15-851 Algorithms for Big Data — Spring 2024

Problem Set 2

Due: Thursday, Februrary 22, 11:59pm

Please see the following link for collaboration and other homework policies:
http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15851-spring24/grading.pdf

Problem 1 (1) Let $A = UR$ where U is an orthonormal basis of A. Then we have

$$\sup_x \frac{\|Ax\|_\infty^2}{\|Ax\|_2^2} = \sup_x \frac{\|URx\|_\infty^2}{\|URx\|_2^2} = \sup_{x:\|Rx\|_2=1} \frac{\|URx\|_\infty^2}{\|URx\|_2^2} = \sup_{y:\|y\|_2=1} \|Uy\|_\infty^2,$$

notice that for each $i \in [n]$, we have $\langle U_i, y\rangle^2 \le \|U_i\|_2^2 = \ell_i(U) = \ell_i(A)$, which means that $\mu(A) \ge \sup_x \frac{\|Ax\|_\infty^2}{\|Ax\|_2^2}$. On the other hand, let $j = \text{argmax}_i \ell_i(U)$, then let $y^T = U_i/\|U_i\|_2$, we have $\|Uy\|_\infty^2 \ge \ell_j(U)$, from which we have that $\mu(A) \le \sup_x \frac{\|Ax\|_\infty^2}{\|Ax\|_2^2}$.

(2) Since we have that $\sum_i \ell_i(A) = d$, we can get that $\max_i \ell_i(A) \ge \left(\sum_i \ell_i(A)\right)/n = d/n$. Consider the example where we stack n/d $d \times d$ identity matrix I_d to form the matrix A. That is,

$$A = \begin{bmatrix} I_d \\ I_d \\ \vdots \\ I_d \end{bmatrix}$$

We can see that A has orthogonal columns, after normalization we have that each row of A has leverage score d/n.

(3) Suppose that A is a matrix that satisfies the condition in (2). Then we construct the matrix B where $B_1 = A_1$ and $B_i = -A_i$ for $i \ge 2$. Then we have that for every x, $\|Ax\|_2 = \|Bx\|_2$ and $\|Ax\|_\infty = \|Bx\|_\infty$, which means that $\mu(A) = \mu(B) = d/n$.

We next consider the matrix $C = [A\ B]$. Note that from part (1) we immediately have $\mu(C) \le 1$ as $\|Ax\|_\infty \le \|Ax\|_2$ for every $x \in \mathbb{R}^d$. On the other hand, let x be the vector has the form $(y, -y)$ where $A_1^T y \ne 0$. We then have $(Cx)_1 \ne 0$ while $(Cx)_i = 0$ for $i \ge 2$. This means that $\mu(C) \ge 1$, from which we have that $\mu(C) = 1$.

4) For a matrix C where C has orthonormal columns, note that we have that $\mu(C) = \max_i \|C_i\|_2^2$. Back to our problem, from the assumption we have that all of the matrices A, B, C have orthonormal columns, which means $\mu(A) = \max_i \|A_i\|_2^2, \mu(B) = \max_i \|B_i\|_2^2, \mu(C) = \max_i \|C_i\|_2^2$. Since $C = [A\ B]$, we have that

$$\max\left(\max_i \|A_i\|_2^2, \max_i \|B_i\|_2^2\right) \le \max_i \|C_i\|_2^2 \le \max_i \|A_i\|_2^2 + \max_i \|B_i\|_2^2$$

which means that

$$\max\left(\mu(A), \mu(B)\right) \le \mu(C) \le \mu(A) + \mu(B)$$

Problem 2 (1) Let $A = U\Sigma V$ be the singular value decomposition of A. Then we have that

$$\begin{aligned} AR &= A(I - V_k V_k^T + V_k \Sigma_k^{-1} V_k^T) \\ &= A - A_k + U_k V_k^T \\ &= U\Sigma' V. \end{aligned}$$

where $\Sigma' = \begin{bmatrix} I_k & \\ & \Sigma_{-k} \end{bmatrix}$ is a diagonal matrix where all the diagonal entries are $O(1)$ (from the assumption we have that $\sigma_{k+1} = O(1)$). From this we have that

$$\|ARx\|_2 = \Theta(1)\|x\|_2 \ ,$$

which is what we need.

