Entropy Matters: Understanding Performance of Sparse Random Embeddings

Maciej Skorski
University of Luxembourg, Luxembourg

Abstract
This work shows how the performance of sparse random embeddings depends on the Renyi entropy-like property of data, improving upon recent works from NIPS’18 and NIPS’19.

While the prior works relied on involved combinatorics, the novel approach is simpler and modular. As the building blocks, it develops the following probabilistic facts of general interest:

(a) a comparison inequality between the linear and quadratic chaos
(b) a comparison inequality between heterogenic and homogenic linear chaos
(c) a simpler proof of Latala’s strong result on estimating distributions of IID sums
(d) sharp bounds for binomial moments in all parameter regimes.

2012 ACM Subject Classification
Mathematics of computing → Probabilistic algorithms; Theory of computation → Random projections and metric embeddings

Keywords and phrases Random Embeddings, Sparse Projections, Renyi Entropy

1 Introduction

The celebrated result due to Johnson and Lindenstrauss [38] states that random linear mappings are perfect embeddings: they nearly preserve distances of input data points, while mapping them into a much lower dimension. This enables accomplishing otherwise computationally demanding tasks, by running on the reduced yet representative data. Formally, the lemma states that for any distortion $\epsilon > 0$ and confidence parameter $0 < \delta < 1$, with the embedding dimension $m = \Theta(\log(1/\delta)\epsilon^{-2})$ and the $m \times n$ matrix $A$ sampled from the appropriately scaled normal or Rademacher distribution, for every vector $x \in \mathbb{R}^n$

$$(1 - \epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \epsilon)\|x\|_2$$

with probability $1 - \delta$. (1)

For modest but practically meaningful distortion and confidence parameters $\epsilon, \delta$ and large data dimensions $n$ we obtain $m \ll n$, that is a significant dimension reduction; on the other hand nearly-preserving distances (up to a relative factor of $\epsilon$) translates into nearly-preserving scalar products and thus the internal data geometry, making it representative for many tasks. Indeed, over the years variants of the Johnson-Lindenstrauss Lemma have found important applications to text mining and image processing [7], approximate nearest neighbor search [35, 3], learning mixtures of Gaussians [22], sketching and streaming algorithms [43, 47], approximation algorithms for clustering high dimensional data [6, 12, 54], speeding up linear algebraic computations [57, 61, 16], analyzing combinatorial properties of graphs [28, 52] and even to privacy [9, 42]; on the pure theory side, it is worth mentioning the importance for understanding Hilbert spaces in functional analysis [39].

Although the embedding dimension $m$ is optimal [40, 37], the costly matrix-vector product can be optimized by the use of sparse matrices. The long line of research [1, 21, 51, 3, 55, 41, 18] have finally established the same guarantees for matrices $A$ with only $s = \Theta(\log(1/\delta)\epsilon^{-1})$
entries per column\(^1\). Optimal in the worst-case, these results were far away from the performance empirically observed on real-world data, particularly the remarkable accuracy of feature hashing \([65]\) which uses only \(s = 1\) (!). This led to the following intriguing question:

\[
\textbf{Why extremely sparse random projections work better-than-expected?}
\]

A careful reader notices that so far we have been speaking of data-oblivious results, that is under no assumption on the data structure. Indeed, the relevant research in \([65, 21, 41, 30, 36]\) has finally established \([30, 36]\) that the certain metric which captures data dispersion, more precisely the ratio

\[
v = \frac{\|x\|_\infty}{\|x\|_2},
\]

allows for setting the matrix sparsity to \(s = \Theta(v^2 \epsilon^{-1}) \cdot \max \left\{ \log \frac{1}{\delta}, \log^2 \frac{1}{\epsilon} \right\} \) while keeping the optimal dimension \(m = \Theta(\log(1/\delta) \epsilon^{-2})\). This offers an additional improvement by a factor of \(1/v^2\). In simple terms: the more data is dispersed, the better matrix sparsity works. This breakthrough result still suffers from the following limitations:

1. **Unsatisfactory definition of data dispersion.** The ratio \(\ell_\infty\text{-to-}\ell_2\) is a crude notion: on the unit sphere \(\|x\|_2 = 1\) it depends on the heaviest element and so is not smooth enough. It suffers particularly from “spikes” that are naturally present in real-world data (such as features produced in text-mining \([4]\)) and due to pairwise vector differences studied in multi-vector setup (uniform guarantees for multiple vectors are obtained by looking at pairwise differences \(x - x'\), which leads to “spikes” for example in images \([48]\)). This motivates further research for a more accurate notion of dispersion.

