CSC2240 Class Project – Oblivious Subspace Embeddings

Aida Ramezani & Marc-Etienne Brunet

December 2022

1 Introduction

Oblivious subspace embeddings (OSEs) have numerous applications in approximation and sketching algorithms. They are also interesting in their own right, challenging our intuitions about information compression, and the geometry of high dimensional space.

1.1 Outline

In this project we cover the following:

1. Review of the Johnson-Lindenstrauss (JL) Lemma seen in class.
2. Introduction to oblivious subspace embeddings (OSEs).
3. Application of OSEs to linear regression.
4. Motivation for sparse OSEs.
5. Review of the trace inequalities proof from [Cohen, 2016].

1.2 Notation

To facilitate readability, we use the following notation throughout.

n	dimension in original space
m	dimension of embedding
d	dimension of subspace
k	number of vectors preserved by JL
Π	Embedding matrix $\in \mathbb{R}^{m \times n}$

2 Review of the JL-Lemma

In class previously, we saw the *Johnson-Lindenstrauss* (JL) Lemma, which showed that a set of k vectors in $\mathbb{R}^n$, can be embedded into $m = \mathcal{O}(\epsilon^{-2} \log k)$ dimensions, while approximately preserving all pairwise distances simultaneously, up to a $1 + \epsilon$ distortion factor. Moreover, it is possible to sample a *linear* embedding function, in polynomial-time, and that function does not depend the set of vectors, i.e., it is *oblivious*.

Theorem 2.1 (Johnson–Lindenstrauss Lemma)**.** *For any $\epsilon \in (0, 1/2]$ and $V = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k\} \subset \mathbb{R}^n$, there exists a linear mapping $\Pi : \mathbb{R}^n \mapsto \mathbb{R}^m$ for $m = \mathcal{O}(\epsilon^{-2} \log k)$ such that*

$$\forall i, j \quad (1-\epsilon)\big\|\boldsymbol{x}_i - \boldsymbol{x}_j\big\| \leq \big\|\Pi\boldsymbol{x}_i - \Pi\boldsymbol{x}_j\big\| \leq (1+\epsilon)\big\|\boldsymbol{x}_i - \boldsymbol{x}_j\big\| \tag{1}$$

While it was not given this name in class, the JL-Lemma follows from another result known as the Distributional Johnson–Lindenstrauss Lemma.

Theorem 2.2 (Distributional Johnson–Lindenstrauss Lemma)**.** *There exists a distribution over random matrices, $\Pi : \mathbb{R}^n \mapsto \mathbb{R}^m$, such that if $m = \mathcal{O}(\epsilon^{-2} \log(1/\delta))$, then for any $\boldsymbol{x} \in \mathbb{R}^n$*

$$\mathbf{Pr}\big[(1-\epsilon)\|\boldsymbol{x}\| \leq \|\Pi\boldsymbol{x}\| \leq (1+\epsilon)\|\boldsymbol{x}\|\big] \geq 1 - \delta.$$

This follows easily from the "main lemma" used in class to prove of the JL-Lemma. We can similarly prove Theorem 2.1 from 2.2 with a union bound over the $\binom{k}{2}$ distances vectors in V. Picking $\delta = \delta'/\binom{k}{2} = \Omega(\delta'/k^2)$, it follows that if $m = \mathcal{O}(\epsilon^{-2} \log(k^2/\delta')) = \mathcal{O}(\epsilon^{-2} \log(k/\delta'))$, then Equation 1 occurs with probability at least $1 - \delta'$. (In class we let $\delta' = 1/k$.)

3 Oblivious Subspace Embeddings

Similar to the JL-Lemma, oblivious subspace embeddings (OSEs), are oblivious, ϵ-distortion, random embeddings. However, rather than just embedding a finite set of k vectors, they are concerned with embedding an entire d-dimensional subspace. In the main work that we covered, [Cohen, 2016], OSEs are defined as follows:

Definition 3.1 (Oblivious Subspace Embedding)**.** A probability distribution over m by n matrices, Π, is defined to be a (d, ϵ, δ)-OSE if, for any d-dimensional subspace S of $\mathbb{R}^n$,

$$\mathbf{Pr}\left[\left(\max_{\boldsymbol{x} \in S, \|x\|=1} \Big|\|\Pi\boldsymbol{x}\|^2 - 1\Big|\right) > \epsilon\right] < \delta. \tag{2}$$

From what we've found, the extension of the JL-Lemma to subspaces was originally done by [Sarlos, 2006], who showed that a d-dimensional *subspace* of $\mathbb{R}^n$, could be embedded into $m = \mathcal{O}(\epsilon^{-2} d \log(d/\epsilon))$ dimensions and similarly approximately preserve the lengths of all vectors in the subspace. The allowable distortion is visualized in Figure 1 (left).

