Simple Analyses of the Sparse Johnson-Lindenstrauss Transform*

Michael B. Cohen[†1], T. S. Jayram[2], and Jelani Nelson[‡3]

1 **MIT, 32 Vassar Street, Cambridge, MA 02139, USA**
`micohen@mit.edu`

2 **IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA**
`jayram@us.ibm.com`

3 **Harvard John A. Paulson School of Engineering and Applied Sciences, 29 Oxford Street, Cambridge, MA 02138, USA**
`minilek@seas.harvard.edu`

Abstract

For every n-point subset X of Euclidean space and target distortion $1+\varepsilon$ for $0 < \varepsilon < 1$, the *Sparse Johnson Lindenstrauss Transform* (SJLT) of [19] provides a linear dimensionality-reducing map $f : X \to \ell_2^m$ where $f(x) = \Pi x$ for Π a matrix with m rows where (1) $m = O(\varepsilon^{-2} \log n)$, and (2) each column of Π is sparse, having only $O(\varepsilon m)$ non-zero entries. Though the constructions given for such Π in [19] are simple, the analyses are not, employing intricate combinatorial arguments. We here give two simple alternative proofs of the main result of [19], involving no delicate combinatorics. One of these proofs has already been tested pedagogically, requiring slightly under forty minutes by the third author at a casual pace to cover all details in a blackboard course lecture.

1998 ACM Subject Classification F.2.0 General

Keywords and phrases dimensionality reduction, Johnson-Lindenstrauss, Sparse Johnson-Lindenstrauss Transform

Digital Object Identifier 10.4230/OASIcs.SOSA.2018.15

1 Introduction

A widely applied technique to gain speedup and reduce memory footprint when processing high-dimensional data is to first apply a dimensionality-reducing map which approximately preserves the geometry of the input in a pre-processing step. One cornerstone result along these lines is the following Johnson-Lindenstrauss (JL) lemma [16].

▶ **Lemma 1** (JL lemma). *For all $0 < \varepsilon < 1$, integers $n, d > 1$, and $X \subset \mathbb{R}^d$ with $|X| = n$, there exists $f : X \to \mathbb{R}^m$ with $m = O(\varepsilon^{-2} \log n)$ such that*

$$\forall y, z \in X,\ (1-\varepsilon)\|y-z\|_2 \le \|f(y)-f(z)\|_2 \le (1+\varepsilon)\|y-z\|_2. \tag{1}$$

* The last two authors dedicate this work to the first author, who in this work and other interactions was a constant source of energy and intellectual enthusiasm.

† M. B. Cohen is supported by NSF grants CCF-1111109 and CCF-1553428, and by a National Defense Science and Engineering Graduate Fellowship.

‡ J. Nelson is supported by NSF grant IIS-1447471 and CAREER award CCF-1350670, ONR Young Investigator award N00014-15-1-2388 and DORECG award N00014-17-1-2127, an Alfred P. Sloan Research Fellowship, and a Google Faculty Research Award.

1st Symposium on Simplicity in Algorithms (SOSA 2018).
Editor: Raimund Seidel; Article No. 15; pp. 15:1–15:9
Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The target dimension m given by the JL lemma is optimal for nearly the full range of n, d, ε; in particular, for any n, d, ε, there exists a point set $X \subset \mathbb{R}^d$ with $|X| = n$ such that any $(1+\varepsilon)$-distortion embedding of X into $\mathbb{R}^m$ under the Euclidean norm must have $m = \Omega(\min\{n, d, \varepsilon^{-2}\log(\varepsilon^2 n)\})$ [21, 5]. Note that an isometric embedding (i.e. $\varepsilon = 0$) is always achievable into dimension $m = \min\{n-1, d\}$, and thus the lower bound is optimal except potentially for ε close to $1/\sqrt{n}$.

All known proofs of the JL lemma instantiate f as a linear map. The original proof in [16] picked $f(x) = \Pi x$ where $\Pi \in \mathbb{R}^{m\times d}$ was an appropriately scaled orthogonal projection onto a uniformly random m-dimensional subspace. It was then shown that as long as $m = \Omega(\varepsilon^{-2}\log(1/\delta))$,

$$\forall x \in \mathbb{R}^d \text{ such that } \|x\|_2 = 1, \quad \mathop{\mathbb{P}}_{\Pi}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta. \tag{2}$$

The JL lemma then followed by setting $\delta < 1/\binom{n}{2}$ and considering $x = (y-z)/\|y-z\|_2$ for each pair $y, z \in X$, and adjusting ε by a constant factor. It is known that this bound of m for attaining (2) is tight; that is, m must be $\Omega(\min\{d, \varepsilon^{-2}\log(1/\delta)\})$ [15, 17].

