CS 15-859: Algorithms for Big Data Fall 2019

Lecture 2.1 — September 12

Prof. David Woodruff *Scribe: Anubhav Baweja*

1 Recap: Subspace Embeddings for Regression

Given an $n \times d$ matrix A where $n >> d$ and n-dimensional vector b, we want to find a vector x such that $|Ax - b|_2$ is minimized. Since the deterministic method takes $O(nd^2)$ time, we turn to Subspace Embeddings. The general outline is as follows

1. Want to find x such that $|Ax - b|_2 \leq (1 + \varepsilon) \min_y |Ay - b|_2$.
2. Consider the $d + 1$-dimensional subspace L spanned by the columns of A and b.
3. A matrix S is called a sketch if $\forall y \in L$ we have $|Sy|_2 = (1 \pm \varepsilon)|y|_2$. This gives us that for all x we have $|S(Ax - b)|_2 = (1 \pm \varepsilon)|Ax - b|_2$.
4. Solve $argmin_y |SAy - Sb|_2$ using deterministic methods.
5. This takes $poly(d/\varepsilon)$ time.

However computing SA itself takes $O(nd^2)$ time so this is not an improvement unless we find a way to efficiently compute it. To solve this problem, we select a special matrix S as described in the next section.

2 Subsampled Randomized Hadamard Transform

We define our sketching matrix as $S = PHD$ where P, H and D are matrices that can be efficiently applied to vectors.

2.1 Description of Matrices

1. D is an $n \times n$ diagonal matrix with randomly generated entries. Every element on the diagonal has an equal probability of being 1 or -1. This matrix can be applied to a vector in $O(n)$ time.
2. H is an $n \times n$ dense matrix called the Hadamard matrix and the entries are given by

$$H_{i,j} = \frac{(-1)^{\langle i,j \rangle}}{\sqrt{n}}$$

   WLOG $n = 2^k$ for some $k \in \mathbb{N}$ (otherwise A and b can be extended with 0s to get to the smallest power of 2 bigger than n). So we can define $\langle i, j \rangle = (\sum_{l=0}^{k-1} i_l \cdot j_l) \mod 2$ where i_l is the lth bit in the k-bit binary representation of i.

Claim: H is orthogonal.
Proof: We have

- $\forall i \in [n]$ we have
$$|H_{i*}|_2^2 = \sum_{l=0}^{n-1} H_{i,l}^2 = \sum_{l=0}^{n-1} \frac{1}{n} = 1$$
- $\forall i, j \in [n]$ such that $i \neq j$ we have
$$\langle H_{i*}, H_{j*} \rangle = \sum_{l=0}^{n-1} \frac{1}{n} (-1)^{\langle i,l \rangle} (-1)^{\langle j,l \rangle} = \sum_{l=0}^{n-1} \frac{1}{n} (-1)^{\langle i+j,l \rangle}$$
Now if $i \neq j$ then there is at least one index t where these are different. Therefore at t the binary representation of $i + j \mod 2$ is 1. Now we can pair up every number α with the number that only differs at index t, which we can call β. Now we have $(-1)^{\langle i+j,\alpha \rangle} + (-1)^{\langle i+j,\beta \rangle} = 1 + (-1) = 0$ since if one is 1 the other one has to be (-1) and vice-versa. This gives us that the total sum above is also 0 as desired.

Even though this is a dense matrix, it can be applied to any vector in $O(n \log n)$ time using an algorithm similar to the Fast Fourier Transform.

3. P is a $s \times n$ matrix which selects a random subset of s rows of its input. This can be applied to a vector in $O(s)$ time. As we will see later, s is about d.

Therefore the limiting step is applying the Hadamard matrix. If the number of columns of A is d, the overall complexity of computing SA comes out to be $O(nd \log n)$, which is better than our earlier complexity of $O(nd^2)$.

