NYU CS-GY 6763: Algorithmic Machine Learning and Data Science Fall 2023

Homework 3

Name: Solution key

Problem 1

We will solve this problem by bounding $\mathbb{E}[\|\mathbf{s}-\boldsymbol{\mu}\|_2^2]$ and applying Markov's inequality. We have:

$$\|\mathbf{s}-\boldsymbol{\mu}\|_2^2 = \|\frac{1}{k}\sum_{i=1}^{k}(\mathbf{x}_i-\boldsymbol{\mu})\|_2^2 = \frac{1}{k^2}\sum_{i=1}^{k}\|\mathbf{x}_i-\boldsymbol{\mu}\|_2^2 + \sum_{i\neq j}(\mathbf{x}_i-\boldsymbol{\mu})^T(\mathbf{x}_j-\boldsymbol{\mu}).$$

We want to apply linearity of expectation to the expression above to compute $\mathbb{E}[\|\mathbf{s}-\boldsymbol{\mu}\|_2^2]$. In particular, we are given that

$$\mathbb{E}\left[\sum_{i=1}^{k}\|\mathbf{x}_i-\boldsymbol{\mu}\|_2^2\right] = k\sigma^2$$

and claim that for all $i \neq j$,

$$\mathbb{E}\left[(\mathbf{x}_i-\boldsymbol{\mu})^T(\mathbf{x}_j-\boldsymbol{\mu})\right] = 0.$$

To see why this is the case note that, since $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent,

$$\mathbb{E}\left[(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbb{E}\left[(\mathbf{x}_j-\boldsymbol{\mu})\right]\right] = \mathbb{E}\left[(\mathbf{x}_i-\boldsymbol{\mu})^T 0\right] = 0.$$

So, overall we conclude that:

$$\mathbb{E}\left[\|\mathbf{s}-\boldsymbol{\mu}\|_2^2\right] = \frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^{k}\|\mathbf{x}_i-\boldsymbol{\mu}\|_2^2\right] = \frac{1}{k}\sigma^2.$$

Applying Markov's inequality, we have that :

$$\Pr[\|\mathbf{s}-\boldsymbol{\mu}\|_2^2 \geq \frac{1}{\delta}\frac{1}{k}\sigma^2] \leq \delta.$$

Plugging in $k = \frac{1}{\delta\epsilon^2}$ and taking square roots on the both sides of the inequality inside the probability yields the bound.

Problem 2

We first prove a general result on the strong convexity and smoothness of the sum of two functions. Consider functions $f(\mathbf{x})$ and $h(\mathbf{x})$ that are α_1-strongly convex, β_1 smooth, and α_2-strongly convex,

β_2 smooth, respectively. We have that:

$$\frac{\alpha_1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \leq \frac{\beta_2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$
$$\frac{\alpha_2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq h(\mathbf{y}) - h(\mathbf{x}) - \nabla h(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \leq \frac{\beta_2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

Adding the inequalities, we have:

$$\frac{\alpha_1+\alpha_2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq f(\mathbf{y}) + h(\mathbf{y}) - (f(\mathbf{x}) + h(\mathbf{x})) - (\nabla f(\mathbf{x}) + \nabla h(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \leq \frac{\beta_1+\beta_2}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

So, we conclude that the function $g(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$ is $(\alpha_1+\alpha_2)$-strongly convex.