One should typically think of applying dimensionality reduction techniques for applications as being a two-step process: first (1) one applies the dimension-reducing map f to the data, then (2) one runs some algorithm on the lower dimensional data $f(X)$. While reducing m typically speeds up the second phase, in order to speed up the first phase it is necessary to give an f which can be both found and applied to data quickly. To this end, Achlioptas showed Π can be chosen with i.i.d. entries where $\Pi_{i,j} = 0$ with probability 2/3, and otherwise $\Pi_{i,j}$ is uniform in $\pm 1/\sqrt{m/3}$ [1]. This was accomplished without increasing m by even a constant factor over previous best analyses of the JL lemma. Thus essentially a 3x speedup in step (2) is obtained without any loss in the quality of dimensionality reduction. Later, Ailon and Chazelle developed the FJLT [2] which uses the Fast Fourier Transform to implement a JL map Π with $m = O(\varepsilon^{-2}\log n)$ supporting matrix-vector multiplication in time $O(d\log d + m^3)$. Later work of [3] gave a different construction which, for the same m, improved the multiplication time to $O(d\log d + m^{2+\gamma})$ for arbitrarily small $\gamma > 0$. More recently, a sequence of works give embedding time $O(d\log d)$ but with a suboptimal embedding dimension $m = O(\varepsilon^{-2}\log n \cdot \mathrm{poly}(\log\log n))$ [4, 20, 22, 6, 12].

Note that the line of work beginning with the FJLT requires $\Omega(d\log d)$ embedding time per point, which is worse than the $O(m \cdot \|x\|_0)$ time to embed x using a dense Π if x is sufficiently sparse. Here $\|x\|_0$ denotes the number of non-zero entries in x. Motivated by speeding up dimensionality reduction further for sparse inputs, Kane and Nelson in [19], following [10, 18, 7], introduced the SJLT with $m = O(\varepsilon^{-2}\log n)$, and with $s = O(\varepsilon m)$ non-zero entries per column. This reduced the embedding time to compute Πx from $O(m \cdot \|x\|_0)$ to $O(s \cdot \|x\|_0) = O(\varepsilon m \cdot \|x\|_0)$. The original analysis of the SJLT in [19] showed Equation (2) for $m = O(\varepsilon^{-2}\log(1/\delta))$, $s = O(\varepsilon^{-1}\log(1/\delta))$ via the moment method. Specifically, the analysis there for $\|x\|_2 = 1$ defined

$$Z = \|\Pi x\|_2^2 - 1 \tag{3}$$

then used Markov's inequality to yield $\mathbb{P}(|Z| > \varepsilon) < \varepsilon^{-q} \cdot \mathbb{E}\, Z^q$ for some large even integer q (specifically $q = \Theta(\log(1/\delta))$). The bulk of the work was in bounding $\mathbb{E}\, Z^q$, which was accomplished by expanding Z^q as a polynomial with exponentially many terms, grouping terms with similar combinatorial structure, then employing intricate combinatorics to achieve a sufficiently good bound.

Our Main Contribution. We give two new analyses of the SJLT of [19], both of which avoid expanding Z^q into many terms and employing intricate combinatorics. As mentioned in the abstract, one of these proofs has already been tested pedagogically, requiring slightly under forty minutes by the third author at a casual pace to cover all details in a blackboard lecture.

2 Preliminaries

We say $f(x) \lesssim g(x)$ if $f(x) = O(g(x))$, and $f(x) \simeq g(x)$ denotes $f(x) = \Theta(g(x))$. For random variable X and $q \in \mathbb{R}$, $\|X\|_q$ denotes $(\mathbb{E}\,|X|^q)^{1/q}$. Minkowski's inequality, which we repeatedly use, states that $\|\cdot\|_q$ is a norm for $q \geq 1$. If X depends on many random sources, e.g. $X = X(a, b)$, we use $\|X\|_{L_q(a)}$, say, to denote the q-norm over the randomness in a (and thus the result will be a random variable depending only on b). A *Bernoulli-Rademacher* random variable $X = \eta\sigma$ with parameter p is such that η is a Bernoulli random variable (on $\{0, 1\}$) with $\mathbb{E}\,\eta = p$ and σ is a Rademacher random variable, i.e. uniform in $\{-1, 1\}$. Overloading notation, a random vector X whose coordinates are i.i.d. Bernoulli-Rademacher with parameter p will also be called by the same name. For a square real matrix A, let A° be obtained by zeroing out the diagonal of A. Throughout this paper we use $\|\cdot\|_F$ to denote Frobenius norm, and $\|\cdot\|$ to denote $\ell_2 \to \ell_2$ operator norm.