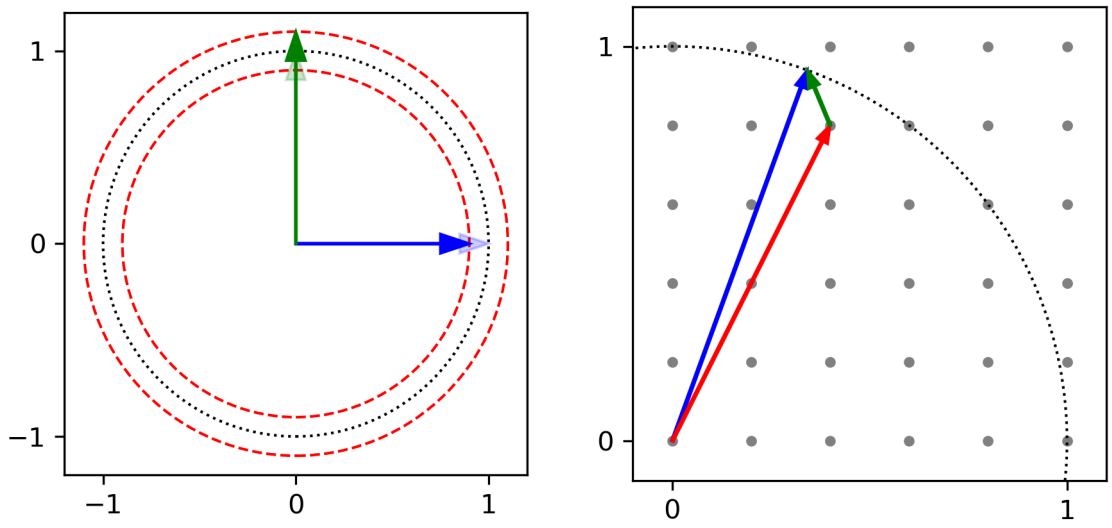

Figure 1: Left – Visualization of allowable distortion vectors in the preserved subspace S. Unit vectors must not be distorted out of the $\pm\epsilon$ shell. Right – Visualization of the epsilon net used to prove Lemma 3.2. The blue vector is an arbitary vector on the unit sphere. Using the triangle inequality, its magnitude is less than the sum of the red and green vectors. Bounding the distortion of the red vector is easy, since it is an element of the net. The distortion of the green vector requires a little more work, but with a sufficiently fine net, this component is very small, and the sum of their distortions are shown to be within the $(1 \pm \epsilon)$ tolerance.

Lemma 3.2. *Let S be an arbitrary d dimensional subspace of $\mathbb{R}^n$ and $\epsilon \in (0, 1/2]$, $\delta < 1$. If Π is a JL transform from $\mathbb{R}^n$ to $m = \mathcal{O}(\epsilon^{-2} d \log(d/\epsilon) \cdot f(\delta))$ dimensions for some function f, then*

$$\mathbf{Pr}\left[\forall \boldsymbol{x} \in S \quad (1-\epsilon)\|\boldsymbol{x}\| \leq \|\Pi \boldsymbol{x}\| \leq (1+\epsilon)\|\boldsymbol{x}\|\right] \geq 1 - \delta.$$

Proof Sketch: The key idea of the proof is then to cover the unit sphere with an epsilon-net that is sufficiently fine that every $\boldsymbol{x} \in S$ is close enough to a point in the net. This is visualized in Figure 1(right). The epsilon net used in the proof has $\mathcal{O}((k/\epsilon)^k)$ elements. Plugging that many elements into the union bound used to prove 2.1, in place of $\binom{k}{2}$, we get to the stated complexity. For more details, see Lemma 10 in [Sarlos, 2006]. □

4 Application to linear regression

4.1 Motivation and setup

Consider linear regression. We want to find a vector $w^* \in \mathbb{R}^d$ such that

$$w^* = \arg\min \|Aw - b\|^2. \tag{3}$$

With $A \in \mathbb{R}^{n \times d}$ being the data matrix (n observations of d features), and $n \gg d$. The solution has a closed form expression, $w^* = (A^T A)^{-1}(A^T b)$. This solution requires roughly $\mathcal{O}(nd^2)$ time for the

A^TA multiplication, $\mathcal{O}(nd)$ time for the A^Tb multiplication, and $\mathcal{O}(d^3)$ time to solve the remaining linear system.

Since we assume $n \gg d$, this is dominated by $\mathcal{O}(nd^2)$. It would therefore be very interesting to reduce n in A, at the expense of obtaining an inexact solution $\tilde{w}$, so long as we could bound the relative error of the approximation. An obvious way to reduce n is to simply sub-sample the rows of A, i.e., the observations. But it is unclear whether we can sub-sample those rows with guarantees on the error any faster than just solving the original problem [Sarlos, 2006].