2. **High proof complexity and lack of modern toolkit.** Proofs in prior works \([30, 36]\) suffer from being lengthy and convoluted, mostly in supplementary materials, which results in numerical mistakes as well as gaps not immediately fixable (see Appendix A). These works did admirable efforts on presenting the self-contained proof, yet did not utilize the modern probability toolkit to the full extent. Their strategy is to see Equation (1) as the concentration of the quadratic form \(x \rightarrow \|Ax\|_2^2\), and quantify its tails by controlling high-order moments estimated via multinomial expansions coupled with combinatorial arguments. However, this does not leverage tools to control quadratic random forms, namely the modern techniques of the Hanson-Wright inequality \([31, 60, 67]\) such as decoupling of quadratic forms \([63, 24]\). Furthermore, it re-develops a variant of the sharp result from \([49]\) on moment estimation and certain known facts from high-dimensional probability on sub-gaussian distributions \([11, 10]\). Finally, while \([36]\) develops its technical lemmas for symmetric random variables, this condition is not satisfied which leaves a gap. Thus, further effort in revisiting and modernizing the toolkit used in recent state-of-art works \([30, 36]\) is well-motivated. Indeed, simplifying proofs and developing novel techniques for the JL Lemma is an independent and valued line of research \([28, 29, 23, 19]\), as these have been historically difficult (the original result used sophisticated geometric approximations, while the sparse variant \([21]\) relied on correlation inequalities \([27]\)).

\(^1\) As shown by \([18]\) one can reduce further sparsity \(s\) by \(B > 1\) at the cost of exponentially increasing the dimension \(m\) by a factor of \(2^\Theta(B)\). However, in practice, sub-optimal dimension is less interesting.

\(^2\) The formula arises from rearranging Theorem 1.5 in \([36]\).
2 Our Contribution

This work offers a solution to the two problems discussed above: we strengthen and to a great extent simplify the state-of-the-art results from prior works.

2.1 Performance of Sparse Random Projections

We introduce the following (novel) notion of the data dispersion:

\[
v_d(x) \triangleq \sup_{|I|<d/2} \left( \frac{\sum_{i \in I} |x_i|^d}{\sum_{i \in I} x_i^2} \right)^{1/2} / \|x\|_2, \quad d > 2.
\]  

(3)

where \( I \) are taken as strict subsets of the support of \( x \).

The matrix \( A \) is sampled from the sparsified Rademacher distribution, as in prior works:

\begin{algorithm}
\caption{Sparse Random Projections: Matrix Sampler.}
\begin{itemize}
\item Data: data dimension \( n \), embedding dimension \( m \), matrix sparsity \( s \)
\item Result: \( A \in \mathbb{R}^{n \times m} \)
\begin{itemize}
\item for every column \( i \), select \( s \) positions at random (without replacement)
\item set randomly \( \pm 1 \) on the selected positions
\item scale the matrix by \( 1/\sqrt{s} \)
\end{itemize}
\end{itemize}
\end{algorithm}

For the matrix as in Algorithm 1 above, we prove the following result.

\textbf{Theorem 1.} Let \( d = \log(1/\delta) \), then the JL Lemma, that is (1), holds for the dimension

\[
m = \Theta(d\varepsilon^{-2})
\]  

and any sparsity \( s \) such that

\[
v_d(x) \leq \Theta(se)^{1/2} \min(\log(m\varepsilon/d)/d, 1/d^{1/2}).
\]  

(5)

We now discuss the result in detail in the series of remarks below.

