Problem 1: Sparse Regression  (12 points)

For any $1 \leq i_1 < i_2 < \cdots < i_k \leq d$, let $U_{i_1,i_2,\ldots,i_k} = \{ Ax - b \mid x_i = 0 \text{ if } i \neq i_1, \ldots, i_k \}$. Note that since the $x$ here has at most $k$ non-zeros entries lie on $i_1, i_2, \ldots, i_k$, we have $U_{i_1,i_2,\ldots,i_k}$ is a $(k+1)$-dimensional vector space. Hence, from the lecture we have that a Gaussian matrix $S$ of $s \times n$ i.i.d Gaussian random variables $N(0,1/s)$ where $s = O((k + \log(1/\delta))/\varepsilon^2)$ will be a $(1 + \varepsilon)$-subspace embedding of $U_{i_1,i_2,\ldots,i_k}$ with probability at least $1 - \delta$.

Let $U = \{ Ax - b \mid x \in \mathbb{R}^d, \|x\|_0 \leq k \}$, then we have $U = \bigcup_{i=1}^{d \choose k} U_i$ where each $U_i$ corresponds to one choice of $1 \leq i_1 < i_2 < \cdots < i_k \leq d$ among all $d \choose k$ choices. By setting $\delta = 1/(10 \cdot (d \choose k))$ and taking a union bound over all $U_i$, we get that with probability at least 0.9, $S$ is a $(1 \pm \varepsilon)$-subspace embedding of $U$, which means we have

$$(1 - \varepsilon)\|Ax - b\|_2 \leq \|S(Ax - b)\|_2 \leq (1 + \varepsilon)\|Ax - b\|_2$$

for any $k$-sparse $x$.

Suppose that $x' = \text{argmin}_x \text{ is } k\text{-sparse } \|S(Ax - b)\|_2$ and $x^* = \text{argmin}_x \text{ is } k\text{-sparse } \|Ax - b\|_2$. From the above we have that

$$\|Ax' - b\|_2 \leq (1 + \varepsilon)\|SAx' - Sb\|_2 \leq (1 + \varepsilon)\|SAx^* - Sb\|_2 \leq (1 + O(\varepsilon))\|Ax^* - b\|_2.$$

Finally we compute the number of rows needed for $S$. Since $\delta = 1/(10 \cdot (d \choose k))$ we have

$$\frac{k + \log(1/\delta)}{\varepsilon^2} \leq O\left(\frac{k + \log(d/k)}{\varepsilon^2}\right) \leq O\left(\frac{k\log(d/k)}{\varepsilon^2}\right)$$

which means that $O\left(\frac{k\log(d/k)}{\varepsilon^2}\right)$ is enough.

Problem 2: Gaussian Subspace Embeddings with Exactly $d$ Rows  (13 points)

(1) Suppose that $S$ has fewer than $d$ rows. Since $SA$ has $d$ columns and fewer than $d$ rows, we have that $\text{rank}(A) < d$. Then we have that there must exist some $y \in \mathbb{R}^d$ such that $SAy = 0$. However, since $A$ is a $n \times d$ matrix with $\text{rank}(A) = d$. Then we have that $Ay \neq 0$, which is a contradiction.
(2) Without loss of generality, we can assume that $A$ has orthonormal columns. Then from the property of Gaussian random variables, we have that each entry of $SA$ is also drawn from standard Gaussian distribution $N(0, 1)$.

Now, as mentioned in the hint, for every diagonal entry of $(SA)_{ii}$, we have that
\[ \Pr[|(SA)_{ii} - 1| \leq 1/poly(d)] \geq \Omega(1/poly(d)) = e^{-\Theta(\log d)} \]
and for every off-diagonal entry $(SA)_{ij}$, we have that
\[ \Pr[|(SA)_{ij}| \leq 1/poly(d)] \geq \Omega(1/poly(d)) = e^{-\Theta(\log d)} \]

Recall that in lecture 1 we have shown that the entries of $SA$ are independent. Hence, we have that with probability at least $e^{-\Theta(\log d)}$ we can write $SA = I + T$, where $I$ is a $d \times d$ identity matrix and all the entries in $T$ are at most $1/poly(d)$. Under this condition, we have that for any unit vector $x \in \mathbb{R}^d$, $SAx = Ix + Tx = x + Tx$ and
\[ \|x\|_2 - \|T\|_2 \leq \|x\|_2 - \|Tx\|_2 \leq \|SAx\|_2 \leq \|x\|_2 + \|Tx\|_2 \leq \|x\|_2 + \|T\|_2 , \]

Note that since the entries of $T$ are all in $[-1/poly(d), 1/poly(d)]$, we have that $\|T\|_2 \leq \|T\|_F = 1/poly(d)$. From this we have
\[ 1 - 1/poly(d) \leq \|SAx\|_2 \leq 1 + 1/poly(d) , \]
which means that $S$ is a $(1 \pm 1/poly(d))$-subspace embedding.

