Fast, Deterministic and Sparse Dimensionality Reduction

Daniel Dadush *[1], Cristóbal Guzmán †[2], and Neil Olver [1,3]

[1]Centrum Wiskunde & Informatica, The Netherlands
[2]Pontificia Universidad Católica de Chile, Chile
[3]Vrije Universiteit Amsterdam, The Netherlands

Abstract

We provide a deterministic construction of the sparse Johnson-Lindenstrauss transform of Kane & Nelson (J.ACM 2014) which runs, under a mild restriction, in the time necessary to apply the sparse embedding matrix to the input vectors. Specifically, given a set of n vectors in $\mathbb{R}^d$ and target error ε, we give a deterministic algorithm to compute a $\{-1, 0, 1\}$ embedding matrix of rank $O((\ln n)/\varepsilon^2)$ with $O((\ln n)/\varepsilon)$ entries per column which preserves the norms of the vectors to within $1 \pm \varepsilon$. If NNZ, the number of non-zero entries in the input set of vectors, is $\Omega(d^2)$, our algorithm runs in time $O(\mathrm{NNZ} \cdot \ln n/\varepsilon)$.

One ingredient in our construction is an extremely simple proof of the Hanson-Wright inequality for subgaussian random variables, which is more amenable to derandomization. As an interesting byproduct, we are able to derive the essentially optimal form of the inequality in terms of its functional dependence on the parameters.

1 Introduction

Dimensionality reduction is an important and widely used technique in areas such as computer science, optimization and machine learning. Arguably, the most important result in dimensionality reduction is the (distributional) Johnson-Lindenstrauss (J-L) Lemma.

LEMMA 1.1. ([18]) *Let d be any positive integer, and let $\varepsilon, \delta \in (0, 1/2)$. Then there exists a random $m \times d$ matrix Π, with $m = \Theta(\varepsilon^{-2} \ln(1/\delta))$, such that for any $v \in \mathbb{R}^d$,*

$$\mathbb{P}\big((1-\varepsilon)\|v\|_2^2 \le \|\Pi v\|_2^2 \le (1+\varepsilon)\|v\|_2^2\big) \ge 1-\delta.$$

In particular, a union bound implies that for any set V of n vectors in $\mathbb{R}^d$, there exists a linear map into $\mathbb{R}^m$, where $m = \Theta(\varepsilon^{-2} \ln n)$, which preserves all lengths to within $1 \pm \varepsilon$.

This result lies at the heart of many state-of-the-art algorithms for problems including approximate nearest neighbors [17], mixtures of Gaussians [10], sketching [33] and fast algorithms for numerical linear algebra [11].

It is very desirable to be able to implement the dimensionality reduction as efficiently and as conveniently as possible, and there have been a number of works in this direction. The original construction involved projecting onto a random subspace [18]. Indyk and Motwani [17] showed that a projection matrix consisting of i.i.d. Gaussian entries also works, and Achlioptas [1] showed that even i.i.d. Rademacher random variables suffice, which is particularly convenient in some applications.

All these constructions produce dense (with at best a constant fraction of zero entries), unstructured projection matrices. Ailon and Chazelle [3] showed how to construct a projection that can be applied to a vector substantially faster than an arbitrary projection matrix. The idea is to first apply a random Hadamard transform that can be very efficiently applied using the fast Fourier transform, followed by projecting by a random matrix with i.i.d. $\{-1, 0, 1\}$ entries which is very sparse. The initial Hadamard transform ensures that with high probability, the weight of the transformed vector is well spread amongst its coordinates; this guarantees that the sparse projection is effective. A number of other results in this direction have followed [4, 24, 5, 23, 15].

One disadvantage of these results is that while they speed up the embedding of dense vectors, they cannot exploit sparsity of the input vectors. This motivated work of Dasgupta et al. [9], refined by Kane and Nelson [20], on choosing projection matrices that are directly sparse. Kane and Nelson [20] achieve the same guarantees as the usual J-L Lemma, with a projection matrix containing only $O(m\varepsilon)$ nonzero entries per column. This can be multiplied by a vector u in time $O(\mathrm{NNZ}(u) \cdot m\varepsilon)$, achieving a speedup of a factor of ε compared to multiplying u by an unstructured dense matrix. For sufficiently low sparsity this can be faster than the construction of Ailon-Chazelle.

A second line of work has been on reducing the amount of randomness required to implement it, or eliminating it altogether. There are two distinct threads here.

Distributional J-L with few random bits. In one thread, the goal is a distributional J-L Lemma, of the form given in Lemma 1.1, but where the number of random bits required to describe the distribution of the random matrix Π is as small as possible. This goal may be alongside ensuring other desirable properties of Π, such as sparsity. This has implications to streaming applications.

Representative recent results include independent work of Kane and Nelson [20] and Meka [25] that achieve a construction requiring $O(\ln d + \ln(1/\varepsilon)\ln(1/\delta) + \ln(1/\delta)\ln\ln(1/\delta))$ random bits (the result of Meka is restricted to $\delta = \Omega(1/\operatorname{poly} d)$). Other works include Clarkson and Woodruff [8] and Karnin et al. [22]. It is an open question to give an explicit construction requiring $O(\ln d + \ln(1/\delta))$ bits; probabilistic arguments show that such distributions do exist.

*Supported by the NWO Veni grant 639.071.510.

†Partially supported by FONDECYT Iniciación project 11160939. This work was partially done during a visit to Centrum Wiskunde & Informatica.

Full input-dependent derandomization. In a second thread, the goal is true derandomization: given the input vectors explicitly, deterministically give an embedding matrix which has low distortion on all of the input vectors. This will be our goal. One application of these results is to derandomize approximation algorithms based on rounding of semidefinite programs, for example, MAX-CUT.

Sivakumar [31] gives a polynomial-time derandomization based on a quite general complexity-theoretic tool due to Nisan [26, 27]. He does not state an explicit bound on the running time. Engebretsen et al. [12] give a derandomization that achieves a running time of $O(mn(\ln n+1/\varepsilon)^{O(1)})$. This matches the time required to project with an unstructured dense matrix, up to logarithmic factors. These logarithmic factors result from the somewhat brute-force nature of their approach. They derandomize the version of J-L with gaussian entries using the method of conditioned expectations. In order to do this, they fix the bits in the binary representation of the projection matrix one at a time (with a logarithmic number of bits needed per entry), and numerically approximate the integrals needed to evaluate various conditioned expectations.

Bhargava and Kosaraju [6] give a derandomization of J-L with $\{-1,1\}$ entries, using the method of pessimistic estimators [29]. We also use pessimistic estimators, but as discussed below, our approach has a number of advantages; in particular, our algorithm is faster due to the use of sparse projection matrices.

Contributions and techniques. Our main result is a full derandomization that takes time proportional to the "embedding time", i.e., the time required to apply the projection matrix to the input vectors. Moreover, we achieve this with a sparse projection matrix: our produced matrix will have roughly $m\varepsilon$ nonzero entries per column, leading to an embedding time of $O(m\varepsilon\mathrm{NNZ}(V))$.

Throughout, we will work with the real model of computation [7], but where in addition we allow the computation of exponentials in constant time. In order to actually implement these computations, sufficiently accurate polynomial approximations of $\exp$ must be used. We defer the details of how this can be achieved, with a logarithmic increase in the running time, to the full version of the paper.

THEOREM 1.1. *Let V be a set of $n \geq d$ vectors in $\mathbb{R}^d$, with $\mathrm{NNZ}(V)=\Omega(d^2)$, $\varepsilon \in (0,1)$ and $m \geq 14\ln(2n)/\varepsilon^2$. Then in time $O(m\varepsilon\mathrm{NNZ}(V))$, a matrix $\Pi \in \mathbb{R}^{m\times d}$ can be found, with $\lceil m\varepsilon \rceil$ nonzero entries per column, for which*

$$(1-\varepsilon)\|v\|_2^2 \leq \|\Pi v\|_2^2 \leq (1+\varepsilon)\|v\|_2^2 \qquad \forall v \in V.$$

This matches the result of Kane and Nelson [20] for the important case $\mathrm{NNZ}(V) \geq d^2$, both with respect to the size and sparsity of the resulting projection matrix, and in terms of the time required to project the vectors. (We also match their result in full generality, except we do not obtain the desired running time.)

Broadly speaking, our approach is based on the use of pessimistic estimators. A pessimistic estimator is nothing more than a quantity that upper bounds the failure probability of an event under some distribution. In this method, one begins with a distribution for which the event occurs with nonzero probability, and iteratively specifies the random variables in such a way that the chosen pessimistic estimator never increases. If this can be done, the event must occur in the final deterministic result with positive probability, and hence probability one. This is a standard technique, but the challenge is to find an appropriate pessimistic estimator that can be updated fast enough for our purposes, and behaves well (is concave in an appropriate sense) under updates. In this respect, our approach is similar to that of Bhargava and Kosaraju [6]. The main advantage of our result is that we obtain an $O(\varepsilon)$ speedup attributable to the sparsity of the projection matrix. Moreover the pessimistic estimators that we use are rather natural, coming from moment generating functions associated with Gaussian random variables.

In order to obtain a sparse projection matrix, we adapt the arguments of Kane and Nelson [21] for obtaining a random sparse projection matrix. Their approach has two parts; the first produces a (random) sparsity pattern, describing which entries of the projection matrix will be nonzero; the second chooses the signs for these entries (in their case, i.i.d. Rademachers). We derandomize both of these steps separately, using the method of pessimistic estimators.

- For the sparsity pattern (mask) of the matrix, Kane and Nelson choose, for each column independently, precisely s nonzero entries uniformly at random, where $s = \Theta(m\varepsilon)$. Thus their sampling procedure involved dependence amongst the entries in a column, and indeed this is crucial. They argue via negative dependence to obtain the required concentration bounds.

  Although our final algorithm eschews its use in order to run as quickly as possible, our derandomization is inspired by *pipage rounding* [2, 32], a general rounding technique for rounding a fractional point in the base polytope of a matroid to a vertex. The relevant matroid in our case is the partition matroid, where we must choose precisely s entries per column; in fact, we later use a more refined partition matroid to obtain our claimed running time. We begin with an initial uniform fractional point, which corresponds to the uniform distribution over sparsity patterns with s nonzero entries per row; the final, integral, output from the pipage rounding procedure corresponds to a deterministic choice of a mask. Harvey and Olver [14] suggest how pipage rounding can be combined with the use of pessimistic estimators, as long as the pessimistic estimator is well-behaved (more precisely, concave in the directions relevant to the pipage rounding scheme). Our pessimistic estimator, which is essentially the moment generating function after an application of Hölder's inequality, satisfies these requirements.

  We have to be very efficient in order to obtain the desired running time: for example, if $\varepsilon \ll 1/\mathrm{NNZ}(V)$ then the number of entries md in the projection matrix is larger than the desired running time $m\varepsilon\mathrm{NNZ}(V)$. We adapt an approach proposed by Kane and Nelson [21] (they named it the *code construction*). For $n \geq d$, there is a "universal" sparsity pattern that satisfies the required properties and depends only on the size of the projection matrix and not on the input vectors. The requirements of this universal sparsity pattern turn out to be convenient when it comes to updating the pessimistic estimator quickly.

- In order to show that randomly signing the chosen nonzero entries works, Kane and Nelson use the Hanson-Wright inequality [13] on a matrix determined by the mask. This inequality provides concentration of a quadratic form $\boldsymbol{z}^\mathsf{T} A\boldsymbol{z}$, where $\boldsymbol{z}$ is a random vector with independent subgaussian entries, in terms of parameters of the matrix A.

  We thus proceed by providing a derandomized version of the Hanson-Wright inequality (to our knowledge, the first such). Existing proofs of the inequality do not seem to yield pessimistic estimators that can be quickly evaluated. We give a simpler proof, based on, but simplifying, a proof of Rudelson and Vershynin [30], that is very convenient for our purposes. The crucial idea is that the moment generating function of the quadratic form can be bounded in terms of the moment generating function of $\boldsymbol{g}^\mathsf{T} A\boldsymbol{g}$, where $\boldsymbol{g}$ is *Gaussian*, which can be handled much more easily. Further work is needed once again to get the desired running time of $O(m\varepsilon \cdot \mathrm{NNZ}(V))$ and for this we exploit some of the structure of the specific matrix obtained in the application of the Hanson-Wright inequality to dimension reduction.