Now, consider the function $h(x) = \lambda\|\mathbf{x}\|_2^2$. The function has gradient $2\lambda\mathbf{x}$. So we have that:

$$\begin{aligned}
h(\mathbf{y}) - h(\mathbf{x}) - \nabla h(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) &= \lambda\|\mathbf{y}\|_2^2 - \lambda\|\mathbf{x}\|_2^2 - 2\lambda\mathbf{x}^T(\mathbf{y}-\mathbf{x}) \\
&= \lambda\|\mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 - 2\lambda\mathbf{x}^T(\mathbf{y} \\
&= \lambda\|\mathbf{x}-\mathbf{y}\|_2^2.
\end{aligned}$$

In other words, the function is 2λ-strongly convex and 2λ-smooth, so has condition number 1. We conclude that $g(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$ is $(\alpha_1 + 2\lambda)$-strongly convex, so is convex. Furthermore, g's condition number is:

$$\frac{\beta_1 + 2\lambda}{\alpha_1 + 2\lambda} \leq \frac{\beta_1 + 2\frac{\beta_1}{\alpha_1}\lambda}{\alpha_1 + 2\lambda} = \frac{\beta_1}{\alpha_1}.$$

In the first inequality, we used that $\beta_1/\alpha_1 \geq 1$.

Problem 3

(a) Call the separation oracles for A and B. If the element is in both sets, it is in $A \cap B$. Otherwise, at least one of the oracles returned a separating hyperplane; return either of these. That hyperplane separates $\mathbf{x}$ from either A or B, so clearly separates it from $A \cap B$.

(b) We can check in $O(n)$ time if the point $\mathbf{x}$ is in the ℓ_1 ball by adding up the magnitude of it's entries. If the point is outside the ℓ_1 ball, return the vector $\mathbf{s} = \text{sign}(\mathbf{x})$ which has a 1 in index i if x_i is positive and otherwise has a negative 1. We have that $\mathbf{s}^T\mathbf{x} = \|\mathbf{x}\|_1 > 1$ since $\mathbf{x}$ is not in the ℓ_1 ball. On the other hand, we always that $\mathbf{s}^T\mathbf{y} \leq \|\mathbf{y}\|_1$. Specifically, $\mathbf{s}^T\mathbf{y} = \sum_{i=1} \pm 1 \cdot |y_i| \leq \sum_{i=1} |y_i| = \|\mathbf{y}\|_1$. So, $\mathbf{s}^T\mathbf{y} \leq 1$ for all u in $\mathcal{A}$. So $(\mathbf{s}, 1)$ determines a valid separating hyperplane. And $\mathbf{s}$ can be computed in $O(n)$ time.

(c) Let $\mathbf{y} = \text{Proj}_A(\mathbf{x})$. If $\mathbf{y} = \mathbf{x}$ then we are in the set and don't need to return anything. Otherwise, return $(\mathbf{a}, \mathbf{a}^T\mathbf{x})$ as our separating hyperplane, where $\mathbf{a} = \mathbf{x} - \mathbf{y}$

Our goal is to prove that, for all $\mathbf{z} \in \mathcal{A}$,

$$\mathbf{z}^T(\mathbf{x}-\mathbf{y}) < \mathbf{x}^T(\mathbf{x}-\mathbf{y})$$

To do so, first observe that, for any $\mathbf{z} \in \mathcal{A}$, the angle between the vectors $\mathbf{x}-\mathbf{y}$ and $\mathbf{z}-\mathbf{y}$ must be greater than or equal to 90 degrees. I.e. that $(\mathbf{x}-\mathbf{y})^T(\mathbf{z}-\mathbf{y}) \leq 0$. To see that this is the case, draw a triangle between $\mathbf{x}, \mathbf{y}, \mathbf{z}$. If the angle was acute, we could draw a line from $\mathbf{x}$ that meets the line between $\mathbf{y}$ and $\mathbf{z}$ at a 90 degree angle. This line would both be in the convex set (since by definition every point on the line between $\mathbf{y}$ and $\mathbf{z}$ is) and also be closer to $\mathbf{x}$ than $\mathbf{y}$, contradicting the fact that $\mathbf{y}$ is $\mathbf{x}$'s projection on to the set.