Both our SJLT analyses in this work show Eq. (2) by analyzing tail bounds for the random variable Z defined in Eq. (3). We continue to use the same notation, where $x \in \mathbb{R}^d$ of unit norm is as in Eq. (3). Our first SJLT analysis uses the following moment bounds for the binomial distribution and for quadratic forms with Rademacher random variables.

▶ **Lemma 2** ([14]). *For Y distributed as* $\mathsf{Binomial}(N, \alpha)$ *for integer $N \geq 1$ and $\alpha \in (0, 1)$, let $1 \leq p \leq N$ and define $B := p/(\alpha N)$. Then*

$$\|Y\|_p \lesssim \begin{cases} \frac{p}{\log B} & \text{if } B \geq e \\ \frac{p}{B} & \text{if } B < e \end{cases}$$

A more modern, general proof of the below Hanson-Wright inequality can be found in [23].

▶ **Theorem 3** (Hanson-Wright inequality [11]). *For $\sigma_1, \ldots, \sigma_n$ independent Rademachers and $A \in \mathbb{R}^{n\times n}$, for all $q \geq 1$*

$$\|\sigma^T A\sigma - \mathbb{E}\,\sigma^T A\sigma\|_q \lesssim \sqrt{q} \cdot \|A\|_F + q \cdot \|A\|.$$

Our second analysis uses a standard decoupling inequality; a proof is in [25, Remark 6.1.3]

▶ **Theorem 4** (Decoupling). *Let $A \in \mathbb{R}^{n\times n}$ be arbitrary, and $X_1, \ldots, X_n$ be independent and mean zero. Then, for every convex function $F : \mathbb{R} \to \mathbb{R}$*

$$\mathbb{E}\, F\Big(\sum_{i\neq jj} A_{i,j} X_i X_j\Big) \leq \mathbb{E}\, F\Big(4 \cdot \sum_{i,j} A_{i,j} X_i X'_j\Big)$$

where the X'_i are independent copies of the X_i.

Before describing the SJLT, we describe the related CountSketch of [8], which was shown to satisfy Eq. (3) in [24]. In this construction for Π, one picks a hash function $h : [d] \to [m]$ from a pairwise independent family, and a function $\sigma : [d] \to \{-1, 1\}$ from a 4-wise independent family. Then for each $i \in [d]$, $\Pi_{h(i),i} = \sigma(i)$, and the rest of the ith column is 0. It was shown

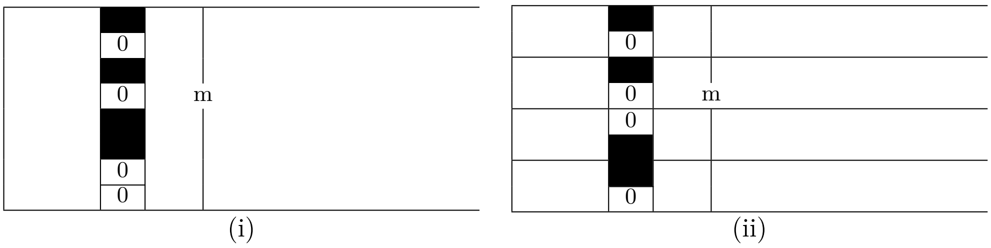

Figure 1 Both distributions have s non-zeroes per column, with each non-zero being independent in $\pm 1/\sqrt{s}$. In (i), they are in random locations, without replacement. (ii) is the CountSketch (with $s > 1$), whose rows are grouped into s blocks of size m/s each, with one non-zero per block per column in a uniformly random location, independent of other blocks; in this example, $m = 8, s = 4$.

in [24] that this distribution satisfies Eq. (3) for $m = \Omega(1/(\varepsilon^2\delta))$. Note that the column sparsity s equals 1. The analysis is simply via Chebyshev's inequality, i.e. bounding the second moment of Z.

