15-851 Algorithms for Big Data — Spring 2025

Problem Set 1 Solutions

Problem 1: Subspace Embeddings via Random Sign Matrices

1. Like in lecture, we can assume that A has orthonormal columns and x is a unit vector. Therefore, we have $\|Ax\|_2^2 = 1$. So, we want to show that $E_S[\|SAx\|_2^2] = 1$.

   We have that $\|SAx\|_2^2 = \sum_{i\in[k]} < r_i, x >^2$ where r_i is the i-th row of SA. By the linearity of expectation, if we can show that $E[< r_i, x >^2] = 1/k$, then we will have the result.

   We have that $< r_i, x >^2 = < (SA)_i, x >^2 = (S_i Ax)^2$. So,

$$\begin{aligned}
E[< r_i, x >^2] &= E[(S_i Ax)^2] \\
&= E\left[\left(\sum_{j=1}^{n} S_{ij}(Ax)_j\right)^2\right] \\
&= E\left[\sum_{j=1}^{n}\sum_{k=1}^{n} S_{ij}(Ax)_j S_{ik}(Ax)_k\right] \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} E[S_{ij}S_{ik}](Ax)_j(Ax)_k.
\end{aligned}$$

   When $j \neq k$, we have that $E[S_{ij}S_{ik}] = 0$ since S_{ij} and S_{ik} are independent both with mean 0. When we have $j = k$, we have that $E[S_{ij}S_{ik}] = 1/k$. Therefore we have that

$$\begin{aligned}
E[< r_i, x >^2] &= \sum_{j=1}^{n}\sum_{k=1}^{n} E[S_{ij}S_{ik}](Ax)_j(Ax)_k \\
&= \sum_{j=1}^{n} E[S_{ij}^2](Ax)_j^2 \\
&= \frac{1}{k}\sum_{j=1}^{n}(Ax)_j^2 = \frac{1}{k}.
\end{aligned}$$

2. Since $Y \in \{-1, 1\}$ is chosen uniformly at random, we have that $E[e^{tY}] = \frac{1}{2}e^{-t} + \frac{1}{2}e^{t}$. Observe that this is the $cosh(t)$ function. To finish the problem, we want to show that $E[e^{tY}]$ is upperbounded for all t by $e^{t^2/2}$. To do this, we will compare the Taylor series.

   The Taylor series for $e^{t^2/2}$ is

$$1 + \frac{t^2}{2} + \frac{(t^2/2)^2}{2!} + \frac{(t^2/2)^3}{3!} + \dots.$$

   In particular, the n-th term in the taylor series is $(t^2/2)^{n-1}/(n-1)!$. The Taylor series for $cosh(t)$ (note that equivalently you can just use the Taylor expansion for e^x twice) is

$$1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \dots.$$

   In particular, the n-th term in this taylor expansion is $t^{2(n-1)}/(2(n-1))!$.

   We need to prove that $2^{n-1} \cdot (n-1)! \leq (2(n-1))!$. We can proceed via induction. This is clearly true for $n = 1$. Now let's assume that it is true for all $n \leq k$, and we will prove it for $n = k + 1$. So we want to prove that

$$2^k \cdot k! \leq (2(k))!.$$

Let us expand the left side. We have

$$2^k \cdot k! = 2 \cdot 2^{k-1} \cdot k \cdot (k-1)! \leq (2(k-1))! \cdot 2k \leq (2(k))!.$$

Therefore, we have the result.

3. Here we want to show that $E[e^{t\sum_i \alpha_i Y_i}] \leq e^{\sigma^2 \sum_i \alpha_i^2 t^2/2}$ for all t. We have

$$E[e^{t\sum_i \alpha_i Y_i}] = E[e^{t\alpha_1 Y_1} \cdot e^{t\alpha_2 Y_2} \cdot \ldots].$$

Since we have that $Y_1, Y_2, \ldots, Y_n$ are independent, we can re-write this as $\prod_{i=1}^n E[e^{t\alpha_i Y_i}]$. From the problem we know that $Y_1, \ldots, Y_n$ are sub-Gaussian with parameter σ^2, so finally we have by the previous part that

$$\prod_{i=1}^n E[e^{t\alpha_i Y_i}] \leq \prod_{i=1}^n e^{\sigma^2 (t\alpha_i)^2/2}$$

which gives us the result.