\textbf{Remark 2 (Intuition).} We give the following rationale for one could conjecture a result like the one above: the analysis of sparse random projections establishes that the performance depends on the \( d \)-th moment of the error expression, where \( d = \log(1/\delta) \) is relatively small; it seems reasonable to expect that the assumptions on the data should not include moments higher than that of order \( d \), particularly bounding \( \|x\|_\infty \) seems to be an overshooting.

\textbf{Remark 3 (Comparison with previous bounds).} Since \( v_d(x) \leq \|x\|_\infty/\|x\|_2 \), we obtain the previous state-of-the-art bounds from [36], by rearranging Equation (5) to Equation (2). This approximation is however rather crude, as it merely replaces the \( d \)-th norm \( \|x\|_d \) by \( \|x\|_\infty \), and our bound can do much better. Consider the more explicit example where \( x_i^2 = (n/d)^{-1/d} \) for \( d \) values of \( i \) and \( x_i^2 = 1 - (n/d)^{-1/d}/(n - d) \) otherwise. We then have \( v_d(x) = \Theta(n^{-1/d}) \) while \( \|x\|_\infty/\|x\|_2 = \Theta(n^{-1/d}) \). Since the best possible sparsity \( s \) is roughly proportional to \( v_d(x)^{-2} \), our gain over the previous approach is by a factor of \( n^{1-2d^{-2}} \) which is huge for moderate values of \( d \) and large \( n \) (that is, in a typical application regime).

\textbf{Remark 4 (Relation to Renyi Entropy).} Let’s introduce the probability measure \( w_i \sim x_i^2 \), then \( (\sum_{i} |x_i|^d/\sum_{i} x_i^2) \sum_{i} w_i^{\frac{d}{2}} / \|x\|_2 = (\sum_{i} w_i^{\frac{d}{2}}) \sum_{i} w_i^{\frac{d}{2}} = 2^H_{d/2}((w_i))/2 \) where the Renyi entropy [58] of the distribution \( w \) is defined as \( H_d(w) \triangleq \frac{1}{1-d} \sum_{i} w_i^d \) and \( H_\infty(w) \triangleq -\log \max_i w_i \) when
Sparse Random Projections

\[ d = \infty. \] Under the mild assumption that \( x \) such that \( \sum_{i \notin I} x_i^2 = \Theta(\|x\|_2^2) \) for all \( |I| \leq d \) we can thus compare the sparsity achieved in Theorem 1 and the result in [36] as low-order Renyi entropy versus min-entropy. More precisely, our bound on \( s \) is better by a factor of \( 2^{H_d/2((w_i)) - H_{\infty}((w_i))} \) that is the gain is exponential in entropy deficiency understood as \( H_d/2((w_i)) - H_{\infty}((w_i)) \). The well-known bounds from information-theory [14] show that this gap can be as big as \( \frac{1}{d/2} - 1 \) which is unbounded without some restrictions on \( x \).

\[ \text{Remark 5 (Dimension-Sparsity Tradeoffs). It is possible to improve the sparsity parameter } s \text{ by a factor of } B \text{ at the expense of making the dimension worse by a factor of } e^{\Theta(B)}, \text{ exactly as in [36]. However, this tradeoff does not seem to be interesting from the application-oriented point of view (the whole idea of random projections is to keep the low dimension).} \]

2.2 Techniques of Independent Interest

2.2.1 From Quadratic to Linear Chaos

One important novelty in our approach is that we get rid of analyzing quadratic forms, which appear due to considering the expression \( \|Ax\|_2^2 \), by an elegant reduction to their linear analogues. Although quadratic chaoses of symmetric random variables have been studied in the past [49, 46], the generic bounds were found intractable to analyze by the authors of prior works [30, 36] and other workaround have been proposed. It has been not clear if one can get rid of these complicated methods. Indeed, we show that we can:

\[ \text{Lemma 6. Let } X_i \text{ be independent zero-mean random variables, with possibly different distributions. Then for even } d \geq 2 \text{ we have} \]

\[ \| \sum_{i \neq j} X_i X_j \|_d \leq 32 \| \sum_i X_i \|_d^2. \]

\[ \text{Remark 7. The result is fairly general, not requiring symmetry or identical distributions. In fact, the constant reduces to 4 if } X_i \text{ are already symmetric.} \]

