| | |
|---|---|
| **CS 15-859: Algorithms for Big Data** | Fall 2022 |

Lecture 2 Part 1 — September 15, 2022

*Prof. David Woodruff* | *Scribe: Ashley Wang*

# Back to Regression

Previously, we showed that $S$ is a subspace embedding, that is, simultaneously for all $x$,

$$\|SAx\|_2 = (1 \pm \varepsilon)\|Ax\|_2$$

Bringing it back to regression, we want $x$ so that $\|Ax-b\|_2 \leq (1+\varepsilon)\min_y \|Ay-b\|_2$. By considering the subspace $L$ spanned by columns of $A$ together with $b$, we get that for all $y \in L$, $\|Sy\|_2 = (1 \pm \varepsilon)\|y\|_2$. Hence, $\|S(Ax-b)\|_2 = (1 \pm \varepsilon)\|Ax-b\|_2$ for all $x$. Given $SA$ and $Sb$, we can solve

$$\arg\min_y \|(SA)y - (Sb)\|_2$$

in poly$(d/\varepsilon)$ time using the normal equations. However, the problem that we are faced with is computing $SA$ takes $O(nd^2/\varepsilon^2)$ time, so we need to reconsider what sketching matrix we choose.

## How to Choose the Right Sketching Matrix $S$

**Definition.** $S$, a $k \times n$ matrix, is a Subsampled Randomized Hadamard Transform such that:

- $S = P * H * D$
- $D$ is an $n \times n$ diagonal matrix with $\pm 1$ on the diagonals. Diagonal entries are either $+1$ or $-1$ with equal probability.
- $H$ is the $n \times n$ Hadamard matrix such that

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

  where $\langle i,j \rangle$ denotes the inner product between the binary representations of $i$ and $j$. Note that $H$ is orthonormal.
- $P$ chooses a random (small) subset of rows from $H * D$.

### Important properties to know about $H$

- The matrix $H$ has a recursive structure: Let $H_n$ denote the $n \times n$ Hadamard matrix. Let $H'_n = \sqrt{n} H_n$. Since $H'_n$ has a recursive structure, so does $H$.

  $$H'_n = \begin{bmatrix} H'_{n/2} & H'_{n/2} \\ H'_{n/2} & -H'_{n/2} \end{bmatrix}$$

- $H_n$ can be applied to a vector $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ where $x_1, x_2 \in \mathbb{R}^{n/2}$ in $O(n \log n)$ time.

  We do this by computing $H_n = \left(\frac{1}{\sqrt{n}}\right) H'_n$ like so:

$$
\begin{aligned}
H_n x &= \left(\frac{1}{\sqrt{n}}\right) H'_n x \\
&= \left(\frac{1}{\sqrt{n}}\right) \cdot \begin{bmatrix} H'_{n/2} & H'_{n/2} \\ H'_{n/2} & -H'_{n/2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
&= \left(\frac{1}{\sqrt{n}}\right) \cdot \begin{bmatrix} H'_{n/2} x_1 & H'_{n/2} x_2 \\ H'_{n/2} x_1 & -H'_{n/2} x_2 \end{bmatrix}
\end{aligned}
$$

  In order to apply $H_n$ to vector $x$, we perform two recursive calls to inputs of size $n/2$, and then combine the results in $O(n)$ time. The recurrence is:

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

- Any two columns (and any two rows) of $H$ are orthogonal to each other. We can look at the inner product of two columns of $H$:

$$
\begin{aligned}
\langle H_{*j}, H_{*j'} \rangle &= \sum_i H_{ij} H_{ij'} \\
&= \frac{1}{n} \sum (-1)^{\langle i,j \rangle} \cdot (-1)^{\langle i,j' \rangle} \\
&= \frac{1}{n} \sum (-1)^{\langle i,j+j' \rangle \mod 2}
\end{aligned}
$$

  We can set up a bijection, e.g. WLOG $j$ and $j'$ differ at the $k$th entry only. This allows us to write the above expression as:

$$
= \frac{1}{n} \sum_{\substack{i \text{ s.t.} \\ i_k = 1}} (-1)^{\langle i,j+j' \rangle \mod 2} + \frac{1}{n} \sum_{\substack{i \text{ s.t.} \\ i_k = 0}} (-1)^{\langle i,j+j' \rangle \mod 2}
$$

  One of these terms will be positive, and the other negative, so they sum to 0. From this, we conclude that columns (and rows) of $H$ are orthogonal.

- If $n$ is not a power of 2, we can add padding to our matrices.

**Important properties to know about $S$**

- $S = PHD$ can be applied to a vector $x$ in $O(n \log n)$ time.

  Applying $D$ to $x$ takes $O(n)$ time, applying $H$ takes $O(n \log n)$ time, and applying $P$ takes $O(\leq n)$ time because we select a subset of $n$ rows. In total, this is in $O(n \log n)$ time.

- $S$ can be applied to a matrix $A$ in $O(nd \log n)$ time. This follows from the above property.

Now, it suffices to show that $\|SAx\|_2^2 = \|PHDAx\|_2^2 = 1 \pm \varepsilon$ for all unit $x$.

**Remark 1.** $\|HDAx\|_2^2 = \|Ax\|_2^2$ for unit $x$, since $H * D$ is a rotation matrix ($H$ and $D$ are both orthonormal).

