15-859 : Homework 2 Solutions

1 Question 1

As we assumed A to be full rank, we have that $A^\mathsf{T}A$ is full rank and therefore invertible which implies that $x^* = (A^\mathsf{T}A)^{-1}b$ is an optimal solution of cost 0. Unfortunately computing the matrix $A^\mathsf{T}A$ is slow. As in the problem of high precision regression, we first compute a preconditioner matrix R as follows. Let $\varepsilon_0 > 0$ be a small constant. Let S be a countsketch matrix with $O(d^2/\varepsilon_0^2) = O(d^2)$ rows. Therefore we have $\|SAx\|_2 = (1 \pm \varepsilon_0)\|Ax\|_2$ with constant probability.

Let $SA = QR^{-1}$ be the the decomposition of SA which can be computed in $O(d^4)$ time and R can be inverted in $O(d^3)$ time. Therefore $SAR = Q$. We now have the following properties:

- $\sigma_{\min}(R^{-1}) = \sigma_{\min}(QR^{-1}) = \sigma_{\min}(SA) \geq (1-\varepsilon_0)\sigma_{\min}(A)$ and similarly $\sigma_{\max}(R^{-1}) \leq (1+\varepsilon_0)\sigma_{\max}(A)$. These properties can be seen by noting that for any matrix A, $\sigma_{\max}(A) = \max \|Ax\|_2/\|x\|_2$ and $\sigma_{\min}(A) = \min \|Ax\|_2/\|x\|_2$.
- For any vector x, $\frac{1}{1+\varepsilon_0}\|x\|_2 = \frac{1}{1+\varepsilon_0}\|SARx\|_2 \leq \|ARx\|_2 \leq \frac{1}{(1-\varepsilon_0)}\|SARx\|_2 = \frac{1}{(1-\varepsilon_0)}\|x\|_2$. Therefore, $\sigma_{\max}(AR) < 1/(1-\varepsilon_0)$ and $\sigma_{\min}(AR) \geq 1/(1+\varepsilon_0)$ which implies that $\kappa(AR) \leq (1+\varepsilon_0)/(1-\varepsilon_0)$.

Now let us relate $\|R^\mathsf{T}A^\mathsf{T}ARx - R^\mathsf{T}b\|_2$ with $\|A^\mathsf{T}ARx - b\|_2$ and show that solution to one would be a good solution to other problem. For any vector x, we have the following

$$\begin{aligned}\|A^\mathsf{T}ARx - b\|_2 &\leq \frac{\|R^\mathsf{T}A^\mathsf{T}ARx - R^\mathsf{T}b\|_2}{\sigma_{\min}(R^\mathsf{T})} = \sigma_{\max}(R^{-1})\|R^\mathsf{T}A^\mathsf{T}ARx - R^\mathsf{T}b\|_2 \\ &\leq (1+\varepsilon_0)\sigma_{\max}(A)\|R^\mathsf{T}A^\mathsf{T}ARx - R^\mathsf{T}b\|_2.\end{aligned}$$

So a good solution for $\min_x \|R^\mathsf{T}A^\mathsf{T}ARx - R^\mathsf{T}b\|_2$ is also a good solution for our original problem. Let $x_0 \leftarrow 0$ and

$$x_{m+1} \leftarrow x_m + R^\mathsf{T}A^\mathsf{T}AR(R^\mathsf{T}b - R^\mathsf{T}A^\mathsf{T}ARx_m).$$

This is the same update rule as in high-precision regression from the class. Now we bound the error of x_{m+1} in terms of x_m. Let x^* be the optimal solution for the problem.

$$\begin{aligned}R^\mathsf{T}A^\mathsf{T}AR(x_{m+1} - x^*) &= R^\mathsf{T}A^\mathsf{T}AR(x_m + R^\mathsf{T}A^\mathsf{T}AR(R^\mathsf{T}b - R^\mathsf{T}A^\mathsf{T}ARx_m)) - R^\mathsf{T}A^\mathsf{T}ARx^* \\ &= (R^\mathsf{T}A^\mathsf{T}AR - (R^\mathsf{T}A^\mathsf{T}AR)^2)x_m + R^\mathsf{T}A^\mathsf{T}ARR^\mathsf{T}b - R^\mathsf{T}A^\mathsf{T}ARx^*.\end{aligned}$$

By normal equations, $(R^\mathsf{T}A^\mathsf{T}AR)^2x^* = R^\mathsf{T}A^\mathsf{T}ARR^\mathsf{T}b$. Using this fact, we obtain that

$$R^\mathsf{T}A^\mathsf{T}AR(x_{m+1} - x^*) = (R^\mathsf{T}A^\mathsf{T}AR - (R^\mathsf{T}A^\mathsf{T}AR)^2)(x_m - x^*).$$