The reason for the poor dependence in m on the failure probability δ is that we use Chebyshev's inequality. This is avoided by bounding a higher moment (as in [19], or our first analysis in this work), or by analyzing the moment generating function (MGF) (as in our second analysis in this work). To improve the dependence of m on $1/\delta$, we allow ourselves to increase s.

Now we describe the SJLT. This is a JL distribution over Π having exactly s non-zero entries per column where each entry is a scaled Bernoulli-Rademacher. Specifically, in the SJLT, the random $\Pi \in \mathbb{R}^{m\times d}$ satisfies $\Pi_{r,i} = \eta_{r,i}\sigma_{r,i}/\sqrt{s}$ for some integer $1 \le s \le m$. The $\sigma_{r,i}$ are independent Rademachers and jointly independent of the Bernoulli random variables $\eta_{r,i}$ satisfying:

(a) For any $i \in [d]$, $\sum_{r=1}^m \eta_{r,i} = s$. That is, each column of Π has *exactly* s non-zero entries.
(b) For all $r \in [m], i \in [d]$, $\mathbb{E}\,\eta_{r,i} = s/m$.
(c) The $\eta_{r,i}$ are negatively correlated: $\forall\ S \subset [d] \times [n]$, $\mathbb{E}\prod_{(r,i)\in S}\eta_{r,i} \le \prod_{(r,i)\in S}\mathbb{E}\,\eta_{r,i} = (s/m)^{|S|}$.

See Figure 1 for at least two natural distributions satisfying the above requirements. Thus

$$\|\Pi x\|_2^2 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i,j=1}^{d}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j.$$

Using (a) above we have $(1/s)\cdot\sum_r\sum_i \eta_{r,i}x_i^2 = \|x\|_2^2 = 1$, so that

$$Z = \|\Pi x\|_2^2 - 1 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j. \tag{4}$$

▶ Remark. In both our analyses, item (a) above is only used to remove the diagonal $i = j$ terms from eq. (4). Thenceforth, it turns out in both analyses of SJLT that (b) and (c) imply we can assume the $\eta_{r,i}$ are fully independent, i.e., the entries of Π are fully independent. This is not the same as saying we can replace the sketch matrix Π with fully independent entries because then part (a) would be violated and it is important for only the "cross" terms in the quadratic form representing Z to be present. In the analysis we justify this assumption by considering the integer moments of Z which we show here cannot decrease by replacement with fully independent entries. For each integer q, each monomial in the expansion of Z^q has expectation equal to $s^{-q}x_{\alpha_1}^{d_1}\cdots x_{\alpha_t}^{d_t}\cdot(\mathbb{E}\prod_{(r,i)\in S}\eta_{r,i})$ whenever all the d_j are even, and S contains all the distinct (r,i) such that $\eta_{r,i}$ appears in the monomial; otherwise the expectation equals 0. Now, $s^{-q}x_{\alpha_1}^{d_1}\cdots x_{\alpha_t}^{d_t}$ is nonnegative, and $\mathbb{E}\prod_{(r,i)\in S}\eta_{r,i} \le (s/m)^{|S|}$. Thus monomials' expectations are term-by-term dominated by the case that all $\eta_{r,i}$ are i.i.d. Bernoulli with expectation s/m.

3 Proof Overview

Hanson-Wright analysis. Note Z can be written as the quadratic form $\sigma^T A_{x,\eta}\sigma$, where $A_{x,\eta}$ is block diagonal with m blocks, where the rth block is $(1/s)x^{(r)}(x^{(r)})^T$ but with the diagonal zeroed out. Here $x^{(r)}$ is the vector with $(x^{(r)})_i = \eta_{r,i}x_i$. To apply Hanson-Wright, we must then bound $\|\|A_{x,\eta}\|_F\|_p$ and $\|\|A_{x,\eta}\|\|_p$, over the randomness of η. This was done in [19], but suboptimally, leading to a simple proof there but of a weaker result (namely, the bound on s proven there was suboptimal by a $\sqrt{\log(1/\varepsilon)}$ factor). As already observed in [19], a simple calculation shows $\|A_{x,\eta}\| \le 1/s$ with probability 1. In this work we improve the analysis of $\|\|A_{x,\eta}\|_F\|_p$ by a simple combination of the triangle and Bernstein inequalities to yield a tight analysis.