Sampling wouldn't work to preserve the norm if there is a relatively large entry and the others are zero. At a high level, what $HDAx$ does is a structured random rotation to spread out the Euclidean norm across the vector so that row sampling works well at preserving the norm (scaled).

**Lemma 1.** *(Azuma-Hoeffding Bound). Let $Z_1, Z_2, \ldots, Z_k$ be zero-mean independent random variables where $|Z_j| < \beta_j$ with probability 1. Then*

$$\mathbb{P}\left[\left|\sum_{j=1}^{k} Z_j\right| > t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{j=1}^{k}\beta_j^2}\right)$$

**Lemma 2.** *(Flattening Lemma). For any fixed unit vector $y$ and some constant $C$,*

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

**Proof of Lemma 2.** Let $C > 0$ be a constant. We apply the Azuma-Hoeffding bound on $(HDy)_i$ for all $i$. We know that:

- $Z_j := H_{ij}D_{jj}y_j$ is zero-mean in expectation because $D_{jj}$ is zero-mean ($\pm 1$ with equal probability).
- $(HDy)_i = \sum_j H_{ij}D_{jj}y_j = \sum_j Z_j$.
- With probability 1, $|Z_j| \leq \frac{|y_j|}{\sqrt{n}} =: \beta_j$.

Applying these observations,

$$\sum_{j=1}^{d}\beta_j^2 = \sum_{j=1}^{d}\frac{y_j^2}{n} = \frac{\|y\|_2^2}{n} = \frac{1}{n}$$

Applying Azuma-Hoeffding bound gives:

$$\mathbb{P}\left[\left|\sum_{j=1}^{d} Z_j\right| > C\sqrt{\frac{\log(nd/\delta)}{n}}\right] \leq 2e^{-\frac{C^2\log(nd/\delta)}{2}}$$

By choosing an appropriate constant $C$, we can get

$$\mathbb{P}\left[\left|\sum_{j=1}^{d} Z_j\right| > C\sqrt{\frac{\log(nd/\delta)}{n}}\right] \leq \frac{\delta}{2nd}$$

Now, we union bound over all $i$ to get the Flattening Lemma.

**Consequences of the Flattening Lemma**

**Remark 2.** $HDA$ has orthonormal columns, since $A$ has orthonormal columns and both $H$ and $D$ are rotation matrices.

Let $x = e_i$ for some fixed $i \in [d]$. Applying the Flattening Lemma tells us that

$$\|HDAe_i\|_\infty \leq C\sqrt{\frac{\log(nd/\delta)}{n}} \qquad \text{with probability } 1 - \frac{\delta}{2d}$$

By applying the union bound over all $d$ columns, the probability that all entries in all columns of $HDA$ have abslute value at most $C\sqrt{\frac{\log(nd/\delta)}{n}}$ is at least $1 - \frac{\delta}{2d}$.

Hence, we achieve the upper bound on the norm of a row of $HDA$:

$$\|e_j^T HDA\|_2 \leq C\sqrt{\frac{d\log(nd/\delta)}{n}}$$

# Matrix Chernoff Bound

Here, $|X|_2$ represents the operator norm of matrix $X$, or the maximum singular value of $X$.

**Theorem 3.** *Let $X_1, X_2, \ldots, X_s$ 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 $\left|\mathbb{E}[X^T X]\right|_2 \leq \sigma^2$. Let $W = \frac{1}{s}\sum_{i\in[s]} X_i$. For any $\varepsilon > 0$,*

$$\mathbb{P}\left[|W|_2 > \varepsilon\right] \leq 2d \cdot e^{-s\varepsilon^2/(\sigma^2+\gamma\varepsilon/3)}$$

*Here, $|W|_2 = \sup_x \frac{|Wx|}{|x|}$.*

Now, let's define our matrix $X$.

Let $V = H * D * A$, where $V$ has orthonormal columns. Suppose $P$ in the $S = PHD$ definition samples $s$ rows uniformly with replacement and scales them by $\frac{\sqrt{n}}{\sqrt{s}}$. There is a proof version of doing so without replacement, but turns out you don't gain much from doing so. If row $j$ is sampled in the $i$th sample, then $P_{i,j} = \frac{\sqrt{n}}{\sqrt{s}}$ and 0 otherwise.

Let $Y_i$ be the $i$th sampled row of $V = H * D * A$.

Let $X_i = I_d - nY_i^T Y_i$. To show that $X_i$ satisfies the assumptions for the Matrix Chernoff Bound, first let's show that $\mathbb{E}[X_i] = 0$.

- Since $Y_i$ is sampled uniformly, we get

$$\mathbb{E}[Y_i^T Y_i] = \sum_{j=1}^{n} \mathbb{P}[Y_i = v_j] \cdot v_j^T v_j = \sum_{j=1}^{n} \frac{1}{n} \cdot v_j^T v_j = \frac{1}{n} V^T V$$

Notice that $V^T V = I_d$ because $V$ has orthogonal columns. From this,

$$\begin{aligned}
\mathbb{E}[X_i] &= \mathbb{E}[I_d - nY_i^T Y_i] \\
&= I_d - n\mathbb{E}[Y_i^T Y_i] \\
&= I_d - n \cdot \frac{1}{n} V^T V \\
&= I_d - I_d \\
&= 0
\end{aligned}$$