Here we use the fact that $R^\mathsf{T}A^\mathsf{T}AR$ is well conditioned as the matrix AR is well conditioned to obtain that $\sigma_{\max}(R^\mathsf{T}A^\mathsf{T}AR - (R^\mathsf{T}A^\mathsf{T}AR)^2) = O(\varepsilon_0)$. Which therefore implies that

$$\|R^\mathsf{T}A^\mathsf{T}AR(x_{m+1} - x^*)\|_2 \leq O(\varepsilon_0)\|x_m - x^*\|_2.$$

But now again using the fact that $\sigma_{\min}(R^\mathsf{T}A^\mathsf{T}AR) \geq (1-\varepsilon_0)^2$, we obtain that

$$\|R^\mathsf{T}A^\mathsf{T}AR(x_{m+1}-x^*)\|_2 \leq \frac{O(\varepsilon_0)}{(1-\varepsilon_0)^2}\|R^\mathsf{T}A^\mathsf{T}AR(x_m-x^*)\|_2 \leq O(\varepsilon_0)\|R^\mathsf{T}A^\mathsf{T}AR(x_m-x^*)\|_2$$

for small enough constant ε_0. We therefore obtain $\|R^\mathsf{T}A^\mathsf{T}AR(x_t-x^*)\|_2 \leq O(\varepsilon_0)^t\|R^\mathsf{T}A^\mathsf{T}AR(x_0-x^*)\|_2 = O(\varepsilon_0)^t\|R^\mathsf{T}b\|_2$. Here we used the fact that $x_0=0$ and for optimum x^*, $R^\mathsf{T}A^\mathsf{T}ARx^* = R^\mathsf{T}b$. But $\|R^\mathsf{T}b\|_2 \leq \|R\|_2\|b\|_2 \leq \frac{1}{\sigma_{\min}(R^{-1})}\|b\|_2 \leq \frac{1}{(1-\varepsilon_0)\sigma_{\min}(A)}\|b\|_2$. Using all these we obtain

$$\begin{aligned}\|A^\mathsf{T}ARx_t - b\|_2 &\leq (1+\varepsilon_0)\sigma_{\max}(A)\|R^\mathsf{T}A^\mathsf{T}ARx_t - R^\mathsf{T}b\|_2\\ &\leq \frac{1+\varepsilon_0}{1-\varepsilon_0}\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}O(\varepsilon_0)^t\|b\|_2\\ &\leq 2\kappa\|b\|_2 O(\varepsilon_0)^t.\end{aligned}$$

Therefore for $t=\log(\kappa/\varepsilon)$, we obtain that $\|A^\mathsf{T}ARx_t - b\|_2 \leq \varepsilon\|b\|_2$. This implies Rx_t is a solution of additive error at most $\varepsilon\|b\|_2$.

Note that preconditioning the matrix to have condition number close to 1 allowed us to obtain that the error decreases in every iteration which we couldn't have if we applied the "descent method" directly on matrix A.

Time Complexity We can compute SA in time $\text{nnz}(A)$ and its QR factorization in $O(d^4)$ time. We can compute the vector $R^\mathsf{T}A^\mathsf{T}ARR^\mathsf{T}b$ in time $O(\text{nnz}(A)+d^2)$. Every iteration then involves computing $(R^\mathsf{T}A^\mathsf{T}AR)^2x$ which can be done in $O(\text{nnz}(A)+d^2)$ by multiplying the vector x from left to right. So the total time complexity is

$$O((\text{nnz}(A)+d^2)\log(\kappa/\varepsilon)+d^4).$$

Question 2

Let $A_i \in \mathbb{R}^d$ be the ith row of the matrix A. We have $A^\mathsf{T}A = \sum_i A_iA_i^\mathsf{T}$. If $r(j)$ be the jth sampled row. Define $\boldsymbol{X}_j = (1/q_{r(j)})A_{r(j)}A_{r(j)}^\mathsf{T} - A^\mathsf{T}A$. It is easy to see that $(1/r)\sum_j \boldsymbol{X}_j = A^\mathsf{T}S^\mathsf{T}SA - A^\mathsf{T}A$. Now we use Matrix Chernoff bounds to upperbound the probability that $\|(1/r)\sum_j \boldsymbol{X}_j\|_2 \geq \varepsilon\|A\|_2^2$.

$$\begin{aligned}\mathbf{E}[\boldsymbol{X}_j] &= \mathbf{E}[\frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T} - A^\mathsf{T}A] = \mathbf{E}[\frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T}] - A^\mathsf{T}A\\ &= \sum_{i=1}^d \frac{1}{q_i}A_iA_i^\mathsf{T}\,\mathbf{Pr}[r(j)=i] - A^\mathsf{T}A = \sum_{i=1}^d \frac{1}{q_i}A_iA_i^\mathsf{T}q_i - A^\mathsf{T}A = 0.\end{aligned}$$