4.2 OSEs for approximate regression

Reducing the dimensionality of A with an OSE is a promising approach, and is used in practice. Define

$$\tilde{w} = \arg\min_{w \in \mathbb{R}^n} \|\Pi A w - \Pi b\|^2.$$

It was shown by [Sarlos, 2006] that:

Claim 4.1. *If $m = \Omega(\epsilon^{-2})$, then with probability at least $2/3$,*

$$\|\Pi b - \Pi A\tilde{w}\| \leq (1+\epsilon)\big\|b - Aw^*\big\| \tag{4}$$

Claim 4.2. *If $m = \Omega(\epsilon^{-1} d \log(d))$, then with probability at least $1/3$,*

$$\|b - A\tilde{w}\| \leq (1+\epsilon)\big\|b - Aw^*\big\| \tag{5}$$

Claim 4.1 follows directly from the Distributional JL Lemma.

Proof. (of Claim 4.1) Simply apply the JL transform Π to $b - Aw^*$.

$$\begin{aligned}\|\Pi b - \Pi A\tilde{w}\| = \big\|\Pi(b - A\tilde{w})\big\| &\leq \big\|\Pi(b - Aw^*)\big\| \\ &\leq (1+\epsilon)\big\|b - Aw^*\big\|\end{aligned}$$

□

However, Claim 4.1 is not a terribly interesting result. It simply states that the error in the embedded space, is close to the error in the original space. From the standpoint of bounding the error of the approximate solution, Claim 4.2 is much more useful. Since this tells us about the error of the approximate solution when applied in the original space, which is the problem we actually care about.

Intuitively, if we embed the whole $(d+1)$ dimensional subspace spanned by the columns of A and the vector b, then we preserve the optimization problem. This is confirmed formally in [Woodruff, 2014, p31], although this author questions his proof approach in Appendix A. According to Lemma 3.2, embedding this subspace should require $m = \mathcal{O}(\epsilon^{-2} d \log(d/\epsilon)$ dimensions. The relaxation to the ϵ^{-1} dependence in Claim 4.2 is not immediately obvious. The key to its proof is to use the *normal equations*, i.e., the exact solution to the linear regression mentioned above,

$$w^* = (A^TA)^{-1}A^Tb. \tag{6}$$

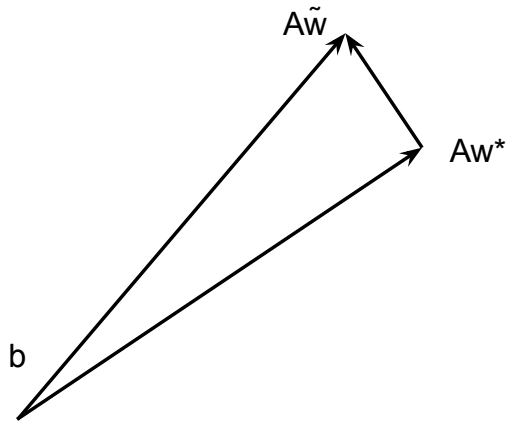

Figure 2: Visualization of the vectors in Equation 7.

From these equations we note that

$$A^T Aw^* - A^T b = 0$$
$$A^T(Aw^* - b) = 0,$$

i.e., $(Aw^* - b)$, the residual vector, is orthogonal to the columns of A. As such, $(A\tilde{w} - b)$ can be written as the sum of two orthogonal vectors, $(Aw^* - b)$ and $(A\tilde{w} - Aw^*)$. We can then use the Pythagorean theorem to note,

$$\|A\tilde{w} - b\|^2 = \big\|Aw^* - b\big\|^2 + \big\|A\tilde{w} - Aw^*\big\|^2. \tag{7}$$

The proof thus only needs to show that $\|A\tilde{w} - Aw^*\|^2 = \mathcal{O}(\epsilon)\|Aw^* - b\|^2$. After an appropriate constant factor rescaling of epsilon, this implies

$$\begin{aligned}\|A\tilde{w} - b\| &= \sqrt{(1+\epsilon)}\big\|Aw^* - b\big\| \\ &\leq (1+\epsilon)\big\|Aw^* - b\big\|,\end{aligned}$$

i.e., we get Claim 4.2.

By taking this approach, we do not need the embedding to preserve the entire subspace of $\{\text{colspan}(A) \cup b\}$, but rather just need it to bound a few specific terms. One such term requires only a fixed error embedding, the other term involves an embedding with error that is square root in ϵ, thus reducing the bound to a ϵ^{-1} requirement. For the complete proof, we direct the reader to [Sarlos, 2006] or [Woodruff, 2014, p33].

4.3 Why do we want *sparse* subspace embeddings?

At a high level, we are interested in sparse subspace embeddings because the procedure for sampling a dense embedding matrix Π, then multiplying it through to embed the problem, may offset the computational benefit of embedding the problem in the first place.

