COMPSCI 690RA: Problem Set 3

Due: 4/15 by 8pm 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. Tighter Bounds for Trace Estimation (4 points)

Consider any matrix $A \in \mathbb{R}^{n \times n}$. Use the Hanson-Wright inequality to show that if $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \{-1,1\}^n$ are chosen to have independent and uniformly distributed ± 1 entries, then for $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, $\bar{\mathbf{T}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^T A \mathbf{x}_i$ satisfies,

$$\Pr\left[|\bar{\mathbf{T}} - \operatorname{tr}(A)| > \epsilon \|A\|_F\right] \leq \delta.$$

How does this compare to the bound proven in class using Chebyshev's inequality?

2. Matrix Concentration from Scratch (8 points)

Consider a random symmetric matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ where $\mathbf{M}_{ij} = \mathbf{M}_{ji}$ is set independently to 1 with probability 1/2 and -1 with probability 1/2. Let $\|\mathbf{M}\|_2 = \max_{x:\|x\|=1} \|\mathbf{M}x\|_2$ be the spectral norm of $\mathbf{M}$. Recall that $\|\mathbf{M}\|_2$ is equal to the largest singular value of $\mathbf{M}$, which equals the largest magnitude of one of its eigenvalues.

1. (2 points) Give upper and lower bounds on $\|\mathbf{M}\|_2$ that hold deterministically – i.e., for any random choice of the entries of $\mathbf{M}$. **Hint:** You'll probably want to use $\|\mathbf{M}\|_F$, and its relation to the singular values to derive your bounds.

2. (2 points) Observe that you can also write $\|\mathbf{M}\|_2 = \max_{x:\|x\|=1} |x^T \mathbf{M} x|$. Show that for any $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, with probability $\geq 1 - \delta$, $|x^T \mathbf{M} x| = c\sqrt{\log(1/\delta)}$ for some constant c.

   Hint: Use Hoeffding's inequality, which is a useful variant on the Bernstein inequality. For independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and scalars $a_1, \ldots, a_n, b_1, \ldots, b_n$ with $\mathbf{X}_i \in [a_i, b_i]$, $\Pr\left[|\sum_{i=1}^n \mathbf{X}_i - \mathbb{E}[\sum_{i=1}^n \mathbf{X}_i]| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$.

3. (4 points) Prove that with probability $1 - \frac{1}{n^{c_1}}$, $\|\mathbf{M}\|_2 \leq c_2\sqrt{n \log n}$ for some fixed constants c_1, c_2. **Hint:** Use an ϵ-net and part (1).

3. Randomized Preconditioning (12 points)

One way that subspace embeddings are often used in practice are within *preconditioned iterative methods* for linear regression. Here we will see how to analyze one such method. Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, the goal is to find an approximate minimizer $x \in \mathbb{R}^d$ of the least squares loss function $\|Ax - b\|_2^2$.

1. (2 points) Assume that $\mathbf{S} \in \mathbb{R}^{m \times n}$ is an 1/4-subspace embedding for $A \in \mathbb{R}^{n \times d}$. I.e., for all $x \in \mathbb{R}^d$, $\frac{3}{4}\|Ax\|_2 \leq \|\mathbf{S}Ax\|_2 \leq \frac{5}{4}\|Ax\|_2$. Prove that all eigenvalues of $(A^T\mathbf{S}^T\mathbf{S}A)^{-1}A^TA$ lie in the range $[1/2, 2]$.

   Hint: You may assume that A^TA has full rank. You may also want to use that for any two matrices $M, N \in \mathbb{R}^{d \times d}$, the non-zero eigenvalues of MN are equal to those of NM.

2. (2 points) Consider solving least squares regression iteratively, starting with some guess $x_0 \in \mathbb{R}^d$ and repeatedly applying the iteration $x_{i+1} = x_i - \eta A^T(Ax_i - b)$, where $\eta \in (0, 1)$ is some step size. Let $x_* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$. Prove that this iteration is equivalent to:

   $$x_{i+1} = (I - \eta A^TA)(x_i - x_*) + x_*.$$

   Hint: Prove that $A^TAx_* = A^Tb$.

