15-859 : Homework 1 Solutions

1 Question 1

1.1 Part 1

Let U be an orthonormal basis for matrix A. As S is a subspace embedding, we have that

$$\|SUx\|_2^2 = (1 \pm \varepsilon)\|Ux\|_2^2 = (1 \pm \varepsilon)\|x\|_2^2.$$

Therefore $\min_{x:\|x\|_2=1} \|SUx\|_2 \geq 1 - O(\varepsilon) \geq 3/4$. Therefore, $\sigma_{\min}(U^\mathsf{T} S^\mathsf{T} SU) = \sigma_{\min}(SU)^2 \geq 9/16$. Let $Uz = A\hat{x} - Ax_{\text{opt}}$. We have

$$\begin{aligned}
\|A\hat{x} - Ax_{\text{opt}}\|_2 &= \|Uz\|_2 \\
&= \|z\|_2 \\
&\leq \frac{16}{9}\|U^\mathsf{T} S^\mathsf{T} SUz\|_2 \\
&\leq \frac{16}{9}\|U^\mathsf{T} S^\mathsf{T} S(A\hat{x} - Ax_{\text{opt}})\|_2 \\
&\leq \frac{16}{9}\|U^\mathsf{T} S^\mathsf{T} S(A\hat{x} - b) - U^\mathsf{T} S^\mathsf{T} S(Ax_{\text{opt}} - b)\|_2
\end{aligned}$$

Now we claim that $U^\mathsf{T} S^\mathsf{T} S(A\hat{x} - b) = 0$. The vector $S(A\hat{x} - b) = SA(SA)^\dagger Sb - Sb = ((SA)(SA)^\dagger - I)Sb$ is Sb projected away from colspan of SA. Therefore, $(SA)^\mathsf{T} S(A\hat{x} - b) = 0$ which implies $A^\mathsf{T} S^\mathsf{T} S(A\hat{x} - b) = 0$. As $U = AQ$ for some matrix Q, $U^\mathsf{T} S^\mathsf{T} S(A\hat{x} - b) = Q^\mathsf{T} A^\mathsf{T} S^\mathsf{T} S(A\hat{x} - b) = 0$. Therefore,

$$\|A\hat{x} - Ax_{\text{opt}}\|_2 \leq (16/9)\|U^\mathsf{T} S^\mathsf{T} S(Ax_{\text{opt}} - b)\|_2.$$

But we also have that $U^\mathsf{T}(Ax_{\text{opt}} - b) = 0$ since $Ax_{\text{opt}} - b$ is orthogonal to colspan(A) and therefore to colspan of U. Therefore,

Now we prove an important Approximate Matrix Multiplication property satisfied by a subspace embedding.

Lemma 1.1. *If S is an ε subspace embedding matrix for colspan($[U, V]$), then*

$$\|U^\mathsf{T} S^\mathsf{T} SV - U^\mathsf{T} V\|_2 \leq \varepsilon\|U\|_2\|V\|_2.$$

Proof. Without loss of generality we can assume that $\|U\|_2 = \|V\|_2 = 1$. We also have the following

$$\|U^\mathsf{T} S^\mathsf{T} SV - U^\mathsf{T} V\|_2 = \max_{x,y:\|x\|_2=\|y\|_2=1} |x^\mathsf{T}(U^\mathsf{T} S^\mathsf{T} SV - U^\mathsf{T} V)y| = \max_{x,y:\|x\|_2=\|y\|_2=1} |\langle SUx, SVy\rangle - \langle Ux, Vy\rangle|.$$

For any unit vectors x, y,

$$\begin{aligned}
\langle SUx, SVy\rangle &= \frac{\|SUx + SVy\|_2^2 - \|SUX - SVy\|_2^2}{4} \\
\langle Ux, Vy\rangle &= \frac{\|Ux + Vy\|_2^2 - \|Ux - Vy\|_2^2}{4}.
\end{aligned}$$