MGF analysis. We apply the Chernoff-Rubin bound $\mathbb{P}(|Z|>\varepsilon) \le 2e^{-t\varepsilon}\,\mathbb{E}\cosh(tZ)$, so that we must bound $\mathbb{E}\cosh(tZ)$ (for t in some bounded range) then optimize the choice of t. We accomplish our analysis by writing $Z = X^{\mathsf{T}}A^{\circ}X$ for an appropriate matrix A where X *is* a Bernoulli-Rademacher vector, by Taylor expansion of cosh and considerations similar to Remark 2. We then bound $\mathbb{E}\cosh(tX^{\mathsf{T}}A^{\circ}X)$ using decoupling followed by arguments similar to [13, 23]. We note one can also recover an MGF-based analysis by specializing the analysis of [9] for analyzing sparse oblivious subspace embeddings to the case of "1-dimensional subspaces", though the resulting proof would be quite different from the one presented here. We believe the MGF-based analysis we give in this work appeals to more standard arguments, although the analysis in [9] does provide the advantage that it yields tradeoff bounds for s, m.

4 Our SJLT analyses

4.1 A first analysis: via the Hanson-Wright inequality

▶ **Theorem 5.** *For Π coming from an SJLT distribution, as long as $m \simeq \varepsilon^{-2}\log(1/\delta)$ and $s \simeq \varepsilon m$,*

$$\forall x : \|x\|_2 = 1,\ \mathop{\mathbb{P}}_{\Pi}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta.$$

Proof. As noted, we can write Z as a quadratic form

$$Z = \|\Pi x\|_2^2 - 1 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j \stackrel{\text{def}}{=} \sigma^T A_{x,\eta}\sigma,$$

Set $q = \Theta(\log(1/\delta)) = \Theta(s^2/m)$. By Hanson-Wright and the triangle inequality,

$$\|Z\|_q \le \|\sqrt{q}\cdot\|A_{x,\eta}\|_F + q\cdot\|A_{x,\eta}\|\|_q \le \sqrt{q}\cdot\|\|A_{x,\eta}\|_F\|_q + q\cdot\|\|A_{x,\eta}\|\|_q,$$

where $A_{x,\eta}$ is defined in Section 3. Since $A_{x,\eta}$ is block-diagonal, its operator norm is the largest operator norm of any block. The eigenvalue of the rth block is at most $(1/s)\cdot$

$\max\{\|x^{(r)}\|_2^2, \|x^{(r)}\|_\infty^2\} \leq 1/s$, and thus $\|A_{x,\eta}\| \leq 1/s$ with probability 1. Next, define $Q_{i,j} = \sum_{r=1}^m \eta_{r,i}\eta_{r,j}$ so that

$$\|A_{x,\eta}\|_F^2 = \frac{1}{s^2}\sum_{i\neq j} x_i^2 x_j^2 \cdot Q_{i,j}.$$

Suppose $\eta_{r_1,i}, \ldots, \eta_{r_s,i} = 1$ for distinct r_t and write $Q_{i,j} = \sum_{t=1}^s Y_t$, where Y_t is an indicator random variable for the event $\eta_{r_t,j} = 1$. By Remark 2 we may assume the Y_t are independent, in which case $Q_{i,j}$ is distributed as $\mathsf{Binomial}(s, s/m)$. Then by Lemma 2, $\|Q_{i,j}\|_q \lesssim q$. Thus,

$$\begin{aligned}
\|\|A_{x,\eta}\|_F\|_q &= \|\|A_{x,\eta}\|_F^2\|_{q/2}^{1/2}\\
&\leq \|\frac{1}{s^2}\sum_{i\neq j} x_i^2 x_j^2 \cdot Q_{i,j}\|_q^{1/2}\\
&\leq \frac{1}{s}\left(\sum_{i\neq j} x_i^2 x_j^2 \cdot \|Q_{i,j}\|_q\right)^{1/2} \quad \text{(triangle inequality)}\\
&\leq \frac{1}{\sqrt{m}}
\end{aligned}$$

Then by Markov's inequality and the settings of q, s, m,

$$\mathbb{P}(|\|\Pi x\|_2^2 - 1| > \varepsilon) = \mathbb{P}(|\sigma^T A_{x,\eta}\sigma| > \varepsilon) < \varepsilon^{-q} \cdot C^q(m^{-q/2} + s^{-q}) < \delta. \qquad ◀$$

▶ Remark. Less general bounds than Lemma 2 would have still sufficed for our purposes. For example, Bernstein's inequality and the triangle inequality together imply $\|Y\|_p \lesssim \alpha N + p$ for any $p \geq 1$, which suffices for our application since we were interested in the case $p = \alpha N$.