4. We have that $E[e^{tV}] = e^{t^2\sigma^2/2}$. So we can use this to say that

$$E_V[e^{\sqrt{2t}|Y|V/\sigma}] = e^{\frac{2t|Y|^2}{\sigma^2} \cdot \frac{\sigma^2}{2}} = e^{tY^2}.$$

Therefore, as per the hint, we have $E_Y[e^{tY^2}] = E_{Y,V}[e^{(\sqrt{2t}Y/\sigma)V}]$.

Now let us work with $E_{Y,V}[e^{(\sqrt{2t}Y/\sigma)V}]$. We want to show that

$$E_{Y,V}[e^{(\sqrt{2t}Y/\sigma)V}] \leq E_V[e^{tV^2}].$$

We can rewrite $E_{Y,V}[e^{(\sqrt{2t}Y/\sigma)V}]$ as $E_V[E_Y[e^{(\sqrt{2t}Y/\sigma)V}]]$. Now we use the fact that Y is mean zero σ^2-sub-gaussian and get that

$$E_V[E_Y[e^{(\sqrt{2t}Y/\sigma)V}]] \leq E_V[e^{tV^2}].$$

5. We first show the upper tail, or that $\mathbf{Pr}[Y \geq 1 + \varepsilon] \leq e^{-\Theta(d)}$. Per the hint, using Markov's inequality we get that

$$\mathbf{Pr}[Y \geq 1 + \epsilon] = \mathbf{Pr}[e^{tkY} \geq e^{tk(1+\epsilon)}] \leq \frac{\mathbf{E}[e^{tkY}]}{e^{tk(1+\epsilon)}} = \prod_{i=1}^k \frac{\mathbf{E}[e^{tY_i^2}]}{e^{t(1+\epsilon)}}$$

for $t > 0$. Using part 4 and the hint again, we get that

$$\mathbf{Pr}[Y \geq 1 + \varepsilon] \leq \prod_{i=1}^k \frac{E[e^{tV_i^2}]}{e^{t(1+\varepsilon)}} \leq \prod_{i=1}^k \frac{1}{\sqrt{1-2t} \cdot e^{t(1+\varepsilon)}} = \frac{1}{\sqrt{1-2t}^k} \cdot \frac{1}{e^{tk(1+\varepsilon)}}$$

for $t < 1/2$.

Now, let us set $t = \varepsilon/100$. Let us prove a lower bound on $\sqrt{1-2t} \cdot e^{t(1+\varepsilon)}$. We have that

$$\ln(\sqrt{1-2t} \cdot e^{t(1+\varepsilon)}) = \ln(\sqrt{1-2t}) + \ln(e^{t(1+\varepsilon)}) = \frac{1}{2}\ln(1-2t) + t(1+\varepsilon).$$

Using the taylor expansion for $\ln(1-x)$, we get that $\ln(1-2t) \geq -2t - 4t^2/2 - O(\varepsilon^3)$. So we have that

$$\ln(\sqrt{1-2t} \cdot e^{t(1+\varepsilon)}) = \frac{1}{2}\ln(1-2t) + t(1+\varepsilon) \geq \varepsilon t - t^2 + O(\varepsilon^3) = \Theta(\varepsilon^2).$$

So therefore we have

$$\frac{1}{(\sqrt{1-2t} \cdot e^{t(1+\varepsilon)})^k} \leq \frac{1}{(e^{\Theta(\varepsilon^2)})^k}.$$

Plugging in $k = O(d/\varepsilon^2)$ gives the result.

We will now do the lower tail, or show that $\mathbf{Pr}[Y \le (1-\varepsilon)] \le e^{-\Theta(d)}$. Similarly to the above,

$$\mathbf{Pr}[Y \le (1-\varepsilon)] = \mathbf{Pr}[e^{-tkY} \ge e^{-tk(1-\varepsilon)}] \le \frac{E[e^{-tkY}]}{e^{-tk(1-\varepsilon)}} = \prod_{i=1}^{k} \frac{E[e^{-tY_i^2}]}{e^{-t(1-\varepsilon)}}.$$

As per the hint, we have that

$$E[e^{tY_i^2}] \le 1 + t + (ct)^2 E[e^{Y_i^2/c}]$$

for any $|t| < 1/c$.