By the fact that S is an ε subspace embedding for span of U and V, we get

$$\begin{aligned}|\langle SUx, SVy\rangle - \langle Ux, Vy\rangle| &= |\frac{\|SUx+SVy\|_2^2 - \|SUX-SVy\|_2^2}{4} - \frac{\|Ux+Vy\|_2^2 - \|Ux-Vy\|_2^2}{4}| \\ &\leq \frac{1}{4}|\|SUx+SVy\|_2^2 - \|Ux+Vy\|_2^2| + \frac{1}{4}|\|SUX-SVy\|_2^2 - \|Ux-Vy\|_2^2| \\ &\leq \frac{1}{4}\varepsilon\|Ux+Vy\|_2^2 + \frac{1}{4}\varepsilon\|Ux-Vy\|_2^2 \\ &\leq \frac{1}{4}\varepsilon(2\|Ux\|_2^2 + 2\|Vy\|_2^2) \leq \frac{1}{4}\varepsilon(2+2) = \varepsilon.\end{aligned}$$

Here we used the fact that $\|U\|_2 = \|V\|_2 = 1$ and that x, y are unit vectors. Therefore we have for any U, V

$$\|\frac{U^\mathsf{T}}{\|U\|_2}S^\mathsf{T}S\frac{V}{\|V\|_2} - \frac{U^\mathsf{T}}{\|U\|_2}\frac{V}{\|V\|_2}\|_2 \leq \varepsilon$$

which implies

$$\|U^\mathsf{T}S^\mathsf{T}SV - U^\mathsf{T}V\|_2 \leq \varepsilon\|U\|_2\|V\|_2. \qquad \square$$

Using the above lemma

$$\|A\hat{x}-Ax_{\text{opt}}\|_2 \leq (16/9)\|U^\mathsf{T}S^\mathsf{T}S(Ax_{\text{opt}}-b)\|_2 = (16/9)\|U^\mathsf{T}S^\mathsf{T}S(Ax_{\text{opt}}-b) - U^\mathsf{T}(Ax_{\text{opt}}-b)\|_2 \leq (16/9)\varepsilon\|Ax-b\|_2.$$

Finally,

$$\|\hat{x} - x_{\text{opt}}\|_2 \leq \frac{\|A\hat{x} - Ax_{\text{opt}}\|_2}{\sigma_{\min}(A)} = (16/9)\varepsilon\|Ax-b\|_2\|A^\dagger\|_2.$$

1.2 Part 2

For any vector v, we have that $|\langle\hat{x}, v\rangle - \langle x_{\text{opt}}, v\rangle| = |\langle\hat{x} - x_{\text{opt}}, v\rangle| \leq \|\hat{x} - x_{\text{opt}}\|_2\|v\|_2$ by Cauchy-Schwartz. Therefore,

$$|\langle\hat{x}, v\rangle - \langle x_{\text{opt}}, v\rangle| \leq O(\varepsilon)\|v\|_2\|Ax-b\|_2\|A^\dagger\|_2.$$

1.3 Part 3

$\hat{x} = (SA)^\dagger(Sb)$ and $x_{\text{opt}} = A^\dagger b$. As A is full rank matrix, we obtain from subspace embedding guarantee by S that SA is also full rank as if S is a subspace embedding nullspace(SA) = nullspace$(A) = \{0\}$. Therefore rowspan$(SA) = \mathbb{R}^d$ which implies that $(SA)^\dagger SA = I$. Therefore, $x_{\text{opt}} = A^\dagger b = (SA)^\dagger(SA)Sb$ which implies that $\hat{x} - x_{\text{opt}} = (SA)^\dagger Sb - (SA)^\dagger(SA)A^\dagger b = (SA)^\dagger S(b - AA^\dagger b)$. Therefore

$$|\langle\hat{x}, v\rangle - \langle x_{\text{opt}}, v\rangle| = |\langle(SA)^\dagger S(b - AA^\dagger b), v\rangle|.$$

