CS 15-859: Algorithms for Big Data Fall 2022

Lecture 2.2 — 9/15/2022

Prof. David Woodruff *Scribe: Max Li*

1 Matrix Chernoff Bound

Definition (Operator norm)**.** Given a matrix $W \in \mathbb{R}^{m\times n}$, we denote the operator norm of W as $\|W\|_2$, and define it as

$$\|W\|_2 = \sup \frac{\|Wx\|_2}{\|x\|_2}$$

across all existing $x \in \mathbb{R}^n$, where sup is the supremum operator.

The operator norm of W can also be thought of in terms of the SVD decomposition of W. If $W = U\Sigma V^T$ where $\Sigma \in \mathbb{R}^{d\times d}$, then $\|W\|_2 = \max(\|\Sigma_{1,1}\|, \|\Sigma_{d,d}\|)$.

Theorem 1 (Matrix Chernoff Bound)**.** Let $X_1, \ldots, X_s$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d\times d}$ with $\mathbb{E}[X] = 0$, $\|X\|_2 \leq \gamma$, and $\|\mathbb{E}[X^T X]\|_2 \leq \sigma^2$. Let $W = \frac{1}{s}\sum_{i\in[S]} X_i$. Then, for any $\epsilon > 0$,

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

We can use the Matrix Chernoff Bound to prove that the SRHT matrix $S = PHD$ is a valid sketching matrix. In order to do this, we first define the characteristics of P, which we assume samples s rows uniformly and with replacement, so we may end up sampling the same row twice. We will define the elements of matrix P such that $P_{i,j} = \frac{\sqrt{n}}{\sqrt{s}}$ if P samples row j on its i-th sample, and $P_{i,j} = 0$ otherwise. Notice that each row contains exactly one nonzero element.

Next we define Y_i to be the i-th sampled row of $V = HDA$, and we also define $X_i = I_d - n\cdot Y_i^T Y_i$. However, before we can use the Matrix Chernoff Bound, we must verify that our variables satisfy the required properties.

Remark 1. The matrix I_d is symmetric since it is the identity matrix, and $n \cdot Y_i^T Y_i$ is also symmetric since we are taking the outer product of a vector with itself. Since the difference of two symmetric matrices is another symmetric matrix, we conclude that $X_i = I_d - n \cdot Y_i^T Y_i$ is symmetric as well.

Claim 1. All X_i matrices satisfy the property that $\mathbb{E}[X_i] = 0$.

Proof. By linearity of expectation, $\mathbb{E}[X_i] = \mathbb{E}[I_d] - n\cdot\mathbb{E}[Y_i^T Y_i]$. The expectation of an identity matrix is itself, but since we are sampling uniformly from V, there is a $\frac{1}{n}$ chance of selecting

each row v_i in V, so

$$\mathbb{E}[Y_i^T Y_i] = \sum_j \left(\frac{1}{n}\right) v_i^T v_i = V^T V$$

because $\sum_j \left(\frac{1}{n}\right) v_i^T v_i$ is equivalent to calculating $V^T V$ using outer products. This tells us that $\mathbb{E}[X_i] = I_d - V^T V$. Then, since we assumed that V has orthonormal columns, we know that $V^T V = I_d$, so we conclude that

$$\mathbb{E}[X_i] = I_d - V^T V = I_d - I_d = 0$$

■

Claim 2. For $\gamma = \Theta(d \log\left(\frac{nd}{\delta}\right))$, we have that $\|X_i\|_2 \leq \gamma$.

Proof. Since the operator norm is a norm function, it obeys the triangle inequality, so we can write that

$$\begin{aligned} \|X_i\|_2 &= \|I_d - n \cdot Y_i^T Y_i\|_2 && (1)\\ &\leq \|I_d\|_2 + n \cdot \|Y_i^T Y_i\|_2 && (2)\\ &\leq \|I_d\|_2 + n \cdot \max_j \|e_j HDA\|_2^2 && (3) \end{aligned}$$

The previous step follows from the fact that

$$\begin{aligned} nY_i^T Y_i &= \left(\frac{Y_i^T}{\|Y_i\|_2}\right) n\|Y_i\|_2^2 \left(\frac{Y_i}{\|Y_i\|_2}\right) && (4)\\ &\implies \|nY_i^T Y_i\| = n\|Y_i\|_2^2 && (5) \end{aligned}$$

along with the fact that Y_i is the i-th row of HDA, and therefore $\|Y_i\|_2^2$ is equal to the squared norm of the row Y_i, so we can upper bound all $\|Y_i\|_2^2$ by the largest row.