For example, remember that in the Least Squares Regression problem, using subspace embedding matrix $\Pi \in \mathbb{R}^{m\times n}$, reduces the dimension to $\mathcal{O}(md^2)$. If we use a dense Gaussian matrix from JL-lemma of dimension $\mathcal{O}(\epsilon^{-2}d\log(d))$ for Π, it would still take $\mathcal{O}(nd^2\log(d))$ to compute ΠA. Therefore, if $n \gg d$, our running time is still dominated by n, which counteracts the computational benefit of solving the linear system in a smaller dimension.

We can speed up this process by working with a **sparse** random matrix Π. Such matrix would have exactly s non-zero entries on each column. Multiplying Π by any matrix, such as $A \in \mathbb{R}^{n\times d}$, takes $\mathcal{O}(snd)$ time. If A is also a sparse matrix, it would only take $\mathcal{O}(s \times Non - zero(A))$ to compute ΠA, where $Non - zero(A)$ is the number of non-zero entries in A.

To construct such matrix Π, one needs to randomly pick s entries per column and set them to $\pm\frac{1}{\sqrt{s}}$, randomly and uniformly as well. The rest of the entries should be zero. In fact, there is a trade-off between the sparsity s and the row count m. In the next section, we will show how matrix trace inequalities can give us an almost tight bound for (d, ϵ, δ)-OSE, where for any $B > 2$

$$m = \mathcal{O}(\epsilon^{-2}Bd\log(d/\delta))$$

and

$$s = \mathcal{O}(\epsilon^{-1}\log_B(d/\delta)).$$

Note, that here the parameter B controls the trade-off between s and m. If s is small, then the time complexity to compute ΠA depends only on the non-zero entries of the matrix A. However, the reduced dimension m could get so large that storing and working with any vector in this dimension becomes inefficient and costly.

4.4 Proof sketch of bounds

To construct the matrix $\Pi \in \mathbb{R}^{m\times n}$ we randomly select s entries on each column to be non-zero and we set them to be $+\frac{1}{\sqrt{s}}$ or $-\frac{1}{\sqrt{s}}$ with equal probabilities. To formalize this we can use two sets of random variables $\delta_{r,i}$ and $\sigma_{r,i}$, where

$$\delta_{r,i} = \begin{cases} 1 & \Pi_{r,i} \neq 0 \\ 0 & o.w \end{cases}$$

and

$$P(\sigma_{r,i} = 1) = P(\sigma_{r,i} = -1) = \frac{1}{2}.$$

Using this notation, we can write the entries in Π as $\Pi_{r,i} = \frac{1}{\sqrt{s}}\delta_{r,i}\sigma_{r,i}$, and proceed with the proof.

The next step is to quantify the subspace embedding distortion error, which we want to bound by $\mathcal{O}(\epsilon)$.

Fact 4.3. *The OSE matrix Π successfully embeds a subspace S with an orthonormal basis U if and only if*

$$\left\|(\Pi U)^T(\Pi U) - U^TU\right\| \leq \epsilon \tag{8}$$

Proof Sketch:

The reader can note that the spectral norm of a symmetric matrix A is its largest eigenvalue, which can be written as

$$\|A\| = \max_{\|e\|=1} e^T A e.$$

Therefore, the spectral norm of the symmetric matrix $(\Pi U)^T(\Pi U) - U^T U$ would be

$$\left\|(\Pi U)^T(\Pi U) - U^T U\right\| = \max_{\|e\|=1} e^T((\Pi U)^T(\Pi U) - U^T U)e = \max_{\|e\|=1} \|\Pi e\|^2 - \|e\|^2$$

where, $\|\Pi e\|^2 - \|e\|^2$ is the distortion error of the subspace embedding, which we want to bound by $\mathcal{O}(\epsilon)$. [Nelson and Nguyên, 2013] provide an alternative way to think about this problem. It suffices to write S as

$$S = \{x : \exists y \in \mathbb{R}^d, x = Uy\}.$$

Then Π is an OSE if and only if $\|\Pi x\| = (1 \pm \epsilon)\|x\|$ with high probability. Note that $\|\Pi x\| = \|\Pi U y\|$, and since U is a unitary matrix it preserves the norm, thus we have

$$\|\Pi U y\| = \|\Pi y\| = (1 \pm \epsilon)\|y\|.$$

This property is equivalent to having all the singular values of ΠU in the range of $(1-\epsilon, 1+\epsilon)$, meaning that the eigenvalues of $(\Pi U)^T(\Pi U)$ would be in the range of $((1-\epsilon)^2, (1+\epsilon)^2)$. Now since $U^T U = I$, then

$$(\Pi U)^T(\Pi U) = I + ((\Pi U)^T(\Pi U) - U^T U).$$

Following Weyl's inequality [Nelson and Nguyên, 2013] we can show that the eigenvalues of $(\Pi U)^T(\Pi U)$ are $1 \pm \left\|(\Pi U)^T(\Pi U) - U^T U\right\|$, which we need them to be in the range of $(-2\epsilon + \epsilon^2 + 1, 2\epsilon + \epsilon^2 + 1)$. Consequently if ϵ is small enough, then $\left\|((\Pi U)^T(\Pi U) - U^T U)\right\| \leq 3\epsilon = \mathcal{O}(\epsilon)$

□

Now, we can continue the proof by reducing the spectral norm $\left\|(\Pi U)^T(\Pi U) - U^T U\right\|$ and find an upper bond on m and s. One possible approach is to decompose $(\Pi U)^T(\Pi U)$ to its rows and use Matrix Chernoff concentration bound.

