Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities

Michael B. Cohen
S.B. Mathematics, MIT, 2014

Submitted to the Department of Electrical Engineering and Computer Science on May 20, 2016 in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 2016

Signature of Author: ______________________________
Department of Electrical Engineering and Computer Science
May 20, 2016

Certified by: ______________________________
Jonathan Kelner
Associate Professor of Applied Mathematics
Thesis Supervisor

Accepted by: ______________________________
Leslie A. Kolodziejski
Chair of the Committee on Graduate Students

Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities

Michael B. Cohen

Submitted to the Department of Electrical Engineering and Computer Science on May 20, 2016 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Electrical Engineering and Computer Science

Abstract

We present a new analysis of sparse oblivious subspace embeddings, based on the "matrix Chernoff" technique. These are probability distributions over (relatively) sparse matrices such that for any d-dimensional subspace of R^n, the norms of all vectors in the subspace are simultaneously approximately preserved by the embedding with high probability–typically with parameters depending on d but not on n. The families of embedding matrices considered here are essentially the same as those in [NN13], but with better parameters (sparsity and embedding dimension). Because of this, this analysis essentially serves as a "drop-in replacement" for Nelson-Nguyen's, improving bounds on its many applications to problems such as as least squares regression and low-rank approximation.

This new method is based on elementary tail bounds combined with matrix trace inequalities (Golden-Thompson or Lieb's theorem), and does not require combinatorics, unlike the Nelson-Nguyen approach. There are also variants of this method that are even simpler, at the cost of worse parameters. Furthermore, the bounds obtained are much tighter than previous ones, matching known lower bounds up to a single log(d) factor in embedding dimension (previous results had more log factors and also had suboptimal tradeoffs with sparsity).

Thesis Supervisor: Jonathan Kelner
Title: Associate Professor of Applied Mathematics

1 Introduction

Recently there has been substantial interest in the algorithms community in oblivious subspace embeddings (OSEs) and in particular *sparse* OSEs which can be applied efficiently. 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 R^n,

$$P[(\max_{\boldsymbol{x} \in S, \|\boldsymbol{x}\|=1} |\|\Pi x\|^2 - 1|) > \epsilon] < \delta.$$

That is, oblivious subspace embeddings must, with some given probability, *simultaneously* approximately preserve the norms of all vectors in a d-dimensional subspace, and furthermore they must be *oblivious*, with no dependence on the specific subspace being embedded. Note that for $d = 1$ the OSE property is essentially equivalent to the *distributional Johnson-Lindenstrauss* property.

These embeddings, and their applications to randomized numerical linear algebra algorithms, were popularized by [Sar06], which showed that Johnson-Lindenstrauss matrices of dimension about d/ϵ^2–notably, with no dependence on n–satisfied the property. However, using standard, dense Johnson-Lindenstrauss matrices (such as i.i.d. Gaussian or sign matrices) often does not lead to efficient algorithms, as multiplying by these matrices is slow. This can often be improved by substituting "fast Johnson-Lindenstrauss transform" variants ([AC09, AL13])–but even these methods fail to exploit sparsity in their running time: multiplying by any n-dimensional vector takes on the order of $n \log n$ time, even if it is very sparse.

An alternative approach was introduced in [CW13, MM13, NN13]. These papers proposed defining Π to be a *sparse* variant of random sign matrices: placing exactly s nonzero entries in each column, sampled randomly without replacement (and independently across columns), with each nonzero entry a random sign times a normalizer of $\frac{1}{\sqrt{s}}$. The simplest and most extreme version of this approach is $s = 1$: choosing a single random nonzero entry (with a random sign) for each column. This has a simple analysis based on the matrix second moment method, showing that $m = O\left(\frac{d^2}{\epsilon^2 \delta}\right)$ suffices. Unfortunately, the d^2 dependence is known to be tight ([NN14]). Multiplying a vector (or another matrix) by such a matrix is also extremely efficient, with runtime proportional only to its number of nonzero entries. In the more general case, the runtime is proportional to s times this number of nonzero entries.

[NN13] also examined the case where s is small but larger than 1: in particular, between $\Theta(1/\epsilon)$ and $\Theta(\text{polylog}(d)/\epsilon)$. In this range, the paper established a tradeoff between row count and sparsity, obtaining a result showing that for any $B > 2$, m could be set to about $Bd \log(d/\delta)^8/\epsilon^2$ with s about $\log_B(d/\delta)^3/\epsilon$. Notably, if B is set to d^γ for any fixed $\gamma > 0$, and

δ to any inverse polynomial in d, the $\log_B(d/\delta)$ factor is constant. Thus, the result shows that if one is willing to tolerate a column sparsity of $O(1/\epsilon)$ rather than 1, the $\frac{d^2}{\epsilon^2}$ result can be replaced with any power of d larger than 1.

However, if B is set to grow slower than any fixed power of d, this tradeoff is unappealing, with a cubic dependence in $\log_B(d/\delta)$. Furthermore, the row count has a large number of logarithmic factors that are present no matter how small B is. Lower bounds ([NN14]), on the other hand, suggest that the true tradeoff allows $m = O(B(d + \log(1/\delta))/\epsilon^2)$ with $s = O(\log_B(d/\delta)/\epsilon)$.

An alternative analysis of these types of matrices was given in [BDN15], Section 4. That obtained m of about $d \log^2 d/\epsilon^2$ and s about $\log^4 d/\epsilon^2$. This is much better than [NN13] in terms of row count with small s, but allows no tradeoff between row count and sparsity and does not come close to matching the lower bound.

In this paper, we provide an alternative analysis in the same setting. Unlike the approach in [NN13], which involved applying careful combinatorics to count the terms in an expanded moment polynomial, or [BDN15], which used deep results on Gaussian processes, this requires only simple probabilistic tools combined with a matrix trace inequality, the Golden-Thompson inequality. It is a variant of the "matrix Chernoff bound" approach introduced by [AW02] and largely popularized by [Tro12]. This result comes close to matching the lower bound, obtaining $m = O(Bd \log(d/\delta)/\epsilon^2)$ and $s = O(\log_B(d/\delta)/\epsilon)$. The result fails to be tight because of the extra multiplicative $\log(d/\delta)$ on the row count, which should not be necessary but appears to come from a fundamental limitation of the matrix Chernoff approach.

We note that our approach also gives a novel analysis of sparse Johnson-Lindenstrauss embeddings if one plugs in $d = 1$; the result matches [KN12]. This analysis would be completely elementary, not requiring Golden-Thompson or any matrix trace inequalities.

