CS 15-851: Algorithms for Big Data	Spring 2024
Lecture 2 — 01/25/2024	
Prof. David Woodruff	*Scribe: Junwei Huang*

1 Subsampled Randomized Hadamard Transform cont'd

Definition. Matrix Chernoff bound is such that if we have $X_1, ..., X_d$ be i.i.d copies of a symmetric random matrix $X \in \mathbb{R}^{d\times d}$ with $\mathbb{E}[X] = 0$ and $\|X\| \le \gamma$ and $\|\mathbb{E}[X^t X]\|$ is bounded by σ^2. Let $W = \frac{1}{s}\sum_{i\in[S]} X_i$ for any $\epsilon > 0$,

$$Pr[\|W\| \ge \epsilon] \le 2d\exp(-\frac{s\epsilon^2}{\sigma^2 + \gamma\epsilon/3})$$

.

Continuing our investigation of $S = PHD$.

- P the matrix can be considered as a sampling matrix that **uniformly** sampling s rows. Specifically,

$$P_{i,j} = \frac{\sqrt{n}}{\sqrt{s}}$$

  if row j is sampled and 0 otherwise.

- H is the Hadamard Matrix, where each entry

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

  . i and j are binary vectors.

- D is the diagonal matrix with $D_{i,i} = \pm 1$ with equal probability.

This Hadamard matrix H has interesting properties.

- **Note:** H is not a random matrix, and can be recursively defined as

$$H_1 = \begin{bmatrix} 1 \end{bmatrix}$$

$$H_{2n} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}$$

- H is orthonormal, i.e. $H^T H = I$.

Proof.

$$\begin{aligned}\langle H_{*j}, H_{*k}\rangle &= \frac{1}{n}\sum_{i=1}^{n}(-1)^{\langle i\cdot j\rangle}(-1)^{\langle i\cdot k\rangle} \\ &= \frac{1}{n}\sum_{i=1}^{n}(-1)^{\langle i\cdot(j+k)\rangle} \\ &= \begin{cases}1 & \text{if } j=k \\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

■

- We can apply the matrix H to any vector $\begin{bmatrix}x_1\\x_2\end{bmatrix}$ $(x_1, x_2 \in R^n)$ in $O(n\log n)$ time.

  Proof. We can apply the submatrix H_n to x_1 and x_2 recursively. Denoting the running time of applying H_n to x_1 and x_2 as $T(n)$. We can combine the results in $O(n)$. we have

  $$\begin{aligned}T(n) &= 2T(\frac{n}{2}) + O(n) \\ &\in O(n\log n)\end{aligned}$$

  ■

- $S = PHD$ can be applied to any vector in $O(n\log n)$ time. Since we only need $O(n\log n)$ time to apply H to any vector, and P and D are diagonal matrices, we can apply them in $O(n)$ time. (Better yet we can apply P in $O(s)$ time, since P is a sampling matrix.)

  Remark 1. Note that HD is a rotation matrix and thus we have that $|HDAx|_2 = |Ax|_2$. (H, D both orthonormal).

Theorem 1 (Azuma-Hoeffding Bound)**.** *Let $X_1, ..., X_d$ be independent random variables with $|X_i| \leq c_i$ and $E[X_i] = 0$. Let $X = \sum_{i=1}^{d} X_i$. Then for any $\epsilon > 0$, we have*

$$Pr\left[|X| > \epsilon\right] \leq 2\exp(-\frac{\epsilon^2}{2\sum_i c_i^2})$$

Lemma 1 (Flattening lemma)**.** *For any fixed vector $y \in \mathbb{R}^n$ and constant C, we have*

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

Proof. We have the following observation: Let C be a constant, we apply the Azuma-Hoeffding bound to the random variable HDy_i.

