CS 15-851: Algorithms for Big Data Spring 2024

Lecture 4 (Part 2) — February 8

Prof. David Woodruff *Scribe: Cheng Ma*

1 Leverage Score Sampling

We now introduce **leverage score sampling** as a way to form a subspace embedding based on sampling and rescaling a small number of rows of A based on their *importance*. By forming a subspace embedding this way, we can obtain a low dimension subspace embedding in a way that preserves the sparsity of A and can be computed efficiently.

Before introducing leverage score sampling, we will first consider why a simple sampling method such as sampling rows uniformly will likely not provide a good subspace embedding. By sampling the rows of A and b uniformly, in the worst case, it is likely that we will miss the important rows needed to effectively solve the regression problem, and therefore lead to large errors.

1.1 Leverage Score

Let us formalize the notion of the *importance of a row* with its leverage score.

Definition (Leverage Score)**.** Let the matrix $A \in \mathbb{R}^{n\times d}$ be a rank d matrix. Let the SVD of $A = U\Sigma V^T$.
We can define the ith leverage score as the following, where $U_{i,*}$ is the ith row of U.

$$\ell(i) = ||U_{i,*}||_2^2$$

We can observe that $\sum_{i=1}^n \ell(i) = \sum_{i=1}^n ||U_{i,*}||_2^2 = ||U||_F^2 = d$. We can create define a distribution from the leverage scores to sample from, which can be defined as $(q_1, q_2, ..., q_n)$, where $q_i \geq \frac{\beta\ell(i)}{d}$ and β is a parameter.

Remark 1. When the matrix A is not full rank, we have $A = U\Sigma V^T$ where $U \in \mathbb{R}^{n\times r}$ and $r = \text{rank}(A)$. In this case $\sum_{i=1}^n \ell(i) = \sum_{i=1}^n ||U_{i,*}||_2^2 = ||U||_F^2 = r$.

Definition (Sampling Matrix)**.** Let us define the sampling matrix $S = D\Omega^T$, where $D \in \mathbb{R}^{k\times k}$ is the (diagonal) rescaling matrix and $\Omega \in \mathbb{R}^{n\times k}$ is the sampling matrix.

Formally, for each column $j \in [k]$ fo Ω and D, we independently sample with replacement from index $i \in [n]$ corresponding to the row of A with probability q_i. Then, we set $\Omega_{i,j} = 1$ and $D_{j,j} = \frac{1}{\sqrt{q_i k}}$

The Ω^T matrix is essentially composed of zeros on each row, except for 1 indicating which row of U that's been sampled. The role of the D matrix is just to rescale each row.

Claim 1 (Leverage Scores do not depend on the choice of orthonormal basis U for columns of A)**.** *Proof:* Let U and U' be two orthonormal bases for the columns of A. We want to show that the leverage scores from U and U' are the same, namely that $||e_i U||_2^2 = ||e_j U||_2^2$ for all $i \in [n]$.

Since U and U' have the same column space which is equal to that of A, we can always express $U = U'Z$ for some change of basis matrix Z eg. the first column of U is some linear combination of the columns of U' given by the first column of Z.

Since U and U' have orthonormal columns, Z must be a rotation matrix — it has orthonormal rows and columns and all singular values have to be 1.

AFSOC Z is not a rotation matrix. This means there exists some singular value of Z that is not equal to 1. Now consider the unit vector $\vec{y}$. If Z is not a rotation matrix, then $||Zy||_2 = \sigma_i \neq 1$, which is a contradiction since $U = U'Z$ and U has orthonormal columns and therefore preserves norms.

Therefore, $|e_i U|_2^2 = |e_i U'Z|_2^2 = |e_i U'|_2^2$. The leverage scores are a property of the columnspace of A and are independent of orthonormal basis we use to represent that subspace.

2 Leverage Score Sampling gives a Subspace Embedding

We want to show that the sampling matrix $S = D\Omega^T$ is a subspace embedding for the column span of A. That is, with high probability, for all $\mathbf{x}$, $||SAx||_2^2 = (1 \pm \varepsilon)||Ax||_2^2$.