Since the matrices in question are just the ones from [NN13], apart from being smaller and sparser, the analysis improves bounds on the many algorithmic applications in that paper.

2 Background

We define a distribution over m by n matrices Π as *sparse embedding matrices* with sparsity s, as depending on random variables $\delta_{r,i}$ and $\sigma_{r,i}$ ($1 \leq r \leq m$; $1 \leq i \leq n$). Exactly s $\delta_{r,i}$ in each column are uniformly selected without replacement to be 1, independently across columns; the remainder of the $\delta_{r,i}$ are set to 0. Each $\sigma_{r,i}$ is equal to an independent random sign.

Π is then defined as having its (r, i) entry equal to $\frac{1}{\sqrt{s}}\delta_{r,i}\sigma_{r,i}$.

Note that there are other, similar distributions of random sparse matrices–for instance, rather than uniformly selecting s entries per column without replacement, one may group the rows into s "blocks" and randomly select a single nonzero from each block in each column. The proofs in this paper can be straightforwardly adapted to that case. Unfortunately, they still fall short of the generality of [NN13], which provides a systematic rule for analyzing various such distributions, requiring only a simple anticorrelation property (which they called the "OSNAP property"). It is not clear whether this paper's analysis allows such a clean, systematic approach for analyzing new distributions (although the two just mentioned are the most important).

A key fact in proving subspace embedding properties is that they are equivalent to matrix spectral norm error bounds. In particular, Π successfully embeds a subspace S if and only if

$$\|(\Pi \boldsymbol{U})^T(\Pi \boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U}\| \leq \epsilon$$

where $\boldsymbol{U}$ is an orthonormal basis for S. This follows from the fact that the spectral norm of a symmetric matrix $\boldsymbol{A}$ is the maximum absolute value of $\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$ over all unit vectors $\boldsymbol{e}$, and for a vector $\boldsymbol{e}$ in the subspace,

$$\boldsymbol{e}^T((\Pi \boldsymbol{U})^T(\Pi \boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U})\boldsymbol{e} = \|\Pi\boldsymbol{e}\|^2 - \|\boldsymbol{e}\|^2.$$

For vectors $\boldsymbol{e}$ outside S, the evaluation is equivalent to that on the orthogonal projection of $\boldsymbol{e}$ on to S.

Thus, we will aim to establish this spectral norm result.

We will use the notation $\boldsymbol{A}_r$ to refer to the rth row of $\boldsymbol{A}$, treated as a column vector.

3 Our Approach

As stated above, the bounds are obtained in the "matrix Chernoff" framework. In order to motivate the full methods of this paper, we will begin by briefly sketching a simpler matrix Chernoff-based analysis, and then sketching the additional techniques used in the proof of the full result. The actual proof of the main result will be in the next section.

First, consider a standard, "black-box" matrix Chernoff bound, as can be found in e.g. [Tro12]:

Theorem 3.1 (Matrix Chernoff)**.** *Let $\boldsymbol{A}_i$ be independent random positive semidefinite matrices satisfying $\mathrm{E}\sum_i \boldsymbol{A}_i = \boldsymbol{I}$ and, for all i, $\|\boldsymbol{A}_i\| \leq \alpha$ with probability 1. Then for any*

$\epsilon < 1$,

$$\| \sum_i \boldsymbol{A}_i - \boldsymbol{I} \| \leq \epsilon$$

with probability at least $1 - 2d \exp\left(-\frac{\epsilon^2}{3\alpha}\right)$.

In particular, for a sum of independent random positive semidefinite matrices, a uniform spectral norm bound for the summands of $\alpha = O\left(\frac{\log(d/\delta)}{\epsilon^2}\right)$ suffices to obtain spectral norm error no worse than ϵ with probability at least $1 - \delta$.

Unfortunately, sparse embedding matrices, as defined above, do not seem to naturally fit into this framework of sums of independent random matrices. Instead, we will consider a slightly different family of random matrices: instead of randomly selecting s entries per column to make nonzero, make each entry independently nonzero with probability $\frac{s}{m}$. We call these *i.i.d. sparse embedding matrices*. This family of matrices turns out to be worse than the true sparse embedding matrices, but is simpler to analyze. Now, the rows of Π, and thus of $\Pi\boldsymbol{U}$, are independent from each other, so the matrix

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

is a sum of indepedent random matrices, making it an attractive target for matrix Chernoff. There is still a small technical issue that we do not have a uniform upper bound on $\|(\Pi\boldsymbol{U})_r(\Pi\boldsymbol{U})_r^T\| = \|(\Pi\boldsymbol{U})_r\|^2$ which holds with probability 1. This turns out not to be a major obstruction; we may simply take a bound that holds with high probability and slightly tweak the probability distribution of matrices to truncate those matrices with larger norms. Using this kind of argument, we can show that when ϵ and δ are not worse than inverse-polynomial in d, $m = O(d \log d/\epsilon^2)$ and $s = O(\log^2 d/\epsilon^2)$ suffices. Note that this s both contains two log factors and an ϵ^2 rather than ϵ dependence.

The truncation argument, though, is intuitively lossy, since it effectively always assumes the worst case scenario for the row norms. Since matrix Chernoff bounds are proved via estimates of the matrix exponentials of the $\boldsymbol{A}_i$, one may reasonably expect to be able to avoid these truncation losses by directly estimating those matrix exponentials, rather than just using an upper bound on the norm. One may construct a complete argument of this sort by using the matrix exponential estimate lemma 4.5 and general matrix concentration framework lemma 4.2 discussed later in this paper without any further tricks. This obtains bounds of $m = O(d \log(d/\delta)/\epsilon^2)$ and $s = O(\log(d/\delta)/\epsilon^2)$. This is actually the optimal sparsity that can be obtained for the i.i.d. sparse embeddng matrices. We need to analyze the true sparse embedding matrices instead, handling the fact that the rows are no longer independent and actually obtaining a win from the fact that the number of nonzeroes per column is fixed.