Again take $t = \varepsilon/100$. So we have $|t| < 1/c$ for $c = 100$. So, we have that

$$\mathbf{Pr}[Y \le (1-\varepsilon)] \le \prod_{i=1}^{k} \frac{E[e^{-tY_i^2}]}{e^{-t(1-\varepsilon)}} \le \prod_{i=1}^{k} \frac{1 - t + (ct)^2 E[e^{Y_i^2/c}]}{e^{-t(1-\varepsilon)}}.$$

Applying the same logic as the upper bound we have

$$\prod_{i=1}^{k} \frac{1 - t + (ct)^2 E[e^{Y_i^2/c}]}{e^{-t(1-\varepsilon)}} \le \prod_{i=1}^{k} \frac{1 - t + (ct)^2 E[e^{V_i^2/c}]}{e^{-t(1-\varepsilon)}} \le \prod_{i=1}^{k} \frac{1 - t + (ct)^2/\sqrt{1-2/c}}{e^{-t(1-\varepsilon)}}.$$

So we have

$$\mathbf{Pr}[Y \le 1-\varepsilon] \le \prod_{i=1}^{k} \frac{1 - t + (ct)^2/\sqrt{1-2/c}}{e^{-t(1-\varepsilon)}} \le \prod_{i=1}^{k} \frac{1 - t + c^3 t^2}{e^{-t(1-\varepsilon)}} = \frac{e^{k \ln(1-t+c^3t^2)}}{e^{-tk(1-\varepsilon)}}.$$

Using the taylor expansion, we have that

$$\mathbf{Pr}[Y \le (1-\varepsilon)] \le \frac{e^{k\cdot\Theta(-t+t^2)}}{e^{-tk(1-\varepsilon)}} = e^{\Theta(-tk+t^2k)+tk(1-\varepsilon)} = e^{\Theta(t^2k-\varepsilon tk)} = e^{-\Theta(\varepsilon^2 k)}.$$

Plugging in $k = O(d/\varepsilon^2)$ gives the result.

6. This directly follows from the net argument proof in class and parts 1 and 5.

Problem 2: Multiplying Gaussian Matrices

1. Let us consider the $d \times d$ matrix $V^T G_2$. First, we can see that each entry of $V^T G_2$ is $< v_r, g_c >$ where v_r is a row of V^T and g_c is a column of G_2. Recall that the rows of V^T are orthonormal. So, each row has length 1. From class we know that adding $X + Y$ where $X = N(0, a^2)$ and $Y = N(0, b^2)$ gives random variable $Z = N(0, a^2 + b^2)$. So, we have that $< v_r, g_c >$ is a Gaussian random variable with variance $|v_r|_2^2 = 1$.

   Now, we only need that the entries of $V^T G$ are independent. The rows of V^T are orthonormal and therefore for each pair of rows v_{r_1}, v_{r_2} we have $< v_{r_1}, v_{r_2} >= 0$. Therefore by rotational invariance (from class), we have that the entries are independent.

2. Let us consider $M = V^T G_2$. From Part 1 we saw that M is a matrix of i.i.d. $N(0,1)$ entries. We can re-write this as $\sqrt{t}M'$ where M' is a matrix of i.i.d. $N(0, 1/t)$ entries.

   We can now use the result of Jiang. Let us set the parameter $z = t$. Therefore, we have that M' is indistinguishable from $\tilde{M}$ where $\tilde{M}$ is a $d \times d$ submatrix of a random $t \times t$ matrix with orthonormal rows and columns. Note that we have $d^2 = o(t)$.

3. So we have $G_1G_2 = \sqrt{t}U\Sigma\tilde{M}$ where $\tilde{M}$ is a $d \times d$ submatrix of a random $t \times t$ matrix R with orthonormal rows and columns. We can rewrite $\tilde{M}$ as LP where L is a $d \times t$ submatrix of R and $P = [I_d, 0]$. Note that this means L has orthonormal rows.