By expressing $A = U\Sigma V^T$, this is equivalent to showing that $||SUy||_2^2 = (1 \pm \varepsilon)||y||_2^2$ for all $\mathbf{y}$. Similar to when we analyzed the SHRT, we can show $|U^T S^T SU - I|_2 \leq \varepsilon$ and analyze $U^T S^T SU$ with Matrix Chernoff.

2.1 Matrix Chernoff

Definition (Matrix Chernoff)**.** Let $X_1, ..., X_k$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $E[X] = 0, |X|_2 \leq \gamma, |E[X^T X]|_2 \leq \sigma^2$. Let $W = \frac{1}{k}\sum_{j \in [k]} X_j$. For any $\varepsilon > 0$

$$\mathbf{Pr}[|W|_2 > \varepsilon] \leq 2de^{\left(\frac{-k\varepsilon^2}{\sigma^2 + \frac{\gamma\varepsilon}{3}}\right)}$$

Define $i(j)$ to be the index of the row of U (which corresponds to the row of A) sampled in the jth trial.

Define $X_j = I_d - \dfrac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}$ where $U_{i(j)}$ is the jth sampled row of U.

Let's show each of the properties required for Matrix Chernoff.

1. $E[X_j] = I_d - E\left[\dfrac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}\right] = \sum_i q_i \left(\dfrac{U_i^T U_i}{q_i}\right) = I_d - U^T U = I_d - I_d = 0$
2. $|X_j| \leq 1 + \frac{d}{\beta}$

$$
\begin{aligned}
|X_j|_2 &\leq |I_d|_2 + \frac{|U_{i(j)}^T U_{i(j)}|_2}{q_{i(j)}} \\
&\leq 1 + \max_i \frac{|U_i|_2^2}{q_i} (U_{i(j)}^T U_{i(j)} \text{ can be written as normalized rank 1 matrix}) \\
&\leq 1 + \frac{d}{\beta} (\text{By definition, } q_i \geq \frac{\beta \ell(i)}{d} \text{ and } \ell(i) \text{ cancels})
\end{aligned}
$$

3. $|E[X^T X]|_2 \leq \frac{d}{\beta} - 1$

$$
\begin{aligned}
E[X^T X] &= I_d - 2E\left[\frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}\right] + E\left[\frac{U_{i(j)}^T U_{i(j)} U_{i(j)}^T U_{i(j)}}{q_{i(j)}^2}\right] \\
&= \sum_i \frac{U_{i(j)}^T U_{i(j)} U_{i(j)}^T U_{i(j)}}{q_{i(j)}} - I_d \\
&\leq \left(\frac{d}{\beta}\right) \sum_i U_i^T U_i - I_d && (\tfrac{d}{\beta} \geq \tfrac{\ell(i)}{q_i}) \\
&\leq \left(\frac{d}{\beta} - 1\right) I_d
\end{aligned}
$$

Where $A \leq B$ for matrices A, B when $x^T A x \leq x^T B x$ for all x.

2.2 Applying Matrix Chernoff

Using the values from section 2.1, we have that $\gamma = 1 + \frac{d}{\beta}$ and $\sigma^2 = \frac{d}{\beta} - 1$. Furthermore, using X_j defined above, we can see that similar to to SHRT where we're looking at the sum of outer products,

$$
W = \frac{1}{k} \sum_{j \in [k]} X_j = I_d - \frac{1}{k} \sum_{j \in [k]} \frac{U_{i(j)}^T U_{i(j)}}{q_i(j)} = I_d - U^T S^T S U
$$

Hence, by the Matrix Chernoff bound, we have that

$$
\mathbf{Pr}[|I_d - U^T S^T S U|_2 > \varepsilon] \leq 2de^{-k\varepsilon^2 \Theta(\frac{\beta}{d})}
$$

By setting $k = \Theta(\frac{d \log d}{\beta \varepsilon^2})$, we can get the above $\mathbf{Pr}[|I_d - U^T S^T S U|_2 > \varepsilon] \leq 2e^{-\Theta(1)}$ to guarantee the following with a large enough constant probability.

