15-859: Algorithms for Big Data Fall 2020

Lecture 3.2 — September 24

Prof. David Woodruff *Scribe: Dongyu Li and Arvind Mahankali*

1 Affine Embeddings

Consider the generalized regression problem

$$\min_X \|AX - B\|_F^2$$

where A is a tall and thin matrix with d columns, and B is a matrix with a large number of columns potentially much larger than d.

We can try solving the regression problem for each column of B separately and form the matrix X, but it's slow due to large number of columns of B.

Note that we also can't directly apply subspace embeddings, because the dimension of subspace increases by the large number of columns of B.

Let's try to show $\|SAX - SB\|_F = (1 \pm \epsilon)\|AX - B\|_F$ for all X and see what properties we would require of S. Just as before, we can assume A has orthonormal columns. Let $B^* = AX^* - B$, where X^* is the optimum.

Let's first analyze the expression $\|S(AX - B)\|_F^2 - \|SB^*\|_F^2$ and see what we can get

$$\begin{aligned}
&\|S(AX - B)\|_F^2 - \|SB^*\|_F^2 \\
&= \|SA(X - X^*) + S(AX^* - B)\|_F^2 - \|SB^*\|_F^2 \\
&= \|SA(X - X^*)\|_F^2 + \|S(AX^* - B)\|_F^2 + 2\text{tr}[(X - X^*)^T A^T S^T S(AX^* - B)] - \|SB^*\|_F^2 \quad (1) \\
&= \|SA(X - X^*)\|_F^2 + 2\text{tr}[(X - X^*)^T A^T S^T SB^*] \quad (2) \\
&\in \|SA(X - X^*)\|_F^2 \pm 2\|X - X^*\|_F \|A^T S^T SB^*\|_F \quad (3) \\
&\in \|SA(X - X^*)\|_F^2 \pm 2\epsilon\|X - X^*\|_F \|B^*\|_F \quad (4) \\
&\in \|A(X - X^*)\|_F^2 \pm \epsilon(\|A(X - X^*)\|_F^2 + 2\|X - X^*\|_F \|B^*\|_F) \quad (5)
\end{aligned}$$

Facts used for steps of the above derivation:

- (1) $\|C + D\|_F^2 = \|C\|_F^2 + \|D\|_F^2 + 2tr(C^T D)$, which we will prove later.
- (2) $\|S(AX^* - B)\|_F^2 = \|SB^*\|_F^2$, because we defined $B^* = AX^* - B$.
- (3): $\text{tr}(CD) \leq \|C\|_F \|D\|_F$, which we will prove later.
- (4) Note that $\|A^T S^T SB^* - A^T B^*\|_F^2 = \|A^T S^T SB^* - 0\|_F^2 = \|A^T S^T SB^*\|_F^2$, because columns of B^* are orthogonal to column span of A. If we have approximate matrix product guarantee

for S, then $\|A^T S^T S B^*\|_F^2 \leq O(\frac{1}{\#\text{ of rows of } S})\|A^T\|_F^2 \cdot \|B\|_F^2$. Since A is an $n \times d$ orthonormal matrix, $\|A\|_F^2 = d$. Therefore, we can bound $\|A^T S^T S B^*\|_F^2 \leq \epsilon^2 \|B\|_F^2$, if the number of rows in $S \geq \frac{d}{\epsilon^2}$. Hence, under this assumption $\|A^T S^T S B^*\|_F \leq \epsilon \|B^*\|_F$

- (5) if S is a subspace embedding for the column span of A, for which $A(X - X^*)$ is in, the Frobenius norm is preserved for a multiplicative error up to ϵ.

Let's now look at the normal equation that's analogous to the one for least square regression we looked at during Lecture 1.

$$\|AX - B\|_F^2 = \|A(X - X^*)\|_F^2 + \|B^*\|_F^2$$

Note that in this case columns of B^* are orthogonal to the column span of A.