1.4 Part 4

We have that $b - AA^\dagger b$ is orthogonal to span of A. As independent Gaussians multiplied to orthonormal vectors are independent, we have that $S(b - AA^\dagger b)$ is independent of the matrix SA and hence independent of the vector $v^\mathsf{T}(SA)^\dagger$ for a fixed vector v. Given $(SA)^\dagger$, the vector $S(b - AA^\dagger b)$ is distributed as a vector of independent gaussians of mean 0 and variance $\|b - AA^\dagger b\|_2^2/s$. So, given $(SA)^\dagger$, $v^\mathsf{T}(SA)^\dagger S(b - AA^\dagger b)$ is distributed as a $N(0, \|v^\mathsf{T}(SA)^\dagger\|_2^2\|b - AA^\dagger b\|_2^2/s)$. We have with $1 - 1/d^2$ probability that $|v^\mathsf{T}(SA)^\dagger S(b - AA^\dagger b)| \leq O(\sqrt{\log(d)})\|v^\mathsf{T}(SA)^\dagger\|_2\|b - AA^\dagger b\|_2/\sqrt{s} \leq O(\varepsilon\sqrt{\log d}/\sqrt{d})\|v^\mathsf{T}(SA)^\dagger\|_2\|b - AA^\dagger b\|_2$ given $(SA)^\dagger$. Here we used the subgaussianity of gaussians i.e., $\mathbf{Pr}[X > t] \leq e^{-t^2/2}$ for a normal random variable.

Considering $v = e_1, \ldots, e_d$, by union bound we obtain that, given $(SA)^\dagger$, for all i,

$$|e_i^\mathsf{T}(SA)^\dagger S(b - AA^\dagger b)| \leq O(\varepsilon\sqrt{\log d}/\sqrt{d})\|e_i^\mathsf{T}(SA)^\dagger\|_2\|b - AA^\dagger b\|_2 \leq O(\varepsilon\sqrt{\log d}/\sqrt{d})\|(SA)^\dagger\|_2\|b - AA^\dagger b\|_2$$

with probability $\geq 1 - 1/d$.

Now we have $\sigma_{\min}(SA) = \min_{x:\|x\|_2=1} \|(SA)x\|_2 \geq (1-\varepsilon)\min_{x:\|x\|_2=1} \|Ax\|_2 \geq (1-\varepsilon)\sigma_{\min}(A)$. Therefore, $\|(SA)^\dagger\|_2 = 1/\sigma_{\min}(SA) \leq 1/(1-\varepsilon)\sigma_{\min(A)} \leq \|A^\dagger\|_2/(1-\varepsilon)$ with probability $\geq 4/5$. Using union bound, we have with probability $\geq 2/3$ that for all i,

$$|\langle e_i, \hat{x} - x_{\text{opt}}\rangle| = e_i^\mathsf{T}(SA)^\dagger S(b - AA^\dagger b) \leq O(\varepsilon\sqrt{\log d}/\sqrt{d})\|(SA)^\dagger\|_2\|b - AA^\dagger b\|_2.$$

Therefore,

$$\|\hat{x} - x_{\text{opt}}\|_\infty = \max_i |\langle e_i, \hat{x} - x_{\text{opt}}\rangle| \leq O(\varepsilon\sqrt{\log d}/\sqrt{d})\|(SA)^\dagger\|_2\|b - AA^\dagger b\|_2.$$

2 Question 2

2.1 Part 1

Let $U\Sigma V^\mathsf{T}$ be the full singular value decomposition of the matrix A. Let $\hat{A} = \begin{bmatrix} A & b \\ I_d & 0 \end{bmatrix}$. As in the hint, consider the matrix $\hat{U} = \begin{bmatrix} U\Sigma(\Sigma^\mathsf{T}\Sigma + I_d)^{-1/2} & b_1 \\ V(\Sigma^\mathsf{T}\Sigma + I_d)^{-1/2} & b_2 \end{bmatrix}$ where $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ is a unit vector perpendicular to first d columns of $\hat{A}$ and is in the direction of projection of $(b^\mathsf{T}\ 0)^\mathsf{T}$ away from the first d columns of $\hat{A}$. Therefore, we have that

$$\begin{bmatrix} b \\ 0 \end{bmatrix} = \begin{bmatrix} A \\ I_d \end{bmatrix} \alpha + \beta \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

for some vector α and a scalar β. First we have that

$$\hat{U}\begin{bmatrix} (\Sigma^\mathsf{T}\Sigma + I_d)^{1/2}V^\mathsf{T} \\ 0 \end{bmatrix} = \begin{bmatrix} U\Sigma V^\mathsf{T} \\ VV^\mathsf{T} \end{bmatrix} = \begin{bmatrix} A \\ I_d \end{bmatrix}.$$