**Problem 3: Active Regression** (13 points)
We first define our sampling matrix $S$.

**Definition 1** Given a parameter number $k$, the sampling matrix $S \in \mathbb{R}^{k \times n}$ that samples $k$ rows of a matrix $A$ is defined as follows. For each row of $S$, we independently and uniformly pick an index $i \in [n]$ and set the value of this entry is $\sqrt{n}/k$, then set the values of the other entries in this row as 0.

We will use the matrix Chernoff’s bound to show that if $k = O(d \log(d)/\varepsilon^2)$, $SA$ is actually a $(1 \pm \varepsilon)$-subspace embedding of the matrix $A$. Let $i(j)$ denote the index of the sampled row in the $j$-th trial and $X_j = I_d - nA^T_{i(j)}A_{i(j)}$. Then, we have that
\[ \mathbb{E}[X_j] = I_d - \sum_i \frac{1}{n} \cdot nA^T_i A_i = 0 \]
since $A$ has orthonormal columns.

Next, by triangle inequality we have that
\[ \|X_j\|_2 \leq \|I_d\| + n\|A^T_{i(j)} A_{i(j)}\|_2 = O(d) \]
from the assumption that $\|A_i\|_2^2 = O(d/n)$.

Lastly, for every $j$, we have that

$$\mathbb{E} \left[ X_j^T X_j \right] = I_d - 2n\mathbb{E} \left[ A_{i(j)}^T A_{i(j)} \right] + n^2\mathbb{E} \left[ A_{i(j)}^T A_{i(j)} A_{i(j)}^T A_{i(j)} \right]$$

$$= I_d - 2n \cdot \frac{1}{n} I_d + n^2\mathbb{E} \left[ \|A_{i(j)}\|_2^2 A_{i(j)}^T A_{i(j)} \right]$$

$$\leq I_d - 2I_d + dI_d \leq dI_d .$$

Note that $1/k \cdot (\sum_j X_j) = I_d - A^T S^T S A$. Hence, from the matrix Chernoff’s bound we have that

$$\Pr \left[ \|I_d - A^T S^T S A\|_2 > \varepsilon \right] \leq 2d \cdot \exp \left( \frac{-k\varepsilon^2}{d + d \varepsilon} \right) \leq 1/10$$

when $k = O(d \log(d)/\varepsilon^2)$. Recall that for a symmetric matrix $W$ we have that $\|W\|_2 = \max_{\|x\|_2=1} x^T W x$. Hence we get that it means $\|SA\|_2 = (1 \pm \varepsilon)\|Ax\|_2$ for all $x \in \mathbb{R}^d$.

Suppose that $S$ is the sampling matrix that uniformly samples $O(d \log d)$ rows of $A$. Then we can see that to solve the regression problem $\min_{x \in \mathbb{R}^d} \|SA - Sb\|_2$, we only need to read $O(d \log d)$ entries of $b$. And from the above process we have that $S$ is a $(1 + O(1))$-subspace embedding of $A$ with probability at least 0.95. Now, let $x_c = \arg \min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$, we have

$$\|Ax_c - b\|_2 \leq \|Ax_c - Ax^*\|_2 + \|Ax^* - b\|_2 \leq \|Ax^* - b\|_2 + O(\|SAx_c - SAx^*\|_2) .$$

Also we have that

$$\|SAx_c - SAx^*\|_2 \leq \|SAx_c - Sb\|_2 + \|Sb - SAx^*\|_2 \leq 2\|Sb - SAx^*\|_2 ,$$

The only remaining thing is to bound $\|Sb - SAx^*\|_2$. In fact, let $z = S(Ax^* - b)$, we have that

$$\mathbb{E} \left[ \|S(Ax^* - b)\|_2^2 \right] = \sum_i \mathbb{E}[z_i^2] = \frac{n}{k} \sum_{i=1}^k \frac{1}{n} \sum_{j=1}^n (Ax_j^* - b)^2 = \|Ax^* - b\|_2^2$$

Since we have that $\mathbb{E} \left[ \|Sb - SAx^*\|_2^2 \right] = \|Ax^* - b\|_2^2$, then by Markov’s inequality we have that with probability at least 0.95, $\|Sb - SAx^*\|_2^2 \leq 20 \|Ax^* - b\|_2^2$, which means that $\|SAx_c - SAx^*\|_2 \leq O(\|Ax^* - b\|_2)$.

Put everything together and by taking a union bound, we have that with probability at least 0.9

$$\|Ax_c - b\|_2 \leq C\|Ax^* - b\|_2$$

for some constant $C$, which is what we need.