We now bound $\|\boldsymbol{X}_j\|_2$.

$$\begin{aligned}\|\boldsymbol{X}_j\|_2 &= \|\frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T} - A^\mathsf{T}A\|_2 \leq \|\frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T}\|_2 + \|A^\mathsf{T}A\|_2\\ &\leq \frac{\|A_{r(j)}\|_2^2}{q_{r(j)}} + \|A\|_2^2 \leq \frac{\|A\|_\mathsf{F}^2}{\beta} + \|A\|_2^2.\end{aligned}$$

We finally bound $\| \mathbf{E}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]\|_2$.

$$\begin{aligned}
\mathbf{E}[\boldsymbol{X}_j^\mathsf{T}\boldsymbol{X}_j] &= \mathbf{E}[\frac{1}{q_{r(j)}^2}A_{r(j)}A_{r(j)}^\mathsf{T}A_{r(j)}A_{r(j)}^\mathsf{T} - \frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T}A^\mathsf{T}A - \frac{1}{q_{r(j)}}A^\mathsf{T}AA_{r(j)}A_{r(j)}^\mathsf{T} + A^\mathsf{T}AA^\mathsf{T}A] \\
&= \mathbf{E}[\frac{\|A_{r(j)}\|_2^2}{q_{r(j)}}\frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T}] - 2A^\mathsf{T}AA^\mathsf{T}A + A^\mathsf{T}AA^\mathsf{T}A \\
&\preceq \frac{\|A\|_\mathsf{F}^2}{\beta}\,\mathbf{E}[\frac{1}{q_{r(j)}}A_{r(j)}A_{r(j)}^\mathsf{T}] - A^\mathsf{T}AA^\mathsf{T}A \\
&= \frac{\|A\|_\mathsf{F}^2}{\beta}A^\mathsf{T}A - A^\mathsf{T}AA^\mathsf{T}A.
\end{aligned}$$

Therefore $\| \mathbf{E}[\boldsymbol{X}_j^\mathsf{T}\boldsymbol{X}_j]\|_2 \le \|A\|_\mathsf{F}^2\|A\|_2^2/\beta + \|A\|_2^4 \le 2\|A\|_\mathsf{F}^2\|A\|_2^2/\beta$. Plugging all these bounds into Matrix-Chernoff, we obtain

$$\mathbf{Pr}[\|(1/r)\sum_i \boldsymbol{X}_j\|_2 \ge \varepsilon\|A\|_2^2] \le 2d\exp(-\frac{r\varepsilon^2\|A\|_2^4}{2\|A\|_\mathsf{F}^2\|A\|_2^2/\beta + \frac{2\|A\|_\mathsf{F}^2\varepsilon\|A\|_2^2}{\beta}}) = 2d\exp(-\frac{r\varepsilon^2}{2\rho/\beta + 2\rho\varepsilon/\beta}).$$

By choosing $r = O(\frac{\rho}{\beta\varepsilon^2}\ln(d))$ we obtain that $\|A^\mathsf{T}S^\mathsf{T}SA - A^\mathsf{T}A\|_2 \le \varepsilon\|A\|_2^2$ with probability $\ge 9/10$.

2 Question 3

Let $U^* \in \mathbb{R}^{n\times k}, V^* \in \mathbb{R}^{k\times d}$ be such that

$$\|U^*\cdot V^* - A\|_{1,2} = \min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}}\|U\cdot V - A\|_{1,2} = \text{Opt}.$$

Consider the optimization problem

$$\min_{U\in\mathbb{R}^{n\times k}}\|U\cdot V^* - A\|_{1,2}.$$

Clearly the above problem has the optimum value of Opt. Suppose there is a matrix R with $O((k+\log(d))/\varepsilon^2)$ columns such that for all matrices U,

$$\|UV^*R - AR\|_{1,2} = (1\pm\varepsilon)\|UV^* - A\|_{1,2}. \tag{1}$$

Suppose $\tilde{U} = \arg\min_U \|UV^*R - AR\|_{1,2}$. Now we claim that $\|AR(\tilde{U}V^*R)^\dagger(\tilde{U}V^*R) - AR\|_{1,2} \le \|\tilde{U}V^*R - AR\|_{1,2}$.