If $\mathrm{NNZ}(V) = o(d^2)$, we can still obtain the indicated sparse projection matrix, but with running time $O(m(\mathrm{NNZ}(V) + d))$. The second term is comparable to the time required to even write down the projection matrix, and so is quite justified. But we do not obtain an improvement over just using a dense projection matrix. The difficulty comes in the selection of the sparsity pattern (Section 4.2); if this can be improved to time $O(s(\mathrm{NNZ}(V) + d))$ then the existing algorithm for choosing the signs is sufficiently fast.

As a further implication of our simplified proof of the Hanson-Wright inequality, we are able to substantially sharpen it, and give essentially the correct functional dependence on the parameters. We think that this is of independent interest.

The paper is structured as follows. In Section 2 we introduce our notation and some preliminary results. In Section 3 we present our new Hanson-Wright inequality, together with its proof, and the proposed pessimistic estimators for derandomizing this inequality. In Section 4 we exploit the connection between Hanson-Wright and Johnson-Lindenstrauss dimensionality reduction to obtain pessimistic estimators for the latter: this includes the signing algorithm (Section 4.1) and the masking algorithm (Section 4.2), which can be analyzed separately. In Section 5 we focus on efficient computation of the various pessimistic estimators we use; in particular, for the Hanson-Wright inequality in general (Section 5.1), and for the signing (Section 5.2) and masking (Section 5.3) components of our new fast deterministic algorithm for sparse J-L. Finally, in Section 6, we discuss further technical aspects of our sharp Hanson-Wright inequality.

2 Notation and Preliminaries

We denote the natural logarithm by $\ln$. Random variables are denoted by bold characters. For any positive integer k, $[k]$ denotes the set $\{1, 2, \ldots, k\}$.

Given a matrix A, we use $\|A\|_{\mathrm{op}}$ and $\|A\|_F$ to denote the operator and Frobenius norms, respectively. Given two vectors u, v in the same vector space, $\langle u, v\rangle$ denotes the standard inner product. Given two matrices (or vectors), their Hadamard product corresponds to the component-wise multiplication, and this operation is denoted by $\odot$. The direct sum, $\bigoplus_{i=1}^k A_i$, is used to denote a block diagonal matrix with blocks $A_1, \ldots, A_k$. Given a matrix A, $\mathrm{Diag}(A)$ is the diagonal matrix that coincides with A on the diagonal. Given a vector v, $\mathrm{Diag}(v)$ is a diagonal matrix with diagonal coefficients from v. We denote by $\mathrm{supp}(A)$ the support of A, where A may be a vector or matrix. Given a vector $u \in \mathbb{R}^k$ and $S \subseteq [k]$, $u_S \in \mathbb{R}^{|S|}$ denotes the restriction of u to the coordinates in S. Similarly, given $M \in \mathbb{R}^{k\times k}$ and $S, T \subseteq [k]$, $M_{S,T}$ denotes the submatrix of M indexed by rows S and columns T.

The following well-known result allows to obtain fast computation of the determinant for a rank one perturbation of a matrix.

LEMMA 2.1. *Let $A \in \mathbb{R}^{k\times k}$ be an invertibe matrix, and $u, v \in \mathbb{R}^k$ vectors, then*

1. *Matrix determinant lemma:* $\det(A + uv^\mathsf{T}) = \det(A) \cdot (1 + v^\mathsf{T} A^{-1}u)$.
2. *Sherman-Morrison formula:* $(A + uv^\mathsf{T})^{-1} = A^{-1} - \dfrac{A^{-1}uv^\mathsf{T} A^{-1}}{1 + v^\mathsf{T} A^{-1}u}$.

We record the following simple inequalities regarding the logarithm.

LEMMA 2.2. *For $x \geq 0$,*

1. $\ln(1 + x) \leq \frac{x}{2} + \frac{x}{2(1+x)}$.
2. $\ln(1 + x) \geq \frac{x}{x+1}$.
3. $x \ln(1 + 1/x)$ *is increasing.*

We consider an ambient space $\mathbb{R}^d$ where our set of n input vectors V lie. Given $\varepsilon > 0$ and a matrix $\Pi \in \mathbb{R}^{m\times d}$ we will say it is a low-distortion projection for V if for all $v \in V$

$$(1 - \varepsilon)\|v\|_2^2 \leq \|\Pi v\|_2^2 \leq (1 + \varepsilon)\|v\|_2^2.$$

Given such a projection matrix Π we say its column sparsity is s if for all columns it has at most s nonzero coordinates.

We recall the definition of the moment generating function (mgf) of a random variable $\boldsymbol{x}$, $\psi_{\boldsymbol{x}}(\lambda) = \mathbb{E}_{\boldsymbol{x}}[e^{\lambda \boldsymbol{x}}]$. Given two real-valued random variables $\boldsymbol{y}, \boldsymbol{z}$ on the same probability space, we write $\boldsymbol{y} \preceq_{\mathrm{m}} \boldsymbol{z}$ if $\psi_{\boldsymbol{y}}(\lambda) \leq \psi_{\boldsymbol{z}}(\lambda)$ for all $\lambda \in \mathbb{R}$.

A random variable $\boldsymbol{x}$ is defined to be *ν-subgaussian*, $\nu > 0$, if $\boldsymbol{x} \preceq_{\mathrm{m}} \boldsymbol{z}$ where $\boldsymbol{z} \sim \mathcal{N}(0, \nu^2)$, namely $\mathbb{E}[e^{\lambda \boldsymbol{x}}] \leq e^{\lambda^2\nu^2/2}$ $\forall \lambda \in \mathbb{R}$. It is well known that up to a constant factor in the parameter, the above is equivalent to $\boldsymbol{x}$ being centered, $\mathbb{E}[\boldsymbol{x}] = 0$, and having subgaussian tails $\mathbb{P}[|\boldsymbol{x}| \geq \nu t] \leq$

$2e^{-t^2/2}$ $\forall t \geq 0$. The former definition will however be more convenient for our purposes.

We recall that Rademacher random variables (i.e. uniform on $\{-1, 1\}$) are 1-subgaussian.

LEMMA 2.3. *For $\boldsymbol{x}$ Rademacher and $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \boldsymbol{x}}] \leq e^{\lambda^2/2}$.*

The following formula, which comes from direct computation, will be frequently used.

LEMMA 2.4. *Let $\boldsymbol{z}$ be a standard Gaussian. Then for any $\alpha \in [0, 1/2)$ and $\beta \in \mathbb{R}$*

$$\mathbb{E}_{\boldsymbol{z}}[\exp\left(\alpha \boldsymbol{z}^2 + 2\beta \boldsymbol{z}\right)] = \frac{1}{\sqrt{1-2\alpha}} \exp\left(\frac{2\beta^2}{1-2\alpha}\right).$$

3 The Hanson-Wright Inequality and Derandomization

We state below the Hanson-Wright inequality in the form described by Rudelson and Vershynin [30], restricting ourselves to the case of zeros on the diagonal, as this case suffices for dimensionality reduction. It is well known that the general case can be handled by separately analyzing the contribution of diagonal terms, which can be easily bounded as a sum of independent random variables.

THEOREM 3.1. (HANSON-WRIGHT [13, 30]) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be independent ν-subgaussian random variables, and let $A = (a_{ij}) \in \mathbb{R}^{k \times k}$ be a symmetric matrix with $a_{ii} = 0$ for all $i \in [k]$. Then there exists a universal constant $C_{HW} > 0$ so that for every $t > 0$,*

$$\mathbb{P}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > \nu^2 t) \leq \exp\left(-C_{HW} \min\left\{\tfrac{t^2}{\|A\|_F^2}, \tfrac{t}{\|A\|_{\text{op}}}\right\}\right).$$

As one of our contributions, we give a simple proof that provides an essentially sharp version of this result. The bound we achieve is stated below.

THEOREM 3.2. (SHARP HANSON-WRIGHT) *Under the conditions of Theorem 3.1,*

$$(3.1) \quad \mathbb{P}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > \nu^2 t) \\ \leq \exp\left(-\tfrac{1}{2}\left(\tfrac{t}{\|A\|_{\text{op}}} - \tfrac{\|A\|_F^2}{\|A\|_{\text{op}}^2} \cdot \ln(1 + t\tfrac{\|A\|_{\text{op}}}{\|A\|_F^2})\right)\right).$$

To prove this result, our main new and simple observation is that the moment generating function of the quadratic form $\boldsymbol{x}^\mathsf{T} A \boldsymbol{x}$ only gets larger when the entries of $\boldsymbol{x}$ are replaced by corresponding Gaussians. We then derive (3.1) by computing an essentially optimal bound on the moment generating function for Gaussian entries as a function of the operator and Frobenius norm. The theorem is therefore optimal with respect to the Chernoff-Cramér method, and hence we expect that it cannot in general be improved.

Comparing to the proof of Rudelson and Vershynin [30], they also reduce to the Gaussian case however with the use of an additional "decoupling step", which replaces the mgf by an average of mgfs obtained by restricting A to a randomly chosen set of rows and columns. We note that decoupling weakens the achievable bound, in particular, as far as we are aware, the best computed Hanson-Wright constant C_{HW} using their method is $1/64$ [28] whereas we achieve $\frac{1-\ln(2)}{2} \approx 0.153 > 3/20$ (see Section 6 for a full proof and discussion). Aside from this, decoupling leads to a more computationally expensive pessimistic estimator from the perspective of derandomization, crucial to the context of this paper, namely an average of many estimators (naïvely an exponential number; using pairwise independence, this can be reduced to $O(k^2)$) versus the single estimator provided by the proof of Theorem 3.2.

As an interesting consequence of the Theorem 3.2, we give a slight quantitative improvement on the concentration bounds given in [1] for Rademacher Johnson-Lindenstrauss projections as a function of sparsity.

LEMMA 3.1. *Let $v \in \mathbb{R}^d$ be a unit vector with s non-zero entries. For $\Pi \in \{-1, 1\}^{m \times d}$, an $m \times d$ random matrix with coefficients distributed i.i.d. Rademacher, and $\varepsilon > 0$, we have that*

$$\max\{\mathbb{P}(\|\Pi v\|_2^2/m \geq 1 + \varepsilon), \mathbb{P}(\|\Pi v\|_2^2/m \leq 1 - \varepsilon)\} \\ \leq e^{-\frac{m}{2}\frac{s}{s-1}(\varepsilon - \ln(1+\varepsilon))}.$$

We note that this is better than the error exponent achieved by Gaussian random projections (at least for the harder upper tail estimate) by a small but explicit $s/(s-1)$ factor. The proof of the lemma follows by using sparsity dependent bounds on the operator and Frobenius norm of the relevant matrix, namely $vv^\mathsf{T} - \text{Diag}(vv^\mathsf{T})$, and applying Theorem 3.2. We defer the proof of the above lemma to Section 6.

We now begin with our proof of Theorem 3.2. The crux of the proof is the following simple lemma which allows us to compare the moment generating functions of diagonal 0 quadratic forms.

LEMMA 3.2. *Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ all be independent, with $\boldsymbol{y}_i \preceq_{\text{m}} \boldsymbol{z}_i$ for all i. Then for a zero diagonal matrix $A \in \mathbb{R}^{k \times k}$, $\boldsymbol{y}^\mathsf{T} A \boldsymbol{y} \preceq_{\text{m}} \boldsymbol{z}^\mathsf{T} A \boldsymbol{z}$.*

Proof. It suffices to prove the lemma in the case that $\boldsymbol{y}_i = \boldsymbol{z}_i$ for all $i \neq \ell$ for some $\ell \in [k]$. The full lemma then follows by combining the inequalities

$$\mathbb{E}_{\boldsymbol{z}^{(i)}}[e^{\lambda(\boldsymbol{z}^{(i)})^\mathsf{T} A \boldsymbol{z}^{(i)}}] \leq \mathbb{E}_{\boldsymbol{z}^{(i+1)}}[e^{\lambda(\boldsymbol{z}^{(i+1)})^\mathsf{T} A \boldsymbol{z}^{(i+1)}}],$$

where $\boldsymbol{z}_j^{(i)} = \boldsymbol{z}_j$ for $i \leq j$ and $\boldsymbol{z}_j^{(i)} = \boldsymbol{y}_j$ for $j > i$. For $\lambda \in \mathbb{R}$, we have that

$$\begin{aligned}
&\mathbb{E}[e^{\lambda \boldsymbol{y}^\mathsf{T} A \boldsymbol{y}}] \\
&\quad = \mathbb{E}_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k} e^{\lambda \sum_{i,j} a_{ij} \boldsymbol{y}_i \boldsymbol{y}_j} \\
&\quad = \mathbb{E}_{\boldsymbol{y}_i : i \neq \ell}\left[\mathbb{E}_{\boldsymbol{y}_\ell} e^{\lambda(\boldsymbol{y}_\ell \sum_{i \neq \ell}(a_{i\ell} + a_{\ell i})\boldsymbol{y}_i + \sum_{i \neq j \neq \ell} a_{ij} \boldsymbol{y}_i \boldsymbol{y}_j)}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(since } a_{\ell\ell} = 0\text{)} \\
&\quad \leq \mathbb{E}_{\boldsymbol{y}_i : i \neq \ell}\left[\mathbb{E}_{\boldsymbol{z}_\ell} e^{\lambda(\boldsymbol{z}_\ell \sum_{i \neq \ell}(a_{i\ell} + a_{\ell i})\boldsymbol{y}_i + \sum_{i \neq j \neq \ell} a_{ij} \boldsymbol{y}_i \boldsymbol{y}_j)}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by } \boldsymbol{y}_\ell \preceq_m \boldsymbol{z}_\ell\text{)} \\
&\quad = \mathbb{E}_{\boldsymbol{z}_i : i \neq k}\left[\mathbb{E}_{\boldsymbol{z}_\ell} e^{\lambda(\boldsymbol{z}_\ell \sum_{i \neq \ell}(a_{i\ell} + a_{\ell i})\boldsymbol{z}_i + \sum_{i \neq j \neq \ell} a_{ij} \boldsymbol{z}_i \boldsymbol{z}_j)}\right] \\
&\quad = \mathbb{E}_{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k}[e^{\lambda \boldsymbol{z}^T A \boldsymbol{z}}].
\end{aligned}$$

□

This result allows us to replace all subgaussian variables by Gaussians.