To handle lack of independence, we do a conceptually simple trick: we look at the contribution of the first and second halves of the rows separately, and argue that each half is, in a sense, almost as good as independent. The final bound can be obtained by combining bounds on error incurred from the first and second halves. This trick alone would suffice to extend the previously discussed analyses from i.i.d. sparse embedding matrices to standard sparse embedding matrices. On its own, though, it would still leave the $\log(d/\delta)/\epsilon^2$ dependence in the sparsity. The other key trick we need is to separate out the contribution of a row into "diagonal terms" and "cross-terms". The fixed number of nonzero entries per column turns out to guarantee that the diagonal terms add up to exactly the identity, allowing one to look at the contribution only of the cross-terms of each row. This trick is essentially identical to one used in [KN12] for analyzing sparse Johnson-Lindenstraus transforms. Once we have these pieces, we can use standard tools such as decoupling to reduce the problem to proving elementary inequalities.

4 Proof of main results

This section contains the proof of our main result, though an auxiliary inequality, lemma 4.5, is deferred to appendix A.

First, we split $(\Pi \boldsymbol{U})^T(\Pi \boldsymbol{U})$ as follows:

$$\begin{aligned}
(\Pi \boldsymbol{U})^T(\Pi \boldsymbol{U}) &= \sum_r (\Pi \boldsymbol{U})_r (\Pi \boldsymbol{U})_r^T \\
&= \frac{1}{s}\sum_r \left(\sum_i \delta_{r,i}\sigma_{r,i}\boldsymbol{u}_i\right)\left(\sum_i \delta_{r,i}\sigma_{r,i}\boldsymbol{u}_i^T\right) \\
&= \frac{1}{s}\sum_r \left(\sum_i \delta_{r,i}\boldsymbol{u}_i\boldsymbol{u}_i^T\right) + \left(\sum_{i\neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}\boldsymbol{u}_i\boldsymbol{u}_j^T\right) \\
&= \frac{1}{s}\left(\sum_i \left(\sum_r \delta_{r,i}\right)\boldsymbol{u}_i\boldsymbol{u}_i^T\right) + \frac{1}{s}\left(\sum_r\sum_{i\neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}\boldsymbol{u}_i\boldsymbol{u}_j^T\right) \\
&= \boldsymbol{U}^T\boldsymbol{U} + \frac{1}{s}\sum_r\sum_{i\neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}\boldsymbol{u}_i\boldsymbol{u}_j^T.
\end{aligned}$$

This corresponds to separating the contribution of the diagonal and off-diagonal parts of $\Pi^T\Pi$; that the first term is simply $\boldsymbol{U}^T\boldsymbol{U}$ reflects that the the diagonal has all ones due to the fixed number of nonzero entries per column. Since the first term is $\boldsymbol{U}^T\boldsymbol{U}$, the error

$(\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U}$ is just the remainder. For convenience, we define

$$\boldsymbol{Z}_r = \sum_{i \neq j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}\boldsymbol{u}_i\boldsymbol{u}_j^T \tag{1}$$

representing the contribution of the off-diagonal part from a single row r of Π, so that

$$(\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U} = \frac{1}{s}\sum_r \boldsymbol{Z}_r \tag{2}$$

Note that this expression is a matrix analogue of Equation 5 of [KN12].

An advantage of this expression is that the variation in the $\boldsymbol{Z}_r$ comes only from the "collisions" (i.e. multiple nonzeroes in the same row of Π) and not from individual nonzero entries.

We might then hope to be able to apply matrix concentration machinery to the $\boldsymbol{Z}_r$. However, there are two substantial obstructions. First, the definition of the $\boldsymbol{Z}_r$, as a sum over $i \neq j$, is difficult to deal with. Second, and more seriously, the $\boldsymbol{Z}_r$ are not independent.

The first issue can be dealt with by a standard trick, *decoupling*. Consider a new set of $\{0,1\}$-valued variables, w_i. Now define

$$\boldsymbol{Z}'_r = 2\sum_{i,j \mid w_i \neq w_j} \delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}\boldsymbol{u}_i\boldsymbol{u}_j^T \tag{3}$$

Now, let the w_i be i.i.d. random variables, with equal probability of choosing 0 or 1. Then

Lemma 4.1. *Let f be any convex function. Then*

$$\mathrm{E}_{\delta,\sigma}[f((\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U})] \leq \mathrm{E}_{\delta,\sigma,w}\left[f\left(\frac{1}{s}\sum_r \boldsymbol{Z}'_r\right)\right].$$

In particular, if there is a uniform bound K such that

$$\mathrm{E}_{\delta,\sigma}\left[f\left(\frac{1}{s}\sum_r \boldsymbol{Z}'_r\right)\right] \leq K$$

for any fixed choice of w, then

$$\mathrm{E}_{\delta,\sigma}[f((\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U})] \leq K.$$

Proof. For any $i \neq j$, $P[w_i \neq w_j] = \frac{1}{2}$. Thus, $\mathrm{E}_w[\boldsymbol{Z}'_r] = \boldsymbol{Z}_r$, so

$$\mathrm{E}_w\left[\frac{1}{s}\sum_r \boldsymbol{Z}'_r\right] = (\Pi \boldsymbol{U})^T(\Pi \boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U}.$$

The claim then follows from the convexity of f. □

This then reduces our problem to obtaining bounds for $\frac{1}{s}\sum_r \boldsymbol{Z}'_r$ for any fixed w. This is substantially more tractable, since we may express

$$\boldsymbol{Z}'_r = 2(\boldsymbol{x}_r\boldsymbol{y}_r^T + \boldsymbol{y}_r\boldsymbol{x}_r^T) \tag{4}$$

where

$$\boldsymbol{x}_r = \sum_{i|w_i=0} \delta_{r,i}\sigma_{r,i}\boldsymbol{u}_i \tag{5}$$

$$\boldsymbol{y}_r = \sum_{i|w_i=1} \delta_{r,i}\sigma_{r,i}\boldsymbol{u}_i. \tag{6}$$

Note that with w_i fixed, $\boldsymbol{x}_r$ is independent from $\boldsymbol{y}_r$.

However, we still need a strategy for dealing with the non-independent $\boldsymbol{Z}'_r$. Recall that independence is not strictly necessary for concentration bounds: for instance, there are Chernoff-type bounds on martingales as well as sums of independent random variables. However, they have strict requirements on the *conditional* distribution of each random variable given the previous. Here, we will use the following simple form of matrix concentration bound for non-independent random variables; it is similar to bounds given in and follows from the same methods [AW02, Tro12, Tro11]:

Lemma 4.2. *Let $\boldsymbol{A} = \sum_i \boldsymbol{A}_i$ be a sum of m random symmetric matrices such that for all i and all allowable values $\boldsymbol{A}'_1, \boldsymbol{A}'_2, ..., \boldsymbol{A}'_{i-1}$,*

$$\|\mathrm{E}[\exp(c\boldsymbol{A}_i) \mid \boldsymbol{A}_1 = \boldsymbol{A}'_1, \boldsymbol{A}_2 = \boldsymbol{A}'_2, \ldots \boldsymbol{A}_{i-1} = \boldsymbol{A}'_{i-1}]\| \leq C.$$

Then

$$\mathrm{E}[\mathrm{tr}(\exp(c\boldsymbol{A}))] \leq dC^m.$$

Note that this requires a uniform bound for the distribution of $\boldsymbol{A}_i$ conditioned on all previous terms. Furthermore, it requires spectral norm bounds on the exponential moments of each $\boldsymbol{A}_i$, but only gives the trace of the exponential moment of $\boldsymbol{A}$.