(2) Let $A = U\Sigma V^T$. Then we have for a $x \in \mathbb{R}^n$ where $\|x\|_2 = 1$

$$\|ARx\|_2^2 = \|U\Sigma V^T Rx\|_2^2 = \|\Sigma V^T Rx\|_2^2 = \sum_i (\Sigma V^T Rx)_i^2$$

Let $x = \alpha z + \beta w$, where $\langle z, w\rangle = 0$, $\|w\|_2 = 1$, and $\alpha^2 + \beta^2 = 1$. Then, for $i = 1$, we have

$$(\Sigma V^T Rx)_1 = \sigma_1(A) v_1^T (I - zz^T)\beta w + \sigma_1(A)\frac{1}{\lambda} v_1^T z z^T z$$

from the assumption that $\langle v_1, z\rangle \geq 1 - 1/(10\sigma_1(A)^2)$ and $\|a-b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 - 2\langle a, b\rangle$ we have that

$$\|v_1^T - \langle v_1, z\rangle z^T\|_2^2 = \|v_1\|_2^2 + \langle v_1, z\rangle^2 \|z\|_2^2 - 2\langle v_1, z\rangle^3 = O\left(\frac{1}{\sigma_1(A)^2}\right)$$

from which we have that

$$\begin{aligned} |(\Sigma V^T Rx)_1| &= |\sigma_1(A) v_1^T (I - zz^T)\beta w + \sigma_1(A)\frac{1}{\lambda} v_i^T z z^T z| \\ &\leq O(1) \cdot |\beta| + \Theta(1) \cdot |\alpha|. \end{aligned}$$

Next, for $i \geq 2$ we have

$$\begin{aligned} |(\Sigma V^T Rx)_i| &= \left|\sigma_i(A) v_i^T (I - zz^T)\beta w + \sigma_i(A)\frac{1}{\lambda} v_i^T z z^T z\right| \\ &= \left|\sigma_i(A) v_i^T \beta w + \sigma_i(A)\frac{1}{\lambda}\langle v_i, z\rangle \alpha\right| \\ &= \left|\sigma_i(A)\langle v_i, w\rangle \beta + \sigma_i(A)\frac{1}{\lambda}\langle v_i, z\rangle \alpha\right| \\ &\leq \Theta(1) \cdot (|\langle v_i, w\rangle||\beta| + |\langle v_i, z\rangle||\alpha|) \end{aligned}$$

Using the fact that $(a+b)^2 \le 2(a^2+b^2)$ we have

$$\begin{aligned}\|ARx\|_2^2 &= \sum_i (\Sigma V^T Rx)_i^2 \\ &\le O(1)\cdot \beta^2 + \Theta(1)\cdot \alpha^2 + \sum_{i\ge 2} \Theta(1)\cdot (\langle v_i, w\rangle^2 \beta^2 + \langle v_i, z\rangle^2 \alpha^2) \\ &= \Theta(1).\end{aligned}$$

We next consider the other direction. If $|\beta| \le 2|\alpha|$, from the above we have

$$\begin{aligned}|(\Sigma V^T Rx)_1| &= |\sigma_1(A) v_1^T (I - zz^T)\beta w + \sigma_1(A)\frac{1}{\lambda} v_i^T zz^T z| \\ &\ge 0.9\cdot|\alpha| - O(1)\cdot|\beta| \ge \Theta(1).\end{aligned}$$

since $\alpha^2+\beta^2=1$, and on the other case $|\beta| > 2|\alpha|$, from $\sum_i \langle v_i, z\rangle^2 = \sum_i \langle v_i, w\rangle^2 = 1$ and $|\langle v_1, z\rangle| > 1 - O\left(\frac{1}{\sigma_1(A)^2}\right)$, $|\langle v_1, w\rangle| \le O\left(\frac{1}{\sigma_1(A)}\right)$ we have that

$$\sum_i |(\Sigma V^T Rx)_i|^2 = \sum_i \left|\sigma_i(A)\langle v_i, w\rangle \beta + \sigma_i(A)\frac{1}{\lambda}\langle v_i, z\rangle \alpha\right|^2 \ge \Theta(1)\cdot(\beta-\alpha)^2 \ge \Theta(1)\ .$$

from which we can get that in both cases, $\|ARx\|_2^2 \ge \Theta(1)$, which is what we need.

(3) In each iteration of the gradient descent we need to compute $R^T A^T (b - ARx_i)$. Since we have $R = I - zz^T + \frac{1}{\lambda} zz^T$, which means that we can compute Rx_i in $O(n)$ time, $b - ARx_i$ in $O(n)$ time, and then $A^T(b - ARx_i)$ in $\text{nnz}(A)$ time and finally $R^T(A^T(b-ARx_i))$ in $O(n)$ time, from which we can get that the per-iteration time we can have is $O(\text{nnz}(A)+n)$.