Now, let's analyze the expression that subtracts the non-sketched difference between $\|AX - B\|_F^2$ and $\|B^*\|_F^2$ from the sketeched difference

$$\begin{aligned}
&\|S(AX - B)\|_F^2 - \|SB^*\|_F^2 - (\|AX - B\|_F^2 - \|B^*\|_F^2) \\
&\in \|A(X - X^*)\|_F^2 \pm \epsilon(\|A(X - X^*\|_F^2 + 2\|X - X^*\|_F \|B^*\|_F) - \|A(X - X^*)\|_F^2 \\
&\in \pm\epsilon(\|A(X - X^*)\|_F^2 + 2\|X - X^*\|_F \|B^*\|_F) \\
&\in \pm\epsilon(\|A(X - X^*)\|_F^2 + 2\|A(X - X^*)\|_F \|B^*\|_F + \|B^*\|_F^2) \qquad (1) \\
&\in \pm\epsilon(\|A(X - X^*)\|_F + \|B^*\|_F)^2 \\
&\in \pm 2\epsilon(\|A(X - X^*)\|_F^2 + \|B\|_F^2) \qquad (2) \\
&\in \pm 2\epsilon \|AX - B\|_F^2
\end{aligned}$$

Facts used for steps of the above derivation:

- (1) A is orthonormal, preserving Frobenius norm
- (2) $(a + b)^2 \leq 2a^2 + 2b^2$, because $ab \leq \frac{a^2 + b^2}{2}$

We will also assume

$$\|SB^*\|_F^2 = (1 \pm \epsilon)\|B^*\|_F^2$$

holds for our sketching matrix S with constant probability. Note that $B^* = AX^* - B$, which does not depend on X. Hence, we just need this to hold for a fixed B^*, which is not much to ask.

Then, we get

$$\begin{aligned}
\|S(AX - B)\|_F^2 &= \|AX - B\|_F^2 + (\|SB^*\|_F^2 - \|B^*\|_F^2) \pm 2\epsilon\|AX - B\|_F^2 \\
&= (1 \pm 2\epsilon)\|AX - B\|_F^2 + \epsilon\|B^*\|_F^2 \\
&= (1 \pm 2\epsilon)\|AX - B\|_F^2 + \epsilon\|AX^* - B\|_F^2 \\
&= (1 \pm 3\epsilon)\|AX - B\|_F^2
\end{aligned}$$

The last step is because at optimum X^*, $\|AX - B\|_F^2$ is smaller than any other possible X, by definition of being optimal.

Therefore, we have shown that S is an affine embedding, if it satisfies these properties:

- S is a subspace embedding for columns of A.
- S has the approximate matrix product result.
- S preserves Frobenious norm up to ϵ error for a fixed matrix B^* with constant probability.

Now we show some basic results about the Frobenius norm, which we used while constructing an affine embedding.

Lemma 1. *For two matrices $A, B \in \mathbb{R}^{m\times n}$,*

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\mathrm{tr}(A^T B).$$

Proof. For i between 1 and n, let A_i and B_i be the i^{th} columns of A and B respectively. Then,

$$\begin{aligned}
\|A + B\|_F^2 &= \sum_{i=1}^{n} \|A_i + B_i\|_2^2 \\
&= \sum_{i=1}^{n} (\|A_i\|_2^2 + \|B_i\|_2^2 + 2\langle A_i, B_i\rangle) \\
&= \sum_{i=1}^{n} \|A_i\|_2^2 + \sum_{i=1}^{n} \|B_i\|_2^2 + 2\sum_{i=1}^{n} \langle A_i, B_i\rangle \\
&= \|A\|_F^2 + \|B\|_F^2 + 2\mathrm{tr}(A^T B)
\end{aligned} \tag{1}$$

Note that $\sum_{i=1}^{n} \langle A_i, B_i\rangle = \mathrm{tr}(A^T B)$ since the entry of $A^T B$ in row i and column i is $\langle A_i, B_i\rangle$. ■

Lemma 2. *For $A, B \in \mathbb{R}^{m\times n}$*

$$|\mathrm{tr}(AB)| \le \|A\|_F \|B\|_F$$

Proof. Observe that

$$\begin{aligned}
|\mathrm{Tr}(AB)| &= |\sum_{i=1}^{n} \langle A^i, B_i\rangle| \\
&\le \sum_{i=1}^{n} \|A^i\|_2 \|B_i\|_2 \\
&\le \Big(\sum_{i=1}^{n} \|A^i\|_2^2\Big)^{\frac{1}{2}} \Big(\|B_i\|_2^2\Big)^{\frac{1}{2}}
\end{aligned} \tag{2}$$

where the first inequality follows from the triangle inequality and Cauchy-Schwarz (applied to each of the inner products), and the second inequality follows from Cauchy-Schwarz as well, this time applied to the vectors $(|A^1|_2, \ldots, |A^n|_2)$ and $(|B_1|_2, \ldots, |B_n|_2)$. ■

In addition, recall that in order for a sketching matrix $S \in \mathbb{R}^{k \times n}$ to be an affine embedding, it must satisfy the condition that for any fixed $n \times d$ matrix B^*, with constant probability,

$$\|SB^*\|_F^2 = (1 \pm \varepsilon)\|B^*\|_F^2$$

This condition is met by the CountSketch matrix with $O(1/\varepsilon^2)$ rows. Following is an elementary proof of this fact based.