3 Fast Computation of Leverage Scores

Naively, we need to do an SVD to compute the leverage scores which would be too costly. We just need a fast method to compute leverage scores for the matrix A, and we can do this with good approximate leverage scores.

Let's use sketching to help us by first computing SA with a subspace embedding S like count sketch with constant error like ε_0.

Let's compute the QR decomposition of $SA = QR^{-1}$ such that Q has orthonormal columns. From the first part of the lecture, let's use AR from the first part of lecture, and we can set $\ell_i' = |e_i AR|_2^2$ as the approximate leverage score. Note $\kappa(AR) = \frac{1+\varepsilon_0}{1-\varepsilon_0}$

Since AR has the same column span as A, we can write $AR = UT^{-1}$ for some matrix T.

Because S is a subspace embedding for A, we can apply the subspace embedding property and R's preconditioning, which gives us the following

- $(1-\varepsilon)|ARx|_2 \leq |SARx|2 = |x|_2$
- $(1+\varepsilon)|ARx|_2 \geq |SARx|2 = |x|_2$

Using the above two statements, we therefore have that 1) $(1 \pm O(\varepsilon))|x|_2 = |ARx|_2$ and substituting $AR = UT^{-1}$) gives us $(1 \pm O(\varepsilon))|x|_2 = |ARx|_2 = |UT^{-1}x|_2$. Since U has orthonormal columns and therefore norms are preserved, $(1 \pm O(\varepsilon))|x|_2 = |ARx|_2 = |UT^{-1}x|_2 = |T^{-1}x|_2$

From this chain of equalities, we can deduce that $T \in \mathbb{R}^{d \times d}$ (since it holds for all x) has singular values that are all $(1 \pm O(\varepsilon))$

Looking at the actual leverage scores ℓ_i,

$$\ell_i = |e_i U|_2^2 = |e_i ART|_2^2 = (1 \pm O(\varepsilon))|e_i AR|_2^2 = (1 \pm O(\varepsilon))\ell_i'$$

, the approximate leverage score.

For the final step, we need to consider how we compute AR which we want in $nnz(A)$ time. Instead of computing the exact squared row norms of AR, we can look at the approximate squared row norm of AR and the error can go into β.

From our approximation of leverage scores $\ell_i = (1 \pm O(\varepsilon))\ell_i$, we can set ε to some constant and pay for it in β when determining the rows needed for matrix chernoff $k = \Theta(\frac{d \log d}{\beta \varepsilon^2})$.

Because computing $\ell_i' = |e_i AR|_2^2$ is too expensive, we can utilize the Johnson-Lindenstrauss Lemma from lecture 1.

When $G \in \mathbb{R}^{d \times O(\log(n))}$ matrix of iid normal random variables, we have that for any vector z, $\mathbf{Pr}[|zG|_2^2 = (1 \pm \frac{1}{2})|z|^2] \geq 1 - \frac{1}{n^2}$

Now, instead of directly computing $\ell_i' = |e_i AR|_2^2$, let's compute $\ell_i' = |e_i (A(RG))|_2^2$ where the brackets indicate the order of multiplication.

This is a useful trick because RG costs $O(d^2 \log(n))$, so $A(RG)$ costs $O(nnz(A) \log(n))$, allowing the total cost to only be $(nnz(A) + d^2) \log(n)$ because we've collapsed the number of columns of the matrix multiplying A on the right.

Finally, the approximation to the leverage score is through this method is $\ell_i = |e_i ARG|_2^2 = (1 \pm \frac{1}{2})|e_i AR|_2^2 = (1 \pm \frac{1}{2})(1 \pm O(\varepsilon))\ell_i$, which is really fast.

This let's us solve regression in $nnz(A) \log(n) + poly(\frac{d \log(n)}{\varepsilon})$ time.