COROLLARY 3.1. *Under the assumptions of Theorem 3.1,*

$$\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \preceq_{\mathrm{m}} \nu^2 \boldsymbol{g}^\mathsf{T} A \boldsymbol{g},$$

where $\boldsymbol{g}$ is a standard k-dimensional Gaussian random vector.

Given the above, to be able to apply the Chernoff bound it suffices to get strong bounds on the moment function of $\boldsymbol{g}^\mathsf{T} A \boldsymbol{g}$. We recall the form of the mgf below. Let $\mu_1 \geq \cdots \geq \mu_k$ denote the eigenvalues of A. For $\lambda \in \mathbb{R}$, $|\lambda| < \frac{1}{2\|A\|_{\mathrm{op}}}$,

$$\begin{aligned} \mathbb{E}[\exp(\lambda \boldsymbol{g}^\mathsf{T} A \boldsymbol{g})] &= \mathbb{E}\Big[\exp\Big(\sum_{i=1}^{k} \lambda \mu_i \boldsymbol{g}_i^2\Big)\Big] \\ &\qquad \text{(by rotational symmetry of } \boldsymbol{g} \text{)} \\ &= \prod_{i=1}^{k} (1 - 2\lambda\mu_i)^{-\frac{1}{2}} \quad \text{(by Lemma 2.4)} \\ &= \det(I - 2\lambda A)^{-1/2}. \end{aligned} \tag{3.2}$$

From here, computing an optimal mgf bound corresponds to finding the minimum value of the determinant $\det(I - 2\lambda A)$ subject to fixed upper bounds on $\|A\|_{\mathrm{op}}$ and $\|A\|_F$. In the following lemma, we give a bound on this problem which is optimal in the asymptotic regime, i.e., for $k \to \infty$, under the condition that $\|A\|_{\mathrm{op}}^2$ evenly divides $\|A\|_F^2$. We defer the proof to Section 6.

LEMMA 3.3. (DETERMINANT LOWER BOUND) *For $0 \leq \alpha < 1$, $\beta \geq \alpha$, we have that*

$$\begin{aligned} &\inf\{\det(I - M) : M \in \mathbb{R}^{k \times k} \textit{ symmetric}, \\ &\quad M_{ii} = 0 \, \forall i \in [k], k \in \mathbb{N}, \|M\|_{\mathrm{op}} \leq \alpha, \|M\|_F \leq \beta\} \\ &\qquad \geq (1 - \alpha)^{\beta^2/\alpha^2} \exp(\beta^2/\alpha), \end{aligned} \tag{3.3}$$

where equality is attained when β^2/α^2 is an integer.

Applying the above lemma to the matrix $M = 2\lambda A$, we get the following mgf bound:

$$\begin{aligned} \mathbb{E}[\exp(\lambda \boldsymbol{g}^\mathsf{T} A \boldsymbol{g})] &= \det(I - 2\lambda A)^{-1/2} \\ &\leq \exp(-\lambda \tfrac{\|A\|_F^2}{\|A\|_{\mathrm{op}}} - \tfrac{1}{2} \tfrac{\|A\|_F^2}{\|A\|_{\mathrm{op}}^2} \ln(1 - 2\lambda\|A\|_{\mathrm{op}})). \end{aligned} \tag{3.4}$$

Using the above bound with the Chernoff-Cramér method, we obtain Theorem 3.2. We state it in the following form for future reference.

LEMMA 3.4. *Under the conditions of Theorem 3.1,*

$$\begin{aligned} \mathbb{P}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > t) &\leq \inf_{\lambda > 0} e^{-\lambda t} \cdot \mathbb{E}[\exp(\lambda \boldsymbol{g}^\mathsf{T} A \boldsymbol{g})] \\ &\leq \exp\left(-\tfrac{1}{2}\left(\tfrac{t}{\|A\|_{\mathrm{op}}} - \tfrac{\|A\|_F^2}{\|A\|_{\mathrm{op}}^2} \cdot \ln(1 + t\tfrac{\|A\|_{\mathrm{op}}}{\|A\|_F^2})\right)\right). \end{aligned}$$

Proof. By homogeneity, it suffices to prove the claim when $\boldsymbol{x}$ is 1-subgaussian. We have that

$$\begin{aligned} \mathbb{P}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > t) &\leq \inf_{\lambda > 0} \mathbb{P}\big(\exp(\lambda \boldsymbol{x}^\mathsf{T} A \boldsymbol{x}) > \exp(\lambda t)\big) \\ &\leq \inf_{\lambda > 0} \mathbb{E}[\exp(\lambda \boldsymbol{x}^\mathsf{T} A \boldsymbol{x})] \exp(-\lambda t) \\ &\leq \inf_{\lambda > 0} e^{-\lambda t} \cdot \mathbb{E}[\exp(\lambda \boldsymbol{g}^\mathsf{T} A \boldsymbol{g})] \quad \text{(by Cor. 3.1)} \\ &\leq \inf_{0 < \lambda < 1/(2\|A\|_{\mathrm{op}})} e^{-\lambda t} \cdot \det(I - 2\lambda A)^{-1/2} \\ &\leq \inf_{0 < \lambda < 1/(2\|A\|_{\mathrm{op}})} \exp\Big(-\lambda(t + \tfrac{\|A\|_F^2}{\|A\|_{\mathrm{op}}}) \\ &\qquad - \tfrac{1}{2} \tfrac{\|A\|_F^2}{\|A\|_{\mathrm{op}}^2} \ln(1 - 2\lambda\|A\|_{\mathrm{op}})\Big) \quad \text{(by (3.4))} \\ &= \exp\Big(-\tfrac{1}{2}\Big(\tfrac{t}{\|A\|_{\mathrm{op}}} - \tfrac{\|A\|_F^2}{\|A\|_{\mathrm{op}}^2} \ln(1 + t\tfrac{\|A\|_{\mathrm{op}}}{\|A\|_F^2})\Big)\Big). \end{aligned}$$

The last equality follows by setting

$$\lambda = \frac{1}{2} \frac{t}{\|A\|_{\mathrm{op}}(t + \|A\|_F^2/\|A\|_{\mathrm{op}})}, \tag{3.5}$$

which is a minimizer since the function in the exponent is convex and this choice sets the derivative to zero. □

3.1 Derandomization of Hanson-Wright. As opposed to other approaches for the Hanson-Wright inequality, our proof can be easily derandomized. For this we will restrict ourselves to the Rademacher case, where $\boldsymbol{x}$ is an n-dimensional vector of i.i.d. Rademacher random variables, so all components are 1-subgaussian. The derandomization is based on using the Gaussian mgf as a pessimistic estimator, and then exploiting the concavity of the mgf in order to fix the coordinates of $\boldsymbol{x}$ one by one, ensuring that the estimator does not increase.

For any $r = 0, \ldots, k$ and $\sigma \in \{-1, 1\}^r$, let $\mathcal{D}(\sigma)$ denote the distribution of a random vector in $\mathbb{R}^k$ whose i-th coordinate is deterministically σ_i for $i \leq r$, and an independent Rademacher for $i > r$; this describes a partial fixing of the variables $\boldsymbol{x}_i$. Define $\mathcal{G}(\sigma)$ similarly, except that for all $i > r$, the i-th coordinate is an independent standard Gaussian.

Our pessimistic estimator will be

$$\psi(\sigma) := e^{-\lambda t} \cdot \mathbb{E}_{\boldsymbol{g} \sim \mathcal{G}(\sigma)} \exp(\lambda \boldsymbol{g}^\mathsf{T} A \boldsymbol{g}). \tag{3.6}$$

By Lemma 3.4,

$$\mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}(\emptyset)}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > t) \leq \psi(\emptyset),$$

and moreover if we choose λ optimally, $\psi(\emptyset) < 1$ whenever probability bound guaranteed by Theorem 3.2 is itself below 1.

All that remains is to confirm that ψ has the required concavity properties to be usable as a pessimistic estimator.

LEMMA 3.5. *Let $0 \leq r < k$ and $\sigma \in \{-1, 1\}^r$. Let $\sigma^- = (\sigma, -1)$ and $\sigma^+ = (\sigma, 1)$. Then*

$$\psi(\sigma) \geq \tfrac{1}{2}(\psi(\sigma^-) + \psi(\sigma^+)).$$

Proof. Let $\boldsymbol{y} \in \mathbb{R}^k$ be a random vector where $\boldsymbol{y}_{r+1}$ is an independent Rademacher, $\boldsymbol{y}_i = \sigma_i$ for $i \leq r$, and $\boldsymbol{y}_i$ is an independent standard Gaussian for $i > r + 1$. Define

$\boldsymbol{z} \in \mathbb{R}^k$ identically, except that $\boldsymbol{z}_{r+1}$ is an independent standard Gaussian. Notice that

$$\begin{aligned} \tfrac{1}{2}(\psi(\sigma^-) + \psi(\sigma^+)) &= e^{-\lambda t} \cdot \mathbb{E}_{\boldsymbol{y}_{r+1}}\big[\mathbb{E}_{\boldsymbol{y}_i : i \neq r+1}[e^{\lambda \boldsymbol{y}^\mathsf{T} A \boldsymbol{y}}]\big] \\ &\leq e^{-\lambda t} \cdot \mathbb{E}_{\boldsymbol{z}}[e^{\lambda \boldsymbol{z}^\mathsf{T} A \boldsymbol{z}}] \qquad \text{by Lemma 3.2} \\ &= \psi(\sigma). \end{aligned}$$

□

The algorithm (modulo computation of the pessimistic estimator) is now straightforward. We sequentially choose signs $\sigma_1, \ldots, \sigma_k$, in such a way that at every step we choose the $\sigma_i \in \pm 1$ so that the pessimistic estimator does not increase. So we obtain $\sigma \in \{-1, 1\}^k$ with $\psi(\sigma) < 1$, which immediately implies that $\sigma^\mathsf{T} A \sigma \leq t$.

All that remains is to check that we can actually compute ψ; we postpone this to Section 5.1.

4 Derandomizing Sparse J-L via Pessimistic Estimators

Kane and Nelson [19, 21] demonstrated how sparse dimensionality reduction follows from the Hanson-Wright inequality. Here we will give a different proof of their randomized result, following the framework that they lay out, and show how this can be easily derandomized using pessimistic estimators derived naturally from the moment-generating functions that we use in our proof. The only aspect we will not yet discuss is the efficient computation of the pessimistic estimators, which will follow in Section 5.

We will assume from now on that $\|v\| = 1$ for all $v \in V$; this is clearly without loss of generality. Recall that $m \geq 14\ln(2n)/\varepsilon^2$. We will assume for simplicity that $m\epsilon$ is an integer, so our goal is a projection matrix with $s := m\epsilon$ nonzero entries per column. More precisely, we will use a J-L matrix with the following form:

$$\Pi = \frac{1}{\sqrt{s}} (\delta_{rj}\sigma_{rj})_{r \in [m], j \in [d]}.$$

Here, $\Delta := (\delta_{rj})_{r,j}$ is a masking 0-1 matrix, and its role is to sparsify the projection; and $\Sigma := (\sigma_{rj})_{r,j}$ is a ± 1 matrix, which (with proper scaling) can be seen as a standard (Rademacher) projection matrix. Notice that due to the masking most σ_{rj}'s will not be used at all in the final projection (nor in the algorithm), but keeping them makes the analysis cleaner.