Lemma 3. *Suppose $B^* \in \mathbb{R}^{n \times d}$ is a fixed matrix, and $S \in \mathbb{R}^{k \times n}$ is the CountSketch matrix. If $k = O\left(\frac{1}{\varepsilon^2}\right)$, then*

$$|SB^*|_F^2 = (1 \pm \varepsilon)|B^*|_F^2$$

with constant probability.

This lemma was Problem #3 in HW 1 of Fall 2017. The key idea of the proof is to use Chebyshev's inequality to bound the error probability. This can be done by first computing the expectation and variance of $\|SB^*\|_F^2$:

$$\begin{aligned} E\left[\|SB^*\|_F^2\right] &= \sum_{i=1}^{n} E\left[\|SB_i^*\|_2^2\right] \\ &= \sum_{i=1}^{n} \|B_i^*\|_2^2 \\ &= \|B^*\|_F^2 \end{aligned} \tag{3}$$

where the second equality was shown at the beginning of the first half of the lecture. To bound the variance of $\|SB^*\|_F^2$, it is enough to compute $E[\|SB^*\|_F^4]$. This computation is similar to the analysis done in the first half of today's lecture when computing $\|Sx\|_2^4$ where x is a unit vector.

The full proof is given below for reference.

Proof. We give an elementary argument based on Chebyshev's inequality. Let A_i denote the i-th column of A, for $i \in [d]$. For each of the d rows i of S, let $h(i) \in [r]$ denote the location of the single non-zero entry of S in the i-th row, and let $\sigma_i \in \{-1, 1\}$ be this entry. Then

$$\|AS\|_F^2 = \sum_{j \in [r]} \| \sum_{i \in [d] \text{ such that } h(i)=j} \sigma_i A_i \|_2^2 = \sum_{j \in [r]} \sum_{i,i' \in [d] \text{ such that } h(i)=j} \sigma_i \sigma_{i'} \langle A_i, A_i \rangle.$$

For any fixed h, taking expectation over σ we have that $\mathbf{E}[\sigma_i \sigma_{i'}] = 0$ unless $i = i'$, in which case $\mathbf{E}[\sigma_i \sigma_{i'}] = 1$. It follows by linearity of expectation that

$$\mathbf{E}[\|AS\|_F^2] = \sum_{j \in [r]} \sum_{i \text{ such that } h(i)=j} \|A_i\|_2^2 = \|A\|_F^2.$$

We also have

$$\|AS\|_F^4 = \sum_{j_1, j_2 \in [r]} \sum_{i_1, i_2 \text{ such that } h(i_1)=h(i_2)=j_1} \sigma_{i_1} \sigma_{i_2} \langle A_{i_1}, A_{i_2} \rangle \sum_{i_3, i_4 \text{ such that } h(i_3)=h(i_4)=j_2} \sigma_{i_3} \sigma_{i_4} \langle A_{i_3}, A_{i_4} \rangle.$$

Let $\delta(h(i_1) = j_1)$ be 1 if $h(i_1) = j_1$, and be 0 otherwise. Then we can write $\mathbf{E}[\|AS\|_F^4]$ as

$$\sum_{j_1, j_2 \in [r], i_1, i_2, i_3, i_4 \in [d]} \mathbf{E}[\delta(h(i_1) = j_1)\delta(h(i_2) = j_1)\delta(h(i_3) = j_2)\delta(h(i_4) = j_2)\sigma_{i_1}\sigma_{i_2}\sigma_{i_3}\sigma_{i_4}]$$

$$\cdot \langle A_{i_1}, A_{i_2} \rangle \langle A_{i_3}, A_{i_4} \rangle$$

Taking expectation only with respect to σ, to have a non-zero expectation, we must be able to partition $\{i_1, i_2, i_3, i_4\}$ into equal pairs. This drives the analysis behind the following cases.