As is typical with matrix Chernoff-type bounds, this can be proved using an exponential trace inequality, the Golden-Thompson inequality (it can alternatively be proved using Lieb's concavity theorem).

Lemma 4.3 (Golden-Thompson)**.** *For any symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,*

$$\operatorname{tr}(\exp(\boldsymbol{A}+\boldsymbol{B})) \leq \operatorname{tr}(\exp(\boldsymbol{A})\exp(\boldsymbol{B})). \tag{7}$$

Note that this clearly holds with equality for scalars and for commuting matrices, while the general inequality is substantially more difficult to prove.

We can then prove lemma 4.2:

Proof of lemma 4.2. Define $\boldsymbol{S}_i = \sum_{j=1}^{i} \boldsymbol{A}_i$, and without loss of generality assume $c = 1$ (since this effectively just scales $\boldsymbol{A}$). Now, for any $i > 0$,

$$\begin{aligned}
\mathrm{E}[\operatorname{tr}(\exp(\boldsymbol{S}_i))] &= \mathrm{E}_{\boldsymbol{S}_{i-1}}[\mathrm{E}_{\boldsymbol{A}_i}[\operatorname{tr}(\exp(\boldsymbol{S}_{i-1}+\boldsymbol{A}_i))]] \\
&\leq \mathrm{E}_{\boldsymbol{S}_{i-1}}[\mathrm{E}_{\boldsymbol{A}_i}[\operatorname{tr}(\exp(\boldsymbol{S}_{i-1})\exp(\boldsymbol{A}_i))]] \\
&= \mathrm{E}_{\boldsymbol{S}_{i-1}}[\operatorname{tr}(\exp(\boldsymbol{S}_{i-1})\mathrm{E}_{\boldsymbol{A}_i}[\exp(\boldsymbol{A}_i)])] \\
&\leq \mathrm{E}_{\boldsymbol{S}_{i-1}}[\|\mathrm{E}_{\boldsymbol{A}_i}\exp(\boldsymbol{A}_i)\| \operatorname{tr}(\exp(\boldsymbol{S}_{i-1}))] \\
&\leq C\mathrm{E}[\operatorname{tr}(\exp(\boldsymbol{S}_{i-1}))].
\end{aligned}$$

The second line in the above follows from the Golden-Thompson inequality; the third follows from the linearity of trace and matrix product. Since $\mathrm{E}[\operatorname{tr}(\exp(\boldsymbol{S}_0))] = d$, we have $\mathrm{E}[\operatorname{tr}(\exp(\boldsymbol{S}_i))] \leq dC^i$ and $\mathrm{E}[\operatorname{tr}(\exp(\boldsymbol{A}))] \leq dC^m$, as desired. □

However, our variables $\boldsymbol{Z}'_r$ will not satisfy the hypotheses of lemma 4.2 for any reasonable value of C. The problem is that for certain (highly improbable) cases, the *conditional* distribution of $\boldsymbol{Z}'_r$ could be very bad. For instance, if each column of Π has had exactly $s-1$ nonzero entries before the last row, then the conditional distribution of $\delta_{m,i}$ is always 1, producing a much larger row of $\Pi\boldsymbol{U}$ than average.

However, this problem only becomes serious after one has already chosen a large number of rows. In particular, up to $r = m/2$, the conditional probability of $\delta_{r,i}$ can be no larger than $2\frac{s}{m}$ (compared to the unconditional probability of $\frac{s}{m}$. That suggests a simple solution:

simply split out our quantity to be bounded into two matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$:

$$\boldsymbol{R}_1 = \sum_{r=1}^{m/2} \boldsymbol{Z}'_r \tag{8}$$

$$\boldsymbol{R}_2 = \sum_{r=m/2+1}^{m} \boldsymbol{Z}'_r. \tag{9}$$

$\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are not independent, but with only two terms this is not a serious problem and only costs us constant factors. We now may actually apply lemma 4.2 directly to $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$.

Finally, we will use a helpful matrix inequality to analyze exponential moments for matrices of the form of $\boldsymbol{Z}_r$:

Lemma 4.4. *For any vectors $\boldsymbol{x}$, $\boldsymbol{y}$,*

$$\begin{aligned}&\cosh(c(2(\boldsymbol{x}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{x}^T))) - \boldsymbol{I} \preceq \\ &\qquad (\exp(2c\|\boldsymbol{y}\|^2) - 1)(\exp(2c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I}) + (\exp(2c\|\boldsymbol{x}\|^2) - 1)(\exp(2c\boldsymbol{y}\boldsymbol{y}^T) - \boldsymbol{I}).\end{aligned}$$

Here $\preceq$ represents the Louwner ordering. That is, for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we say that $\boldsymbol{A} \preceq \boldsymbol{B}$ if and only if $\boldsymbol{B} - \boldsymbol{A}$ is a positive semidefinite matrix. Additionally, we define the hyperbolic cosine of a matrix by plugging the matrix into the Taylor series for cosh*, or equivalently,* $\cosh(\boldsymbol{A}) = \frac{\exp(\boldsymbol{A})+\exp(-\boldsymbol{A})}{2}$.