The Kane-Nelson analysis, and ours as well, breaks into two parts. One is the choice of mask matrix Δ, which will be required to satisfy certain properties. The other is the choice of the sign matrix Σ once Δ has been determined. We discuss each of these parts separately.

4.1 The sign matrix. For the moment, let Δ be a given, deterministic 0-1 matrix; we will elucidate what properties we require from it. Let δ_r denote the r'th row of Δ, and for $v \in V$ let

$$A_v = \frac{1}{s} \bigoplus_{r=1}^{m} [B_{r,v} - \mathrm{Diag}(B_{r,v})],$$

with $B_{r,v} = (\delta_r \odot v)(\delta_r \odot v)^\mathsf{T}$. So each block of A_v is a symmetric rank-1 matrix less its diagonal.

Definition 1. *We say that $\Delta \in \{0,1\}^{m \times d}$ is* ***good*** *if*

(i) each column of Δ has precisely s nonzero entries; and

(ii) for each $v \in V$, the resulting A_v satisfies

$$\|A_v\|_{\mathrm{op}} \leq \frac{1}{s} \quad \textit{and} \quad \|A_v\|_F^2 \leq \frac{2}{m}.$$

The rationale for this definition will become clear shortly. So suppose Δ is a good matrix. Let $\sigma \in \mathbb{R}^{md}$ be a vectorized version of Σ, obtained by concatenating its rows. Then for any input vector $v \in V$,

$$\begin{aligned} &\|\Pi v\|_2^2 - 1 \\ &\quad = \frac{1}{s} \sum_{r=1}^{m} \Big(\sum_{j=1}^{d} \delta_{rj}\sigma_{rj} v_j \Big)^2 - 1 \\ &\quad = \frac{1}{s} \sum_r \sum_{j \neq k} \delta_{rj}\delta_{rk}\sigma_{rj}\sigma_{rk} v_j v_k + \frac{1}{s} \sum_r \sum_j \delta_{rj} v_j^2 - 1 \\ &\quad = \sigma^\mathsf{T} A_v \sigma + \sum_j \Big(\frac{1}{s} \sum_r \delta_{rj} \Big) v_j^2 - 1 \\ &\quad = \sigma^\mathsf{T} A_v \sigma. \end{aligned}$$

The final step uses that Δ has precisely s nonzero entries per column, and that each $v \in V$ is a unit vector.

Let the entries of Σ be i.i.d. Rademachers. Then we obtain by Hanson-Wright (Theorem 3.2) and a union bound that

(4.1)
$$\begin{aligned} &\mathbb{P}\big(\exists v \in V : \big|\|\Pi v\|_2^2 - 1\big| > \varepsilon\big) \\ &\quad \leq 2n \cdot \exp\Big(-\tfrac{1}{2} \Big(\tfrac{\varepsilon}{\|Av\|_{\mathrm{op}}} - \tfrac{\|A_v\|_F^2}{\|A_v\|_{\mathrm{op}}^2} \cdot \ln\big(1 + \varepsilon \tfrac{\|A_v\|_{\mathrm{op}}}{\|A_v\|_F^2}\big) \Big) \Big) \\ &\quad \leq 2n \exp\big(-\tfrac{1}{2} \big(\varepsilon s - 2\varepsilon s \ln(1 + 1/2) \big) \big) \\ &\quad < 2n \exp(-\varepsilon^2 m/14) \\ &\quad \leq 1. \end{aligned}$$

Existence of a good sign matrix, being a positive probability event, is thus guaranteed. The rationale for the definition of a good mask matrix should now be apparent.

Derandomization. Since the argument for the existence of a good choice of Σ (given a mask matrix with the desired properties) was nothing more than an application of Hanson-Wright, the derandomization used in Section 3.1 applies. Our pessimistic estimator is simply

$$\Phi(\sigma) := \sum_{v \in V} \big(\psi_v^+(\sigma) + \psi_v^-(\sigma) \big),$$

where $\psi_v^+(\sigma) = e^{-\lambda t} \cdot \mathbb{E}_{\boldsymbol{g} \sim \mathcal{G}(\sigma)} \exp\{\lambda \boldsymbol{g}^\mathsf{T} A_v \boldsymbol{g}\}$ and $\psi_v^-(\sigma) = e^{-\lambda t} \cdot \mathbb{E}_{\boldsymbol{g} \sim \mathcal{G}(\sigma)} \exp\{-\lambda \boldsymbol{g}^\mathsf{T} A_v \boldsymbol{g}\}$, are pessimistic estimators for the events $\boldsymbol{x}^\mathsf{T} A_v \boldsymbol{x} > t$ and $\boldsymbol{x}^\mathsf{T} A_v \boldsymbol{x} < -t$, respectively. We have already seen in (4.1) that under the conditions of Theorem 1.1, and with the appropriate choice of λ (namely, as given by (3.5) with $\|A\|_{\mathrm{op}} = 1/s$ and $\|A\|_F^2 = 2/m$), we have $\Phi(\emptyset) < 1$. By Lemma 3.5, for each $v \in V$

$$\psi_v^+(\sigma) \geq \tfrac{1}{2}(\psi_v^+(\sigma^-) + \psi_v^+(\sigma^+))$$

(and similarly for ψ_v^-). Thus

$$\Phi(\sigma) \geq \tfrac{1}{2}(\Phi(\sigma^-) + \Phi(\sigma^+)),$$

and once again we can sequentially choose signs (of course, we only need to consider entries in the support of Δ) until we obtain some $\sigma \in \{-1, +1\}^{\mathrm{supp}(\Delta)}$ for which $\Phi(\sigma) < 1$. This implies that

$$\|\Pi v\|_2^2 - 1 \;=\; \sigma^{\mathsf{T}} A_v \sigma \;\in\; [-\varepsilon, +\varepsilon] \qquad \forall v \in V.$$

In order to implement this algorithm, we need to be able to compare $\Phi(\sigma^+)$ with $\Phi(\sigma)$. We postpone the discussion of how to do this efficiently (in time comparable to applying the sparse projection) to Section 5.2.

4.2 The mask matrix.

To show the existence of a good mask matrix, we follow a randomized block construction of Kane and Nelson [21]. They partition the rows into s parts of size $m/s = 1/\varepsilon$ (for simplicity, we will assume that $1/\varepsilon$ is an integer): so let $B_1 = \{1, 2, \ldots, 1/\varepsilon\}$, $B_2 = \{1/\varepsilon{+}1, \ldots, 2/\varepsilon\}$, etc. For any column j and any q, we call the entries in column j and rows in B_q the entries of *block* (q, j). The distribution Kane and Nelson choose is that for each block, a single 1 entry is chosen uniformly at random, with each block being chosen independently. Clearly this yields a random matrix Δ that always has precisely s nonzero entries per column. The main difficulty in showing that Δ is good with some positive probability is bounding the Frobenius norm. Kane and Nelson do this by controlling its moments: the moments of

$$\|A_v\|_F^2 \;=\; \frac{1}{s^2} \sum_{j \neq k} \boldsymbol{q}_{jk} v_{ij}^2 v_{ik}^2, \tag{4.2}$$

where $\boldsymbol{q}_{jk} := \sum_{r \in [m]} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk}$ is the number of entries that column j and k have in common. The difficulty of bounding the moments of $\|A_v\|_F^2$ comes precisely from the fact that the $\boldsymbol{q}_{jk}$'s are not independent. However, it is easy to see these variables are negatively associated, and they exploit this fact to bound the moments as if these variables were independent. We do this differently in order to provide a convenient pessimistic estimator.

Let us begin by bounding the operator norm of A_v; this requires nothing more than the block structure of A_v, and holds surely.

$$\begin{aligned}
\|A_v\|_{\mathrm{op}} &= \frac{1}{s} \cdot \max_{r \in [m]} \Big\| (\delta_r \odot v)(\delta_r \odot v)^\top \\
&\qquad\qquad - \mathrm{Diag}((\delta_r \odot v)(\delta_r \odot v)^\top) \Big\|_{\mathrm{op}} \\
&\leq \frac{1}{s} \cdot \max_{r \in [m]} \max \Big\{ \|(\delta_r \odot v)(\delta_r \odot v)^\top\|_{\mathrm{op}}, \\
&\qquad\qquad \|\,\mathrm{Diag}((\delta_r \odot v)(\delta_r \odot v)^\top)\|_{\mathrm{op}} \Big\} \\
&\leq \frac{1}{s} \cdot \max_{r \in [m]} \max \Big\{ \|\delta_r \odot v\|_2^2, \max_{j \in [d]} v_j^2 \Big\} \\
&\leq \frac{1}{s},
\end{aligned}$$

where the first inequality holds because $\|M - N\|_{\mathrm{op}} \leq \max\{\|M\|_{\mathrm{op}}, \|N\|_{\mathrm{op}}\}$ when M and N are both positive semidefinite.

Now we consider $\|A_v\|_F^2$. For any $\lambda > 0$ we have

$$\begin{aligned}
&\mathbb{P}\big(\|A_v\|_F^2/2 > t\big) \qquad (4.3)\\
&\quad = \mathbb{P}\Big(\frac{1}{2s^2} \sum_{r \in [m]} \sum_{j \neq k} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk} v_j^2 v_k^2 > t\Big) \\
&\quad = \mathbb{P}\Big(\frac{1}{s^2} \sum_{r \in [m]} \sum_{j < k} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk} v_j^2 v_k^2 > t\Big) \\
&\quad \leq e^{-\lambda t} \cdot \mathbb{E}\Big[\exp\Big\{ \frac{\lambda}{s^2} \sum_{j<k} \sum_{q=1}^{s} \big(\sum_{r \in B_q} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk} \big) v_j^2 v_k^2 \Big\} \Big] \\
&\quad = e^{-\lambda t} \cdot \mathbb{E}\Big[\prod_{q=1}^{s} \prod_{j<k} \exp\Big\{ \frac{\lambda}{s^2} \big(\sum_{r \in B_q} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk} \big) \Big\}^{v_j^2 v_k^2} \Big] \\
&\quad = e^{-\lambda t} \cdot \prod_{q=1}^{s} \mathbb{E}\Big[\prod_{j<k} \exp\Big\{ \frac{\lambda}{s^2} \big(\sum_{r \in B_q} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk} \big) \Big\}^{v_j^2 v_k^2} \Big], \qquad (4.4)
\end{aligned}$$

where in the last step we have used that by the choice of the distribution, the random variables are independent among blocks. Now we will use the generalized Hölder inequality, noting that

$$\sum_{j<k} v_j^2 v_k^2 = \frac{1}{2} \sum_{j \neq k} v_j^2 v_k^2 \leq \frac{1}{2} \big(\sum_j v_j^2\big)^2 = \frac{1}{2}, \tag{4.5}$$

and hence

$$\begin{aligned}
&\mathbb{P}\big(\|A_v\|_F^2 > 2t\big) \\
&\quad \leq e^{-\lambda t} \cdot \prod_{q=1}^{s} \prod_{j<k} \mathbb{E}\Big[\exp\Big\{ \frac{\lambda}{s^2} \big(\sum_{r \in B_q} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk} \big) \Big\} \Big]^{v_j^2 v_k^2}.
\end{aligned}$$

Choose $\lambda = s^2 \ln 2$. Notice that for any $j < k \in [d]$, $q \in [s]$, $\sum_{r \in B_q} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk}$ is distributed as a Bernoulli with parameter ε. Thus

$$\mathbb{E}\Big[2^{\sum_{r \in B_q} \boldsymbol{\delta}_{rj} \boldsymbol{\delta}_{rk}} \Big] = 1 + \varepsilon.$$

Taking a union bound over all vectors $v \in V$,

$$\begin{aligned}
&\mathbb{P}\big(\exists v \in V : \|A_v\|_F^2 > 2 \cdot \tfrac{1}{m}\big) \qquad (4.6)\\
&\qquad \leq \sum_{v \in V} e^{-(\ln 2) s^2/m} \cdot (1 + \varepsilon)^{s \cdot \sum_{j<k} v_j^2 v_k^2} \\
&\qquad \leq n \cdot e^{-(\ln 2) \varepsilon^2 m} \cdot (e^{\varepsilon})^{s/2} \qquad \text{by (4.5)} \\
&\qquad < n \exp\big(-\varepsilon^2 m/7\big) \qquad (4.7)\\
&\qquad < 1.
\end{aligned}$$

Thus a good mask matrix does exist.