This bound allows for reducing a bulk of technical calculations, and almost directly applying existing tractable bounds for linear forms such as those in [50]. The proof uses decoupling [63] which allows for upper-bounding the moments of the quadratic form \( \sum_{i \neq j} X_i X_j \) by the moments of bilinear form \( \sum_{i \neq j} X_i X'_j \), and symmetrization [64] which allows for replacing \( X_i \) by their symmetrized versions \( X_i - X'_i \) at the expense of a constant factor.

2.2.1.1 Heterogenic Sparse Rademacher Chaos

Although we reduce the problem to studying linear forms, they are not IID sums. More precisely in our case we will be interested in sums of form \( \sum_i x_i X_i \) where \( X_i \) are symmetric and IID, but the given weights \( x_i \) can be very different. Such sums are notoriously difficult to analyze, the best example being probably the classical Khintchine’s inequality which seeks to bound \( \| \sum_i x_i \sigma_i \|_d \) where \( \sigma_i \) are Rademachers, for a given sequence of weights \( (x_i) \); it took a while until the original bounds [44] have been tightened, in a way that explicitly depend on \( x \) [33]. While prior works [30, 36] handle this difficulty in our context implicitly (in combinatorial analyses of multinomial expansions), we use majorization theory to essentially compare the heterogenic and homogenic (easier) setup. We prove
Lemma 8. Let $\|x\|_2 = 1$ and $X_i \sim \text{IID } \eta_i \sigma_i$ where $\eta_i$ are IID Bernoulli and $\sigma_i$ are IID Rademacher r.v.s. Then for $v = v_d(x)$ where $v_d(x)$ is as in Equation (3), and even $d > 0$

$$\| \sum_i x_i X_i \|_d \leq O(\|K^{-1/2} \sum_{i=1}^{K} X_i\|_d), \quad K = \lceil v^{-2} \rceil.$$  

The result depends on the structure of $x$ captured by $v = v_d(x)$, note that the equality holds when $x_i = v$ for all non-zero weights $x_i$ (note that we normalize $\|x\|_2 = 1$ w.l.o.g.); this is the core of our method, and we can see it as a sparse analogue of Khintchine’s Inequality (Bernoulli variables restrict the summation to a random subset). The result should be considered strong and somewhat surprising; per analogy to the case when there are no Bernoulli variables, results from majorization theory seem to suggest that the moment should be rather minimized for $x_i$ that are nearly uniform. The answer is in the condition $v_d(x)$ which is, to a certain degree, a relaxation of the requirement that $x_i$ is flat and in the constant under $O(1)$. What we prove is not that $(x_i)$ with $K$ elements gives the maximum, but that the value differs from the actual maximum by at most a constant factor. In our proof, we use the assumption in Equation (3) and majorization [17] to compare the behavior of sums $S_k = \sum_{i \neq j \in \pi} x_{i_1}^2 \cdots x_{i_k}^2$ when $x_i$ is uniform over $K$ elements versus over the whole space. Under the normalizing condition $\|x\|_2 = 1$, they can be interpreted as birthday collision probabilities, which makes the comparison easy to evaluate.

2.2.1.2 Moments of IID Sums

We will need a result which provides tight bounds on moments of iid sums. Although this problem has been solved by a characterization due to Latala [50], the result seems to be little known within the TCS community; instead classical bounds due to Hoeffding [34], Chernoff [15], Bernstein [5] or more modern bounds stated sub-gaussian or sub-gamma distributions [11] are used. Since the analysis of sparse random projections involves random variable with little exotic behavior, the classical inequalities are not sufficient.

In hope for popularizing the technique and to make the paper self-consistent, we provide an alternative and simpler proof of Latala’s result [50].

