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 \( \arg\min_y |SAy - Sb|_2 \) using deterministic methods.
5. This takes \( \text{poly}(d/\varepsilon) \) 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
   \[
   H_{i,j} = \frac{(-1)^{<i,j>}}{\sqrt{n}}
   \]
   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 \( <i,j> = (\sum_{l=0}^{k-1} i_l \cdot j_l) \mod 2 \) where \( i_l \) is the \( l \)-th bit in the \( k \)-bit binary representation of \( i \).
**Claim:** $H$ is orthogonal.

**Proof:** We have

- $\forall i \in [n]$ we have
  
  $$|H_i|^2 = \sum_{l=0}^{n-1} H_{i,l}^2 = \sum_{l=0}^{n-1} \frac{1}{n} = 1$$

- $\forall i, j \in [n]$ such that $i \neq j$ we have
  
  $$\langle H_i, H_j \rangle = \sum_{l=0}^{n-1} \frac{1}{n} (-1)^{(i,l)} (-1)^{(j,l)} = \sum_{l=0}^{n-1} \frac{1}{n} (-1)^{(i+j,l)}$$

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 $\alpha$ with the number that only differs at index $t$, which we can call $\beta$. Now we have $(-1)^{(i+j,\alpha)} + (-1)^{(i+j,\beta)} = 1 + (-1) = 0$ 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 $y$:

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

where $C > 0$ is a constant.

#### 2.2.1 Proof

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

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

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 $D_{i,j}$ are independent random variables with 0 mean. Therefore $Z_j$ are independent with 0 mean as well.

• $|Z_j| \leq |H_{i,j}| \cdot |D_{i,j}| \cdot |y_j| \leq \frac{1}{\sqrt{n}} \cdot 1 \cdot |y_j| = \frac{|y_j|}{\sqrt{n}}.$

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:

$$\Pr[|\sum_j Z_j| > t] \leq 2e^{-\frac{t^2}{2\sum_j |y_j|^2}} = 2e^{-\frac{n t^2}{4}}$$

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

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

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 $1 - \frac{\delta}{2}$ 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 \( QAQ^{-1} \) where the \( i \)th column of \( Q \) is given by the \( i \)th eigenvector and \( \Lambda \) is a diagonal matrix where \( \Lambda_{ii} \) is the \( i \)th eigenvalue. If \( W \) is real and symmetric then \( Q \) is orthogonal and therefore the eigendecomposition can be given as \( QAQ^T \).

### 2.3.1 Setup

Let’s define \( V = HDA \) and let \( Y_i \) be the \( i \)th sampled row of \( V \). Also define \( X_i = I_d - nY_i^TY_i \) which gives us

\[
|X_i|_2 \leq |I|_2 + n \max_j |e_jHDA|^2 \quad \text{[Triangle Inequality for operator norm]}
\]

\[
= 1 + nC^2 \left( \frac{d \log \left( \frac{nd}{\delta} \right)}{n} \right) \quad \text{[Flattening Lemma]}
\]

\[
= 1 + C^2d \log(\frac{nd}{\delta}) \in O(d \log(\frac{nd}{\delta}))
\]

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

\[
E[X^TX + I_d] = E[(I_d - nY_i^TY_i)^T(I_d - nY_i^TY_i) + I_d]
\]

\[
= I_d + I_d - 2nE[Y_i^TY_i] + n^2E[Y_i^TY_iY_i^TY_i]
\]

\[
= 2I_d - 2n \left( \frac{1}{n} I_d \right) + n^2E[Y_i^TY_iY_i^TY_i]
\]

\[
= n^2 \sum_i \frac{1}{n} v_i^Tv_i
\]

\[
= n \sum_i v_i^Tv_i \cdot |v_i|^2
\]

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

**Claim:** All eigenvalues of \( E[X^TX + I_d] \) and \( Z \) are non-negative. Also for all \( x \) we have \( x^TE[X^TX + I_d]x \geq x^TZx \).

**Proof:** Since both matrices are real and symmetric, their eigendecomposition is given by \( QAQ^T \). Therefore if \( x \) is an eigenvector of \( E[X^TX + I_d] \) then the corresponding eigenvalue \( \lambda \) is

\[
\lambda = x^T(n \sum_i v_i^Tv_i \cdot |v_i|^2) x
\]

\[
= n \sum_i (v_i^T x)^2 \cdot |v_i|^2 \geq 0
\]

Similarly if \( y \) is an eigenvector of \( Z \) then the corresponding eigenvalue \( \lambda \) is

\[
\lambda = y^T(n \sum_i v_i^Tv_i \cdot C^2 \frac{n}{d} \log(\frac{nd}{\delta})) y
\]

\[
= n \sum_i (v_i^T y)^2 \cdot C^2 \frac{n}{d} \log(\frac{nd}{\delta}) \geq 0
\]
Therefore all the eigenvalues of $\mathbb{E}[X^TX + I_d]$ and $Z$ are non-negative. Also since $C^2 n^2 \log(\frac{nd}{\delta})$ is an upper bound for $|v_i|^2$, the second part of our claim is immediate.

**Claim:** $|\mathbb{E}[X^TX + I_d]|_2 \leq |Z|_2$.

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

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

### 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

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

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

\[
W = \frac{1}{s} \sum_{i=0}^{s-1} X_i
\]
\[
= \frac{1}{s} \sum_{i=0}^{s-1} (I_d - nY_i^TY_i)
\]
\[
= I_d - \frac{n}{s} \sum_{i=0}^{s-1} Y_i^TY_i
\]
\[
= I_d - \sum_{i=0}^{s-1} (Y_i^T \sqrt{\frac{n}{s}} \sqrt{\frac{n}{s}} Y_i)
\]
\[
= I_d - (PHDA)^T (PHDA)
\]

Since $Y_i$ is the $i$th 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

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

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

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

which is equivalent to

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

### 2.4 Satisfying preconditions for subspace embeddings

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

$$
\varepsilon \geq |I_d - (PHDA)^T (PHDA)|_2 \geq |x^T (I_d - (PHDA)^T (PHDA)) x| = |x^T x - (PHDAx)^T (PHDAx)| = |1 - |PHDAx|_2^2|
$$

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.