4.2 A second analysis: bounding the MGF

In this analysis we show the following bound on the symmetrized MGF of the error:

$$\mathbb{E}\cosh(tZ) \leq \exp\big(\tfrac{K^2t^2}{m}\big), \quad \text{for } |t| \leq \tfrac{s}{K}, \text{ where } K = 4\sqrt{2} \tag{5}$$

Using the above, we obtain tail estimates in a standard manner. By the generic Chernoff-Rubin bound:

$$\mathbb{P}(|Z| > \varepsilon) \leq 2e^{-t\varepsilon}\,\mathbb{E}\cosh(tZ) \leq 2\exp\big(\tfrac{K^2t^2}{m} - t\varepsilon\big), \quad \text{for all } 0 \leq t \leq \tfrac{s}{K}$$

Optimizing over the choice of t, we obtain the tail bound:

$$\mathbb{P}(|Z| > \varepsilon) \leq 2\max\big\{\exp(-C^2\varepsilon^2 m), \exp(-C\varepsilon s)\big\}, \qquad \text{where } C = \tfrac{1}{8\sqrt{2}}$$

▶ Remark. The cross-over point for the two bounds is when $\frac{s}{m} = \Theta(\varepsilon)$. To obtain a failure probability of δ, this yields the desired $s = O\big(\frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\big)$ and $m = O\big(\frac{1}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\big)$.

Our goal now is to prove eq. (5) for t satisfying $|t| \leq \frac{s}{K}$. Now by Taylor expansion, we have $\mathbb{E}\cosh(tZ) = \sum_{\text{even } q} \frac{|t|^q}{q!} \cdot \mathbb{E}\, Z^q$. Therefore, by section 2, we may assume that the $\eta_{r,i}$ are fully independent to bound $\mathbb{E}\cosh(tZ)$ from above. Now $\mathbb{E}\cosh(tZ) = \frac{1}{2}(\mathbb{E}\exp(tZ) + \mathbb{E}\exp(-tZ)) \leq \max\big\{\mathbb{E}\exp(tZ), \mathbb{E}\exp(-tZ)\big\}$, for all $t \in \mathbb{R}$. Let $B \stackrel{\text{def}}{=} \frac{1}{s}xx^\mathsf{T}$. Let $\Pi = \frac{1}{s}H$ and let $Y_1, Y_2, \ldots, Y_m$ denote the rows of H. Then $Z = \sum_{r=1}^m Y_r^\mathsf{T} B^\circ Y_r$. By the independence assumption, Y_i are i.i.d. Bernoulli-Rademacher vectors. Letting Y denote an identical copy of a single row of H,

$$\mathbb{E}\exp(\pm tZ) = \prod_r \mathbb{E}\exp(\pm tY_r^\top B^\circ Y_r) = \big(\mathbb{E}\exp(\pm tY^\top B^\circ Y)\big)^m, \quad \text{for all } t \in \mathbb{R} \tag{6}$$

Let Y' be an independent copy of Y. By decoupling (Theorem 4),

$$\mathbb{E}\exp(tY^\top B^\circ Y) \le \mathbb{E}\exp(4tY^\top BY') = \mathbb{E}\exp(Y^\top \tilde{B}Y'), \quad \text{for all } t \in \mathbb{R}, \text{ where } \tilde{B} \stackrel{\text{def}}{=} 4tB \tag{7}$$

We show below that

$$\mathbb{E}\exp(Y^\top \tilde{B}Y') \le 1 + \tfrac{K^2t^2}{m^2}, \quad \text{provided } |t| \le \tfrac{s}{K}, \text{ where } K = 4\sqrt{2} \tag{8}$$

Substituting this bound in eq. (7) and combining with eq. (6), we obtain:

$$\mathbb{E}\exp(\pm tZ) \le \big(1 + \tfrac{K^2t^2}{m^2}\big)^m \le \exp\big(\tfrac{K^2t^2}{m}\big), \quad \text{provided } |t| \le \tfrac{s}{K}, \text{ where } K = 4\sqrt{2},$$

which completes the proof of (5) as desired. It remains to prove eq. (8).