Lemma 9. For zero-mean r.v.s. $X_i \sim \text{IID } X$ and even $d > 0$

$$\| \sum_{i=1}^{n} X_i \|_d \leq 2e \cdot \max_k \left[ \left( \frac{d}{k} \right)^{1/k} \left( \exp(d/n) - 1 \right)^{-1/k} \|X\|_k : \max(2, d/n) \leq k \leq d \right]$$ (6)  

which implies the following simpler bound

$$\| \sum_{i=1}^{n} X_i \|_d \leq \frac{2e^2}{(1 - e^{-1})^{1/2}} \cdot \max_k \left[ \frac{d}{k} \cdot (n/d)^{1/k} \cdot \|X\|_k : \max(2, d/n) \leq k \leq d \right].$$ (7)  

Remark 10. In addition to simplifying the proof, we provide an explicit constant (not given in the original proof). For non-symmetric distributions, our numerical constant is better than the one implied by symmetrizing the original proof. We also note that there is the same matching, up to a constant, lower bound [49], so that in the result above we have the equality up to a constant.

The map $(x_i) \rightarrow \| \sum_i x_i \sigma_i \|_d^2$ is Schur-concave in variables $x_i^2$ [26].
2.2.1.3 Sharp Bounds for Binomial Moments

Having reduced the problem to studying moments of \( \sum_i \eta_i \sigma_i \), we face the problem of estimating binomial moments. Somewhat surprisingly, the literature does not offer good bounds for binomial moments. What we know are combinatorial formulas \([45]\) not in a closed asymptotic form, and nearly perfect estimates (up to \(o(1)\) relative error) for binomial probabilities \([62]\) as well as the tails \([20, 53, 56]\) (see also the survey in \([2]\)) — these tails unfortunately lead to intractable integrals expressing moments (with Kullback-Leibler terms).

Since the question is foundational with clear potential for applications beyond our problem, we give the following general and detailed answer

▶ **Lemma 11.** Let \( S \sim \text{Binom}(K, p) \) where \( p \leq \frac{1}{2} \), and \( d > 0 \) be even. Then

\[
\|S - \mathbb{E}S\|_d = \Theta(1) \begin{cases} 
(dKp)^{1/2} & \log(d/Kp) < d/K \leq 2 \\
Kp^{K/d} & \log(d/Kp) < 2 \leq d/K \\
\frac{d}{\log(d/Kp)} & \max(2, d/K) \leq \log(d/Kp) \leq d.
\end{cases}
\]  

(8)

▶ **Remark 12.** The bound has up to 4 regimes, in which we provide an estimate sharp up to a constant. The upper bound (sufficient for our needs) follows from Lemma 9, while the lower bound holds because the bound in Lemma 9 is sharp up to an absolute constant \([49]\).

2.3 Proof Outline

We actually prove that

\[
(1 - \epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \epsilon)\|x\|_2^2
\]
with probability \( 1 - \delta \)

from which Equation (1) follows by taking the square roots and using the elementary inequalities \( \sqrt{1 + \epsilon} \leq 1 + \epsilon \), \( 1 - \epsilon \leq \sqrt{1 - \epsilon} \). Denoting \( Z = \|Ax\|_2^2 - \|x\|_2^2 \) we find \([36]\)

\[
Z = \frac{1}{s} \sum_{r=1}^m Z_r, \quad Z_r \triangleq \sum_{i \neq j} x_i x_j \eta_i \eta_j \sigma_i \sigma_j.
\]  

(10)

It can be shown that \( Z_r \) are negatively dependent and thus their sum obey moment upper-bounds for independent random variables \([25, 8]\). More precisely we have that

\[
\|Z\|_d \leq \frac{1}{s} \| \sum_{r=1}^m Z_r \|_d, \quad Z_r \sim_{\text{IID}} \sum_{i \neq j} x_i x_j \eta_i \eta_j \sigma_i \sigma_j.
\]  

(11)

The techniques outlined above, namely Lemma 6 and Lemma 8 show that for \( K = \lceil v_d(x)^{-2} \rceil \)

\[
\|Z_r\|_d \leq O(K^{-1} \|S - S'\|_d^2), \quad S, S' \sim_{\text{IID}} \text{Binom}(K, p).
\]  