Theorem 4.4 (MATRIX CHERNOFF)**.** *Let A_i be independent random positive semi-definite matrices satisfying $\mathbb{E}[\sum A_i] = I$ and $\|A_i\| \leq O(\frac{log(d/\delta)}{\epsilon^2})$, then for any $\epsilon < 1$*

$$P(\left\|\sum A_i - I\right\| \leq \epsilon) \geq 1 - \delta.$$

To use the Matrix Chernoff bound, one can write the matrix $(\Pi U)^T(\Pi U)$ as the sum of random matrices:

$$(\Pi U)^T(\Pi U) = \sum_r (\Pi U)_r (\Pi U)_r^T \tag{9}$$

However, the norms of these uniform random matrices ($\left\|(\Pi U)_r^T(\Pi U)_r\right\| = \left\|(\Pi U)_r\right\|^2$) are not bounded, and they are also not independent. To handle this problem, we can show that the

contribution of the off-diagonal terms in $(\Pi U)^T(\Pi U)$ gives us a set of new random matrices that are equivalent to the OSE error matrix $(\Pi U)^T(\Pi U) - U^T U$:

$$\begin{aligned}
(\Pi U)^T(\Pi U) &= \sum_r (\Pi U)_r (\Pi U)_r{}^T \\
&= \frac{1}{s}\sum_r (\sum_i \delta_{r,i}\sigma_{r,i}u_i)(\sum_i \delta_{r,i}\sigma_{r,i}u_i{}^T) \\
&= \frac{1}{s}\sum_r (\sum_i \delta_{r,i}u_i u_i^T) + (\sum_{i\neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i u_j{}^T) \\
&\frac{1}{s}(\sum_i (\sum_r \delta_{r,i})u_i u_i{}^T) + \frac{1}{s}(\sum_r (\sum_{i\neq j} \delta_{r,i}\delta_{r,jr,i}\sigma_{r,j}u_i u_j{}^T) \\
&= U^T U + \frac{1}{s}\sum_r \sum_{i\neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i u_j{}^T
\end{aligned}$$

To find a bound on the spectral norm of $(\Pi U)^T(\Pi U) - U^T U$ we only need to find a matrix concentration bound on the contribution of rows in the off-diagonal entries. For each row, we now have to consider the collision incidents only, i.e., when multiple non-zero entries exist in a row of Π. We refer to these matrices by Z_r:

$$(\Pi U)^T(\Pi U) - U^T U = \frac{1}{s}\sum_r \sum_{i\neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i u_j^T = \frac{1}{s}\sum_r Z_r. \tag{10}$$

Even though the Z_r variables are still not independent, we can use an alternative form of Matrix Chernoff bound with a requirements on the conditional distribution of each random variable given the previous ones.

Lemma 4.5. *Let $A = \sum_i A_i$ be a sum of m random symmetric matrices such that for all i and all allowable values $A_1^{'}, A_2^{'}, \dots A_{i-1}^{'}$,*

$$\left\| \mathbb{E}[cA_i | A_1^{'}, A_2^{'}, \dots A_{i-1}^{'}] \right\| \leq C$$

then,

$$\mathbb{E}[\,Tr\big(\exp(cA)\big)] \leq dC^m.$$

Proof. Define S_i as the partial sum of A_j matrices: $S_i = \sum_{j=1}^{i} A_j$. Without loss of generality we can assume $c = 1$, since $cA_i = B_i$, where B_i's are our new random symmetric matrices. Now, for $i > 0$

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}\big(\exp(S_i)\big)] &= \mathop{\mathbb{E}}_{S_{i-1}}[\mathop{\mathbb{E}}_{A_i}[\mathrm{Tr}\big(\exp(A_i + S_{i-1})\big)]] \\
&\leq \mathop{\mathbb{E}}_{S_{i-1}}[\mathop{\mathbb{E}}_{A_i}[\mathrm{Tr}\big(\exp(A_i)\exp(S_{i-1})\big)]]
\end{aligned}$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{S_{i-1}}[\mathrm{Tr}\left(\mathop{\mathbb{E}}_{A_i}[\exp(A_i)]\exp(S_{i-1})\right)] \\
&\leq \mathop{\mathbb{E}}_{S_{i-1}}[\left\|\mathop{\mathbb{E}}_{A_i}[\exp(A_i)\right\| \mathrm{Tr}\big(\exp(S_{i-1})\big)] \\
&\leq C\,\mathbb{E}[\mathrm{Tr}\big(\exp(S_{i-1})\big)]
\end{aligned}
$$