Case: $j_1 \neq j_2$**.** Then the set $\{i_1, i_2\}$ must be disjoint from $\{i_3, i_4\}$ since we cannot have $h(i) = j_1$ and $h(i) = j_2$ for some $j_1 \neq j_2$. It follows that $i_1 = i_2$ and $i_3 = i_4$ and $i_1 \neq i_3$ are the only terms which contribute to the expectation. It follows that the total contribution from terms for which $j_1 \neq j_2$ is

$$\sum_{j_1 \neq j_2 \in [r], i_1 \neq i_3 \in [d]} \frac{1}{r^2} \|A_{i_1}\|_2^2 \|A_{i_3}\|_2^2 \leq \|A\|_F^4 - \sum_i \|A_i\|_2^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_2 = i_3 = i_4$**.** The total contribution from these terms is

$$\sum_{j_1 \in [r], i_1 \in [d]} \frac{1}{r} \|A_{i_1}\|_2^4 = \sum_i \|A_i\|_2^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_2$**,** $i_3 = i_4$**,** $i_1 \neq i_3$**.** The total contribution from these terms is

$$\sum_{j_1 \in [r], i_1 \neq i_3 \in [d]} \frac{1}{r^2} \|A_{i_1}\|_2^2 \|A_{i_3}\|_2^2 = O(1/r)\|A\|_F^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_3$**,** $i_2 = i_4$**,** $i_1 \neq i_2$**.** The total contribution from these terms is

$$\sum_{j_1 \in [r], i_1 \neq i_2 \in [d]} \frac{1}{r^2} \langle A_{i_1}, A_{i_2} \rangle^2 = O(1/r)\|A\|_F^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_4$**,** $i_2 = i_3$**,** $i_1 \neq i_2$**.** This case is the same as the previous case, and contributes $O(1/r)\|A\|_F^4$.

In total, we have $\mathbf{E}[\|AS\|_F^4] = \|A\|_F^4 + O(1/r)\|A\|_F^4$. Hence, $\mathbf{Var}[\|AS\|_F^2] = \mathbf{E}[\|AS\|_F^4] - \mathbf{E}^2[\|AS\|_F^2] = O(1/r)\|A\|_F^4$. By Chebyshev's inequality,

$$\mathbb{P}[|\|AS\|_F^2 - \|A\|_F^2| \geq \epsilon \|A\|_F^2] = \frac{O(1/r)\|A\|_F^4}{\epsilon^2 \|A\|_F^4} \leq \frac{1}{10},$$

for suitably chosen $r = \Theta(1/\epsilon^2)$. ■

2 Low-Rank Approximation Using Affine Embeddings

We now consider an application of affine embeddings which arises often when dealing with large datasets. Consider a matrix $A \in \mathbb{R}^{n \times d}$, where n and d may both be large. In many cases, A may be approximated by a low-rank matrix UV, where $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{k \times d}$, and $k \ll n, d$ (k is an upper bound on the rank of U and V).

This offers several advantages when k is small. First, the amount of space needed to store A decreases from $O(nd)$ to $O(nk + kd)$. In addition, multiplication of A by a vector $x \in \mathbb{R}^d$ can be done in $O(nk + kd)$ time, through first multiplying x by V and then by U. Finally, this may remove noise which had artificially increased the rank of A, and can improve the interpretability of the data.

2.1 Exact Algorithms with SVD

Consider the singular value decomposition $U\Sigma V^T$ of A. If this can be computed, then we can obtain a good rank k approximation of A as follows.

First, suppose $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ are the (nonzero) singular values of A, and r is the rank of A. Then, define Σ_k to be the diagonal matrix with $\sigma_1, \sigma_2, \ldots, \sigma_k$ on its diagonal. In addition, take V_k^T to be the matrix consisting of the first k rows of V^T (in other words, the first k singular vectors). Similarly, take U_k to be the matrix consisting of the k leftmost columns of U.

Then, $A_k := U_k \Sigma_k V_k^T$ is a matrix of rank k, and is in fact the best rank k approximation to A in the sense that

$$A_k = \underset{\text{rank k matrices } B}{\arg\min} \|A - B\|_F$$

(This holds under other norms, in addition to the Frobenius norm).

However, recall that computing the SVD of A will take time $O(nd^2)$. To obtain faster algorithms, we will relax the problem as we did in lecture 1. More specifically, our goal is to compute a rank k matrix A' such that

$$\|A - A'\|_F \leq (1+\varepsilon)\|A - A_k\|_F$$

with high probability. This can be done in time $O(\text{nnz}(A) + (n+d)\text{poly}(k/\varepsilon))$, as proposed in [2] and [1]. This is a significant improvement, as even if A is dense, $\text{nnz}(A) = O(nd)$.