(12)

Since \( \|S - S'\|_d \leq 2\|S - \mathbb{E}S\|_d \) (the triangle inequality), by Lemma 11 we obtain

▶ **Corollary 13.** For any even \( d > 0 \) we have

\[
\|Z_r\|_d \leq O(1) \begin{cases} 
\frac{dp}{\log(d/Kp)} & \log(d/Kp) < d/K \leq 2 \\
Kp^{2K/d} & \log(d/Kp) < 2 \leq d/K \\
\frac{k^{-1}d^2}{\log(d/Kp)} & \max(2, d/K) \leq \log(d/Kp) \leq d \\
K^{-1}(Kp)^{2/d} & d < \log(d/Kp)
\end{cases}
\]  

(13)
It now suffices to plug this bound in Lemma 8 (it applies for negatively dependent r.v.s.) and analyze the 4 different regimes, to obtain moment bounds for $Z$ from Equation (10); then Theorem 1 follows by Markov’s inequality. The work has been mostly finalized at this point, due to our modular approach; the application of Lemma 8 is discussed in the appendix.

Remark 14. At the final stage [36] also obtains analogous bounds (with $K$ defined in terms of $v = \|x\|_\infty/\|x\|_2$). They are however not derived via a single application of a lemma, but rather a mixture of three techniques (direct bounds on quadratic forms, linear forms, and the reproved result on the sub-gaussian norm of a binary random variable [13]).

2.4 Organization

The rest of the paper is organized as follows: in Section 3 we introduce basic notation and some simple auxiliary facts that will be used throughout the discussion, in Section 4 we present proofs of the key ingredients of our proof. Details omitted in the proof outline are provided in Appendix B. In Section 5 we conclude the work.

3 Preliminaries

3.1 Basic Notation

For a random variable $X$, we define its $d$-th moment as $E|X|^d$ and its $d$-th norm as $\|X\|_d = (E|X|^d)^{1/d}$ (this is indeed a norm when $d \geq 1$). For the sequence $(x_i)$ we define $\|(x_i)\|_d = (\sum_i |x_i|^d)^{1/d}$ for $0 < d < 1$, $\|x\|_\infty = \max_i |x_i|$ and $\|x\|_0 = \#\{i : x_i \neq 0\}$.

By $\text{Bern}(p)$ we denote the Bernoulli distribution, that is 1 with probability $p$ and zero otherwise. By $\text{Binom}(K,p)$ we denote the binomial distribution with parameters $K$ and $p$ (equal in the distribution to the sum of $K$ independent copies of $\text{Bern}(p)$).

3.2 Auxiliary Functions

We need the elementary properties of the two functions that often appear in our analysis:

- **Proposition 15.** The function $g(d) = 1/q \cdot a^{1/q}$ for $q > 0$ is decreasing when $a \geq 1$ and for $a < 1$ it achieves its local maximum at $q = \log(1/a)$ with the value $g(q) = 1/e \log(1/a)$.

- **Proposition 16.** The function $g(q) = q \cdot a^{1/q}$ for $q > 0$ is increasing when $a \leq 1$ and for $a > 1$ achieves its local minimum at $q = \log a$ with the value $g(q) = e \log a$.

3.3 Probabilistic Techniques

The following fact will allow us to handle non-symmetric distributions.

- **Proposition 17 (Symmetrization trick [64]).** For any norm $\| \cdot \|$ we have

$$\frac{1}{2} \| \sum_i X_i \sigma_i \| \leq \| \sum_i X_i \sigma_i \| \leq 2\| \sum_i X_i \sigma_i \|,$$

for any zero-mean independent $X_i$ and independent Rademacher random variables $\sigma_i$.