Therefore, the first d columns of $\hat{U}$ span first d columns of $\hat{A}$ and as first d columns of $\hat{A}$ has rank d, we conclude that colspan(first d columns of $\hat{A}$) = colspan(first d columns of $\hat{U}$). This implies that the vector $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ is orthogonal to first d columns of $\hat{U}$ which shows

$$\hat{U}^\mathsf{T}\hat{U} = \begin{bmatrix} (\Sigma^\mathsf{T}\Sigma + I_d)^{-1/2}(\Sigma^\mathsf{T}U^\mathsf{T}U\Sigma + V^\mathsf{T}V)(\Sigma^\mathsf{T}\Sigma + I_d)^{-1/2} & 0 \\ 0 & 1 \end{bmatrix} = I_{d+1}$$

as $U^\mathsf{T}U = I$ and $V^\mathsf{T}V = I$. Now,

$$\hat{U}\begin{bmatrix} (\Sigma^\mathsf{T}\Sigma + I_d)^{1/2}V^\mathsf{T} & (\Sigma^\mathsf{T}\Sigma + I_d)^{1/2}V^\mathsf{T}\alpha \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} U\Sigma V^\mathsf{T} & U\Sigma V^\mathsf{T}\alpha + \beta b_1 \\ VV^\mathsf{T} & VV^\mathsf{T}\alpha + \beta b_2 \end{bmatrix} = \hat{A}.$$

This shows that colspan($\hat{U}$) spans the columns of $\hat{A}$ and as $\hat{A}$ is full rank (since there is an I_d in the bottom part), we conclude that $\hat{U}$ is an orthonormal basis for colspan($\hat{A}$). Now, $\|[U\Sigma(\Sigma^\mathsf{T}\Sigma + I_d)^{-1/2}\ b_1]\|_\mathsf{F}^2 = \|[\Sigma(\Sigma^\mathsf{T}\Sigma + I_d)^{-1/2}\ b_1]\|_\mathsf{F}^2 \leq \sum_i \left(\frac{\sigma_i}{\sqrt{\sigma_i^2+1}}\right)^2 + 1 = \sum_i \frac{1}{1+1/\sigma_i^2} + 1$.

2.2 Part 2

Let $\hat{U} = \begin{bmatrix} U_1 & b_1 \\ U_2 & b_2 \end{bmatrix}$. We showed that $\|[U_1\ b_1]\|_{\mathsf{F}}^2 \leq t$. Let $\hat{S} = \begin{bmatrix} S & 0 \\ 0 & I_d \end{bmatrix}$. We claim that $\hat{S}$ is a subspace embedding for colspan$(\hat{U})$. We have that

$$\begin{aligned}
\|\hat{U}^{\mathsf{T}}\hat{S}^{\mathsf{T}}\hat{S}\hat{U} - \hat{U}^{\mathsf{T}}\hat{U}\|_2 &= \|[U_1\ b_1]^{\mathsf{T}}S^{\mathsf{T}}S[U_1\ b_1] + [U_2\ b_2]^{\mathsf{T}}[U_2\ b_2] - [U_1\ b_1]^{\mathsf{T}}[U_1\ b_1] - [U_2\ b_2]^{\mathsf{T}}[U_2\ b_2]\|_2 \\
&= \|[U_1\ b_1]^{\mathsf{T}}S^{\mathsf{T}}S[U_1\ b_1] - [U_1\ b_1]^{\mathsf{T}}[U_1\ b_1]\|_2 \\
&\leq \|[U_1\ b_1]^{\mathsf{T}}S^{\mathsf{T}}S[U_1\ b_1] - [U_1\ b_1]^{\mathsf{T}}[U_1\ b_1]\|_{\mathsf{F}} \\
&\leq \frac{\|[U_1\ b_1]\|_{\mathsf{F}}^2}{\sqrt{\#\text{ rows in } S}} \\
&\leq \frac{(t)}{\sqrt{\#\text{ rows in } S}} \\
&\leq \varepsilon.
\end{aligned}$$