2.2 Flattening Lemma

Before we proceed to use the Matrix Chernoff Bound to prove that $S = PHD$ is a valid sketching matrix, we need to prove a lemma: for any fixed vector y:

$$\mathbf{Pr}[|HDy|_\infty \geq C\sqrt{\frac{\log(\frac{nd}{\delta})}{n}}] \leq \frac{\delta}{2d}$$

where $C > 0$ is a constant.

2.2.1 Proof

We shall prove that for any $i \in [n]$:

$$\mathbf{Pr}[|HDy|_i \geq C\sqrt{\frac{\log(\frac{nd}{\delta})}{n}}] \leq \frac{\delta}{2nd}$$

If we show the above then we can union bound over all the values of i and obtain the Flattening Lemma.

We have $|HDy|_i = \sum_j H_{i,j} D_{j,j} y_j$. Let's define $Z_j = H_{i,j} D_{j,j} y_j$. This gives us 2 facts:

- We know that $D_{j,j}$ are independent random variables with 0 mean. Therefore Z_j are independent with 0 mean as well.
- $|Z_j| \leq |H_{i,j}| \cdot |D_{j,j}| \cdot |y_j| \leq \frac{1}{\sqrt{n}} \cdot 1 \cdot |y_j| = \frac{|y_j|}{\sqrt{n}}$.

Given that Z_j are independent with 0 mean and an upper bound of $\frac{|y_j|}{\sqrt{n}}$, we can use the Azuma-Hoeffding inequality:

$$\mathbf{Pr}[|\sum_j Z_j| > t] \leq 2e^{\frac{-t^2}{2\sum_j \frac{y_j^2}{n}}} = 2e^{\frac{-nt^2}{2}}$$

Putting in $t = C\sqrt{\frac{\log(\frac{nd}{\delta})}{n}}$, we get

$$\mathbf{Pr}[|\sum_j Z_j| > C\sqrt{\frac{\log(\frac{nd}{\delta})}{n}}] \leq 2e^{\frac{-C^2 \log(\frac{nd}{\delta})}{2}} = 2(\frac{\delta}{nd})^{\frac{C^2}{2}} \leq \frac{\delta}{2nd}$$

as desired.

2.2.2 Consequences

The Flattening Lemma tells us that all entries in HDA are small and close to $\frac{1}{\sqrt{n}}$ in absolute value.

Claim: $|e_j HDA|_2 \leq \sqrt{\frac{d\log(\frac{nd}{\delta})}{n}}$ for all j with probability $1 - \frac{\delta}{2}$.
Proof: Columns of A are orthonormal. Since both H and D are rotation matrices, HD is also a rotation matrix. Therefore the columns of HDA are also orthonormal. The Flattening Lemma implies that

$$|HDAe_i|_\infty \leq \sqrt{\frac{\log(\frac{nd}{\delta})}{n}}$$

with probability at least $1 - \frac{\delta}{2d}$ for a fixed $i \in [d]$. Using the union bound, we get that $|e_j HDAe_i| \leq \sqrt{\frac{\log(\frac{nd}{\delta})}{n}}$ with probability at least $1 - \frac{\delta}{2}$. Since $e_j HDAe_i$ is the (i,j)th entry of the matrix HDA, we get that every entry of the matrix HDA is small in absolute value with high probability.

Finally we get that $|e_j HDA|_2 \leq \sqrt{\frac{d\log(\frac{nd}{\delta})}{n}}$ for all j with probability $1 - \frac{\delta}{2}$ by using the definition of the Euclidean norm and this is what we will use in the Matrix Chernoff Bound.

2.3 Matrix Chernoff Bound

In our sketching matrix $S = PHD$, P samples s rows uniformly with replacement. If row i is sampled in sample j we have $P_{j,i} = \sqrt{\frac{n}{s}}$. All other entries of P are zero.

Definition. The **operator norm** of a matrix W is defined as $|W|_2 = \sup_{|x|_2=1} |Wx|_2$. The operator norm is also equal to the maximum singular value of W.