3. (2 points) Let $\lambda_{\max}(A^TA), \lambda_{\min}(A^TA)$ be the largest and small eigenvalues of A^TA respectively, and let $\kappa = \frac{\lambda_{\max}(A^TA)}{\lambda_{\min}(A^TA)}$. Prove that if we set $\eta = \frac{1}{\lambda_{\max}(A^TA)}$, then the t^{th} iterate satisfies:

   $$\|x_t - x_\star\|_2 \leq \left(1 - \frac{1}{\kappa}\right)^t \cdot \|x_0 - x_\star\|_2.$$

   Hint: Bound the eigenvalues of $I - \eta A^TA$.

4. (2 points) Use the above to show for any $\epsilon \geq 0$, after $t = O\left(\kappa \cdot \log(1/\epsilon)\right)$ iterations, the t^{th} iterate satisfies $\|x_t - x_\star\|_2 \leq \epsilon\|x_\star\|_2$, assuming that we initialize with $x_0 = 0$.

5. (2 points) κ is known as the condition number of A^TA, and when it is large, the performance of this, and many other iterative methods for linear regression degrade. To avoid this we will instead consider a *preconditioned* update: let $\mathbf{S} \in \mathbb{R}^{m \times n}$ be random sketching matrix. And update: $x_{i+1} = x_i - \eta(A^T\mathbf{S}^T\mathbf{S}A)^{-1}A^T(Ax_i - b)$. Following the analysis above, and using part (1), show that if $\mathbf{S}$ is an 1/4-subpsace embedding for A, then this preconditioned method with an appropriately chosen η, has $\|x_t - x_\star\|_2 \leq \epsilon\|x_\star\|_2$ after $t = O(\log(1/\epsilon))$ iterations. That is, there is no dependence on κ.

6. (2 points) How large must m be so that $\mathbf{S}$ satisfies the required subspace embedding property with probability at least 99/100? Assuming that $\mathbf{S}A \in \mathbb{R}^{m \times d}$ is already computed, how long does it take to compute $(A^T\mathbf{S}^T\mathbf{S}A)^{-1}$? And how long does each iteration of the preconditioned method take? How does this compare to the non-preconditioned method? How about to directly solving the system using an exact method? Assume that $n \gg d$ in your discussion.

4. Compressed Sensing From Subspace Embedding (6 points)

Given a vector $x \in \mathbb{R}^n$ and a random matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, consider computing $\mathbf{y} = \mathbf{S}x$. If $m < n$, you can in general not determine $x \in \mathbb{R}^n$ from $\mathbf{y} \in \mathbb{R}^m$, since $\mathbf{S}$ is not an invertible map. Here, we will argue that you can recover x, assuming that it is k-sparse for small enough k. I.e., that it has at most k nonzero entries. This is known as *compressed sensing* or *sparse recovery.*

1. (2 points) Assume that $\mathbf{S}$ satisfies the distributional JL lemma/subspace embedding theorem proven in class. I.e., for any $A \in \mathbb{R}^{n \times d}$, if $m = O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$, then with probability at least $1-\delta$, $\mathbf{S}$ is an ϵ-subspace embedding for A. Prove that if $m = O\left(\frac{k\log(n/k)+\log(1/\delta)}{\epsilon^2}\right)$, with probability $\geq 1-\delta$, for all $z \in \mathbb{R}^n$ such that z is k-sparse, $(1-\epsilon)\|z\|_2 \leq \|\mathbf{S}z\|_2 \leq (1+\epsilon)\|z\|_2$.

   Hint: Show that with high probability, $\mathbf{S}$ is an ϵ-subspace embedding simultaneously for $\binom{n}{k}$ different matrices.

2. (2 points) Use the above result, applied with $k' = 2k$, to show that if $m = O\left(k\log(n/k) + \log(1/\delta)\right)$, and $x \in \mathbb{R}^n$ is k-sparse, then with probability $\geq 1-\delta$, x can be recovered exactly from $\mathbf{y} = \mathbf{S}x$.

   Hint: Consider solving the equation $\mathbf{y} = \mathbf{S}x$, under the restriction that x is k-sparse. Show that there is a unique solution.

3. (2 points) Argue that the above result is nearly optimal in terms of how much x is compressed. In particular, prove that for any function $f : \mathbb{R}^n \to \{0,1\}^{o(k\log(n/k))}$, given $f(x)$ for some k-sparse $x \in \mathbb{R}^n$, one cannot recover x uniquely, even under the assumption that all entries of x are either 0 or 1.