Bilinear forms of Bernoulli-Rademacher random variables.

The MGF of a Bernoulli-Rademacher random variable $X = \eta\sigma$ with parameter p equals $\mathbb{E}\exp(tX) = 1 - p + p\,\mathbb{E}\exp(t\sigma) \le 1 - p + p\exp(t^2/2)$, for all $t \in \mathbb{R}$.

Let $\lambda(z) \stackrel{\text{def}}{=} \exp(z) - 1$. Rewriting the above, we have $\mathbb{E}\,\lambda(tX) \le p\,\lambda(t^2/2) = p\,\mathbb{E}\,\lambda(tg)$, where $g \sim \mathcal{N}(0,1)$. We show an analogous replacement inequality for Bernoulli-Rademacher vectors.

▶ **Lemma 6.** *Let Y be a Bernoulli-Rademacher vector with parameter p. Then:*

$$\mathbb{E}\,\lambda(b^\top Y) \le p\,\lambda(\|b\|^2/2) = p\,\mathbb{E}\,\lambda(b^\top g) \quad \textit{for all vectors } b, \textit{ where } g \sim \mathcal{N}(0, I_n)$$

Proof. By stability of Gaussians, $\mathbb{E}\exp(b^\top g) = \exp(\|b\|_2^2/2)$, demonstrating the last equality above. Let $g(t) \stackrel{\text{def}}{=} \sum_{S \neq \emptyset} t^{|S|-1} \prod_{i \in S} \lambda(b_i^2/2)$ for $t \ge 0$. We have $\prod_i \big(1 + t\,\lambda(b_i^2/2)\big) = 1 + tg(t)$. Now:

$$\mathbb{E}\exp(b^\top Y) = \prod_i \mathbb{E}\exp(b_iY_i) = \prod_i \big(1 + \mathbb{E}\,\lambda(b_iY_i)\big) \le \prod_i \big(1 + p\,\lambda(b_i^2/2)\big) = 1 + pg(p)$$

Thus, $\mathbb{E}\,\lambda(b^\top Y) \le pg(p) \le pg(1)$, since $g(t)\uparrow$. To conclude, we claim that $g(1) = \lambda(\|b\|_2^2/2)$. Indeed:

$$1 + g(1) = \prod_i (1 + \lambda(b_i^2/2)) = \prod_i \exp(b_i^2/2) = \exp\big(\sum_i b_i^2/2\big) = 1 + \lambda\big(\|b\|_2^2/2\big)$$ ◀

Let $p \stackrel{\text{def}}{=} \frac{s}{m}$. In the left side of eq. (8), we have $\mathbb{E}\exp(Y^\top \tilde{B}Y') = 1 + \mathbb{E}\,\lambda(Y^\top \tilde{B}Y')$. By the law of total expectation:

$$\mathop{\mathbb{E}}_{Y,Y'} \lambda(Y^\top \tilde{B}Y') = \mathop{\mathbb{E}}_{Y}\mathop{\mathbb{E}}_{Y'}[\lambda((Y^\top \tilde{B})Y') \mid Y] \le \; p \cdot \mathop{\mathbb{E}}_{Y}\mathop{\mathbb{E}}_{g'}[\lambda((Y^\top \tilde{B})g') \mid Y]$$

(by lemma 6, applied to Y')

Exchange the order of expectations of Y and g' via Fubini-Tonelli's theorem. Now apply lemma 6, this time to Y. Finish using the law of total expectation which yields an upper bound of $p^2 \cdot \mathbb{E}\,\lambda(g^\top \tilde{B}g')$. Thus:

$$\mathbb{E}\exp(Y^\top \tilde{B}Y') \le 1 + p^2 \cdot \mathbb{E}\,\lambda(g^\top \tilde{B}g') \tag{9}$$

In order to be self-contained we include a standard proof of the following lemma, though note that the lemma itself is equivalent to the Hanson-Wright inequality for gaussian random variables since it gives a bound on the MGF of decoupled quadratic forms in gaussian random variables.

▶ **Lemma 7.** $\mathbb{E}\exp(g^\mathsf{T} Q g') \le \exp\big(\|Q\|_F^2\big)$ *for independent* $g, g' \sim \mathcal{N}(0, I_n)$*, provided* $\|Q\| \le \frac{1}{\sqrt{2}}$.

