| | |
|---|---|
| **CS 15-851: Algorithms for Big Data** | Spring 2024 |
| Lecture 3 — 2/1/2024 | |
| *Prof. David Woodruff* | *Scribe: Nick Kocurek* |

# 1 Birthday Paradox

**Claim**

CountSketch requires $\Omega(d^2)$ number of rows to be a subspace embedding

Here is a brief sketch:

Think of the $k$ rows as hash buckets. When we multiply $Sx$ each bucket receives some expression of $\pm x_i$, and since there is one non-zero entry per column, each $x_i$ appears in exactly one bucket. For a matrix, this just does this for all columns, which ends up throwing signed rows of $A$ into the $k$ buckets at random.

In order to be a subspace embedding, we need $\text{rank}(SA) = d$. If we take an example where $A$ has rank $d$, then we interpret the above to say that we are throwing $d$ balls (signed rows) randomly into $k$ bins. If we have a collision, this means $< d$ bins are non-zero, which corresponds to $SA$ having $< d$ non-zero rows and $< d$ rank. To avoid collision with decent probability, we need to take $k = \Omega(d^2)$ bins, as seen in the Birhtday Paradox problem.

# 2 Affine Embeddings

Want to solve $\min_X ||AX - B||_F^2$ where $A$ is tall and thin ($n \times d$ with $n >> d$) but $B$ has a lot of columns. We will try to figure out what properties $S$ needs to have to satisfy:

$$||SAX - SB||_F = (1 \pm \varepsilon)||AX - B||_F$$

for all $X$ simultaneously. Once again we can assume $A$ has orthonormal columns and if we set $B^* = AX^* - B$ where $X^*$ is the optimum, this will satisfy the normal equations (just think about it column by column). Observe that:

$$\begin{aligned}
||S(AX - B)||_F^2 - ||SB||_F^2 &= ||SA(X - X^*) + S(AX^* - B)||_F^2 - ||SB^*||_F^2 \\
&= ||SA(X - X^*)||_F^2 + 2\,\text{tr}\left[(X - X^*)^T A^T S^T SB^*\right] && \text{(Fact 1)} \\
&\in ||SA(X - X^*)||_F^2 \pm 2||X - X^*||_F||A^T S^T SB^*||_F && \text{(Fact 2)} \\
&\in ||SA(X - X^*)||_F^2 \pm 2\varepsilon||X - X^*||_F||B^*||_F && \text{(Approx. mat. prod.)} \\
&\leq ||A(X - X^*)||_F^2 \pm \varepsilon(||A(X - X^*)||_F^2 + 2||X - X^*||_F||B^*||_F && (S \text{ s.e.})
\end{aligned}$$

where the facts are basic matrix inequalities that we will postpone proving until later. The third and fourth steps use previous properties of the matrix $S$, which we have shown work if we choose $S$ correctly and with a sufficient number of rows.

In all above we have:

$$||S(AX-B)||_F^2 - ||SB^*||_F^2 \in ||A(X-X^*)||_F^2 \pm \varepsilon(||A(X-X^*)||_F^2 + 2||X-X^*||_F||B^*||F)$$

The normal equations tell us $||AX-B||_F^2 = ||A(X-X^*)||_F^2 + ||B^*||_F^2$. Geometrically we imagine for each column $X_i$ in $X$, $AX_i$ is some point in the column space of $A$. The columns $B_i$ are points (potentially) not in the column space. Like in regression, we have that $AX_i^*$ is the closest point to $B_i$ in the column space. If we calculate the distance between these two points we get:

$$B_i^* = AX_i^* - B_i$$

which makes up the term $||B^*||_F$. We then look at the distance between $AX_i$ and $AX_i^*$ and this makes up $||A(X-X^*)||_F$. Collating Pythagorean theorems gives us in whole the normal equations.

This allows us to say something about the approximation above:

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

which tells us the error from our subspace embedding is approximately $2\varepsilon||AX-B||_F^2$. Using a fact $||SB^*||_F^2 = (1\pm\varepsilon)||B^*||_F^2$ proved below, we can rearrange this further to:

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

which tells us that $S$ is a $(1+3\varepsilon)$-affine embedding for $X$. □

**Cleaning up the missing facts:**

**Fact 1:** $||A+B||_F^2|| = ||A||_F^2 + ||B||_F^2 + 2\,\mathrm{tr}(A^TB)$

**Proof:**

$$\begin{aligned}
||A+B||_F^2 &= \sum_i |A_i + B_i|_2^2 && \text{(Def. of } ||\cdot||_F \text{ and } |\cdot|_2) \\
&= \sum_i |A_i|_2^2 + \sum_i |B_i|_2^2 + 2\langle A_i, B_i\rangle && \text{(Like } (a+b)^2 = a^2+2ab+b^2 \text{ but for } |\cdot|_2) \\
&= ||A||_F^2 + ||B||_F^2 + 2\,\mathrm{tr}(A^TB) && \text{(Def. of norms and trace)}
\end{aligned}$$