**Problem 4: Fast High Probability Matrix Product**  (12 points)

We will use the following lemmas.
Lemma 2 Let $S$ be a $k \times n$ matrix of i.i.d normal random variables drawn from $N(0, 1/k)$ where $k = O(\log(1/\delta)/\varepsilon^2)$. Then given two unit vectors $u, v \in \mathbb{R}^n$, we have with probability at least $1 - \delta$,

$$|\langle Sx, Sy \rangle - \langle x, y \rangle| \leq \varepsilon .$$

We have

$$\langle Sx, Sy \rangle = \frac{\|Sx + Sy\|_2^2 - \|Sx - Sy\|_2^2}{4}$$

and

$$\langle x, y \rangle = \frac{\|x + y\|_2^2 - \|x - y\|_2^2}{4}$$

As we did in class, by Johnson-Lindenstrauss lemma, we have with probability at least $1 - \delta$ we have that $\|S(x + y)\|_2^2 = (1 \pm \frac{1}{2} \varepsilon)\|x + y\|_2^2$ and $\|S(x - y)\|_2^2 = (1 \pm \frac{1}{2} \varepsilon)\|x - y\|_2^2$. Hence we have that

$$|\langle Sx, Sy \rangle - \langle x, y \rangle| = \frac{\|Sx + Sy\|_2^2 - \|x + y\|_2^2}{4} + \frac{\|x - y\|_2^2 - \|Sx - Sy\|_2^2}{4} \leq \frac{1}{2} \varepsilon \left( \frac{\|x + y\|_2^2}{4} + \frac{\|x - y\|_2^2}{4} \right) \leq \varepsilon$$

Lemma 3 Let $S$ be a $k \times n$ matrix of i.i.d normal random variables drawn from $N(0, 1/k)$ where $k = O(\log n/\varepsilon^2)$. Then for any matrix $A, B \in \mathbb{R}^{n \times n}$, we have with probability at least $1 - 1/n$,

$$\|A^T S^T SB - A^T B\|_F \leq \varepsilon \|A\|_F \|B\|_F .$$

let $A_i$ denote the $i$-th column of $A$ and $B_j$ denote the $j$-column of $B$. For a Gaussian matrix $a$ with $O(\log n/\varepsilon^2)$ rows, from Lemma 2 we have that with probability at least $1 - 1/n$,

$$|\langle A_i, SB_j \rangle - \langle A_i, B_j \rangle| \leq \varepsilon \|A_i\|_2 \|B_j\|_2 .$$

Taking a union bound of all $(i, j)$ pair we have that with probability at least $1 - 1/n$,

$$\|A^T S^T SB - A^T B\|_F^2 \leq \sum_i \sum_j \varepsilon^2 \|A_i\|_2^2 \|B_j\|_2^2 = \varepsilon^2 \|A\|_F^2 \|B\|_F^2 ,$$

which means

$$\|A^T S^T SB - A^T B\|_F \leq \varepsilon \|A\|_F \|B\|_F .$$

Lemma 4 Let $S$ be a $k \times n$ Count-Sketch matrix of where $k = O(1/(\delta \varepsilon^2))$. Then for any matrix $A \in \mathbb{R}^{n \times d}$, we have with probability at least $1 - \delta$,

$$\|SA\|_F^2 = (1 \pm \varepsilon)\|A\|_F^2 .$$

The proof was given in the Problem 3 in [https://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall17/ps1sol.pdf](https://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall17/ps1sol.pdf) by replace $r$ with $O(1/(\delta \varepsilon^2))$. 4
Back to the original problem. Now we design the $S = S_1 S_2$ where $S_1$ is the Gaussian matrix with $O(\log n)$ rows, and $S_2$ is the CountSketch matrix with $O(n^{0.99})$ rows (where both correspond to $\varepsilon = 1/100$ and $\delta = 1/(3n^{0.99})$ in Lemma 2 and Lemma 3). We first have with probability at least $1 - 1/(3n^{0.99})$

$$\|A^T S_2^T S_1^T S_1 S_2 B - A^T S_2^T S_2 B\|_F \leq \frac{1}{100} \|A^T S_2^T\|_F \|S_2 B\|_F.$$ 

Since $S_2$ is a Count-Sketch matrix, from Lemma 4 we have that with probability at least $1 - 1/(3n^{0.99})$, $\|A^T S_2^T\|_F^2 = (1 \pm 0.01) \|A\|_F^2$ and $\|S_2 B\|_F^2 = (1 \pm 0.01) \|B\|_F^2$, which means that

$$\|A^T S_2^T S_1^T S_1 S_2 B - A^T S_2^T S_2 B\|_F \leq \frac{1}{100} \|A^T S_2^T\|_F \|S_2 B\|_F \leq \frac{1}{50} \|A\|_F \|B\|_F.$$ 

Next, from Lemma 2 we have that with probability at least $1 - 1/(3n^{0.99})$,

$$\|A^T S_2^T S_2 B - A^T B\|_F \leq \frac{1}{100} \|A\|_F \|B\|_F$$

Putting these two things together and by triangle inequality we have that with probability at least $1 - 1/n^{0.99}$ (after taking a union bound),

$$\|A^T S_2^T S_1^T S_1 S_2 B - A^T B\|_F \leq \frac{1}{10} \|A\|_F \|B\|_F.$$ 

Now we consider the time complexity of the above sketching matrix. First, since $S_2$ is a CountSketch matrix, hence we can use $O(n^2)$ time to get $S_2 A$ and $S_2 B$. Next, since $S_2 A$ and $S_2 B$ have $n^{0.99}$ rows and $S_1$ has $O(\log n)$ rows. Hence we can get $S_1 S_2 A$ and $S_1 S_2 B$ in time $O(\log n \cdot n^{1.99}) = O(n^2)$, which is a total $O(n^2)$ time.