Last inequality by assuming S has $O(t^2/\varepsilon^2)$ rows. Therefore, $\hat{S}$ is a subspace embedding for colspan$(\hat{U}) =$ colspan $\left(\begin{bmatrix} A & b \\ I_d & 0 \end{bmatrix}\right)$ which implies that

$$\|\hat{S}\begin{bmatrix} A & b \\ I_d & 0 \end{bmatrix}\begin{bmatrix} x \\ 0 \end{bmatrix}\|_2^2 \in (1 \pm \varepsilon)\|\hat{S}\begin{bmatrix} A & b \\ I_d & 0 \end{bmatrix}\begin{bmatrix} x \\ 0 \end{bmatrix}\|_2^2$$

for all $x \in \mathbb{R}^d$. The left hand side is equal by $\|SAX - Sb\|_2^2 + \|x\|_2^2$ and the right hand side is equal to $(1 \pm \varepsilon)(\|Ax - b\|_2^2 + \|x\|_2^2)$. Therefore,

$$\|SAx - Sb\|_2^2 + \|x\|_2^2 = (1 \pm \varepsilon)(\|Ax - b\|_2^2 + \|x\|_2^2).$$

3 Question 3

We are given that the matrix C has at most k independent columns. So, the colspan(C) has dimension at most k and therefore a countsketch matrix $S \in \mathbb{R}^{s \times n}$ with $s = \Theta(k^2)$ is a $1/2$-subspace embedding for colspan(C) with probability $\geq 2/3$. Therefore,

$$\|SCx\|_2^2 \in (1 \pm 1/2)\|Cx\|_2^2$$

which implies that $SCx = 0 \Leftrightarrow Cx = 0$. Therefore a set of columns of C are independent iff the corresponding set of columns of SC are independent. The matrix SC can be computed in time nnz$(C) + nk^2$ and as SC is now a $\Theta(k^2) \times n$ matrix with at most k independent columns, any basis finding algorithm (say Gram-Schmidt) would run in time $n\,\mathrm{poly}(k)$.

4 Question 4

As we require the sampling to be finished in time $O(\min(\text{nnz}(A), \text{nnz}(B))) + O(n/\varepsilon^2)$, the only possible way is to sample few columns/rows of the denser matrix without even looking at the entire matrix. But the required running time allows to look at and do some computation with the sparser matrix. Without loss of generality, assume that nnz$(A) <$ nnz(B). We can write

$$AB = \sum_i A_{*i}B_{i*}.$$

Here A_{*i} is a column vector denoting ith column of A and B_{*i} is the row vector corresponding to ith row of B. One natural way to estimate the product AB is to choose an index $i \in \{1, \ldots, n\}$ with probability p_i

and then output $\boldsymbol{X} = (1/p_i)A_{*i}B_{i*}$. We have

$$\mathbf{E}[\boldsymbol{X}] = \sum_{i\in[n]} \left(\frac{1}{p_i}A_{*i}B_{i*}\right) p_i = \sum_{i\in[n]} A_{*i}B_{i*} = AB.$$

Therefore, estimator $\boldsymbol{X}$ is an unbiased estimator of the matrix AB. Now we compute the variance $\boldsymbol{X}_{kl}$, the kl entry of the matrix $\boldsymbol{X}$. We already saw that $\mathbf{E}[\boldsymbol{X}_{kl}] = (AB)_{kl}$.

$$\begin{aligned}
\mathbf{E}[(\boldsymbol{X}_{kl} - (AB)_{kl})^2] &= \mathbf{E}[\boldsymbol{X}_{kl}^2] - (AB)_{kl}^2 \\
&= \sum_{i\in[n]} \left(\frac{1}{p_i}A_{ki}B_{il}\right)^2 p_i - (AB)_{kl}^2 \\
&= \sum_{i\in[n]} \frac{A_{ki}^2 B_{il}^2}{p_i} - (AB)_{kl}^2
\end{aligned}$$