Derandomization. We will now derandomize this argument using pessimistic estimators. We will be interested in distributions over mask matrices Δ where some blocks have had the position of their single nonzero entry chosen, and some have not. We will describe such distribution with a vector $p \in \{0, \varepsilon, 1\}^{m \times d}$, satisfying

$$\sum_{r \in B_q} p_{rj} = 1 \qquad \text{for each} \quad q \in [s], j \in [d]. \tag{4.8}$$

For any such p, let $\mathcal{M}(p)$ denote the distribution of a random matrix $(\boldsymbol{\delta}_{rj})_{r\in[m],j\in[d]}$ where $\mathbb{P}(\boldsymbol{\delta}_{rj}=1)=p_{rj}$, and moreover each block is independent of the others and contains precisely a single 1. Let $p^{(0)}$ denote the distribution used in the existential analysis: $p^{(0)}_{rj}=\varepsilon$ for all $r\in[m]$, $j\in[d]$.

Our pessimistic estimator will be $\Psi(p) := \sum_{v\in V}\Psi_v(p)$, where

$$\Psi_v(p) := 2^{-s^2t}\cdot\prod_{q=1}^{s}\prod_{j<k}\Big(\mathbb{E}_{\boldsymbol{\delta}\sim\mathcal{M}(p)}\big[2^{\sum_{r\in B_q}\boldsymbol{\delta}_{rj}\boldsymbol{\delta}_{rk}}\big]\Big)^{v_j^2v_k^2}.$$

The derivation of (4.4) did not use anything about the distribution of $\boldsymbol{\delta}$ aside from the independence between blocks, and so the argument there, combined with a union bound implies that Ψ is indeed a pessimistic estimator:

$$\mathbb{P}_{\boldsymbol{\delta}\sim\mathcal{M}(p)}(\exists v\in V: \|A_v\|_F^2>2t)\le\Psi(p).$$

Moreover we have already seen that $\Psi(p^{(0)})<1$ under the conditions of Theorem 1.1.

It remains to show that we can gradually fix the blocks one at a time in a way that does not increase the value of the pessimistic estimator. Define a total order on blocks $\prec$ defined by the lexicographical order

$$(q,j)\prec(r,k)\quad\iff\quad\text{either } q<r,\ \text{or } q=r \text{ and } j<k.$$

Our algorithm will fix the position of the single nonzero entry of each block one by one, according to this ordering. The precise ordering of the blocks will not matter here, but will aid in the efficient computation of the pessimistic estimator discussed in Section 5.3.

LEMMA 4.1. *Let $p\in\{0,\varepsilon,1\}^{m\times d}$ satisfy* (4.8)*, and let $q'\in[s]$, $j'\in[d]$ be indices so that $p_{rj'}=\varepsilon$ for $r\in B_{q'}$. Then there exists a $p'\in\{0,\varepsilon,1\}^{m\times d}$, agreeing with p outside the block (q',j') but integral on the block (q',j'), for which $\Psi(p')\le\Psi(p)$.*

Proof. We first notice that for $\boldsymbol{\delta}\sim\mathcal{M}(p)$, and for any $q\in[s]$, $j\neq k$, the indicator for the event that there is a "collision" between the blocks (q,j) and (q,k), namely

$$\boldsymbol{c}_{jk,q} := \sum_{r\in B_q}\boldsymbol{\delta}_{rj}\boldsymbol{\delta}_{rk}$$

is distributed as a Bernoulli random variable with parameter $\rho_{jk,q} := \sum_{r\in B_q}p_{rj}p_{rk}$. This only holds because $p\in\{0,\varepsilon,1\}^{m\times d}$; either both blocks are fixed, in which case the outcome is deterministic ($\rho_{jk,q}\in\{0,1\}$); or at least one is uniform (possibly both), in which case collision occurs with probability ε, which equals $\rho_{jk,q}$. So

$$\mathbb{E}\big[2^{\boldsymbol{c}_{jk,q}}\big]=(1-\rho_{jk,q})+2\rho_{jk,q}=1+\rho_{jk,q}\,.$$

Hence we may simplify the expression for Ψ:

$$\Psi(p)=2^{-s^2t}\sum_{v\in V}\prod_{q=1}^{s}\prod_{j<k}(1+\rho_{jk,q})^{v_j^2v_k^2}\,. \tag{4.9}$$

We now consider $f:[0,1]^{B_{q'}}\to\mathbb{R}$ defined by

$$f(z)=2^{-s^2t}\sum_{v\in V}\prod_{(j,k,q)\in\mathcal{Q}}(1+\rho_{jk,q})^{v_j^2v_k^2}\cdot\prod_{k\neq j'}\Big(1+\sum_{r\in B_{q'}}z_r\cdot p_{rk}\Big)^{v_{j'}^2v_k^2};$$

here, $\mathcal{Q}$ consists of all triplets (j,k,q) with $q\in[s]$, $j<k\in[d]$ except for those where $q=q'$ and $j'\in\{j,k\}$. Let $z_0=(\varepsilon,\varepsilon,\dots,\varepsilon)\in[0,1]^{B_{q'}}$; then $f(z_0)=\Psi(p)$.

Since $\sum_{k\neq j'}v_{j'}^2v_k^2=v_{j'}^2(1-v_{j'}^2)\le 1/4$, each term of f is the composition of an affine function with a function of the form $y\to\prod_i y_i^{\alpha_i}$ with $\sum_i\alpha_i\le 1$. Hence f is concave. Since the extreme points of the set $S=\{z\in[0,1]^{B_{q'}}:\sum_{r\in B_{q'}}z_r=1\}$ are integral, it follows that there is a minimizer z' of f over S which is integral. Choosing p' to agree with z' on block (q',j'), we have

$$\Psi(p')=f(z')\le f(z_0)=\Psi(p).$$

□

A universal masking. So far, our arguments have not required the condition $\mathrm{NNZ}(V)\ge d^2$ (or even the mild $n\ge d$). Unfortunately, we do not know how to compute, in general, the pessimistic estimator just discussed quickly enough to obtain the linear running-time required for Theorem 1.1: we are able to compute it in time $O(m\cdot\mathrm{NNZ}(V))$, rather than the $O(s\cdot\mathrm{NNZ}(V))$ time promised in Theorem 1.1.

We will now show that as long as $n\ge d$, we can construct a good mask matrix without even examining V. It will turn out that we can compute this matrix much faster, in time $O(sd^2)$. If $\mathrm{NNZ}(V)=\Omega(d^2)$, this is comparable to the time required to apply the final sparse projection matrix to V.

The idea, first proposed by Kane and Nelson in [21] as the code construction, is the following: Suppose $\Delta\in\{0,1\}^{m\times d}$ is a matrix, with precisely s nonzero entries per column, satisfying that for all $j\neq k$,

$$\frac{1}{s^2}\sum_r\delta_{rj}\delta_{rk}\le\frac{2}{m}. \tag{4.10}$$

Then for an arbitrary input unit vector $v\in\mathbb{R}^d$

$$\|A_v\|_F^2=\frac{1}{s^2}\sum_{j\neq k}\sum_{r\in[m]}\delta_{rj}\delta_{rk}v_j^2v_k^2\le\frac{2}{m}(\|v\|_2^2-\|v\|_4^4)\le\frac{2}{m}.$$

Let $(e_j)_{j\in[d]}$ denote the canonical basis for $\mathbb{R}^d$. Then $\|A_{e_j+e_k}\|_F^2$ is exactly given by (4.10), and so it follows that we can reduce the task of controlling the Frobenius norms of A_v for $v\in V$ to controlling instead the Frobenius norms of A_v for all v in the particular set

$$\bar{V} := \{e_j+e_k : j\neq k\in[d]\}.$$

The vectors of $\bar{V}$ are not unit vectors, but it is easy to confirm that Lemma 4.1 still holds, since for any $v\in\bar{V}$ and $j'\in[d]$,

$$\sum_{k\neq j'}v_{j'}^2v_k^2\le 1.$$

(Alternatively, we could consider the set of unit vectors $\{(e_j + e_k)/\sqrt{2} : j \neq k \in [d]\}$; this would require loosening the constants in Definition 1 and Theorem 1.1.)

We may thus apply the derandomization via pessimistic estimators described above to $\bar{V}$ to obtain a good mask matrix. This can be done as long as $\Psi(\emptyset) < 1$ when considering the pessimistic estimator for $\bar{V}$ rather than V. Considering (4.7), it suffices that $d^2 \exp(-\varepsilon^2 m/7) < 1$, which does indeed hold since $\varepsilon^2 m \geq 14\ln(2n) > 14\ln(d)$. The resulting sparsity pattern is universal: it does not depend on V.

The main advantage of this approach is that the extreme sparsity of the vectors in $\bar{V}$ make the pessimistic estimator computations easier. In Section 5.3, we show how a good universal mask matrix can be computed in time $O(s \cdot d^2)$.

5 Efficient computation of the pessimistic estimators

5.1 Computation for Hanson-Wright in general. Here we show how to compute the signing pessimistic estimator (3.6) in polynomial time. The method we describe here is not yet fast enough for use in the signing step for sparse J-L to obtain the running times claimed in Theorem 1.1.

We remind the reader of the setting: Given $r = 0, \ldots, k$, we consider a partial assignment vector $\sigma \in \{-1, +1\}^r$ (with the convention $r = 0$ means $\sigma = \emptyset$), and a distribution $\mathcal{G}(\sigma)$ over $\mathbb{R}^k$ whose i-th coordinate is deterministically σ_i for $i \leq r$, and an independent Gaussian for $i > r$. We are interested in computing the mgf of the quadratic form $\boldsymbol{g}^\top A \boldsymbol{g}$, where $\boldsymbol{g} \sim \mathcal{G}(\sigma)$.

The following lemma shows how we can evaluate the pessimistic estimator (3.6).

LEMMA 5.1. *Let $t \in \{0, 1, \ldots, k\}$ and $\sigma \in \{-1, 1\}^t$. Let M be a symmetric $k \times k$ matrix with $\|M\|_{\text{op}} < 1/2$. Let $S = \{1, \ldots, t\}$ and $T = \{t+1, \ldots, k\}$. Then*

$$\mathbb{E}_{\boldsymbol{g} \sim \mathcal{G}(\sigma)}\, e^{\boldsymbol{g}^\mathsf{T} M \boldsymbol{g}} = \det(I - 2M_{T,T})^{-1/2} \cdot \\ \exp\bigl\{2\sigma^\mathsf{T} M_{S,T}(I - 2M_{T,T})^{-1} M_{T,S}\sigma + \sigma^\mathsf{T} M_{S,S}\sigma\bigr\}.$$

Proof. Let $\ell = k - t$. Consider the spectral decomposition $M_{T,T} = \sum_{i=1}^{\ell} \lambda_i u_i u_i^\mathsf{T}$, where $u_1, \ldots, u_\ell$ form an orthonormal basis of $\mathbb{R}^\ell$, and the eigenvalues are in decreasing absolute value, $1/2 > |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_\ell|$. Define

$$b_i := \langle M_{T,S}\sigma, u_i \rangle, \qquad \text{for } i \in [\ell].$$

Also let $\boldsymbol{z}_i := \langle \boldsymbol{g}_T, u_i \rangle$ for $i \in [\ell]$; note that this is a standard Gaussian, and that $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are independent for $i \neq j$.

Then

$$\begin{aligned}
&\mathbb{E}\, e^{\boldsymbol{g}^\mathsf{T} M \boldsymbol{g}} \\
&\quad = \mathbb{E}\bigl[\exp\{\boldsymbol{g}_T^\mathsf{T} M_{T,T}\boldsymbol{g}_T + 2\boldsymbol{g}_T^\mathsf{T} M_{T,S}\sigma + \sigma^\mathsf{T} M_{S,S}\sigma\}\bigr] \\
&\quad = \mathbb{E}\Bigl[\exp\Bigl\{\textstyle\sum_{i=1}^{\ell}(\lambda_i\langle \boldsymbol{g}_T, u_i\rangle^2 + 2\langle \boldsymbol{g}_T, u_i\rangle\langle M_{S,T}\sigma, u_i\rangle) \\
&\qquad\quad + \sigma^\mathsf{T} M_{S,S}\sigma\Bigr\}\Bigr] \qquad \text{by orthogonality} \\
&\quad = \prod_{i=1}^{\ell} \mathbb{E}\bigl[e^{\lambda_i \boldsymbol{z}_i^2 + 2\boldsymbol{z}_i b_i}\bigr] \cdot e^{\sigma^\mathsf{T} M_{S,S}\sigma} \\
&\quad = \prod_{i=1}^{\ell} \frac{1}{\sqrt{1 - 2\lambda_i}} e^{2b_i^2/(1-2\lambda_i)} \cdot e^{\sigma^\mathsf{T} M_{S,S}\sigma} \qquad \text{by Lemma 2.4} \\
&\quad = \det(I - 2M_{T,T})^{-1/2} \cdot \\
&\qquad\quad \exp\{2\sigma^\mathsf{T} M_{S,T}(I - 2M_{T,T})^{-1} M_{T,S}\sigma\} \cdot e^{\sigma^\mathsf{T} M_{S,S}\sigma}.
\end{aligned}$$

□

5.2 Computation for signing. The approach above is too slow to obtain the running time claimed in Theorem 1.1. In particular, the best known algorithms for computing the determinant run in matrix-vector multiplication time, which is superlinear. We will next show how to exploit the simple structure of the quadratic forms of interest in the case of dimensionality reduction to obtain a linear-time algorithm for initializing and updating the pessimistic estimator. We will omit some implementation details needed to obtain the claimed running time; we refer the interested reader to the full version of the paper.

