d \times d}$ is the projection matrix onto the k-dimensional subspace $\mathcal{V}$ spanned by $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k$. Prove this formally. I.e., prove that, for any $\vec{y} \in \mathbb{R}^d$:
$$\mathbf{V}\mathbf{V}^T\vec{y} = \arg\min_{\vec{z} \in \mathcal{V}} \|\vec{y} - \vec{z}\|_2^2.$$
Hint: For any $\vec{z} \in \mathbb{R}^d$, $\|\vec{y}-\vec{z}\|_2^2 = \|\mathbf{V}\mathbf{V}^T(\vec{y}-\vec{z})\|_2^2 + \|\vec{y}-\vec{z}-\mathbf{V}\mathbf{V}^T(\vec{y}-\vec{z})\|_2^2$ by the Pythagorean theorem.
2. (3 points) Use part (1) 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. **Hint:** First show that $\mathcal{V} = \mathbb{R}^d$ and then consider $\arg\min_{\vec{z} \in \mathbb{R}^d} \|\vec{y} - \vec{z}\|_2^2$.

2. Projection Matrix Applied to Matrices (9 points)

1. (3 points) Verify that for any matrix $\mathbf{A}$, $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^T\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}^T)$, where $\text{tr}(\cdot)$ is the trace – the sum of diagonal elements of a matrix.
2. (3 points) For any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and orthonormal $\mathbf{V} \in \mathbb{R}^{d \times k}$ prove that $\|\mathbf{X}\|_F^2 = \|\mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 + \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$ directly by using the identity of part (1). This identity is a matrix analog to the classic Pythagorean theorem. **Hint:** Write $\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T + \mathbf{X}(\mathbf{I} - \mathbf{V}\mathbf{V}^T)$ and use part (1) to rewrite $\|\mathbf{X}\|_F^2$.

3. (3 points) Use part (2) to prove the matrix analog to the result proved in Question 1:

$$\mathbf{X}\mathbf{V}\mathbf{V}^T = \underset{\mathbf{B} \text{ with rows in } \mathcal{V}}{\arg\min} \|\mathbf{X} - \mathbf{B}\|_F^2.$$

In a sentence or two, explain how this claim relates to PCA and optimal low-rank approximation.

3. Low-Rank Approximation (9 points)

1. (3 points) You have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where each row corresponds to a student and each column corresponds to their grade on an assignment. The final column is their cumulative grade. All grades are numerical values. Give an upper bound on the rank of this matrix. Do you think it is well approximated by an even lower rank matrix? Why or why not?

2. (3 points) Consider two matrices $\mathbf{A}$ and $\mathbf{B}$ and their product: $\mathbf{X} = \mathbf{A}\mathbf{B}$. Prove that $\text{rank}(\mathbf{X}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$. **Hint:** Use that $\mathbf{X}$'s rows are spanned by the rows of $\mathbf{B}$ and that $\mathbf{X}$'s columns are spanned by the columns of $\mathbf{A}$.

3. (3 points) Consider n data points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lying in a k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$. Show that if we compress these data points using a random matrix $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ – i.e., to $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m$ with $\tilde{x}_i = \mathbf{\Pi}\vec{x}_i$, that the compressed points still lie in a k-dimensional subspace of $\mathbb{R}^m$. **Hint:** Use part (2). You can assume that $k \leq m$.

4. Clustering (8 points)

Consider n data points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$. Consider any clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ of these data points. I.e., each C_j is a set (a cluster) of points and each data point $\vec{x}_i$ is assigned to one of these sets. Let $\vec{\mu}_j = \frac{1}{|C_j|} \sum_{\vec{x} \in C_j} \vec{x}$ be the centroid of cluster C_j. The k-means clustering objective is to minimize $cost(\mathcal{C}) = \sum_{j=1}^{k} \sum_{\vec{x} \in C_j} \|\vec{x} - \vec{\mu}_j\|_2^2$.

1. (2 points) Let $\mathbf{X}_\mathcal{C}$ be the $n \times d$ matrix whose i^{th} row is equal to $\vec{\mu}_j$ if $\vec{x}_i$ is assigned to cluster C_j in $\mathcal{C}$. Verify that the k-means cost function can be written as $\|\mathbf{X} - \mathbf{X}_\mathcal{C}\|_F^2$.

2. (2 points) Use (a) to prove that for any clustering $\mathcal{C}$, $cost(\mathcal{C}) \geq \min_{\mathbf{B}:\text{rank}(\mathbf{B}) \leq k} \|\mathbf{X} - \mathbf{B}\|_F^2$. **Hint:** What is the rank of $\mathbf{X}_\mathcal{C}$?

3. (2 points) Show that we can write $\mathbf{X}_\mathcal{C} = \mathbf{V}\mathbf{V}^T\mathbf{X}$ where $\mathbf{V} \in \mathbb{R}^{n \times k}$ has orthonormal columns.

4. (2 points) Explain in a few sentences what parts (a)-(c) mean. How is k-means clustering similar to PCA? How is it different?