■

Let Y_i be the ith sampled row of $V = HDA$. Let $X_i = I_d - n \cdot Y_i^T Y_i$, We first note that

$$\begin{aligned} E[Y_i^T Y_i] &= \sum_i Pr[Y_i = v_j] v_j^T v_j \\ &= \frac{1}{n} \sum_i v_i^T v_i \\ &= \frac{1}{n} V^T V \end{aligned}$$

And since by the definition of X_i and that V is orthonormal, we have

$$E[X_i] = E[I_d - n \cdot Y_i^T Y_i] = I_d - I_d = 0^{d \times d}$$

Now we consider:

$$\begin{aligned} E[X^T X + I_d] &= I_d + I_d - 2nE[Y_i^T Y_i] + n^2 E[Y_i^T Y_i Y_i^T Y_i] \\ &= 2I_d - 2I_d + n^2 \sum_i (1/n) v_i^T v_i v_i^T v_i \\ &= n \sum_i v_i^T v_i |v_i|_2^2 \end{aligned}$$

Now that we have derive the expectation, we wish to apply the flattening lemma here.
Define:

$$Z = n \sum_i v_i^T v_i C \log(nd/\delta) \cdot (d/n) = C^2 d \log(nd/\delta) \cdot I_d$$

Note that the $X^T X + I_d$ and Z are real and symmetric with non negative eigenvalues.

Claim 1. for all vectors y, we always have

$$y^T E[X^T X + I_d] y \leq y^T Z y$$

Proof. Just consider that the expectation contains the dot product of v_i and y, we then again apply the flattening lemma to show that we have

$$y^T Z y = d \sum_i \langle v_i, y \rangle^2 C^2 \log(nd/\delta)$$

. ∎

Hence, we have a bound on the operator norm of expectation of the covariance matrix: $\|E[X^T X]\|_2 = O(d \log(nd/\delta))$. We can use the matrix chernoff bound now. We apply the matrix chernoff onto the matrix $I_d - (PHDA)^T (PHDA)$.

$$Pr[|I_d - (PHDA)^T (PHDA)|_2 \geq \epsilon] \leq 2d \exp(-\frac{s\epsilon^2}{\Theta(d \log(nd/\delta))}).$$

We now set δ to be reasonable amount so that we have the probability less than $\delta/2$.

With the operator norm bounded, we can now show that we can construct a subspace embedding now with this setup.

$$
\begin{aligned}
\forall x \text{ unit vector}, |x^T(I_d - (PHDA)^T(PHDA))x| < \epsilon \\
&\iff |x^T x - x^T(PHDA)^T(PHDA)x| < \epsilon \\
&\iff |I - |(SAx)|_2^2| < \epsilon \\
&\implies |(SAx)|_2^2 \in [1-\epsilon, 1+\epsilon]
\end{aligned}
$$

Having shown that we have a subspace embedding, we apply the trick in the case of Guassian sketch matrices S to come up with an answer to the original regression problem.

This technique gives an algorithm with running time

$$
O(nd \log n) + \text{poly}(\frac{d \log n}{\epsilon})
$$

.

2 CountSketch Matrices & even faster subspace embeddings

We now make use of CountSketch matrices to achieve even faster subspace embeddings.

Definition (CountSketch Matrix)**.** Matrix S is a $k \times n$ matrix with $k = O(d^2/\epsilon^2)$. Each column of S has exactly one non-zero entry, which is either $+1$ or -1 with equal probability.

Remark 2. note that we can compute SA in $nnz(A)$ time. Because in reality we can keep track of a list of indices of those non-zero entries and then we can just index into A to get the the product. The rest does not matter and ends up as zero anyway.

Now we show how we can construct a subspace embedding with CountSketch matrices. As usual, we have A to be orthonormal. We wish to show that

$$
|SAx|_2^2 \in [1-\epsilon, 1+\epsilon]
$$

. Suffices to show that

$$
|A^T S^T SA - I|_2 \leq |A^T S^T SA - I|_F \leq \epsilon
$$

with high probability.

Lemma 2 (approximate matrix multiplication)**.**

$$
Pr\left[|CS^T SD - CD|_F^2 \leq \left(\frac{6}{\textbf{\textit{number of rows of }} S}\right)|C|_F^2|D|_F^2\right] \geq 1-\delta
$$

Making use of the above lemma, we can show that if we conveniently let $C = A^T$ and $D = A$, we have $|A|_F^2 = d$ and number of rows of S as $6d^2/(\delta\epsilon^2)$. Thus we have shown, again, S will give us a subspace embedding.

We now shift attention to proving the above lemma.

Lemma 3 (JL property)**.** *A matrix S has the (ϵ, δ, ℓ)-JL moment property if for all unit $x \in R^n$, we have*

$$E_S \left| |Sx|_2^2 - 1 \right|^\ell \leq \epsilon^\ell \cdot \delta$$