Definition. The **eigendecomposition** of a matrix W is given by $Q\Lambda Q^{-1}$ where the ith column of Q is given by the ith eigenvector and Λ is a diagonal matrix where Λ_{ii} is the $i$the eigenvalue. If W is real and symmetric then Q is orthogonal and therefore the eigendecomposition can be given as $Q\Lambda Q^T$.

2.3.1 Setup

Let's define $V = HDA$ and let Y_i be the ith sampled row of V. Also define $X_i = I_d - nY_i^T Y_i$ which gives us

$$\begin{aligned}
|X_i|_2 &\leq |I|_2 + n \max_j |e_j HDA|_2^2 && \text{[Triangle Inequality for operator norm]} \\
&= 1 + nC^2\left(\frac{d\log(\frac{nd}{\delta})}{n}\right) && \text{[Flattening Lemma]} \\
&= 1 + C^2 d\log\left(\frac{nd}{\delta}\right) \in O\left(d\log\left(\frac{nd}{\delta}\right)\right)
\end{aligned}$$

We also have two matrices of interest: $\mathbb{E}[X^T X + I_d]$ and $Z = n\sum_i v_i^T v_i \cdot C^2 \frac{n}{d}\log(\frac{nd}{\delta})$. The first one can be simplied as follows

$$\begin{aligned}
\mathbb{E}[X^T X + I_d] &= \mathbb{E}[(I_d - nY_i^T Y_i)^T(I_d - nY_i^T Y_i) + I_d] \\
&= I_d + I_d - 2n\mathbb{E}[Y_i^T Y_i] + n^2\mathbb{E}[Y_i^T Y_i Y_i^T Y_i] \\
&= 2I_d - 2n\left(\frac{1}{n}I_d\right) + n^2\mathbb{E}[Y_i^T Y_i Y_i^T Y_i] \\
&= n^2 \sum_i \frac{1}{n} v_i^T v_i v_i^T v_i \\
&= n\sum_i v_i^T v_i \cdot |v_i|_2^2
\end{aligned}$$

Note that $C^2 \frac{n}{d}\log(\frac{nd}{\delta})$ is an upper bound for $|v_i|_2^2$.

Claim: All eigenvalues of $\mathbb{E}[X^T X + I_d]$ and Z are non-negative. Also for all x we have $x^T\mathbb{E}[X^T X + I_d]x \geq x^T Z x$.
Proof: Since both matrices are real and symmetric, their eigendecomposition is given by $Q\Lambda Q^T$. Therefore if x is an eigenvector of $\mathbb{E}[X^T X + I_d]$ then the corresponding eigenvalue λ is

$$\begin{aligned}
\lambda &= x^T\left(n\sum_i v_i^T v_i \cdot |v_i|_2^2\right)x \\
&= n\sum_i \langle v_i.x\rangle^2 \cdot |v_i|_2^2 \geq 0
\end{aligned}$$

Similarly if y is an eigenvector of Z then the corresponding eigenvalue λ is

$$\begin{aligned}
\lambda &= y^T\left(n\sum_i v_i^T v_i \cdot C^2\frac{n}{d}\log\left(\frac{nd}{\delta}\right)\right)y \\
&= n\sum_i \langle v_i.y\rangle^2 \cdot C^2\frac{n}{d}\log\left(\frac{nd}{\delta}\right) \geq 0
\end{aligned}$$

Therefore all the eigenvalues of $\mathbb{E}[X^TX + I_d]$ and Z are non-negative. Also since $C^2\frac{n}{d}\log(\frac{nd}{\delta})$ is an upper bound for $|v_i|_2^2$, the second part of our claim is immediate.

Claim: $|\mathbb{E}[X^TX + I_d]|_2 \leq |Z|_2$.
Proof: Let $y^* = \text{argmax}_y y^T\mathbb{E}[X^TX + I_d]y$. Then $y^* = |\mathbb{E}[X^TX + I_d]|_2$. But using the claim above we know that $(y^*)^T\mathbb{E}[X^TX + I_d]y^* \leq (y^*)^TZy^*$. And since $(y^*)^TZy^* \leq \text{argmax}_y y^TZy = |Z|_2$ we obtain the above claim.