As we can only use the matrix A to perform computation, we choose $p_i = \|A_{*i}\|_2^2/\|A\|_\mathsf{F}^2$. Note that these probabilities can be computed in $\mathrm{nnz}(A)$ time. Now,

$$\begin{aligned}
\mathbf{E}[\|\boldsymbol{X} - AB\|_\mathsf{F}^2] &= \sum_k \sum_l \mathbf{E}[(\boldsymbol{X}_{kl} - (AB)_{kl})]^2 \\
&= \sum_k \sum_l \left(\sum_{i\in[n]} \left(\frac{A_{ki}^2 B_{il}^2}{p_i}\right) - (AB)_{kl}^2 \right) \\
&= \sum_k \sum_l \left(\|A\|_\mathsf{F}^2 \sum_{i\in[n]} \left(\frac{A_{ki}^2 B_{il}^2}{\|A_{*i}\|_2^2}\right) - (AB)_{kl}^2 \right) \\
&= \|A\|_\mathsf{F}^2 \sum_{i\in[n]} \frac{1}{\|A_{*i}\|_2^2} (\sum_k A_{ki}^2)(\sum_l B_{il}^2) - \|AB\|_\mathsf{F}^2 \\
&= \|A\|_\mathsf{F}^2 \sum_{i\in[n]} \frac{1}{\|A_{*i}\|_2^2} \|A_{*i}\|_2^2 \|B_{i*}\|_2^2 - \|AB\|_\mathsf{F}^2 \\
&= \|A\|_\mathsf{F}^2 \|B\|_\mathsf{F}^2 - \|AB\|_\mathsf{F}^2.
\end{aligned}$$

This variance is too large. So we instead use the standard way of taking s independent samples and averaging to reduce variance. If we take s independent samples $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(s)}$, for the matrix $\boldsymbol{Y} = (1/s)\sum_i \boldsymbol{X}^{(i)}$

$$\begin{aligned}
\mathbf{E}[\|\boldsymbol{Y} - AB\|_\mathsf{F}^2] = \mathbf{E}[\sum_{k,l} (\boldsymbol{Y}_{k,l} - (AB)_{k,l})^2] &= \sum_{k,l} \mathbf{E}[(\boldsymbol{Y}_{k,l} - (AB)_{k,l})]^2 \\
= (1/s) \sum_{k,l} \mathbf{E}[(\boldsymbol{X}_{k,l} - (AB)_{k,l})]^2 &= (1/s)(\|A\|_\mathsf{F}^2 \|B\|_\mathsf{F}^2 - \|AB\|_\mathsf{F}^2).
\end{aligned}$$

Here we used the standard fact that Variance of mean of s independent copies of a random variable is $(1/s)$ times the variance. By picking $s = 3/\varepsilon^2$, we obtain

$$\mathbf{E}[\|\boldsymbol{Y} - AB\|_\mathsf{F}^2] = (\varepsilon^2/3)(\|A\|_\mathsf{F}^2 \|B\|_\mathsf{F}^2 - \|AB\|_\mathsf{F}^2) \leq (\varepsilon^2/3)\|A\|_\mathsf{F}^2 \|B\|_\mathsf{F}^2.$$

By Markov inequality, $\mathbf{Pr}[\|\boldsymbol{Y} - AB\|_\mathsf{F}^2 \geq \varepsilon^2 \|A\|_\mathsf{F}^2 \|B\|_\mathsf{F}^2] \leq 1/3$. Therefore with probability $\geq 2/3$, $\|\boldsymbol{Y} - AB\|_\mathsf{F} \leq \varepsilon \|A\|_\mathsf{F} \|B\|_\mathsf{F}$. So, the sampling matrix S has $3/\varepsilon^2$ rows, each row has exactly one nonzero value. The nonzero value is at coordinate i with probability $\|A_{*i}\|_2^2/\|A\|_\mathsf{F}^2$ and is equal to $\sqrt{\|A\|_\mathsf{F}^2/\|A_{*i}\|_2^2}$. Sampling probabilities $p_1, \ldots, p_n$ can be computed in time $O(\mathrm{nnz}(A))$ and the matrices AS^T and SB can be computed in $O(n/\varepsilon^2)$ time as we just need to collect entries of $O(1/\varepsilon^2)$ columns and rows of A and B respectively.