If the angle between $\mathbf{x}-\mathbf{y}$ and $\mathbf{z}-\mathbf{y}$ is greater than 90 degrees, we have that:

$$(\mathbf{z}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) < 0 \leq (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}).$$

Adding $\mathbf{y}^T(\mathbf{x}-\mathbf{y})$ to both sides of the inequality proves our goal. I.e., we have shown that for $\mathbf{a}=\mathbf{x}-\mathbf{y}$, for all $\mathbf{z}\in\mathcal{A}$, $\mathbf{a}^T\mathbf{z} < \mathbf{a}^T\mathbf{x}$, so we have a valid separating hyperplane.

(d) We still choose $\mathbf{a}=\mathbf{x}-\mathbf{y}$ as in the previous subproblem. To prove this is a valid separating hyperplane, it suffices to show that for all $\mathbf{z}'$ in the ϵ-neighborhood of $\mathcal{A}$, that $\mathbf{a}^T\mathbf{z}' < \mathbf{a}^T\mathbf{x}$ when $\mathbf{x}$ is not in the ϵ-neighborhood of $\mathcal{A}$. To prove this, first note that $\mathbf{z}'$ can always be written as the sum $\mathbf{z}'=\mathbf{z}+\mathbf{w}$ where $\mathbf{z}$ is a point in $\mathcal{A}$ and $\|\mathbf{w}\|_2\leq\epsilon$.

By the same analysis as above, we have that $(\mathbf{x}-\mathbf{y})^T\mathbf{z} < \mathbf{x}^T(\mathbf{x}-\mathbf{y}) - \|\mathbf{x}-\mathbf{y}\|_2^2$. We then bound:

$$\begin{aligned}
(\mathbf{x}-\mathbf{y})^T\mathbf{z}' &= (\mathbf{x}-\mathbf{y})^T\mathbf{z} + (\mathbf{x}-\mathbf{y})^T\mathbf{w} && (1)\\
&\leq \mathbf{x}^T(\mathbf{x}-\mathbf{y}) - \|\mathbf{x}-\mathbf{y}\|_2^2 + (\mathbf{x}-\mathbf{y})^T\mathbf{w} && (2)\\
&\leq \mathbf{x}^T(\mathbf{x}-\mathbf{y}) - \|\mathbf{x}-\mathbf{y}\|_2^2 + \epsilon\|\mathbf{x}-\mathbf{y}\|_2. && (3)
\end{aligned}$$

The last inequality follows from Cauchy-Schwarz. Finally, since $\mathbf{x}$ is not in the ϵ neighborhood, $\|\mathbf{x}-\mathbf{y}\|_2 > \epsilon$, so $\|\mathbf{x}-\mathbf{y}\|_2^2 > \epsilon\|\mathbf{x}-\mathbf{y}\|_2$ and thus we conclude that :

$$(\mathbf{x}-\mathbf{y})^T\mathbf{z}' < \mathbf{x}^T(\mathbf{x}-\mathbf{y}) - 0,$$

which completes the argument.

Problem 4

We start with basically the same analysis from class, where we showed using convexity that:

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)}-\mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)}-\mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta\|\nabla f(\mathbf{x}^{(i)})\|_2^2}{2}.$$

See e.g. Slide 50 on Lecture 6. They don't need to reprove this from scratch.

Now, plug in $\eta = \frac{f(\mathbf{x}^{(i)})-f(\mathbf{x}^*)}{\|f(\mathbf{x}^{(i)})\|_2}$. We get:

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)}-\mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)}-\mathbf{x}^*\|_2^2}{2(f(\mathbf{x}^{(i)})-f(\mathbf{x}^*))} \cdot \|\nabla f(\mathbf{x}^{(i)})\|_2^2 + \frac{f(\mathbf{x}^{(i)})-f(\mathbf{x}^*)}{2}.$$

Using that $\|\nabla f(\mathbf{x}^{(i)})\|_2^2 \leq G^2$, we have:

$$\frac{f(\mathbf{x}^{(i)})-f(\mathbf{x}^*)}{2} \leq \frac{\|\mathbf{x}^{(i)}-\mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)}-\mathbf{x}^*\|_2^2}{2(f(\mathbf{x}^{(i)})-f(\mathbf{x}^*))} \cdot G^2.$$