Proof. Let $Q = U\Sigma V^\mathsf{T}$, where $\Sigma = \mathsf{diag}(s_1, \ldots, s_n)$. So $\mathbb{E}\exp(g^\mathsf{T} Q g') = \mathbb{E}\exp(g^\mathsf{T} U\Sigma V^\mathsf{T} g')$. Since U is orthonormal, by rotational invariance, $U^\mathsf{T} g \sim \mathcal{N}(0, I_n)$ and is independent of $V^\mathsf{T} g' \sim \mathcal{N}(0, I_n)$. Therefore, $\mathbb{E}\exp(g^\mathsf{T} Q g') = \mathbb{E}\exp(g^\mathsf{T}\Sigma g')$. Now $g^\mathsf{T}\Sigma g' = \sum_i s_i g_i g_i'$, therefore:

$$\mathbb{E}\exp(g^\mathsf{T}\Sigma g') = \prod_i \mathbb{E}\,\mathbb{E}[\exp(s_i g_i g_i') \mid g_i] = \prod_i \mathbb{E}\exp(s_i^2 g_i^2/2) = \prod_i \tfrac{1}{\sqrt{1-s_i^2}}$$

Now $s_i^2 \le \|Q\|^2 \le \frac{1}{2}$ for each i. Use the bound $e^{-x} \le \sqrt{1-x}$ for $x \le \frac{1}{2}$ so that:

$$\mathbb{E}\exp(g^\mathsf{T} Q g') \le \prod_i \exp(s_i^2) = \exp\Big(\sum_i s_i^2\Big) = \exp\big(\|Q\|_F^2\big)$$ ◀

Note that $\|\tilde{B}\|_F = 4t\|B\|_F$ and $\|\tilde{B}\| = 4t\|B\|$. Now $B = \frac{1}{s} x x^\mathsf{T}$, so that $\|B\|_F = \|B\| = \frac{1}{s}$. Using the above proposition in the right side of eq. (9) with $Q = \tilde{B}$, we obtain:

$$\mathbb{E}\exp(Y^\mathsf{T}\tilde{B}Y') \le 1 + p^2 \cdot \lambda\big(\tfrac{K^2 t^2}{2s^2}\big), \quad \text{provided } |t| \le \tfrac{s}{K}, \text{ where } K = 4\sqrt{2}$$

In the right side above, use the bound $\lambda(x) \le 2x$, which holds for $x \le \frac{1}{2}$, and substitute $p = \frac{s}{m}$ so that

$$\mathbb{E}\exp(Y^\mathsf{T}\tilde{B}Y') \le 1 + \tfrac{K^2 t^2}{m^2}, \quad \text{provided } |t| \le \tfrac{s}{K}, \text{ where } K = 4\sqrt{2}$$

This yields the desired bound stated in eq. (8).

References

1 Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

2 Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

3 Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

4 Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, 2013.

5 Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2017.

6 Jean Bourgain. An improved estimate in the restricted isometry problem. *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics Volume 2116:65–70, 2014.

7 Vladimir Braverman, Rafail Ostrovsky, and Yuval Rabani. Rademacher chaos, random Eulerian graphs and the sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1011.2590, 2010.

8 Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

9 Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 278–287, 2016.

10 Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

11 David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42:1079–1083, 1971.

12 Ishay Haviv and Oded Regev. The restricted isometry property of subsampled Fourier matrices. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 288–297, 2016.

13 Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3):31, 2007.

14 Meena Jagadeesan. Simple analysis of sparse, sign-consistent JL. *CoRR*, abs/1708.02966, 2017.

15 T. S. Jayram and David P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):26:1–26:17, 2013.

16 William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

17 Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Proceedings of the 15th International Workshop on Randomization and Computation (RANDOM)*, pages 628–639, August 2011.

18 Daniel M. Kane and Jelani Nelson. A derandomized sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1006.3585, 2010.

19 Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4, January 2014. Preliminary version in SODA 2012.

20 Felix Krahmer and Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

21 Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss lemma. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2017.

22 Jelani Nelson, Eric Price, and Mary Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1515–1528, 2014.

23 Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

24 Mikkel Thorup and Yin Zhang. Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.*, 41(2):293–331, 2012.

25 Roman Vershynin. *High-Dimensional Probability*. May 2017. Last accessed at `http://www-personal.umich.edu/~romanv/papers/HDP-book/HDP-book.pdf` on August 22, 2017.