Both the input vectors V and the mask matrix Δ are assumed to be given in a sparse representation, as a list of the locations and values of the nonzero entries. We will do the computations independently for each ψ_v^+ and ψ_v^-; so we fix a $v \in V$ for the remainder, and also focus only on ψ_v^+ (ψ_v^- is analogous). To obtain an overall running time of $O(s \cdot \mathrm{NNZ}(V))$, we can afford time $O(s \cdot \mathrm{NNZ}(v))$ in computing ψ_v^+.

Recall that $A_v = \frac{1}{s}\bigoplus_{r \in [m]}[B_{r,v} - \mathrm{Diag}(B_{r,v})]$, with $B_{r,v} = (\delta_r \odot v)(\delta_r \odot v)^\mathsf{T}$. Given a partial fixing $\sigma \in \{-1, 1\}^t$ for some $t \leq md$, let $\sigma[r]$ denote the components of σ that correspond with the r'th block $B_{r,v}$ of A_v; it exactly corresponds to the entries of σ that end up in the r'th row of the resulting matrix Σ. Thus $\sigma[r]$ has length 0 if $t \leq (r-1)d$, length d if $t \geq rd$, and length $t - (r-1)d$ otherwise.

Our goal is to efficiently compute

$$\begin{aligned}
\psi_v^+(\sigma) &= \mathbb{E}_{\boldsymbol{g} \sim \mathcal{G}(\sigma)} e^{\lambda \boldsymbol{g}^\mathsf{T} A_v \boldsymbol{g}} \\
&= \prod_{r=1}^{m} \mathbb{E}_{\boldsymbol{g} \sim \mathcal{G}(\sigma[r])} e^{\lambda \boldsymbol{g}^\mathsf{T}[B_{r,v} - \mathrm{Diag}(B_{r,v})]\boldsymbol{g}}. \qquad (5.1)
\end{aligned}$$

An evaluation of each term in this product can be done with the aid of Lemma 5.1. But we may exploit the rank-1 structure of $B_{r,v}$, to speed up the calculations, as captured in the following lemma.

LEMMA 5.2. *Let $u \in \mathbb{R}^k$ a vector given in sparse representation such that $\|u\|_2^2 < 1/2$, and let $M = uu^\mathsf{T} - \mathrm{Diag}(uu^\mathsf{T})$;*

furthermore let $S \subseteq [k]$ and $T = [k] \setminus S$ be a partition of the coordinates, and let $\sigma \in \mathbb{R}^S$.

Then defining $\Theta := \sum_{j\in T} u_j^2/(1+2u_j^2)$ and $\nu := \sum_{j\in S} u_j \sigma_j$, we have

$$\text{(5.2)}\quad \mathbb{E}_{\boldsymbol{g}\sim\mathcal{G}(\sigma)}\, e^{\boldsymbol{g}^\mathsf{T} M \boldsymbol{g}} = \Big((1-2\Theta)\prod_{j\in T}(1+2u_j^2)\Big)^{-1/2} \cdot \exp\Big\{2\nu^2\cdot\Theta/(1-2\Theta) + \nu^2 - \sum_{j\in S} u_j^2\Big\}.$$

We delay the proof to the end of this section.

Computing $\psi_v^+(\emptyset)$. Define, for any $r \in [m]$ and $S \subseteq [d]$,

$$\Theta_r(\sigma) := \tfrac{\lambda}{s} \sum_{j\in[d]\setminus \operatorname{supp}(\sigma[r])} \delta_{rj} v_j^2/(1+\tfrac{\lambda}{s}\delta_{rj}v_j^2).$$

We will initially be interested in $\Theta_r(\emptyset)$ for each $r \in [m]$. Since m could be much larger than $O(s \cdot \mathrm{NNZ}(v))$, we do not necessarily have time to even write down all of these numbers; but it is easy to see that at most $s \cdot \mathrm{NNZ}(v)$ of the $\Theta_r(\emptyset)$'s can differ from 0, so this is not a problem with careful implementation. Determining all nonzero $\Theta_r(\emptyset)$'s and their values can be done in time $O(s \cdot \mathrm{NNZ}(v))$. Applying Lemma 5.2, we obtain

$$\begin{aligned}\psi_v^+(\emptyset) &= \prod_{r=1}^{m}\left(\big(1-2\Theta_r(\emptyset)\big)\prod_{j=1}^{d}(1+2\tfrac{\lambda}{s}\delta_{rj}v_j^2)\right)^{-1/2}\\ &= \prod_{r=1}^{m}\left(\big(1-2\Theta_r(\emptyset)\big)\prod_{j\in[d]:\delta_{rj}=1}(1+2\tfrac{\lambda}{s}v_j^2)\right)^{-1/2}.\end{aligned}$$

Using the sparse representations of v and Δ we can aggregate the necessary product terms in one pass through the data, and hence compute this quantity in time $O(s \cdot \mathrm{NNZ}(v))$.

Updating ψ_v^+. Starting from $\psi_v^+(\emptyset)$, we will gradually update its value as we increase the length of our fixed sign vector σ. In order to do this, we will keep track of not just the value of $\psi_v^+(\sigma)$, but also the values of $\Theta_r(\sigma)$, and also the value of $\nu_r(\sigma) := \sum_{j\in\operatorname{supp}(\sigma[r])}\delta_{rj}v_j\sigma_j$. Again, there are too many values of $\nu_r(\sigma)$ to write down, but initially they are all zero, and all but at most $s \cdot \mathrm{NNZ}(v)$ will remain zero.

Given these values for our current choice of σ, we need to evaluate $\psi_v^+(\sigma^+)$ (as well as $\psi_v^+(\sigma^-)$, which will be completely analogous). Let $t = (r-1)d + \ell \in [md]$ be the index of the new entry of σ^+. Then only the term in the product (5.1) corresponding to r can be affected. We observe that $\nu_s(\sigma^+) = \nu_s(\sigma)$ and $\Theta_s(\sigma^+) = \Theta_s(\sigma)$ for all $s \neq r$. Moreover,

$$\begin{aligned}\nu_r(\sigma^+) &= \nu_r(\sigma) + \delta_{r\ell}v_\ell\\ \Theta_r(\sigma^+) &= \Theta_r(\sigma) - \tfrac{\lambda}{s}\delta_{r\ell}v_\ell^2/(1+\tfrac{\lambda}{s}\delta_{r\ell}v_\ell^2).\end{aligned}$$

We can thus update these values, and hence by (5.2) ψ_v^+, in constant time. Since we can only afford $O(s \cdot \mathrm{NNZ}(v))$ time on updating ψ_v^+ until a complete sign choice has been made, some care must be taken. But the update described is nontrivial only if both $\delta_{r\ell}$ and v_ℓ are both nonzero. Considering then one of the $\mathrm{NNZ}(v)$ nonzero choices of v_j, it will be involved in only s nontrivial updates (the ones where $\delta_{rj} = 1$), and so there are only $s \cdot \mathrm{NNZ}(v))$ nontrivial updates to ψ_v^+ in total.

Proof of Lemma 5.2. Let us first note that $\|M\|_{\mathrm{op}} < 1/2$, as required by Lemma 5.1. Since both uu^T and $\operatorname{Diag}(uu^\mathsf{T})$ are PSD matrices, we can upper bound the operator norm by the maximum of the respective operator norms:

$$\begin{aligned}\|M\|_{\mathrm{op}} &\le \max\{\|uu^\mathsf{T}\|_{\mathrm{op}}, \|\operatorname{Diag}(uu^\mathsf{T})\|_{\mathrm{op}}\}\\ &\le \max\{\|u\|_2^2, \max\{u_j^2 : j\in[k]\}\}\\ &< 1/2,\end{aligned}$$

as required.

Using Lemma 5.1, we have that

$$\text{(5.3)}\quad \mathbb{E}_{\boldsymbol{g}\sim\mathcal{G}(\sigma)}\, e^{\boldsymbol{g}^\mathsf{T} M\boldsymbol{g}} = \det(I-2M_{T,T})^{-1/2}\cdot \exp\big\{2\sigma^\mathsf{T} M_{S,T}(I-2M_{T,T})^{-1}M_{T,S}\sigma + \sigma^\mathsf{T} M_{S,S}\sigma\big\}.$$

Let $D := \operatorname{diag}(u_T u_T^\mathsf{T}) \in \mathbb{R}^{T\times T}$. By the matrix determinant lemma,

$$\begin{aligned}\det(I-2M_{T,T}) &= \det(I+2D)\cdot(1-2u_T^\mathsf{T}(I+2D)^{-1}u_T)\\ &= \prod_{j\in T}(1+2u_j^2)\cdot\Big(1-2\sum_{j\in T}\frac{u_j^2}{1+2u_j^2}\Big)\\ &= \prod_{j\in T}(1+2u_j^2)\cdot(1-2\Theta).\end{aligned}$$

By the Sherman-Morrison formula,

$$(I-2M_{T,T})^{-1} = (I+2D)^{-1} + 2\frac{(I+2D)^{-1}u_T u_T^\mathsf{T}(I+2D)^{-1}}{1-2u_T^\mathsf{T}(I+2D)^{-1}u_T}.$$

On the other hand, since $S\cap T = \emptyset$, $M_{S,T} = u_S u_T^\mathsf{T}$. Thus,

$$\begin{aligned}&\sigma^\mathsf{T} M_{S,T}(I-2M_{T,T})^{-1}M_{T,S}\sigma\\ &\quad= \langle u_S,\sigma\rangle^2\cdot u_T^\mathsf{T}(I-2M_{T,T})^{-1}u_T\\ &\quad= \langle u_S,\sigma\rangle^2\cdot\Big(u_T^\mathsf{T}(I+2D)^{-1}u_T\\ &\qquad\quad + 2\frac{(u_T^\mathsf{T}(I+2D)^{-1}u_T)^2}{1-2u_T^\mathsf{T}(I+2D)^{-1}u_T}\Big)\\ &\quad= \langle u_S,\sigma\rangle^2\cdot\Big(\frac{u_T^\mathsf{T}(I+2D)^{-1}u_T}{1-2u_T^\mathsf{T}(I+2D)^{-1}u_T}\Big)\\ &\quad= \nu^2\Theta/(1-2\Theta).\end{aligned}$$

Finally, we have

$$\begin{aligned}\sigma^\mathsf{T} M_{S,S}\sigma &= -\|D\sigma_S\|_2^2 + \langle u_S,\sigma\rangle^2\\ &= -\sum_{j\in S}u_j^2 + \Big(\sum_{j\in S}u_j\sigma_j\Big)^2\\ &= -\sum_{j\in S}u_j^2 + \nu^2.\end{aligned}$$

Replacing the relevant terms in (5.3) yields the lemma. □

5.3 Computation for masking. Here we provide the details of the efficient computation of a universal mask matrix associated with the vectors $\bar{V} = \{e_j + e_k : j < k\}$, in time $O(s \cdot d^2)$. We are also able to compute the mask matrix for an arbitrary set of vectors V, but with the inferior running time $O(m(\cdot \mathrm{NNZ}(V) + d))$. We postpone this to the full version of the paper.

In the case of $\bar{V}$, the formula (4.9) for the pessimistic estimator simplifies to

$$\Psi(p) = 2^{-s^2 t} \cdot \sum_{j<k} \prod_{q=1}^{s} (1 + \rho_{jk,q}),$$

where recall $\rho_{jk,q} := \sum_{r \in B_q} p_{rj} p_{rk}$. We consider a partial assignment vector p: namely, there exists a block (q, ℓ) such that p is integral for all blocks $(\tilde{q}, j) \prec (q, \ell)$, and p is uniform on all blocks $(\tilde{q}, j) \succeq (q, \ell)$. We are interested in derandomizing block (q, ℓ).