And multiplying through by $2(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)$, we have:

$$(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*))^2 \leq \left(\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2\right) \cdot G^2$$

Now we use a telescoping sum argument as before to get that:

$$\sum_{i=0}^{T-1}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*))^2 \leq \left(\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2\right) \cdot G^2 \leq \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 \cdot G^2 \leq R^2 \cdot G^2.$$

I.e.

$$\frac{1}{T}\sum_{i=0}^{T}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*))^2 \leq \frac{R^2G^2}{T}.$$

This implies that the average squared error is less than or equal to R^2G^2/T, so there must be some i such that $(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*))^2 \leq R^2G^2/T$. We conclude that:

$$f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq RG/\sqrt{T}.$$

Setting $T = R^2G^2/\epsilon^2$ yields the desired result.

Problem 5

1. To solve this problem, we show that $\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 = \frac{1}{2n}\sum_{i,j}\mathbf{D}_{i,j}$.

$$\sum_{i,j}\mathbf{D}_{i,j} = \sum_i\sum_j \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 - 2\mathbf{x}_i^T\mathbf{x}_j = \sum_i\sum_j \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 - 2\mathbf{x}_i^T\sum_j \mathbf{x}_j.$$

By our assumption that our points are centered around the origin, $\sum_j \mathbf{x}_j = \mathbf{0}$, so we conclude

$$\sum_{i,j}\mathbf{D}_{i,j} = \sum_{i,j}\|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 = 2n\sum_i \|\mathbf{x}_i\|_2^2.$$

2. Similarly, using the same assumption, we can have that for any i,

$$\sum_j \mathbf{D}_{i,j} = \sum_j \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 - 2\mathbf{x}_i^T\mathbf{x}_j = n\|\mathbf{x}_i\|_2^2 + \sum_j \|\mathbf{x}_j\|_2^2.$$

So we can compute $\|\mathbf{x}_i\|_2^2$ via the formula $\|\mathbf{x}_i\|_2^2 = \frac{1}{n}\left(\sum_j \mathbf{D}_{i,j} - \sum_j \|\mathbf{x}_j\|_2^2\right)$. The second term in parenthesis we can compute using Part 1.

3. Using the norms calculated from part (b), we can form the a matrix $\mathbf{N}$ where;

$$\mathbf{N}_{i,j} = \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2.$$

Let $\mathbf{M} = (\mathbf{N} - \mathbf{D})/2$. Notice that $\mathbf{M}$'s entries satisfy $\mathbf{M}_{i,j} = \mathbf{x}_i^T\mathbf{x}_j$. $\mathbf{M}$'s i^{th} diagonal entry equals $\|\mathbf{x}_i\|_2^2$. So, it suffices to return any $\mathbf{W} \in \mathbb{R}^{n\times d}$ such that $\mathbf{W}\mathbf{W}^T = \mathbf{M}$. $\mathbf{W}$'s rows will have the same pairwise distances as the points in $\mathbf{X}$, and thus solve the problem.

To find such a $\mathbf{W}$, we can factor $\mathbf{M}$ using the SVD. Since it is symmetric, we will get a factorization of the form $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$. Only d of $\mathbf{\Sigma}$'s diagonal entries will be nonzero since $\mathbf{M}$ is rank d (since $\mathbf{M} = \mathbf{X}\mathbf{X}^T$). So we can we can truncate this factorization to get $\mathbf{V}_d\mathbf{\Sigma}_d\mathbf{V}_d^T = \mathbf{M}$ where $\mathbf{V}_d \in \mathbb{R}^{n\times d}$ $\mathbf{\Sigma}_d \in \mathbb{R}^{d\times d}$. Then we return $\mathbf{W} = \mathbf{V}_d\sqrt{\mathbf{\Sigma}_d}$.