Proof. We may write

$$\begin{aligned}&\cosh(c(2(\boldsymbol{x}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{x}^T))) - \boldsymbol{I} \\ &\quad\preceq \frac{\cosh(4c\|\boldsymbol{x}\|\|\boldsymbol{y}\|) - 1}{4\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2}(\boldsymbol{x}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{x}^T)^2 \\ &\quad\preceq \frac{2(\exp(2c\|\boldsymbol{x}\|^2) - 1)(\exp(2c\|\boldsymbol{y}\|^2) - 1)}{4\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2}(2\|\boldsymbol{y}\|^2\boldsymbol{x}\boldsymbol{x}^T + 2\|\boldsymbol{x}\|^2\boldsymbol{y}\boldsymbol{y}^T) \\ &\quad= (\exp(2c\|\boldsymbol{y}\|^2) - 1)\frac{\exp(2c\|\boldsymbol{x}\|^2) - 1}{\|\boldsymbol{x}\|^2}\boldsymbol{x}\boldsymbol{x}^T + (\exp(2c\|\boldsymbol{x}\|^2) - 1)\frac{\exp(2c\|\boldsymbol{y}\|^2) - 1}{\|\boldsymbol{y}\|^2}\boldsymbol{y}\boldsymbol{y}^T \\ &\quad= (\exp(2c\|\boldsymbol{y}\|^2) - 1)(\exp(2c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I}) + (\exp(2c\|\boldsymbol{x}\|^2) - 1)(\exp(2c\boldsymbol{y}\boldsymbol{y}^T) - \boldsymbol{I}).\end{aligned}$$

The first line here follows from the fact that $\frac{\cosh(x)-1}{x^2}$ is an increasing function in $|x|$; the two sides have the same eigenspaces, and the ratio of their eigenvalues is never larger than $\frac{\cosh(4c\|\boldsymbol{x}\|\|\boldsymbol{y}\|)-1}{4\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2}$ since no eigenvalue of $2(\boldsymbol{x}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{x}^T)$ is larger than $4\|\boldsymbol{x}\|\|\boldsymbol{y}\|$. The second line follows from the scalar inequality $\cosh(4cxy) - 1 \leq (\exp(2cx^2) - 1)(\exp(2cy^2) - 1)$, which follows from direct examination of Taylor series (both have Taylor series with all positive terms, and those of the latter dominate those of the former), plus the matrix inequality $(\boldsymbol{x}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{x}^T)^2 \preceq 2\|\boldsymbol{y}\|^2\boldsymbol{x}\boldsymbol{x}^T + 2\|\boldsymbol{x}\|^2\boldsymbol{y}\boldsymbol{y}^T$. □

Finally, putting all of this together, we can show:

Theorem 4.1. *Let c, C be such that for any probabilities $p_i \leq 2\frac{s}{m}$, if δ_i are independent Bernoulli random variables with probabilities p_i and σ_i are independent Rademacher random variables, if we define*

$$\boldsymbol{x} = \sum_i \delta_i \sigma_i \boldsymbol{u}_i$$

we have

$$\| \exp\left(\frac{4c}{s}\boldsymbol{x}\boldsymbol{x}^T\right) - \boldsymbol{I}\| \leq C.$$

Then we have

$$\mathrm{E}[\mathrm{tr}(\exp(c((\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U}))] \leq d\exp(C^2dm).$$

Proof. First, apply lemma 4.1, using the convexity of the function $\mathrm{tr}(\exp(c\boldsymbol{A}))$, to show that any bound on

$$\mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{c}{s}\sum_r \boldsymbol{Z}'_r\right)\right)\right] = \mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{c}{s}(\boldsymbol{R}_1 + \boldsymbol{R}_2)\right)\right)\right]$$

that holds for arbitrary m also holds for

$$\mathrm{E}[\mathrm{tr}(\exp(c((\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U}))]$$

Now, use the convexity of the trace exponential again to show that

$$\begin{aligned}\mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{c}{s}(\boldsymbol{R}_1 + \boldsymbol{R}_2)\right)\right)\right] &\leq \frac{1}{2}\left(\mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{2c}{s}\boldsymbol{R}_1\right)\right)\right] + \mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{2c}{s}\boldsymbol{R}_2\right)\right)\right]\right) \\ &= \mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{2c}{s}\boldsymbol{R}_1\right)\right)\right].\end{aligned}$$

The latter line here follows from the fact that $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ have the same distribution. Note that this makes no assumption of the independence of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, and is based solely on convexity.

Next, note that

$$\boldsymbol{R}_1 = \sum_{r=1}^{m/2} \boldsymbol{Z}'_r$$

with $\boldsymbol{Z}'_r = 2(\boldsymbol{x}_r\boldsymbol{y}_r^T + \boldsymbol{y}_r\boldsymbol{x}_r^T)$. Furthermore, $\boldsymbol{Z}'_r$ is a symmetric random variable (since the case that flips every $\sigma_{r,i}$ with $w_i = 0$ is equally probable), even when conditioning on previous rows, so the conditional expectation of $\exp\left(\frac{2c}{s}\boldsymbol{Z}'\right) - I$ is the same as that of $\cosh\left(\frac{2c}{s}\boldsymbol{Z}'\right) - \boldsymbol{I}$. Applying lemma 4.4, we get that this expectation is Louwner dominated by

$$\mathrm{E}\left[\begin{array}{c}\left(\exp\left(\frac{4c}{s}\|\boldsymbol{y}_r\|^2\right) - 1\right)\left(\exp\left(\frac{4c}{s}\boldsymbol{x}_r\boldsymbol{x}_r^T\right) - \boldsymbol{I}\right) + \\ \left(\exp\left(\frac{4c}{s}\|\boldsymbol{x}_r\|^2\right) - 1\right)\left(\exp\left(\frac{4c}{s}\boldsymbol{y}_r\boldsymbol{y}_r^T\right) - \boldsymbol{I}\right)\end{array}\right].$$

Using the (conditional) independence of $\boldsymbol{x}_r$ and $\boldsymbol{y}_r$, that is equal to

$$\mathrm{E}\left[\exp\left(\frac{4c}{s}\|\boldsymbol{y}_r\|^2\right) - 1\right]\mathrm{E}\left[\exp\left(\frac{4c}{s}\boldsymbol{x}_r\boldsymbol{x}_r^T\right) - I\right] + \\ \mathrm{E}\left[\exp\left(\frac{4c}{s}\|\boldsymbol{x}_r\|^2\right) - 1\right]\mathrm{E}\left[\exp\left(\frac{4c}{s}\boldsymbol{y}_r\boldsymbol{y}_r^T\right) - I\right].$$

Now, $\boldsymbol{x}_r$ and $\boldsymbol{y}_r$ are each of the form of the $\boldsymbol{x}$ in the theorem statement (indices with the opposite value of w_i just have their p_i set to 0). Thus the matrix expectations are upper bounded by $C\boldsymbol{I}$; the scalar expectations are the traces of such matrices, so they are upper bounded by Cd. Overall, the conditional expectation is Louwner upper bounded by $2C^2d\boldsymbol{I}$.