Given $r \in B_q$, let $p^{(r)}$ be a vector identical to p in all blocks, except for (q, ℓ), where it has an integral assignment choosing $p_{r\ell} = 1$. Then

$$\Psi(p^{(r)}) - \Psi(p) = 2^{-s^2 t} \sum_{j<\ell} \prod_{\tilde{q} \neq q} (1 + \rho_{jk,\tilde{q}}) \Big(1 + p_{rj} - (1+\varepsilon)\Big).$$

Now let $D^q_{j\ell} := \prod_{\tilde{q}<q} (1 + \rho_{j\ell,\tilde{q}})$. Noticing that

$$\prod_{\tilde{q} \neq q} (1 + \rho_{jk,\tilde{q}}) = D^q_{j\ell} \cdot (1+\varepsilon)^{s-q},$$

we obtain

$$\Psi(p^{(r)}) - \Psi(p) = 2^{-s^2 t} \big(1+\varepsilon\big)^{s-q} \Big[\sum_{j<\ell} D^q_{j\ell} \cdot p_{rj} - \varepsilon \sum_{j<\ell} D^q_{j\ell}\Big].$$

Since the last term in this equation is independent of r, we conclude that in order to minimize the pessimistic estimator it suffices to choose $r \in B_q$ to minimize

$$\alpha^{(r)}_\ell := \sum_{j<\ell} D^q_{j\ell} \cdot p_{rj}.$$

With this, the algorithm is straightforward. We proceed block by block (q, ℓ) in the $\prec$ order, we compute at each step $\alpha^{(r)}_\ell$ for all $r \in B_q$, and choose the minimizer r, assigning $p_{r\ell} = 1$. Notice that at each step we compute $1/\varepsilon = O(d)$ values of α, and each of these can be computed in time $O(1)$ by keeping track of the value of $D^q_{j(\ell-1)}$. Moreover, we can compute $D^q_{j\ell}$ from $D^{q-1}_{j\ell}$ by looking at the position $r \in B_{q-1}$ where $p_{rj} = 1$ and comparing it with $p_{r\ell}$: if the latter is one, then $D^q_{j\ell} = 2D^{q-1}_{j\ell}$; otherwise, $D^q_{j\ell} = D^{q-1}_{j\ell}$.

As a conclusion, we are performing $s \cdot d$ steps with $O(d)$ computation in each step; therefore, the running time of the algorithm is $O(s \cdot d^2)$ (which is $O(s \cdot \mathrm{NNZ}(V))$ given the assumption that $\mathrm{NNZ}(V) = \Omega(d^2)$).

6 A Sharp Hanson-Wright Inequality

In the following sections, we prove the technical estimate from Lemma 3.3 and give applications of our functional form of the Hanson-Wright inequality to get an essentially optimal Hanson-Wright constant and to get improved concentration bounds for Rademacher projections.

6.1 The Determinant Lower Bound. In this section, we give the proof of determinant lower bound in Lemma 3.3. To begin, we show that this problem is in fact an eigenvalue optimization problem.

LEMMA 6.1. *For $k \in \mathbb{N}$, $0 \leq \alpha < 1$, $\beta \geq 1$,*

$$\begin{aligned} &\min\Big\{\det(I - M) : M \in \mathbb{R}^{k \times k} \text{ symmetric}, \\ &\qquad M_{ii} = 0 \; \forall i \in [k], \|M\|_{\mathrm{op}} \leq \alpha, \|M\|_F \leq \beta\Big\} \\ &= \min\Big\{\prod_{i=1}^{k} (1 - \lambda_i) : |\lambda_i| \leq \alpha, \forall i \in [k], \\ &\qquad \sum_{i=1}^{k} \lambda_i^2 \leq \beta^2, \sum_{i=1}^{k} \lambda_i = 0\Big\}. \end{aligned}$$

Proof. Firstly, we note that the minimum is clearly achieved for both problems since the feasible regions are compact and the objective functions are continuous.

We now show that a solution for one problem can be converted to a solution for the other of same value. Starting with a $k \times k$ diagonal 0 matrix M as above, we claim that the eigenvalues $\lambda_1, \ldots, \lambda_k$ form a solution for the other side. This follows since $\sum_{i=1}^k \lambda_i = \mathrm{trace}(M) = \sum_{i=1}^k M_{ii} = 0$, $\sum_{i=1}^k \lambda_i^2 = \|M\|_F^2 \leq \beta$, $\max_{i \in [k]} |\lambda_i| = \|M\|_{\mathrm{op}}$ and $\det(I - M) = \prod_{i=1}^k (1 - \lambda_i)$.

For the other direction, starting from $\lambda_1, \ldots, \lambda_k$ and given the above identities, we must only show that there exists a symmetric diagonal 0 matrix $M \in \mathbb{R}^{k \times k}$ having $\lambda_1, \ldots, \lambda_k$ as its eigenvalues. By the Schur-Horn theorem [16, Theorem 5], the set of achievable diagonals for the set of symmetric matrices with eigenvalues $\lambda_1, \ldots, \lambda_k$ corresponds exactly to the permutahedron

$$\text{convex.hull}((\lambda_{\pi[1]}, \ldots, \lambda_{\pi[k]}) : \pi : [k] \to [k] \text{ a permutation}).$$

Since $\sum_{i=1}^k \lambda_i = 0$, by averaging over all permutations, we see that the vector $(0, \ldots, 0)$ is an achievable diagonal, as needed. □

Though the most natural approach would be to find the exact minimizer for any fixed dimension k, we are currently only able to get a sharp bound for the minimum value when we allow $k \to \infty$. This allows us to use a natural continuous relaxation, which is sharp when α^2 evenly divides β^2, where we can show that the optimal solution is supported on at most 2 distinct eigenvalues. This reduces the problem to one on two variables, which we can solve directly.

Proof of Lemma 3.3. By Lemma 6.1 and taking logs, the problem reduces to showing that

$$\text{(6.1)}\quad \inf\Big\{\sum_{i=1}^{k}\ln(1-\lambda_i):\ k\in\mathbb{N}, |x_i|\le\alpha\ \ \forall i\in[k],\ \sum_{i=1}^{k}\lambda_i=0,\ \sum_{i=1}^{k}\lambda_i^2\le\beta^2\ \ \forall i\in[k]\Big\} \ge (\beta^2/\alpha^2)\ln(1-\alpha)+\beta^2/\alpha.$$

Since we can clearly drop any zero λ_is without affecting the constraints or objective, we may assume that all the λ_is are non-zero. By modelling multiplicities and allowing fractionality, we see that program (6.1) has value at least

$$\text{(6.2)}\quad \inf\Big\{\sum_{i=1}^{k} l_i\ln(1-\lambda_i) : k\in\mathbb{N}, l_i\ge 0\ \forall i\in[k],\ \lambda_1,\dots,\lambda_k \text{ distinct non-zero}, |\lambda_i|\le\alpha\ \forall i\in[k],\ \sum_{i=1}^{k}\lambda_i=0,\ \sum_{i=1}^{k}\lambda_i^2\le\beta^2\Big\}.$$

For the above relaxed formulation, we claim that we can assume that $k=2$. Since the eigenvalue sum is 0, we need at least one positive and one negative eigenvalue, so we must clearly have $k\ge 2$. To show $k\le 2$, fix any distinct non-zero $(\lambda_1,\dots,\lambda_k)$ which can be extended to a feasible solution. Note that this reduces the problem to a linear program over $l_1,\dots,l_k$. Given that the LP is feasible by assumption and bounded (since all the $\lambda_i\ne 0$), the optimal value can be obtained at a basic feasible solution. Since there are only 2 non-trivial constraints on $l_1,\dots,l_k$ apart from non-negativity, any basic feasible solution can have support on at most two of $l_1,\dots,l_k$, as needed.

Given the above, program (6.2) reduces to

$$\text{(6.3)}\quad \inf\big\{l_1\ln(1-\lambda_1)+l_2\ln(1-\lambda_2):\ \lambda_1\in(0,\alpha),\ \lambda_2\in(-\alpha,0),\ l_1,\ l_2\ge 0,\ l_1\lambda_1+l_2\lambda_2=0,\ l_1\lambda_1^2+l_2\lambda_2^2\le\beta^2\big\}.$$

Since $l_1\ln(1-\lambda_1)+l_2\ln(1-\lambda_2)\le -l_1\lambda_1-l_2\lambda_2=0$, we can always achieve $l_1\lambda_1^2+l_2\lambda_2^2=\beta^2$ by scaling without increasing the objective value. Given this, for feasible λ_1,λ_2 above, it is easy to check that setting

$$l_1=\frac{\beta^2}{\lambda_1(\lambda_1-\lambda_2)}\quad\text{and}\quad l_2=\frac{\beta^2}{\lambda_2(\lambda_2-\lambda_1)}$$

is both feasible and optimal. Thus, the program (6.3) reduces further to

$$\text{(6.4)}\quad \beta^2\inf\Big\{\frac{1}{\lambda_1-\lambda_2}\Big(\frac{\ln(1-\lambda_1)}{\lambda_1}-\frac{\ln(1-\lambda_2)}{\lambda_2}\Big):\ \lambda_1\in(0,\alpha),\lambda_2\in(-\alpha,0)\Big\}.$$

The following claim shows that the objective function is monotonically decreasing in λ_1 and λ_2 over its domain.

CLAIM 1. *The function*

$$g(x,y):=\frac{1}{x-y}\left(\frac{\ln(1-x)}{x}-\frac{\ln(1-y)}{y}\right)$$

is decreasing in x and y on the domain $(0,1)\times(-1,0)$.

Proof. We will show the integral description

$$g(x,y)=-\int_0^1 z\cdot\frac{1}{1-zx}\cdot\frac{1}{1-zy}dz.$$

The claim is then immediate, since for any fixed $z\in[0,1]$, the integrand is clearly increasing on the domain.

Expanding $\ln(1-x)$ and $\ln(1-y)$, we have

$$\begin{aligned} g(x,y)&=\frac{1}{x-y}\sum_{k\ge 0}\frac{(-x^k+y^k)}{k+1}\\ &=-\sum_{k\ge 0}\frac{1}{k+1}\sum_{i=0}^{k-1}x^iy^{k-i-1}\\ &=-\sum_{i\ge 0}\sum_{k\ge i+1}\frac{1}{k+1}x^iy^{k-i-1}\\ &=-\sum_{i\ge 0}\sum_{j\ge 0}\frac{1}{i+j+2}x^iy^j. \end{aligned}$$

Now we use the identity $\frac{1}{i+j+2}=\int_0^1 z^{i+j+1}dz$. We obtain (using the dominated convergence theorem to justify the exchange of integral and summation)

$$\begin{aligned} g(x,y)&=-\int_0^1 z\cdot\sum_{i\ge 0}\sum_{j\ge 0}(zx)^i(zy)^jdz\\ &=-\int_0^1 z\cdot\frac{1}{1-zx}\frac{1}{1-zy}dz, \end{aligned}$$

as required. □

Given the claim, the infimum of (6.4) is thus obtained by the limit

$$\text{(6.5)}\quad \lim_{\lambda_1\to\alpha,\lambda_2\to 0}\frac{\beta^2}{\lambda_1-\lambda_2}\Big(\frac{\ln(1-\lambda_1)}{\lambda_1}-\frac{\ln(1-\lambda_2)}{\lambda_2}\Big) =\frac{\beta^2}{\alpha^2}\ln(1-\alpha)+\beta^2/\alpha,$$

as needed.

We now show that this bound is sharp for the original program (6.1) as long as α^2 divides β^2. In particular, it suffices to construct a sequence of solutions $\lambda_{1,t},\lambda_{2,t}$ to (6.4) such that $\lambda_{1,t}\to\alpha$ and $\lambda_{2,t}\to 0$ as $t\to\infty$, for which the corresponding multiplicities $l_{1,t}=\frac{\beta^2}{\lambda_{1,t}(\lambda_{1,t}-\lambda_{2,t})}$ and $l_{2,t}=\frac{\beta^2}{\lambda_{2,t}(\lambda_{2,t}-\lambda_{1,t})}$ are positive integers. For this purpose, take the sequence $\lambda_{1,t}=\alpha\sqrt{\frac{t}{t+1}}$ and $\lambda_{2,t}=-\alpha\sqrt{\frac{t}{t+1}}\frac{1}{t}$. A direct computation reveals that $l_{1,t}=\frac{\beta^2}{\alpha^2}$ and $l_{2,t}=\frac{\beta^2}{\alpha^2}t$, which are integral for $t\in\mathbb{N}$ by the assumption that α^2 divides β^2. Since this sequence clearly converges to the desired limit, this proves the claim. □

6.2 The Hanson-Wright Constant. The following lemma uses the functional form of Theorem 3.2 to derive optimal tradeoffs between the linear and quadratic behavior of the Hanson Wright inequality.