This finally gives us the final claim in our setup:

Claim: $|\mathbb{E}[X^TX]|_2 \in O(d\log(\frac{nd}{\delta}))$.
Proof:

$$\begin{aligned}
|\mathbb{E}[X^TX]|_2 &\leq |\mathbb{E}[X^TX] + I_d|_2 + |I_d|_2 && \text{[Triangle inequality]} \\
&= |\mathbb{E}[X^TX] + I_d|_2 + 1 \\
&\leq |Z|_2 + 1 \\
&\leq C^2 d\log(\frac{nd}{\delta}) + 1 \in O(d\log(\frac{nd}{\delta}))
\end{aligned}$$

2.3.2 Application

Theorem (Matrix Chernoff Bound): Let $X_1, ..., X_s$ be s independent copies of the symmetric random matrix $X \in \mathbb{R}^{d\times d}$ with $\mathbb{E}[X] = 0$, $|X|_2 \leq \gamma$, and $|\mathbb{E}[X^TX]|_2 \leq \sigma^2$. Let $W = \frac{1}{s}\sum_{i=0}^{s-1} X_i$. For any $\varepsilon > 0$ we have

$$\mathbf{Pr}[|W|_2 > \varepsilon] \leq 2d \cdot e^{-s\varepsilon^2/(\sigma^2 + \frac{\gamma\varepsilon}{3})}$$

The symmetric matrix X is the same as the one we used in the previous subsection. Also s is equal to the number of rows that the matrix P samples. Therefore we get

$$\begin{aligned}
W &= \frac{1}{s}\sum_{i=0}^{s-1} X_i \\
&= \frac{1}{s}\sum_{i=0}^{s-1}(I_d - nY_i^TY_i) \\
&= I_d - \frac{n}{s}\sum_{i=0}^{s-1} Y_i^TY_i \\
&= I_d - \sum_{i=0}^{s-1}(Y_i^T\sqrt{\frac{n}{s}})(\sqrt{\frac{n}{s}}Y_i) \\
&= I_d - (PHDA)^T(PHDA)
\end{aligned}$$

Since Y_i is the ith row that we sampled from HDA and then P performs this sampling step and scales by a factor of $\sqrt{\frac{n}{s}}$.

Finally, since $\sigma \in \Theta(d \log(\frac{nd}{\delta}))$ we get that

$$\mathbf{Pr}[|I_d - (PHDA)^T(PHDA)|_2 > \varepsilon] \leq 2d \cdot e^{-s\varepsilon^2/\Theta(d\log(\frac{nd}{\delta}))}$$

Therefore setting $s = \Theta(d \log(\frac{nd}{\delta})\frac{\log(d/\delta)}{\varepsilon^2})$, we get that

$$\mathbf{Pr}[|I_d - (PHDA)^T(PHDA)|_2 > \varepsilon] \leq \frac{\delta}{2}$$

which is equivalent to

$$\mathbf{Pr}[|I_d - (PHDA)^T(PHDA)|_2 \leq \varepsilon] \geq 1 - \frac{\delta}{2}$$

2.4 Satisfying preconditions for subspace embeddings

Using the definition of the operator norm, for any unit vector x we get

$$\begin{aligned}\varepsilon \geq |I_d - (PHDA)^T(PHDA)|_2 &\geq |x^T(I_d - (PHDA)^T(PHDA))x| \\ &= |x^Tx - (PHDAx)^T(PHDAx)| \\ &= |1 - |PHDAx|_2^2|\end{aligned}$$

which gives us that $|PHDAx|_2^2 \in (1 \pm \varepsilon)$ for all unit x with probability at least $1 - \frac{\delta}{2}$. This means that $S = PHD$ is a valid sketching matrix and we can use it for regression. Since computing SA now takes $O(nd \log n)$ time, the entire algorithm takes $O(nd\log n) + poly(\frac{d \log n}{\varepsilon})$ time which is nearly optimal if $n >> d$ and the matrix A is dense.