This was a bound on the expectation of the exponential minus the identity; the actual upper bound we have on the conditional expectation of $\exp\left(\frac{2c}{s}\boldsymbol{Z}'\right)$ is $(1 + 2C^2d)\boldsymbol{I}$. This is equivalent to saying that the spectral norm of that conditional expectation is at most

$$1 + 2C^2d \leq \exp(2C^2d).$$

Finally, given this we may apply lemma 4.2; since there are $\frac{m}{2}$ terms, each satisfying the upper bound of $\exp(2C^2d)$, our expected trace matrix exponential is at most $d\exp(C^2dm)$, as desired. □

To get our main result, we will need a basic matrix exponential bound. We will use the inequality

Lemma 4.5. *There exist absolute constants D_1, D_2 such that for any set of vectors $\boldsymbol{u}_i$ satisfying*

$$\sum_i \boldsymbol{u}_i\boldsymbol{u}_i^T \preceq \boldsymbol{I}$$

and nonnegative real numbers $pd \leq 1/10$*,* $p_i \leq p$*, and* $c \leq \frac{\log(1/(pd))}{D_1}$*, if we define*

$$\boldsymbol{x} = \sum_i \delta_i \sigma_i \boldsymbol{u}_i$$

where:

- δ_i *are independent* $0, 1$ *random variables, each 1 with probability* p_i*.*
- σ_i *are independent random signs*

then

$$\|\mathrm{E}[\exp(c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I}]\| \leq (\exp(pd(\exp(D_2 c) - 1)) - 1)/d.$$

This is proved in appendix A.

We can now prove our main result:

Theorem 4.2. *For any* $B > 2$*,* $\delta < 1/2$*,* $\epsilon < 1/2$*, a sparse embedding matrix* Π *with* $m = O\left(\frac{Bd\log(d/\delta)}{\epsilon^2}\right)$ *and* $s = O\left(\frac{\log_B(d/\delta)}{\epsilon}\right)$ *satisfies*

$$\|(\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U}\| \leq \epsilon$$

with probability at least $1 - \delta$*.*

Proof. Apply theorem 4.1 and lemma 4.5 with c proportional to $\log(d/\delta)/\epsilon$.

$$\frac{s}{m} \sim \frac{\epsilon}{Bd}$$

while

$$\frac{c}{s} \sim \log B$$

so we can get a bound from lemma 4.5 of $C \sim \frac{\epsilon}{\sqrt{B}d}$, so that $C^2 dm$ in theorem 4.1 is equal to 1. Thus we have

$$\mathrm{E}\left[\mathrm{tr}\left(\exp\left(\frac{\log(d/\delta)}{\epsilon}((\Pi\boldsymbol{U})^T(\Pi\boldsymbol{U}) - \boldsymbol{U}^T\boldsymbol{U})\right)\right)\right] \leq ed.$$

Applying Markov's inequality then gives the desired result. □

5 Acknowledgements

The author thanks Jonathan Kelner and Joel Tropp for useful advice and discussion, and Jelani Nelson for extremely helpful conversations about sparse oblivious subspace embeddings. This work was supported by an Akamai Presidential Fellowship and by National Science Foundation award 1111109.

References

[AC09] Nir Ailon and Bernard Chazelle. The fast johnson-lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, May 2009.

[AL13] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast johnson-lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, June 2013.

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *Information Theory, IEEE Transactions on*, 48(3):569–579, Mar 2002.

[BDN15] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 499–508, New York, NY, USA, 2015. ACM.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 81–90, New York, NY, USA, 2013. ACM.

[KN12] Daniel M. Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 1195–1206. SIAM, 2012.

[MM13] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 91–100, New York, NY, USA, 2013. ACM.

[NN13] J. Nelson and H.L. Nguyen. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 117–126, Oct 2013.

[NN14] Jelani Nelson and Huy L. Nguyn. Lower bounds for oblivious subspace embeddings. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming*, volume 8572 of *Lecture Notes in Computer Science*, pages 883–894. Springer Berlin Heidelberg, 2014.

[Sar06] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium on*, pages 143–152, Oct 2006.

[Tro11] Joel A Tropp. User-friendly tail bounds for matrix martingales. Technical report, DTIC Document, 2011.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, August 2012.

A Auxiliary inequalities

This section builds up the proof of lemma 4.5.

First, we will attack scalar tail bounds of this form.

Lemma A.1. *For any list of real numbers u_i satisfying*

$$\sum_i u_i^2 \leq 1$$

and nonnegative real numbers $p \leq 1/10$ and $p_i \leq p$, define the random variable

$$x = \sum_i \delta_i \sigma_i u_i$$

where:

- *δ_i are independent $0, 1$ random variables, each 1 with probability p_i.*
- *σ_i are independent random signs.*

Then for any k, the probability that $|x| > k$ is at most $2\exp(-k^2 \log(1/p)/4)$.

Proof. We will estimate exponential moments $\mathrm{E}[\exp(cx)]$ for arbitrary c.

Since the random variables $\delta_i \sigma_i u_i$ for each i are all independent, we have:

$$\begin{aligned}\mathrm{E}[\exp(cx)] &= \mathrm{E}\left[\prod_i \exp(c\delta_i\sigma_i u_i)\right] \\ &= \prod_i \mathrm{E}[\exp(c\delta_i\sigma_i u_i)] \\ &= \exp\left(\sum_i \log(\mathrm{E}[\exp(c\delta_i\sigma_i u_i)])\right).\end{aligned}$$

Thus, it will be sufficient to bound $\sum_i \log(\mathrm{E}[\exp(c\delta_i\sigma_i u_i)])$. Now, we note that

$$\mathrm{E}[\exp(c\delta_i\sigma_i u_i)] = 1 + p(\cosh(cu_i) - 1).$$

This is in turn always at most $\cosh(cu_i) \leq \exp(c|u_i|)$, so its logarithm is always at most $c|u_i|$. Furthermore, $\frac{\cosh(z)-1}{z^2}$ is an increasing function in $|z|$, so whenever $c|u_i| \leq \log(1/p)$ we have:

$$\begin{aligned} 1 + p(\cosh(cu_i) - 1) &= 1 + pc^2u_i^2 \frac{\cosh(cu_i) - 1}{c^2u_i^2} \\ &\leq 1 + pc^2u_i^2 \frac{\cosh(\log(1/p)) - 1}{\log(1/p)^2} \\ &\leq 1 + pc^2u_i^2 \frac{\exp(\log(1/p))}{\log(1/p)^2} \\ &= 1 + \frac{c^2u_i^2}{\log(1/p)^2}. \end{aligned}$$