LEMMA 6.2. *Under the conditions of Theorem 3.1, for any $\gamma > 0$, we have that*

$$\text{(6.6)}\quad \mathbb{P}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > t\nu^2) \leq \exp\left(-\tfrac{1-\ln(1+\gamma)/\gamma}{2\gamma} \min\left\{\frac{t^2}{\|A\|_F^2}, \frac{\gamma t}{\|A\|_{\text{op}}}\right\}\right).$$

Proof. Letting $\alpha := \|A\|_{\text{op}}$, $\beta := \|A\|_F$, by Theorem 3.2 we recall that

$$\mathbb{P}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} > t\nu^2) \leq e^{-\frac{1}{2}f(t)} \text{ where } f(t) = \frac{t}{\alpha} - \frac{\beta^2}{\alpha^2}\ln(1+t\frac{\alpha}{\beta^2}).$$

To prove the lemma, it thus suffices to show that

$$\text{(6.7)}\qquad f(t) \geq \frac{1-\ln(1+\gamma)/\gamma}{\gamma} \min\left\{\frac{t^2}{\beta^2}, \frac{\gamma t}{\alpha}\right\}.$$

For $t = \gamma\beta^2/\alpha$, equality is attained between $\frac{t^2}{\beta^2}$ and $\frac{\gamma t}{\alpha}$, and it can be verified that they both equal $f(t)$. All that remains is to show that $\frac{\alpha}{\gamma t} f(t)$ is increasing and $\frac{\beta^2}{t^2} f(t)$ is decreasing. This can be done by direct computation of derivatives. □

Using the above lemma, we derive the optimal Hanson-Wright constant C_{HW} for Theorem 3.1. In particular, this corresponds to setting $\gamma = 1$ above, which gives $C_{\text{HW}} = \frac{1-\ln 2}{2} \approx 0.153 > 3/20$.

As for the different tradeoffs, letting $\gamma \to \infty$ improves the bound for small t (the quadratic regime), where the constant in front the $t^2/\|A\|_F^2$ terms converges to $1/4$. Letting $\gamma \to 0$ improves the bound for large t (the linear regime), where the constant in front of $t/\|A\|_{\text{op}}$ converges to $1/2$.

6.3 Tighter Concentration for Rademacher Projections. In this section, we prove Lemma 3.1, which gives sparsity dependent concentration bound for Rademacher projections.

We begin with the following lemma, which shows that our functional form of Hanson-Wright has good monotonicity properties.

LEMMA 6.3. *The probability bound in Theorem 3.2 is monotone increasing in $\|A\|_{\text{op}}$ and $\|A\|_F$.*

Proof. Letting $\alpha := \|A\|_{\text{op}}$ and $\beta := \|A\|_F$, we must show that

$$\exp\left(-\tfrac{1}{2}\left(\tfrac{t}{\alpha} - \tfrac{\beta^2}{\alpha^2}\cdot\ln(1+t\tfrac{\alpha}{\beta^2})\right)\right)$$

is monotone increasing in α and β. Equivalently, it suffices to show that the function $\frac{t}{\alpha} - \frac{\beta^2}{\alpha^2}\ln(1+t\frac{\alpha}{\beta^2})$ in the exponent is monotone decreasing in α and β. To prove this for α, replacing t by $t\beta^2$, this is equivalent to showing that $\frac{t}{\alpha} - \frac{1}{\alpha^2}\ln(1+t\alpha)$ is decreasing. Taking a derivative with respect to α, we require that

$$-\frac{t}{\alpha^2} + 2\frac{1}{\alpha^3}\ln(1+t\alpha) - \frac{t}{\alpha^2(1+t\alpha)} \leq 0,$$

or equivalently,

$$\ln(1+t\alpha) \leq \frac{t\alpha}{2} + \frac{t\alpha}{2(1+t\alpha)},$$

which follows from Lemma 2.2 part 1 applied to $x = t\alpha$. To prove it for β, it suffices to show that $\frac{\beta^2}{\alpha}\ln(1+t\frac{\alpha}{\beta^2})$ is increasing in β. After replacing β^2 by $\beta t\alpha$, this is equivalent to showing that $\beta\ln(1+1/\beta)$ is increasing, which follows from Lemma 2.2 part 3. □

The following technical lemma bounds the matrix norms required to analyze Rademacher projections. While straightforward, the computations are a bit tedious and so we defer them to the full version of the paper.

LEMMA 6.4. *Let $v \in \mathbb{R}^d$ be a unit vector with at most s non-zero entries. Then for $H = vv^\mathsf{T} - \text{Diag}(vv^\mathsf{T})$, we have that $\|H\|_{\text{op}} \leq 1 - 1/s$ and $\|H\|_F^2 = \|v\|_2^4 - \|v\|_4^4 \leq 1 - 1/s$. Furthermore, the bounds are uniquely attained at the unit vector $v = (1/\sqrt{s}, \ldots, 1/\sqrt{s})$.*

We now prove the claimed concentration bound.

Proof of Lemma 3.1. Let H be as in Lemma 6.4 and let M denote the $(md) \times (md)$ block diagonal symmetric matrix with m blocks corresponding to H/m. Index the entries of M by $M_{(ki),(lj)}$, $k, l \in [m]$, $i, j \in [d]$, where $M_{(ki),(lj)} = 0$ if $k \neq l$ and equal to H_{ij}/m otherwise. By Lemma 6.4, we have that $\|M\|_{\text{op}} = \|H\|_{\text{op}}/m \leq (1-1/s)/m$ and $\|M\|_F^2 = \|H\|_F^2/m \leq (1-1/s)/m$.

For the Rademacher projection matrix $\Pi \in \{-1,1\}^{m\times d}$, a direct computation reveals that

$$\|\Pi v\|_2^2/m - 1 = \sum_{r\in[m]} \sum_{i,j\in[d]} M_{(r,i),(r,j)} \Pi_{r,i}\Pi_{r,j}\,.$$

Thus applying Theorem 3.2 to right hand side quadratic form, we get that

$$\begin{aligned}\max\Big\{&\mathbb{P}(\|\Pi v\|_2^2/m \geq 1+\varepsilon), \mathbb{P}(\|\Pi v\|_2^2/m \leq 1-\varepsilon)\Big\}\\ &\leq \exp\Big(-\frac{1}{2}\Big(\frac{\varepsilon}{\|M\|_{\text{op}}} - \frac{\|M\|_F^2}{\|M\|_{\text{op}}^2}\ln\big(1+\frac{\varepsilon\|M\|_{\text{op}}}{\|M\|_F^2}\big)\Big)\Big)\\ &\leq \exp\big(-\frac{m}{2}\frac{s}{s-1}(\varepsilon - \ln(1+\varepsilon))\big),\end{aligned}$$

where the last inequality follows by setting $\|M\|_{\text{op}}$ and $\|M\|_F$ to their upper bounds together with Lemma 6.3. □

References

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, June 2003.

[2] A. A. Ageev and M. Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3):307–328, 2004.

[3] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[4] N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615, 2009.

[5] N. Ailon and E. Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):21, 2013.

[6] A. Bhargava and S. R. Kosaraju. Derandomization of dimensionality reduction and SDP based algorithms. In F. Dehne, A. López-Ortiz, and J.-R. Sack, editors, *Proceedings of the 9th International Workshop on Algorithms and Data Structures (WADS)*, pages 396–408, 2005.

[7] L. Blum, M. Shub, and S. Smale. On a theory of computation and complexity over the real numbers: NP- completeness, recursive functions and universal machines. *Bull. Am. Math. Soc., New Ser.*, 21(1):1–46, 1989.

[8] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2009.

[9] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 341–350. ACM, 2010.

[10] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 634–644, 1999.

[11] P. Drineas and M. W. Mahoney. Randnla: Randomized numerical linear algebra. *Commun. ACM*, 59(6):80–90, May 2016.

[12] L. Engebretsen, P. Indyk, and R. O'Donnell. Derandomized dimensionality reduction with applications. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 705–712, 2002.

[13] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 06 1971.

[14] N. J. A. Harvey and N. Olver. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 926–945, 2014.

[15] A. Hinrichs and J. Vybíral. Johnson-Lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms*, 39(3):391–398, 2011.

[16] A. Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *Amer. J. Math.*, 76:620–630, 1954.

[17] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[18] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[19] D. M. Kane and J. Nelson. A derandomized sparse Johnson-Lindenstrauss transform. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:98, 2010.

[20] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1195–1206, 2012.

[21] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4:1–4:23, 2014.

[22] Z. S. Karnin, Y. Rabani, and A. Shpilka. Explicit dimension reduction and its applications. In *Proceedings of the 26th Annual IEEE Conference on Computational Complexity (CCC)*, pages 262–272, 2011.

[23] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[24] E. Liberty, N. Ailon, and A. Singer. Dense fast random projections and lean Walsh transforms. *Lecture Notes in Computer Science*, 5171:512–522, 2008.

[25] R. Meka. Almost optimal explicit Johnson-Lindenstrauss transformations. *arXiv preprint arXiv:1011.6397*, 2010.

[26] N. Nisan. Psuedorandom generators for space-bounded computation. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 204–212, 1990.

[27] N. Nisan. RL $\subseteq$ SC. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing (STOC)*, pages 619–623, 1992.

[28] D. Pollard. Lecture notes on empirical processes. Lecture Notes, 2015.

[29] P. Raghavan. Probabilistic construction of deterministic algorithms: Approximating packing integer programs. *Journal of Computer and System Sciences*, 37, 1988.

[30] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:9 pages., 2013.

[31] D. Sivakumar. Algorithmic derandomization via complexity theory. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 619–626. ACM, 2002.

[32] A. Srinivasan. Distributions on level-sets with applications to approximation algorithms. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 588–597, 2001.

[33] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10:1–157, 2014.

A Proofs from Section 2

Here we record for completeness the proofs of the claims in Section 2.

Proof of Lemma 2.2.

Proof of 1. Note that $\ln(1+x)$ and $\frac{x}{2}+\frac{x}{2(1+x)}$ are equal at $x=0$, so it suffices to compare derivatives. In particular, we must show that $\frac{1}{1+x} \le \frac{1}{2}+\frac{1}{2(1+x)}-\frac{x}{2(1+x)^2}$. Multiplying through by $2(1+x)^2$ and rearranging, this inequality is equivalent to $x^2 \ge 0$, which is clearly true.

Proof of 2.

$$\ln(1+x) = \ln\left(\frac{1}{1-\frac{x}{x+1}}\right) \ge \ln\left(\frac{1}{e^{-\frac{x}{x+1}}}\right) = \frac{x}{x+1}\,,$$

where the first inequality follows from the inequality $1+y \le e^y$ applied to $y = -\frac{x}{x+1}$.

Proof of 3. Taking the derivative, we must show that $\ln(1+1/x) - \frac{1/x}{1+1/x} > 0$ for $x > 0$. Replacing x by $1/x$, this is equivalent to $\ln(1+x) > \frac{x}{1+x}$, which follows from part 2. □

Proof of Lemma 2.3. By taking the power series expansion of e^x, we get

$$\begin{aligned}\mathbb{E}[e^{\lambda \boldsymbol{x}}] &= \frac{1}{2}(e^{\lambda} + e^{-\lambda}) \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2} \,.\end{aligned}$$

□

Proof of Lemma 2.4.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}}[\exp\big(\alpha \boldsymbol{z}^2 + 2\beta \boldsymbol{z}\big)] \\ &\quad= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\alpha x^2 + 2\beta \boldsymbol{z}} e^{-x^2/2} dx \\ &\quad= e^{\frac{2\beta^2}{1-2\alpha}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2\alpha)(x-\frac{2\beta}{1-2\alpha})^2/2} dx \\ &\quad= e^{\frac{2\beta^2}{1-2\alpha}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2\alpha)x^2/2} dx \\ &\quad= e^{\frac{2\beta^2}{1-2\alpha}} \cdot \frac{1}{\sqrt{1-2\alpha}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx \\ &\quad= e^{\frac{2\beta^2}{1-2\alpha}} \cdot \frac{1}{\sqrt{1-2\alpha}} \,.\end{aligned}$$

□