   So, we have $G_1G_2 = \sqrt{t}U\Sigma LP$. The hint tells us that the SVD of a random $d \times t$ matrix G_3 of i.i.d. $N(0,1)$ random variables is equal to $U\Sigma V^T$, where $U, \Sigma \in \mathbb{R}^{d \times d}$ and $V^T \in \mathbb{R}^{d \times t}$ are independent matrices and V^T is a random matrix with orthonormal rows. So, we can conclude that $U\Sigma L$ is a random $d \times t$ matrix G_3 of i.i.d. $N(0,1)$ random variables. $U\Sigma LP$ is simply the first d columns of $U\Sigma L$, and is therefore a $d \times d$ matrix of i.i.d. $N(0,1)$ random variables. Finally multiplying by $\sqrt{t}$ gives the desired result.

Problem 3: Learning the Positions and Values of CountSketch

Proof. Let us take matrix A. This would be easier if A had orthonormal columns. Since we cannot assume that, we instead will do the following.

Lets take $A' = AR$ where R is a $d \times d$ diagonal matrix and R_{jj} for $j \in [d]$ is $\sqrt{\frac{1}{\sum_{i=1}^{n} A_{ij}^2}}$. Notice that all we are doing is normalizing the columns of A. So, we have that A' has orthonormal columns and still preserves the property of only having one entry per row. Note that the columns of A were already independent since each row of A was guaranteed to have only one entry, which means that the dot product between any two columns of A is 0.

So, if we can prove that for all x we have $|SARx|_2^2 = |ARx|_2^2$, then by a change of variable (similar to in class), we have the desired statement.

We will set our sketching matrix $S = (A')^\intercal = (AR)^\intercal$. Since we know that A' has only one entry per row, then S only has one entry per column. This satisfies the desired property of S.

Our final step is showing that $|SARx|_2^2 = |ARx|_2^2$ for all x. Take any arbitrary x. Notice that $SARx = (AR)^\intercal ARx$, and that AR has orthonormal columns. Therefore, we have that $(AR)^\intercal ARx = x$ giving us $|SARx|_2^2 = |x|_2^2$. We have $|x|_2^2 = |ARx|_2^2$ since AR has orthonormal columns.

Note that there are other possible solutions. □

Problem 4: Approximate Matrix Product in Terms of Stable Rank

1. Let us suppose for the purposes of contradiction that S has $r < d$ rows.

   One example for $A = B$ is $[I_d, 0]$. This means that we have the identity matrix with dimensions $d \times d$ with rows of 0 appended to the bottom to form our $n \times d$ matrix $A = B$. Therefore to meet the guarantee we want, it would have to be that

$$\mathbf{Pr}[\|A^\intercal S^\intercal SA - I_d\|_2^2 \geq \frac{1}{4}\|A\|_2^2\|B\|_2^2] \leq 1/\operatorname{poly}(n).$$

   Notice that $\|A\|_2^2 = \|B\|_2^2 = 1$, so this simplifies to

$$\mathbf{Pr}[\|A^\intercal S^\intercal SA - I_d\|_2^2 \geq \frac{1}{4}] \leq 1/\operatorname{poly}(n).$$

   Notice that $A^\intercal S^\intercal SA$ has rank at most that of S, which is at most $r < d$. Therefore, simply consider any vector in the kernel of S which means that the operator norm must be large by definition.

2. Using the hint, we have that $\|A\|_F^2 = \sum_{i=1}^{\min(n,d)} \sigma_i^2 \leq d \cdot \sigma_1^2$. We also have that $\|A\|_2^2 = \sigma_1^2$. This gives us the result.

3. (a) Each R_i is $A^\intercal \frac{1}{\sqrt{p_i}} \cdot e_i \cdot \frac{1}{\sqrt{p_i}} \cdot e_i^\intercal B$. When we take $\frac{1}{m}\sum_{i=1}^{m} \frac{1}{p_i} \cdot A^\intercal e_i e_i^\intercal B$, we get $A^T S^T SB$.

   (b) $E[R] = \sum_i p_i \cdot \frac{1}{p_i} A^T e_i e_i^\intercal B = \sum_i A^\intercal e_i e_i^\intercal B = A^\intercal B$.