5. Faster Randomized Dimensionality Reduction (18 points)

The Johnson-Lindenstrauss lemma gives that, for any set of points $\vec{x}_1, \ldots, \vec{x}_n$, letting $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ be a random projection matrix with each entry chosen independently from $\mathcal{N}(0, \frac{1}{m})$ and $m = O\left(\frac{\log(n/\delta)}{\epsilon^2}\right)$, with probability $\geq 1 - \delta$, for all $i, j \in [n]$:

$$(1 - \epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2 \leq \|\mathbf{\Pi}\vec{x}_i - \mathbf{\Pi}\vec{x}_j\|_2^2 \leq (1 + \epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2.$$

Such a transformation can be slow since $\mathbf{\Pi}$ is a dense matrix. In this problem we will see how to speed it up using, surprisingly, the fast Fourier transform.

1. (2 points) Given $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ distributed as above, what is the runtime of computing the compressed vectors $\tilde{x}_i = \mathbf{\Pi}\vec{x}_i$ for all $\vec{x}_i$? Assume a single basic arithmetic operation (addition, multiplication, etc.) on real numbers takes $O(1)$ time.

2. (2 points) One very fast alternative to using a random projection matrix $\mathbf{\Pi}$ is to just perform dimensionality reduction by sampling. Let $\mathbf{W} \in \mathbb{R}^{m \times d}$ be a matrix that samples m rows of a vector uniformly at random and re-weights them. Specifically, the j^{th} row of $\mathbf{W}$ is equal to $\sqrt{d/m} \cdot \vec{e}_i$ with probability $1/d$ where $\vec{e}_i$ is the i^{th} standard basis vector for any $i \in [d]$. Show that if we let $\tilde{x}_i = \mathbf{W}\vec{x}_i$ then $\mathbb{E}[\|\tilde{x}_i - \tilde{x}_j\|_2^2] = \|\vec{x}_i - \vec{x}_j\|_2^2$. That is, this simple dimensionality reduction method preserves the distance between vectors in expectation.

3. (2 points) When might the above sampling method perform poorly in comparison to random projection even though it preserves distances in expectation?

We will now see how to make sampling work by using a preprocessing step. Let $\mathbf{F} \in \mathbb{R}^{d \times d}$ be a "*Hadamard matrix*". You only need to know three properties of $\mathbf{F}$ to solve this problem:

(i) All entries of $\mathbf{F}$ are either $-1/\sqrt{d}$ or $1/\sqrt{d}$.

(ii) The columns of $\mathbf{F}$ are orthonormal. I.e., $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ where $\mathbf{I}$ is the $d \times d$ identity matrix.

(iii) For any vector $\vec{x}$, $\mathbf{F}\vec{x}$ can be computed in $O(d \log d)$ time using the *fast Hadamard transform*, which is a variant of the fast Fourier transform.

Using these properties, solve the following problems:

4. (2 points) Let $\mathbf{S} \in \mathbb{R}^{d \times d}$ be a diagonal matrix with each entry $\mathbf{S}_{i,i}$ set independently to 1 with probability 1/2 and -1 with probability 1/2. Let $\vec{z} = \mathbf{FS}\vec{x}_i - \mathbf{FS}\vec{x}_j$. Show that $\|\vec{z}\|_2^2 = \|\vec{x}_i - \vec{x}_j\|_2^2$. **Hint:** Use that for any vector $\vec{y}$, $\vec{y}^T\vec{y} = \langle \vec{y}, \vec{y} \rangle = \|\vec{y}\|_2^2$.

5. (2 points) Show that, with probability $\geq 1 - \delta$, $\max_{k \in [d]} |\vec{z}(k)| \leq \frac{c \log(d/\delta) \|\vec{x}_i - \vec{x}_j\|_2}{\sqrt{d}}$ for some constant c. **Hint:** Use Bernstein's inequality to bound each $|\vec{z}(k)|$ and then a union bound to bound the maximum. In applying Bernstein's inequality, you might want to use that the maximum magnitude of an entry in a vector is bounded by its Euclidean norm.

6. (2 points) Together, parts (3) and (4) show that the linear transformation $\mathbf{FS}$ exactly preserves the norm of $\vec{x}_i - \vec{x}_j$, while also spreading out its entries to be nearly uniform in size: so that the maximum is bounded by $O\left(\frac{\log(d/\delta)}{\sqrt{d}}\right)$ times the norm. Intuitively, why is this useful as a preprocessing step for sampling-based dimensionality reduction?

7. (2 points) Let $\mathbf{W} \in \mathbb{R}^{m \times d}$ be a sampling matrix as in part (2). Let $\tilde{x}_i = \mathbf{WFS}\vec{x}_i$. Show that $\mathbb{E}[\|\tilde{x}_i - \tilde{x}_j\|_2^2] = \|\vec{x}_i - \vec{x}_j\|_2^2$. **Hint:** Use parts (2) and (4) here.

8. (2 points) Show that for $m = O\left(\frac{\log(d/\delta)^4 \cdot \log(n/\delta)}{\epsilon^2}\right)$, with probability $\geq 1 - \delta$, for all $i, j \in [n]$,

$$(1 - \epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2 \leq \|\tilde{x}_i - \tilde{x}_j\|_2^2 \leq (1 + \epsilon)\|\vec{x}_i - \vec{x}_j\|_2^2.$$

   Hint: Use Bernstein's inequality again, with the entry upper bound from part (5).

9. (2 points) How long does it take to compute the compressed vectors $\tilde{x}_i = \mathbf{WFS}\vec{x}_i$ for all $\vec{x}_i$, assuming that $m < d$? How does this compare to traditional random projection?