□

**Fact 2:** $\mathrm{tr}(AB) \le ||A||_F||B||_F$

**Proof:**

$$
\begin{aligned}
\mathrm{tr}(AB) &= \sum_i \langle A^i, B_i \rangle && (A^i \text{ are rows, } B_i \text{ cols}) \\
&\leq \sum_i |A^i|_2 |B_i|_2 && \text{(Cauchy-Schwarz)} \\
&\leq \sqrt{\left(\sum_i |A_i|_2^2\right)^{\frac{1}{2}} \left(\sum_i |B_i|_2^2\right)^{\frac{1}{2}}} && \text{(Cauchy-Schwarz)} \\
&= ||A||_F ||B||_F && (\text{Def. of } ||\cdot||_F)
\end{aligned}
$$

□

**Fact 3:** $||SB^*||_F^2 = (1 \pm \varepsilon)||B^*||_F^2$ with constant probability if $S$ is a CountSketch matrix with $k = O\left(\frac{1}{\varepsilon^2}\right)$.

**Proof:**

From the Fall 2017 iteration of this course Homework 1 Problem 3.

In summary, we have the following:

**Theorem 2.1: Affine Embedding**

$S$ satisfies with decent probability for all $X$:

$$||S(AX - B)||_F^2 = (1 \pm \varepsilon)||AX - B||_F^2$$

Given that $S$ satisfies:

1. Subspace embedding for colspace($A$)
2. Approximate matrix product
3. Preserves norm of a fixed matrix

For CountSketch to satisfy these three, we need $O\left(\frac{d^2}{\varepsilon^2}\right)$ rows, which is importantly not dependent on the dimensions of $B$.

## 3 Low Rank Approximation

Suppose $A$ is an $n \times d$ matrix representing data. $A$ might be high rank because of noise in the data, but can really be approximated by a low rank matrix approximating $A$. This will be easier to store and will remove the noise, making the data more interpretable.

Recall the Singular Value Decomposition:

$$A = U\Sigma V$$

where:

- $U$ has orthonormal columns
- $\Sigma$ is diagonal with non-increasing positive entries down the diagonal (singular values)
- $V$ has orthonormal rows

One thing we can do is take $\Sigma$ and take the smallest but $k$ singular values and zero them out. This turns $\Sigma$ (and thus $A$) into a rank $k$ matrix, and since we got rid of the smaller singular values, we imagine this might be a good rank-$k$ approximation. This is called the truncated singular value decomposition, and is equivalent to taking the top $k$ principal components. We can then write:

$$A = U_k \Sigma_k V_k + E$$

where the subscript $k$ tells us the truncation and $E$ is just the error. If we write $A_k = U_k \Sigma_k V_k$ we have a good characterization of how good a low rank approximation this is:

$$A_k = \text{argmin}_{k\text{-rank matrices } B]} ||A - B||_F$$

In the end, SVD is slow to calculate, so in the low rank approximation problem, we set out to find $A'$ so that:

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

and our goal will be the following claim:

**Claim**

There is $(1+\varepsilon)$-approximation algorithm for low rank approximation that runs in $\text{nnz}(A) + (n+d) \cdot \text{poly}\left(\frac{k}{\varepsilon}\right)$ time and succeeds w.h.p.

Here is our approach:

Compute $SA$ where $S$ is a random matrix with $k/\varepsilon << n$ rows, which is thought of as $k/\varepsilon$-dimensional random subspace. If we run SVD on $SA$, it will take $n\left(\frac{k}{\varepsilon}\right)^2$ time rather than $nd^2$ for the $d$-dimensional subspace of $A$. As usual we will hope that the optimal low rank approximation $SA_k$ will be approximate for the large subspace of $A$.

Various matrices work for $S$:

- $k/\varepsilon \times n$ Random Gaussian (i.i.d. normals)
- $\tilde{O}(k/\varepsilon) \times n$ Fast Johnson Lindenstrauss
- $\text{poly}(k/\varepsilon) \times n$ CountSketch

Here is a brief sketch of why this approach might work:

Consider the regression problem $\min_X ||A_k X - A||_F$. The best approximation $A_k X$ is $A_k$ by definition, so $X = I$ solves this. If $S$ is an affine embedding we have:

$$||SA_k X - SA||_F = (1 \pm \varepsilon)||A_k X - A||_F$$

Since the matrix is rank $k$, $S$ will work with rows dependent on $k$ instead of $d$ (can confirm in all proofs that the rank is important, not the latent dimension). By the normal equations:

$$\text{argmin}_X ||SA_k X - SA||_F = (SA_k)^- SA$$

giving us:

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

The trick is that $A_k(SA_k)^- SA$, which we shouldn't hope to know, is a good approximation, but moreover this is a rank $k$ matrix in the row span of $SA$! That means if we find it by SVD, or find a better one, it is at least as good as this approximation here.