2.2 Low-Rank Approximation With Sketching

The idea is as follows: view the rows of A as points in $\mathbb{R}^d$. In addition, let $S \in \mathbb{R}^{k \times n}$ be a sketching matrix (where $k/\varepsilon \ll n$, meaning we are perfectly fine with $\text{poly}(k/\varepsilon)$ terms in our running time). Then, the rows of SA are linear combinations of the rows of A, meaning the row span of SA is a lower-dimensional subspace of the row span of A. From here, the algorithm proceeds as follows:

- Find SA, which takes $O(\text{nnz}(A))$ time if S is the CountSketch matrix. (Note: S can be any of the random matrices we considered earlier — a $\frac{k}{\varepsilon} \times n$ matrix of normals, the Subsampled Randomized Hadamard Transform [2], or the CountSketch matrix [1].)
- Project the rows of A onto SA.
- Find a rank k approximation for the projected rows (in other words, find a k-dimensional subspace that approximates the projected rows of A).

To do this, we solve the optimization problem

$$\min_{\text{rank-}k\, X} \|XSA - A\|_F^2$$

Why is this a useful objective function i.e., why does row span of SA need to contain a good solution? Consider a different objective:

$$\min_X \|A_k X - A\|_F^2$$

Clearly, this is minimized when X is the identity. Now consider the sketched version of this objective (here, S is an affine embedding, for instance, the CountSketch matrix):

$$\text{argmin}_X \|SA_kX - SA\|_F$$

Note that this is $(1 \pm \varepsilon)\|A_kX - A\|_F$ for all matrices X.

We can solve the above objective using the normal equations to find $X = (SA_k)^-SA$. Why does this hold? Observe that the i^{th} column of SA_kX is SA_kX_i, where X_i is the i^{th} column of X. Therefore, we can independently choose $X_i = (SA_k)^-(SA)_i$ for each i (where $(SA)_i$ is the i^{th} column of SA).

Now, since S is an affine embedding, this minimizer is an approximate solution to the objective $\|A_kX - A\|_F^2$ — that is,

$$\|A_k(SA_k)^-(SA) - A\|_F \leq (1+\varepsilon)\|A_k - A\|_F$$

This enables us to show that our original objective

$$\min_{\text{rank-}k\ X} \|XSA - A\|_F^2$$

is a good one. Indeed, $A_k(SA_k)^-(SA)$ is a rank k matrix, and its rows are linear combinations of the rows of SA. Therefore,

$$\begin{aligned} \min_{\text{rank-}k\ X} \|XSA - A\|_F^2 &\leq \|A_k(SA_k)^-SA - A\|_F^2 \\ &\leq (1+\varepsilon)\|A - A_k\|_F^2 \end{aligned} \tag{4}$$

and it is useful to find solutions X to our original objective.

We now solve our original objective. Using the normal equations gives

$$\|XSA - A\|_F^2 = \|XSA - A(SA)^-(SA)\|_F^2 + \|A(SA)^-SA - A\|_F^2$$

meaning

$$\min_{\text{rank-}k\ X} \|XSA - A\|_F^2 = \|A(SA)^-SA - A\|_F^2 + \min_{\text{rank-}k\ X} \|XSA - A(SA)^-(SA)\|_F^2$$

Now, we can write $SA = U\Sigma V^T$ in its *thin* SVD form, meaning that we remove all zero singular values from Σ, and remove the corresponding rows from V and columns from U. The second term of the above objective becomes

$$\begin{aligned} \min_{\text{rank-}k\ X} \|XSA - A(SA)^-(SA)\|_F^2 &= \min_{\text{rank-}k\ X} \|XU\Sigma - A(SA)^-U\Sigma\|_F^2 \\ &= \min_{\text{rank-}k\ Y} \|Y - A(SA)^-U\Sigma\|_F^2 \end{aligned} \tag{5}$$

where the first equality is obtained by replacing SA with its thin SVD (we can ignore V^T because its rows are orthonormal — therefore, this does not affect the singular values of the matrix inside the Frobenius norm, while the Frobenius norm is determined by singular values). Meanwhile, the second equality holds since $U\Sigma$ has full rank, so $Y = XU\Sigma$ has the same rank as X and we can go back and forth between X and Y.

To compute the optimal Y, it suffices to compute the SVD of $A(SA)^-U\Sigma$ and discard all but the k greatest singular values. Even though we can compute the matrix SA in $\text{nnz}(A)$ for a count sketch matrix S, the matrix $(SA)^-$ might be a dense matrix and thereby can make the computation of the matrix $A(SA)^-U\Sigma$ very slow. In the next lecture, we will see that we can use Affine embeddings to solve the problem.

References

[1] Clarkson, Kenneth L., and David P. Woodruff. "Low-rank approximation and regression in input sparsity time." Journal of the ACM (JACM) 63.6 (2017): 54.

[2] Sarlos, Tamas. "Improved approximation algorithms for large matrices via random projections." 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). IEEE, 2006.