$\log(1 + z) \leq z$, so the logarithm of this is at most $\frac{c^2u_i^2}{\log(1/p)^2}$. Now, when $c|u_i| > \log(1/p)$, we can use the bound of $c|u_i| \leq \frac{c^2u_i^2}{\log(1/p)}$, and when it is $\leq \log(1/p)$ we can use the bound $\frac{c^2u_i^2}{\log(1/p)^2} < \frac{c^2u_i^2}{\log(1/p)}$. Thus we always have

$$\log(\mathrm{E}[\exp(c\delta_i\sigma_i u_i)] \leq \frac{c^2u_i^2}{\log(1/p)}.$$

Since $\sum_i u_i^2 \leq 1$, we then have $\sum_i \log(\mathrm{E}[\exp(c\delta_i\sigma_i u_i)]) \leq \frac{c^2}{\log(1/p)}$. We thus have

$$\mathrm{E}[\exp(cx)] \leq \exp(c^2/\log(1/p))$$

Picking $c = k\log(1/p)/2$ and applying a Markov bound (for both $x > k$ and $x < -k$) completes the proof. □

This tail bound can be used to estimate quantities of the form $\mathrm{E}[\exp(cx^2) - 1]$:

Lemma A.2. *For all u_i, p_i, p, with x defined as in lemma A.1, and all nonnegative real $c \leq \log(1/p)/8$, we have*

$$\mathrm{E}[\exp(cx^2) - 1] \leq 5p(\exp(4c) - 1).$$

Proof. We begin by noting that lemma A.1, plus the fact that the mean of x^2 is at most p, implies that that the probability distribution of x^2 is second-order stochastically dominated by the distribution with probability density

$$(\log(1/p)/2)\exp(-l\log(1/p)/4).$$

for $l \geq 4$ which has no probability mass for $0 < u < 4$ (we can use no probability mass for $p < 4$ because the mean of this distribution is already greater than the true mean).

$$\int_4^\infty (\log(1/p)/2)\exp(-l\log(1/p)/4)(\exp(cl)-1)\,dl.$$

This expands out to

$$\int_4^\infty (\log(1/p)/2)(\exp(-l\log(1/p)/4+cl)-\exp(-l\log(1/p)/4))\,dl.$$

We may then explicitly compute the integral as

$$\begin{aligned}
&\frac{\log(1/p)/2}{\log(1/p)/4-c}\exp(-\log(1/p)+4c)-2\exp(-\log(1/p))\\
&\quad\leq (2+8c/\log(1/p))p\exp(4c)-2p\\
&\quad= (2+8c/\log(1/p))\,p(\exp(4c)-1)+(8c/\log(1/p))p\\
&\quad\leq 4p(\exp(4c)-1)+(2p/\log(1/p))(\exp(4c)-1)\\
&\quad\leq 5p(\exp(4c)-1).
\end{aligned}$$

□

Next, we extend this bound to estimating the norm of a vector:

Lemma A.3. *There exist absolute constants D_3, D_4 such that for any set of vectors $\boldsymbol{u}_i$ satisfying*

$$\sum_i \boldsymbol{u}_i\boldsymbol{u}_i^T \preceq \boldsymbol{I}$$

and

$$\sum_i \|\boldsymbol{u}_i\|^2 = q$$

and nonnegative real numbers $p \leq 1/10$, $p_i \leq p$, and $c \leq \frac{\log(1/p)}{D_3}$, if we define

$$\boldsymbol{x} = \sum_i \delta_i\sigma_i\boldsymbol{u}_i$$

with δ_i and σ_i defined as before, then

$$\mathrm{E}[\exp(c\|\boldsymbol{x}\|^2) - 1] \leq \exp(pq(\exp(D_4 c) - 1)) - 1.$$

Proof. First, we note that we may assume that $q \geq 1$; otherwise, we would have $\sum_i \boldsymbol{u}_i \boldsymbol{u}_i^T \preceq q\boldsymbol{I}$, and could thus replace $\boldsymbol{u}_i$ with $\frac{\boldsymbol{u}_i}{\sqrt{q}}$, q with 1, and c with $\frac{1}{q}$, reducing to the $q = 1$ case.

We start by rewriting $\|\boldsymbol{x}\|^2$:

$$\begin{aligned}\|\boldsymbol{x}\|^2 &= \left(\sum_i \delta_i \sigma_i \boldsymbol{u}_i\right)^2 \\ &= \sum_i \sum_j \delta_i \delta_j \sigma_i \sigma_j \boldsymbol{u}_i^T \boldsymbol{u}_j \\ &= \left(\sum_i \delta_i \|\boldsymbol{u}_i\|^2\right) + \left(\sum_{i \neq j} \delta_i \delta_j \sigma_i \sigma_j \boldsymbol{u}_i^T \boldsymbol{u}_j\right).\end{aligned}$$

We let m be the exponential moment we are bounding:

$$m = \mathrm{E}[\exp(c\|\boldsymbol{x}\|^2)]$$

Then

$$\begin{aligned}m &= \mathrm{E}\left[\exp\left(c\sum_i \delta_i \|\boldsymbol{u}_i\|^2 + c\sum_{i \neq j} \delta_i \delta_j \sigma_i \sigma_j \boldsymbol{u}_i^T \boldsymbol{u}_j\right)\right] \\ &\leq \max\left(\mathrm{E}\left[\exp\left(2c\sum_i \delta_i \|\boldsymbol{u}_i\|^2\right)\right], \mathrm{E}\left[\exp\left(2c\sum_{i \neq j} \delta_i \delta_j \sigma_i \sigma_j \boldsymbol{u}_i^T \boldsymbol{u}_j\right)\right]\right) \\ &\leq \max\left(\exp(pq(\exp(2c) - 1)), \mathrm{E}\left[\exp\left(2c\sum_{i \neq j} \delta_i \delta_j \sigma_i \sigma_j \boldsymbol{u}_i^T \boldsymbol{u}_j\right)\right]\right).\end{aligned}$$

The last line follows from bounding the left possibility with a Chernoff bound.