Proof. Let $(AR)_i$ be the ith row of the matrix AR. For any matrix B, we have $\|A_i - A_iB^\dagger B\|_2 \le \|A_i - B_i\|_2$ because $A_iB^\dagger B$ is the closest vector to A_i in the row span of B and B_i is a vector in the row span of B. Therefore, we have for all rows i

$$\|(AR)_i(\tilde{U}V^*R)^\dagger\tilde{U}V^*R - (AR)_i\|_2 \le \|(\tilde{U}V^*R)_i - (AR)_i\|_2.$$

Summing over i, we obtain that

$$\|AR(\tilde{U}V^*R)^\dagger\tilde{U}V^*R - AR\|_{1,2} \le \|\tilde{U}V^*R - AR\|_{1,2}. \qquad \square$$

We finally have

$$\begin{aligned}
\|AR(\tilde{U}V^*R)^\dagger\tilde{U}V^* - A\|_{1,2} &\leq \frac{1}{1-\varepsilon}\|AR(\tilde{U}V^*R)^\dagger\tilde{U}V^*R - AR\|_{1,2} && \text{By (1)}\\
&\leq \frac{1}{1-\varepsilon}\|\tilde{U}V^*R - AR\|_{1,2}\\
&\leq \frac{1}{1-\varepsilon}\|U^*V^*R - AR\|_{1,2} && \text{By optimality of } \tilde{U}\\
&\leq \frac{1+\varepsilon}{1-\varepsilon}\|U^*V^* - A\|_{1,2} && \text{By (1)}\\
&= (1+O(\varepsilon))\text{Opt}.
\end{aligned}$$

Therefore we have

$$\min_Y \|ARY - A\|_{1,2} \leq \|AR(\tilde{U}V^*R)^\dagger\tilde{U}V^* - A\|_{1,2} \leq (1+O(\varepsilon))\text{Opt}.$$

Now we show that if R is a matrix of independent gaussian entries, the guarantee (1) holds. In class, we saw the following guarantee, if A is a matrix of rank at most d, and if R is a $k \times n$ matrix of i.i.d normal entries of mean 0 and variance $1/k$ with $k = O(\frac{d+\log(1/\delta)}{\varepsilon^2})$ rows, then with probability $\geq 1-\delta$, for all vectors x

$$\|RAx\|_2^2 \in (1\pm\varepsilon)\|Ax\|_2^2.$$

We can rewrite the above guarantee for a subspace embedding on the right. Now V^* is a fixed but unknown matrix of rank at most k. Let $\mathcal{V}_i$ be the subspace spanned by rows of V^* and A_i(ith row of A). The subspace $\mathcal{V}_i$ has a dimension at most $k+1$. Let R be a *right* subspace embedding matrix with $O((k+1+\log(1/\delta))/\varepsilon^2)$ columns. Then, with probability $\geq 1-\delta$ for all vectors $v \in \mathcal{V}_i$, we have

$$\|v^\mathsf{T}R\|_2^2 \in (1\pm\varepsilon)\|v^\mathsf{T}\|_2^2.$$

For any vector x, the row vector $x^\mathsf{T}V^* - A_i$ is in subspace $\mathcal{V}_i$. Therefore with probability $\geq 1-\delta$,

$$\|x^\mathsf{T}V^*R - A_iR\|_2 \in (1\pm\varepsilon)\|x^\mathsf{T}V^* - A\|_2 \text{ for all vectors } x.$$

We can now union bound over all $i = 1,\ldots,d$ and have that with probability $\geq 1-\delta\cdot d$,

$$\|x^\mathsf{T}V^*R - A_iR\|_2 \in (1\pm\varepsilon)\|x^\mathsf{T}V^* - A_i\|_2 \text{ for all vectors } x \text{ and } i = 1,\ldots,d.$$

This gives that, with probability $\geq 1-\delta d$, for all matrices U,

$$\begin{aligned}
\|UV^*R - AR\|_{1,2} &= \sum_i \|U_iV^*R - A_iR\|_2\\
&\in \sum_i (1\pm\varepsilon)\|U_iV^* - A_i\|_2\\
&\in (1\pm\varepsilon)\|UV^* - A\|_{1,2}.
\end{aligned}$$

Now picking $\delta = 1/10d$, gives that if R has $O((k+\log(d))/\varepsilon^2)$ columns, with probability $\geq 9/10$,

$$\|UVR - AR\|_{1,2} \in (1\pm\varepsilon)\|UV - A\|_{1,2}$$

for all matrices U and therefore

$$\min_Y \|ARY - A\|_{1,2} \leq (1+\varepsilon)\text{Opt}.$$

It is interesting that a *single* gaussian matrix R with $O((k+\log(d))/\varepsilon^2)$ columns can be simultaneously a *right* subspace embedding for *d different* $k+1$ dimensional subspaces with good probability but on the other hand the gaussian matrix needs to have $O(d/\varepsilon^2)$ columns to be a subspace embedding for a *single* d-dimensional subspace.