The second line is a result of the Golden-Thompson lemma that shows

$$\mathrm{Tr}\big(\exp(A+B)\big) \leq \mathrm{Tr}\big(\exp(A)\exp(B)\big).$$

The third line follows the linearity of trace and matrix product as

$$\mathop{\mathbb{E}}_{S_{i-1}}[\mathop{\mathbb{E}}_{A_i}[\mathrm{Tr}\big(\exp(S_{i-1})\big)]] = \mathop{\mathbb{E}}_{S_{i-1}}[\mathrm{Tr}\big(\exp(S_{i-1})\big)].$$

We use the fact that for two positive semi-definite matrices A and B, $\mathrm{Tr}(AB) \leq \mathrm{Tr}(A)\|B\|$ to get the forth line, and by using lemma 4.6, we show that both $\exp(S_{i-1})$ and $\mathbb{E}_{A_i}[\exp(A_i)]$ are positive semi-definite. Since by definition $\exp(0) = I$, then $\mathbb{E}[\mathrm{Tr}\big(\exp(S_0)\big)] = d$, and, thus $\mathbb{E}[\mathrm{Tr}\big(\exp(A)\big)] = dC^m$. □

Lemma 4.6. *The exponential of a symmetric matrix A is positive semi-definite.*

Proof. (Of lemma 4.6) We have to show that $\exp(A)$ is positive semi-definite. Since A is symmetric it is diagonalizable, and can write it as

$$
\begin{aligned}
A &= PDP^{-1} \\
\exp(A) &= \sum_{k=0}^{\infty} \frac{1}{K!} A^k \\
\exp(A) &= \sum_{k=0}^{\infty} \frac{1}{K!} PD^kP^{-1} = P\exp(D)P^{-1}
\end{aligned}
$$

Now the eigenvalues of $\exp(A)$ are the values in the diagonal matrix $\exp(D)$ which are all positive. □

Because the conditional norms of the Z_r's are not bounded, we cannot use Lemma 4.5 directly. For example, when each column of Π has exactly $s-1$ non-zero entries before the last row, then with the conditional probability of 1, the last row of every column in Π is non-zero, which makes the norm of the last Z_r much larger than any constant value of C. This is because we would have to consider the contribution of **all** u_i and u_j's for any subspace:

$$
\begin{aligned}
Z_r &= \sum_{i \neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_iu_j^T \\
&= \sum_{i \neq j} \sigma_{r,i}\sigma_{r,j}u_iu_j^T
\end{aligned}
$$

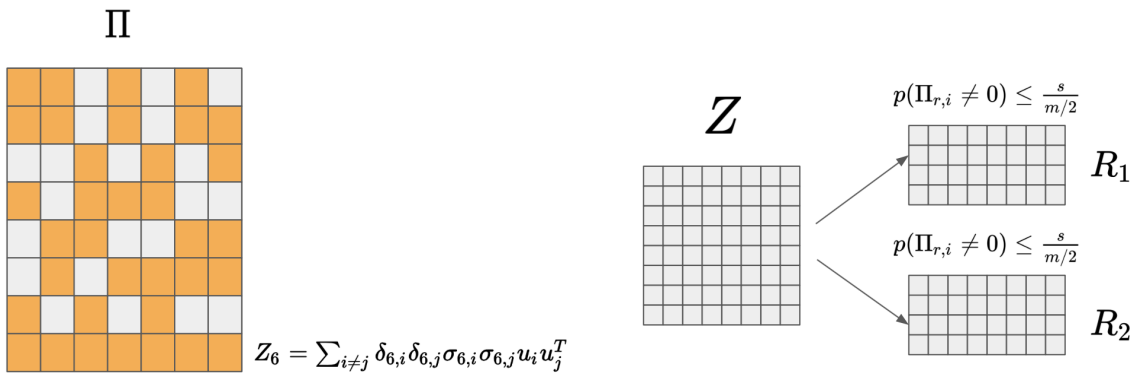

Figure 3: Visualization of splitting the matrix to two halves. The orange boxes in matrix Π specify the non-zero entries. Here $m = 8$ and $s = 5$. Each column of Π has exactly 4 non-zero entries before the last row.