Now, using the results of the Flattening Lemma from earlier to upper bound the value of $\|e_j HDA\|_2$, we have that

$$\begin{aligned} \|X_i\|_2 &\leq \|I_d\|_2 + n \cdot \max_j \|e_j HDA\|_2^2 && (6)\\ &\leq 1 + n \cdot C^2 \log\left(\frac{nd}{\delta}\right)\left(\frac{d}{n}\right) && (7)\\ &= 1 + C^2 \log\left(\frac{nd}{\delta}\right) d && (8)\\ &= \Theta(dlog\left(\frac{nd}{\delta}\right)) && (9) \end{aligned}$$

■

Claim 3. $\|\mathbb{E}[X^T X]\|_2 = O(d \log\left(\frac{nd}{\delta}\right))$.

Proof. First, we will derivative a useful form of $\mathbb{E}[X^T X]$. Specifically, we use the fact that each X_i is a copy of X and the fact that X_i is symmetric. This lets us write

$$\begin{aligned}
\mathbb{E}[X^T X] &= \mathbb{E}[X_i \cdot X_i] && (10)\\
&= \mathbb{E}[(I_d - n_i^T Y_i)^2] && (11)\\
&= \mathbb{E}[I_d - 2n \cdot Y_i^T Y_i + n^2 Y_i^T Y_i Y_i^T Y_i] && (12)\\
&= I_d - 2n\mathbb{E}[Y_i^T Y_i] + n^2 \mathbb{E}[Y_i^T Y_i Y_i^T Y_i] && (13)
\end{aligned}$$

Since Y_i is the i-th row sampled from V, then

$$\begin{aligned}
\mathbb{E}[Y_i^T Y_i] &= \sum_j \frac{1}{n} V_j^T V_j && (14)\\
&= \frac{1}{n} V^T V && (15)\\
&= \frac{1}{n} I_d && (16)
\end{aligned}$$

since V has orthonormal columns.
Now we can plug this value into $\mathbb{E}[X^T X + I_d]$ to get

$$\begin{aligned}
\mathbb{E}[X^T X + I_d] &= I_d + I_d - 2n\mathbb{E}[Y_i^T Y_i] + n^2 \mathbb{E}[Y_i^T Y_i Y_i^T Y_i] && (17)\\
&= 2I_d - 2n\left(\frac{1}{n} I_d\right) + n^2 \sum_j \frac{1}{n} \cdot V_i^T v_i V_i^T v_i && (18)\\
&= n \sum_i V_i^T v_i \cdot \|v_i\|_2^2 && (19)
\end{aligned}$$

Definition (Loewner-John Ordering)**.** Given two positive semidefinite matrices A, B, we say that $A \leq B$ iff for all x it holds that $x^T A x \leq x^T B x$

Theorem 2. *If A, B are positive semidefinite and $A \leq B$ as per Loewner-John ordering, then $\|A\|_2 \leq \|b\|_2$.*

We will define a matrix $Z = n \sum_i V_i^T v_i C^2 \log\left(\frac{nd}{\delta}\right) * \frac{d}{n} = C^2 d \log\left(\frac{nd}{\delta}\right)$.

Claim 4. $\|\mathbb{E}[X^T X] + I_d\| \leq \|Z\|$

Proof. Since $\mathbb{E}[X^T X + I_d]$ and Z are both real, symmetric, and have non-negative eigenvalues, we know that they are positive semidefinite. This means that we can attempt to apply Loewner-John ordering.

It suffices to show that for all vectors y, we have $y^T \mathbb{E}[X^T X + I_d] y \leq y^T Z y$. First we calculate $y^T \mathbb{E}[X^T X + I_d] y$.

$$\begin{aligned}
y^T \mathbb{E}[X^T X + I_d] y &= n \sum_i y^T V_i^T v_i y \|v_i\|_2^2 && (20)\\
&= n \sum_i \langle v_i, y \rangle^2 \|v_i\|_2^2 && (21)
\end{aligned}$$

Next we calculate $y^T Z y$.

$$y^T Z y = n \sum_i y^T v_i^T v_i y C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} \tag{22}$$

$$= d \sum_i \langle v_i, y \rangle^2 C^2 \log\left(\frac{nd}{\delta}\right) \tag{23}$$

Now since $v_i = e_i V = e_i HDA$, we apply the Flattening Lemma to say that $\|v_i\|_2^2 \leq C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n}$.