Problem 3 We define OPT as

$$\text{OPT} = \min_{\text{rank}-k\ A'} \|A' - A\|_F^2.$$

We first consider the following optimization problem,

$$\min_{U_1,\cdots,U_k \in \mathbb{R}^n} \left\| \sum_{i=1}^k U_i \otimes V_i^* \otimes W_i^* - A \right\|_F^2,$$

it is equivalent to

$$\min_{U\in\mathbb{R}^{n\times k}} \left\| UZ_1 - A^1 \right\|_F^2,$$

where $Z_1 = ((V^*)^T \odot (W^*)^T)$. Notice that $\min_{U\in\mathbb{R}^{n\times k}} \|UZ_1 - A^1\|_F^2 = \text{OPT}$ and the optimum solution of U is U^*. Let $(R^1)^T \in \mathbb{R}^{s\times n^2}$ be a Count-Sketch matrix with $s = \text{poly}(k/\varepsilon)$.

We next consider the following optimization problem,

$$\min_{U\in\mathbb{R}^{n\times k}} \|UZ_1R^1 - A^1R^1\|_F^2.$$

Let $\widehat{U} \in \mathbb{R}^{n\times k}$ denote the optimal solution to the above optimization problem. Then $\widehat{U} = A^1 R^1 (Z_1 R^1)^\dagger$. From the class we know that with probability at least 0.99 R^1 is an affine embedding, which means that

$$\|\widehat{U} Z_1 - A^1\| \leq (1+\varepsilon) \min_{U\in\mathbb{R}^{n\times k}} \left\|U Z_1 - A^1\right\|_F^2 = (1+\varepsilon)\mathrm{OPT},$$

which implies

$$\left\|\sum_{i=1}^k \widehat{U}_i \otimes V_i^* \otimes W_i^* - A\right\|_F^2 \leq (1+\varepsilon)\mathrm{OPT}.$$

As our second step, we fix $\widehat{U} \in \mathbb{R}^{n\times k}$ and $W^* \in \mathbb{R}^{n\times k}$, and we convert tensor A into matrix A^2. Let $Z_2 = ((\widehat{U})^T \odot (W^*)^T)$. We consider the following objective function,

$$\min_{V\in\mathbb{R}^{n\times k}} \|V Z_2 - A^2\|_F^2,$$

for which the optimal cost is at most $(1+\varepsilon)\mathrm{OPT}$. We sketch R^2 on the right of the objective function to obtain the new objective function,

$$\min_{V\in\mathbb{R}^{n\times k}} \|V Z_2 R^2 - A_2 R^2\|_F^2.$$

Similarly we have that with probability at least 0.99 the solution $\widehat{V} = A^2 R^2 (Z_2 R^2)^\dagger$ satisfies

$$\|\widehat{V} Z_2 - A^2\|_F^2 \leq (1+\varepsilon) \min_{V\in\mathbb{R}^{n\times k}} \|V Z_2 - A^2\|_F^2 \leq (1+\varepsilon)^2\mathrm{OPT},$$

which implies

$$\left\|\sum_{i=1}^k \widehat{U}_i \otimes \widehat{V}_i \otimes W_i^* - A\right\|_F^2 \leq (1+\epsilon)^2\mathrm{OPT}.$$

As our third step, we fix the matrices $\widehat{U} \in \mathbb{R}^{n\times k}$ and $\widehat{V} \in \mathbb{R}^{n\times k}$ and let $Z_3 = ((\widehat{U})^T \odot (\widehat{V})^T)$. We consider the following objective function,

$$\min_{W\in\mathbb{R}^{n\times k}} \|W Z_3 - A^3\|_F^2,$$

which has optimal cost at most $(1+\epsilon)^2\mathrm{OPT}$. We sketch R^3 on the right of the objective function to obtain a new objective function,

$$\min_{W\in\mathbb{R}^{n\times k}} \|W Z_3 R^3 - A^3 R^3\|_F^2.$$

Similarly we have that with probability at least 0.99 we have the solution $\widehat{W} = A^3 R^3 (Z_3 R^3)^\dagger$ satisfies

$$\|\widehat{V} Z_2 - A^2\|_F^2 \leq (1+\varepsilon) \min_{V\in\mathbb{R}^{n\times k}} \|V Z_2 - A^2\|_F^2 \leq (1+\varepsilon)^3\mathrm{OPT},$$

which means that

$$\left\| \sum_{i=1}^{k} \widehat{U}_i \otimes \widehat{V}_i \otimes \widehat{W}_i - A \right\|_F^2 \leq (1+\epsilon)^2 \mathrm{OPT}.$$

Recall that we have $\widehat{U} \in \mathbb{R}^{n \times k}, \widehat{V} = A^2 R^2 (Z_2 R^2)^\dagger$, and $\widehat{W} = A^3 R^3 (Z_3 R^3)^\dagger$. This means that

$$\min_{X_1, X_2, X_3} \left\| \sum_{i=1}^{k} (A_1 R^1 X_1)_i \otimes (A_2 R^2 X_2)_i \otimes (A_3 R^3 X_3)_i - A \right\|_F^2 \leq (1+\epsilon)^3 \mathrm{OPT}.$$

which is what we need.