This problem becomes particularly noticeable when conditioning on a large number of rows. However, if we look at the first half of rows in Π, the conditional probability of each entry being non-zero does not surpass $\frac{s}{m/2}$, even if all the previous rows in the half matrix are non-zero, so we can simply split the $\sum_r Z_r$ to two halves.

$$\sum_r Z_r = \sum_{r=1}^{\frac{m}{2}} Z_r + \sum_{r=\frac{m}{2}+1}^{m} Z_r = R_1 + R_2 \tag{11}$$

Even though R_1 and R_2 are not independent, we can apply Lemma 4.5 to each. The rest of the proof is straightforward. Using the Golden-Thompson lemma and the convexity of the trace, we can find an upper bound for $\mathbb{E}[\mathrm{Tr}\Big(\exp(c((\Pi U)^T(\Pi U) - U^T U))\Big)]$ by decomposing it as the contribution of the two halves:

$$\begin{aligned}
&\mathbb{E}[\mathrm{Tr}\Big(\exp(c((\Pi U)^T(\Pi U) - U^T U))\Big)] \\
&\qquad = \mathbb{E}[\mathrm{Tr}\left(\exp(\frac{c}{s}\sum_r Z_r)\right)] \\
&\qquad = \mathbb{E}[\mathrm{Tr}\left(\exp(\frac{c}{s}(R_1 + R_2))\right)] \\
&\qquad \leq \mathbb{E}[\mathrm{Tr}\left(\exp(\frac{2c}{s}R_1)\right)]
\end{aligned}$$

We can show that expected exponential value of the Z_r's in the first half are bounded:

$$\mathbb{E}[\exp(\frac{2c}{s}Z_r)] \preceq (2C^2 d + 1)I$$

where $\left\|\exp(\frac{4c}{s}x_r {x_r}^T) - I\right\| \leq C$ and $x_r = \sum_{i|w_i=0} \delta_{r,i}\sigma_{r,i}u_i$ and w_i's are $\{0,1\}$-valued random variables with $P(w_i = 0) = P(w_i = 1) = \frac{1}{2}$. The complete proof for acquiring this bound can be found at [Cohen, 2016, p6].

Now that we have a bound on the expected values of $\exp(\frac{2c}{s}Z_r)$, we can apply Lemma 4.5 on all Z_r variables in R_1 where $r \in 1, \ldots, \frac{m}{2}$:

$$\begin{aligned}\mathbb{E}[\left\|\exp(\frac{2c}{s}Z_r)\right\|] &\leq 2C^2d + 1 \\ &\leq \exp(2C^2d) \\ \mathbb{E}[\mathrm{Tr}\Big(\exp(\frac{2c}{s}R_1)\Big)] &\leq d\exp(C^2dm).\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathrm{Tr}\Big(\exp(c((\Pi U)^T(\Pi U) - U^TU))\Big)] \leq d\exp(C^2dm).$$

The following Theorem uses this bound to find the required number of rows m and sparsity s for Π to be a subspace embedding.

Theorem 4.7. *For any $B > 2$, $\delta < \frac{1}{2}$, $\epsilon < \frac{1}{2}$, a sparse embeddings matrix Π with $m = O(\frac{Bd\log(d/\delta)}{\epsilon^2})$ and $s = O(\frac{\log_B(d/\delta)}{\epsilon})$ satisfies*

$$P(\left\|(\Pi U)^T(\Pi U) - U^TU\right\| \leq \epsilon) \geq 1 - \delta \tag{12}$$

Proof Sketch: It suffices to fix a $c \propto \epsilon^{-1}\log(\frac{d}{\delta})$, such that $s/m \propto \epsilon/Bd$ and $c/s \propto \log B$ and $C^2dm = 1$. Applying the above lemma to these c and C gives us

$$\mathbb{E}[\mathrm{Tr}\Big(\exp(\epsilon^{-1}\log(d/\delta)((\Pi U)^T(\Pi U) - U^TU))\Big)] \leq ed.$$

Now, we can write

$$\begin{aligned}P(\left\|(\Pi U)^T\Pi U - U^TU\right\| \geq \epsilon) &= P(\epsilon^{-1}\log(\frac{d}{\delta})\left\|(\Pi U)^T\Pi U - U^TU\right\| \geq \log(\frac{d}{\delta})) \\ &= P(\exp(\epsilon^{-1}\log(\frac{d}{\delta})\left\|(\Pi U)^T\Pi U - U^TU\right\|) \geq \frac{d}{\delta}) \\ &= P(\mathrm{Tr}\Big(\exp(\epsilon^{-1}\log(\frac{d}{\delta})((\Pi U)^T\Pi U - U^TU))\Big) \geq \frac{d}{\delta}) \\ &\leq \frac{\mathbb{E}[\mathrm{Tr}\Big(\exp(\epsilon^{-1}\log(d/\delta)((\Pi U)^T(\Pi U) - U^TU))\Big)]}{\frac{d}{\delta}} \\ &\leq \frac{ed}{\frac{d}{\delta}} \\ &= \mathcal{O}(\delta) \\ P(\left\|(\Pi U)^T(\Pi U) - U^TU\right\| \leq \epsilon) &\geq 1 - \mathcal{O}(\delta)\end{aligned}$$

We can go from the second line to the third line by using lemma 4.6, which allows us to bound the spectral norm of the positive semi-definite matrix with its trace. Applying Markov's inequality on line 4 then gives us the desired bound.