This allows us to apply Loewner-John and combine the inequalities we have previously established, which tells us that

$$\|\mathbb{E}[X^T X]\|_2 \leq \|\mathbb{E}[X^T X + I_d]\|_2 + \|I_d\|_2 \tag{24}$$

$$= \|\mathbb{E}[X^T X + I_d]\|_2 + 1 \tag{25}$$

$$\leq \|Z\|_2 + 1 \tag{26}$$

$$\leq C^2 d \log\left(\frac{nd}{\delta}\right) + 1 \tag{27}$$

■

Finally, we have verified that $\|\mathbb{E}[X^T X]\|_2 = O(d \log\left(\frac{nd}{\delta}\right))$, so our X_i and W matrices meet all the requirements to apply the Matrix Chernoff Bound. Before that, however, we note that

$$W = \frac{1}{s} \sum_i X_i \tag{28}$$

$$= \frac{1}{s}(sI - n \sum_i Y_i^T Y_i) \tag{29}$$

$$= I - \sum_i Y_i^T \sqrt{\frac{n}{s}} \sqrt{\frac{n}{s}} Y_i \tag{30}$$

$$= I - (PHDA)^T (PHDA) \tag{31}$$

Now when we apply the matrix Chernoff Bound, we get that

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

Additionally, if we set $s = \Theta(d \log\left(\frac{nd}{\delta}\right) \frac{\log\left(\frac{d}{\delta}\right)}{\epsilon^2})$, then we can use union bound to make this probability less than $\frac{delta}{2}$.

2 SRHT Wrapup

In total, we have shown that using the matrix Chernoff bound, if we have $s = d \log\left(\frac{nd}{\delta}\right) \frac{\log\left(\frac{d}{\delta}\right)}{\epsilon^2}$ then $\|I_d - (PHDA)^T (PHDA)\|_2 < \epsilon$. So if we have some matrix Z such that $\|Z\|_2 \leq \epsilon$,

then for all unit vectors x, it follows that $x^T Z x \leq \epsilon$. Therefore, $\|1 - \|PHDAx\|_2^2 = \| = ^T x - x^T (PHDA)^T (PHDA)x\| < \epsilon$, so $\|PHDAx\|_2^2 \in 1 \pm \epsilon$ for all unit vectors x.

We see that SRHT is a subspace embedding, so we can solve the regression problem in the same way as previously by cojoining the column span of A with b. Then, minimizing $\|SAx - Sb\|_2$ takes time $O(nd \log(n))$ to calculate SAx, and only $O(n \log(n))$ time to calculate Sb. Therefore, the total runtime is $O(nd \log(n)) + poly\left(\frac{d \log(n)}{\epsilon}\right)$. ■

3 CountSketch Matrix

We can beat the randomized Hadamard sketching matrix by using CountSketch matrices. A CountSketch is a $k \times n$ matrix, where $k = O(d^2/\epsilon^2)$. Additionally, CountSketch matrices are very sparse since every element is 0, except for a single element in each column that can be either -1 or 1. SA can be computed in nnz(A), where nnz(A) is the number of non-zero entities of A.

Claim 5. To prove that the CountSketch matrix S is a subspace embedding, it suffices to show that $\|A^T S^T SA - I\|_F \leq \epsilon$.

Proof. We can assume the columns of A are orthonormal since we're embedding all vectors in the subspace, so to prove that CountSketch is a subspace embedding we need to show that $\|SAx\|^2 = 1 \pm \epsilon$. One way that this can be accomplished is by taking advantage of Loewner-John, but it can be messy to use the operator norm by itself. Therefore, we will also employ the Frobenius norm. Significantly, the operator norm is less than or equal to the Frobenius norm. Therefore, it suffices to show that $\|A^T S^T SA - I\|_F \leq \epsilon$, because if this is true then we can say that

$$\begin{aligned} &\|A^T S^T SA - I\|_F \leq \epsilon \\ \implies &\|A^T S^T SA - I\|_2 \leq \|A^T S^T SA - I\|_F \leq \epsilon \\ \implies &\|x^T A^T S^T SAx - x^T x\|_2 \leq \epsilon \\ \implies &\|SAx - 1\|_2^2 \leq \epsilon \\ \implies \implies &\|SAx\|_2^2 \leq 1 \pm \epsilon \end{aligned}$$

■