For the other input to the max, we perform decoupling using the convexity of the exponential function, upper-bounding it by

$$\mathrm{E}\left[\exp\left(4c\left(\sum_i \delta_i \sigma_i \boldsymbol{u}_i\right)^T \left(\sum_i \delta_i' \sigma_i' \boldsymbol{u}_i\right)\right)\right].$$

We define $\boldsymbol{y} = \sum_i \delta_i \sigma_i \boldsymbol{u}_i$ and $\boldsymbol{y}' = \sum_i \delta_i' \sigma_i' \boldsymbol{u}_i$. Let $a = \|\boldsymbol{y}\|$ and $b = \frac{\boldsymbol{y}^T \boldsymbol{y}'}{\|\boldsymbol{y}\|}$; then we are upper bounded by

$$\begin{aligned}\mathrm{E}[\exp(4cab)] &\le \mathrm{E}\left[\sqrt{\exp\left(ca^2\right)\exp(16cb^2)}\right]\\ &\le \sqrt{1+5p(\exp(64c)-1)}\mathrm{E}\left[\exp\left(\frac{c}{2}a^2\right)\right].\end{aligned}$$

Here the first line is by the AM-GM inequality applied to the exponent, and the second is by bounding the conditional expectation of the second factor, for any given a, by lemma A.2. Finally, we note that this latter expectation is at most $\sqrt{m}$.

Then assuming $m \ge \exp(pq(\exp(2c)-1))$, we have

$$\begin{aligned}m &\le \sqrt{1+5p(\exp(64c)-1)}\sqrt{m}\\ m &\le 1+5p(\exp(64c)-1).\end{aligned}$$

□

We now have the machinery needed to prove lemma 4.5.

Lemma 4.5. *There exist absolute constants D_1, D_2 such that for any set of vectors $\boldsymbol{u}_i$ satisfying*

$$\sum_i \boldsymbol{u}_i\boldsymbol{u}_i^T \preceq \boldsymbol{I}$$

and nonnegative real numbers $pd \le 1/10$, $p_i \le p$, and $c \le \frac{\log(1/(pd))}{D_1}$, if we define

$$\boldsymbol{x} = \sum_i \delta_i\sigma_i\boldsymbol{u}_i$$

where:

- *δ_i are independent $0, 1$ random variables, each 1 with probability p_i.*
- *σ_i are independent random signs*

then

$$\|\mathrm{E}[\exp(c\boldsymbol{x}\boldsymbol{x}^T)-\boldsymbol{I}]\| \le (\exp(pd(\exp(D_2c)-1))-1)/d.$$

Proof. $\|\mathrm{E}[\exp(c\boldsymbol{x}\boldsymbol{x}^T)-\boldsymbol{I}]\|$ is equal to the max over unit vectors $\boldsymbol{e}$ of

$$\mathrm{E}[\boldsymbol{e}^T(\exp(c\boldsymbol{x}\boldsymbol{x}^T)-\boldsymbol{I})\boldsymbol{e}].$$

We need to upper bound this expectation for any fixed unit vector $\boldsymbol{e}$. To do this, we take the slightly strange step of "splitting" our random coin flips into two steps. We now define

$$\boldsymbol{x} = \sum_i \delta_i^A \delta_i^B \sigma_i \boldsymbol{u}_i$$

where the δ^A, δ^B, and σ are all independent, δ_i^A is a $\{0,1\}$ random variable which is 1 with probability $\frac{1}{d}$, δ_i^B is a $\{0,1\}$ random variable which is 1 with probability $p_i d$, and σ_i is a random sign. This has the correct distribution, as δ_i has the same distribution as $\delta_i^A \delta_i^B$.

We will look at the expectation in question as

$$\mathrm{E}_{\delta^A}[\mathrm{E}_{\delta^B,\sigma}[\boldsymbol{e}^T(\exp(c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I})\boldsymbol{e}]].$$

For a fixed choice of δ^A, we define

$$\begin{aligned} a &= \sum_i \delta_i^A (\boldsymbol{e}^T \boldsymbol{u}_i)^2 \\ q &= \sum_i \delta_i^A \|\boldsymbol{u}_i\|^2. \end{aligned}$$

Given δ^A, δ^B, σ, we define

$$z = \max((\boldsymbol{e}^T\boldsymbol{x})^2/a, \|\boldsymbol{x}\|^2).$$

Then

$$\begin{aligned} \boldsymbol{e}^T(\exp(c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I})\boldsymbol{e} &= (\boldsymbol{e}^T\boldsymbol{x})^2 \frac{\exp(c\|\boldsymbol{x}\|^2) - 1}{\|\boldsymbol{x}\|^2} \\ &\leq az\frac{\exp(cz) - 1}{z} \\ &= a(\exp(cz) - 1). \end{aligned}$$

For fixed δ^A, we then have

$$\begin{aligned} \mathrm{E}[\boldsymbol{e}^T(\exp(c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I})\boldsymbol{e}] &= a\mathrm{E}[\exp(cz) - 1] \\ &\leq a(\mathrm{E}[\exp(c(\boldsymbol{e}^T\boldsymbol{x})^2/a) - 1] + \mathrm{E}[\exp(c\|\boldsymbol{x}\|^2)]) \\ &\leq a\left(5pd(\exp(4c) - 1) + 5\left(\exp\left(qpd(\exp(64c) - 1)\right) - 1\right)\right). \end{aligned}$$

Taking the expectation over δ^A gives

$$\begin{aligned}
&\mathrm{E}_{\delta^A,\delta^B,\sigma}[\boldsymbol{e}^T(\exp(c\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{I})\boldsymbol{e}] \\
&\qquad \leq 5pd(\exp(4c) - 1)E_{\delta^A}[a] + E_{\delta^A}\left[a\left(\exp\left(qpd(\exp(D_4c) - 1)\right) - 1\right)\right] \\
&\qquad \leq 5p(\exp(4c) - 1) + \frac{1}{d}E_{\delta^A}\left[\exp\left((q+1)pd(\exp(D_4c) - 1)\right) - 1\right].
\end{aligned}$$

The second line follows from the fact that $E_{\delta^A}[a] = \frac{1}{d}\boldsymbol{e}^T\sum_i(\boldsymbol{u}_i)^2\boldsymbol{e} \leq \frac{1}{d}$, and from the fact that the expected value of

$$\exp\left(qpd(\exp(D_4c) - 1)\right)$$

conditioned on any particular $\delta_i^A = 1$ is at most

$$E_{\delta^A}\left[\exp\left((q+1)pd(\exp(D_4c) - 1)\right)\right]$$

since $\|u_i\|^2 \leq 1$.

Finally, the desired result follows from a Chernoff bound applied to this expression. □