□

References

[Cohen, 2016] Cohen, M. B. (2016). Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 278–287. SIAM.

[Nelson and Nguyên, 2013] Nelson, J. and Nguyên, H. L. (2013). Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pages 117–126. IEEE.

[Sarlos, 2006] Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pages 143–152. IEEE.

[Woodruff, 2014] Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *CoRR*, abs/1411.4357.

Appendix

A Question about Woodruff's proof

I struggled to understand the proof of a claim in the section about approximate linear regression in the reference that you recommended [Woodruff, 2014, page 31]. Here is a link to the pdf for convenience. He says that if we let S be the span of the columns of A union b, and we have a JL transform, Π, for which

$$\forall y \in S \quad (1-\epsilon)\|y\| \leq \|\Pi y\| \leq (1+\epsilon)\|y\| \tag{13}$$

then it follows that

$$\|b - A\tilde{w}\| \leq (1+\epsilon)\big\|b - Aw^*\big\|, \tag{14}$$

where

$$w^* = \arg\min_w \|b - Aw\| \qquad\qquad \tilde{w} = \arg\min_w \big\|\Pi(b - Aw)\big\|. \tag{15}$$

We do bound the error, however, from what I can tell, if we only invoke the bounds of Equation 13, and don't explicitly consider that Π is a linear transform, we get that

$$\begin{aligned}
\|b - A\tilde{w}\| &\leq \big\|b - Aw^*\big\| + \epsilon\big\|b - Aw^*\big\| + \epsilon\|b - A\tilde{w}\| \\
&\leq (1+\epsilon)\big\|b - Aw^*\big\| + \epsilon\|b - A\tilde{w}\| \\
&\leq \frac{(1+\epsilon)}{(1-\epsilon)}\big\|b - Aw^*\big\| \\
&\leq (1 + 2\epsilon + O(\epsilon^2))\big\|b - Aw^*\big\|
\end{aligned}$$

See the figure on the next page for a visual. I know that with the above bound we could rescale ϵ by a constant and keep the same order of the embedding dimension. My worse case analysis in the figure is also avoided due to the fact that Π is a linear transform. There's no way the $\tilde{w}$ could be in that position without non-linearities in the embedding. (And we know from the latter proof, we can get away with far fewer rows in the embedding.) But I still don't see how the $(1+\epsilon)$ bound simply "follows" from Equation 13. I can't tell if David's proof misses this subtlety, or if he's implicitly invoking something else that I don't see. Is that clear to you?

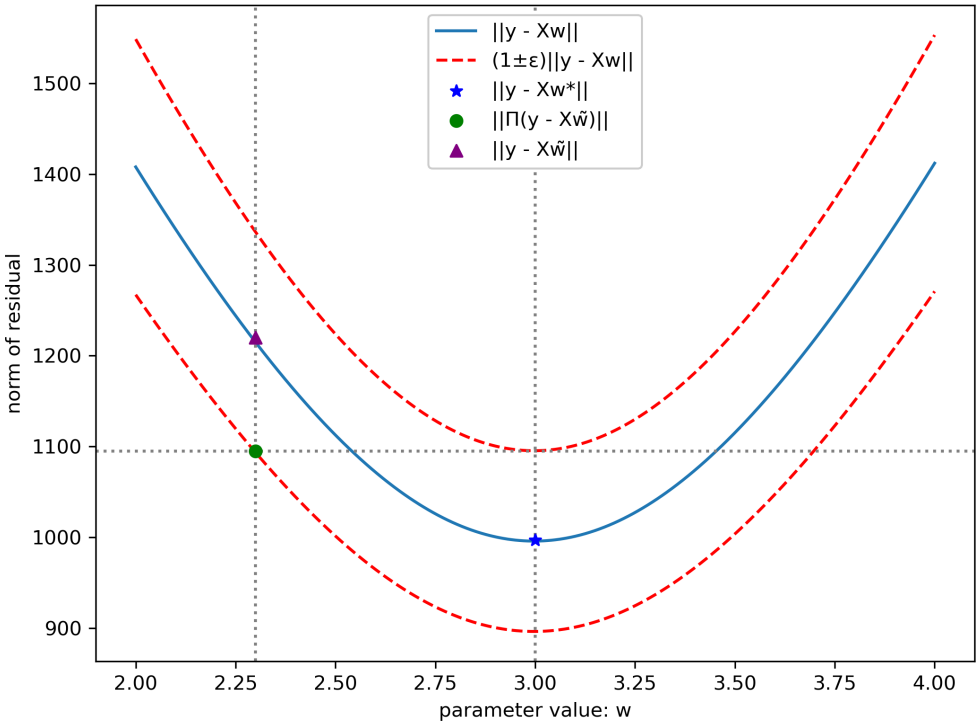

Figure 4: Plot of bounds on the residual.