(c) Recall that we said for some iteration i, we have $R = \frac{1}{p_i} A^\intercal e_i e_i^\intercal B$. So, we get

$$\|R\|_2 \leq \max_i \frac{|A_i^\intercal B_i|_2}{p_i} \leq O(1) \cdot \max_i \frac{|A_i|_2 |B_i|_2 (|A|_F^2 + \gamma |B|_F^2)}{|A_i|_2^2 + \gamma |B_i|_2^2}.$$

Here A_i and B_i denote the i-th row of matrix A and B respectively. Per the hint, we can use the AM-GM inequality to say that $|A_i|_2^2 + \gamma |B_i|_2^2 \geq 2\sqrt{|A_i|_2^2 \cdot \gamma |B_i|_2^2}$. So, we have that the above is at most

$$O(1) \cdot (\frac{1}{\sqrt{\gamma}} \|A\|_F^2 + \sqrt{\gamma} \|B\|_F^2).$$

Plugging in the value of γ gives us

$$O(1) \cdot \frac{\|B\|_2}{\|A\|_2} \|A\|_F^2 + O(1) \cdot \frac{\|A\|_2}{\|B\|_2} \|B\|_F^2 = O(1) \cdot \|A\|_2 \|B\|_2 \cdot srank(A) + O(1) \cdot \|A\|_2 \|B\|_2 \cdot srank(B).$$

(d) To do this, we will calculate $\|E[R^\intercal R]\|_2$ and then $\|E[RR^\intercal]\|_2$. Let us do the first. So, we have that

$$E[R^\intercal R] = \sum_i \frac{|A_i|_2^2 B_i^\intercal B_i}{p_i} \leq O(1) \cdot (\|A\|_F^2 + \gamma \|B\|_F^2) \sum_i \frac{|A_i|_2^2 B_i^\intercal B_i}{|A_i|_2^2 + \gamma |B_i|_2^2}.$$

We can see that $\frac{|A_i|_2^2}{|A_i|_2^2 + \gamma |B_i|_2^2} \leq 1$, so since we have that $B_i^\intercal B_i$ is PSD, we have that this is at most

$$O(1) \cdot (\|A\|_F^2 + \gamma \|B\|_F^2) \sum_i B_i^\intercal B_i = O(1) \cdot (\|A\|_F^2 + \gamma \|B\|_F^2) B^\intercal B.$$

So we have that

$$\begin{aligned} \|E[R^\intercal R]\|_2 &= O(1) \cdot (\|A\|_F^2 + \gamma \|B\|_F^2) \|B^\intercal B\|_2 = O(1) \cdot \|B^\intercal B\|_2 (\|A\|_F^2 + \frac{\|A\|_2^2}{\|B\|_2^2} \|B\|_F^2) \\ &= O(1) \cdot \|B^\intercal B\|_2 (\|A\|_F^2 + \|A\|_2^2 srank(B)) \leq O(1) \cdot \|B\|_2^2 \|A\|_2^2 (srank(A) + srank(B)). \end{aligned}$$

We do a similar process to calculate $\|E[RR^\intercal]\|_2$. We have

$$E[RR^\intercal] = \sum_i \frac{|B_i|_2^2 A_i^\intercal A_i}{p_i} \leq O(1) \cdot (\|A\|_F^2 + \gamma \|B\|_F^2) \sum_i \frac{|B_i|_2^2 A_i^\intercal A_i}{|A_i|_2^2 + \gamma |B_i|_2^2}.$$

Here we can see that $\frac{|B_i|_2^2}{|A_i|_2^2 + \gamma |B_i|_2^2} \leq 1/\gamma$. Since $A_i^\intercal A_i$ is PSD, we have that this is at most

$$O(1) \cdot (\|A\|_F^2 / \gamma + \|B\|_F^2) A^\intercal A.$$

So, we have that

$$\begin{aligned} \|E[RR^\intercal]\|_2 &= O(1) \cdot \|A^\intercal A\|_2 (\frac{\|A\|_F^2}{\gamma} + \|B\|_F^2) \leq O(1) \cdot \|A\|_2^2 (srank(A) \|B\|_2^2 + \|B\|_2^2 srank(B)) \\ &= O(1) \|B\|_2^2 \|A\|_2^2 (srank(A) + srank(B)). \end{aligned}$$

(e) Plugging in the results from the previous 2 parts and plugging in $t = \varepsilon \|A\|_2 \|B\|_2$ to the generalized Matrix Chernoff gives us the result.