We will also need the decoupling inequality, useful in attacking quadratic forms.
Sparse Random Projections

**Proposition 18** (Decoupling inequality [63]). Let $X_i$ be zero-mean independent r.vs. and $X'_i$ be their independent copies. Then for any weights $a_{i,j}$

$$E[f(\sum_{i \neq j} a_{i,j} X_i X_j)] \leq E[f(4 \sum_{i \neq j} a_{i,j} X'_i X'_j)],$$

for any convex function $f$.

**Remark 19.** The summation is over $i \neq j$, e.g. the quadratic form must be off-diagonal!

### 4 Proofs

#### 4.1 Quadratic vs Linear Chaos

**Proof of Lemma 6.** Let $X'_i$ be independent copies of $X_i$. The decoupling inequality gives

$$\| \sum_{i \neq j} X_i X_j \|_d \leq 4 \| \sum_{i \neq j} X'_i X'_j \|_d.$$  \hspace{1cm} (14)

We apply the symmetrization trick to the $d$-th norm twice: first for random variables $X_i$ with any fixed choice of $X'_i$ which gives $\| \sum_{i \neq j} X_i X'_j \|_d \leq 2 \| \sum_{i \neq j} X_i \sigma_i X'_j \|_d$ (here we use the independence of $X_i$ and $X'_i$) and second for random variables $X'_i$ under the fixed values of $X_i \sigma_i$ which gives $\| \sum_{i \neq j} X_i X'_j \|_d \leq 4 \| \sum_{i \neq j} X_i \sigma_i \sigma'_j \|_d$ ($\sigma'_j$ is an independent Rademacher sequence). For simplicity, we denote $X_i := X_i \sigma_i$ and $X'_j := X'_j \sigma'_j$, note that the introduced random variables $X_i \sigma_i$ and $X'_j \sigma'_j$ are also identically distributed.

Consider the sum $\sum_{i,j} X_i X'_j = \sum_{i}(\sum_{j \neq i} X'_j) X_i$ as linear in $X_i$ with coefficients depending on $X'_j$, and apply the multinomial theorem which gives

$$E[(\sum_{i \neq j} X_i X'_j)^d(X'_j)] = \sum_{(d_i)} \left(\frac{d}{2d_1 \ldots 2d_n}\right) \prod_i (\sum_{j \neq i} X'_j)^{2d_i} E_i X_i^{2d_i},$$

where we use the symmetry of $X_i$, so that all odd moments vanish. Again by the multinomial theorem we see that

$$E[(\sum_{j \neq i} X'_j)^d] \leq E[(\sum_{j} X'_j)^d].$$

Combining the last two bounds gives

$$E[\sum_{i \neq j} X_i X'_j]^d \leq E[(\sum_{j} X'_j)^d] E[(\sum_{i \neq j} X'_j)^d](X'_j)]$$

$$\leq \sum_{(d_i)} \left(\frac{d}{2d_1 \ldots 2d_n}\right) E\left[\prod_i (\sum_{j} X'_j)^{2d_i} X_i^{2d_i}\right]$$

$$\leq E[\sum_i (\sum_{j} X'_j) X_i]^d$$

$$= E[\sum_i X_i]^d (\sum_{j} X'_j)^d = E[\sum_i X_i]^{2d},$$

which can be stated as

$$\| \sum_{i \neq j} X_i X'_j \|_d \leq \| \sum_i X_i \|_d^2.$$  \hspace{1cm} (15)
By combining Equation (14) and Equation (15), and keeping in mind that $X_i$ above are symmetrized versions, we obtain for original (only centered) random variables $X_i$

$$\mathbb{E}\|\sum_{j \neq i} X_i X_j\|_d \leq 16\mathbb{E}\|\sum_{j \neq i} X_i \sigma_i\|_d,$$

and the result follows by one more application of the symmetrization trick.

### 4.2 Heterogenic vs Homogenic Chaos

**Proof of Lemma 8.** By the multinomial expansion and the symmetry of $Z_i$ (which implies that the odd moments vanish) we obtain

$$\mathbb{E}(X_1 X_2)^d = \sum_{(d_i)} \left(\begin{array}{c} d \\ 2d_1 \cdots 2d_n \end{array}\right) p^{[(d_i)]} \prod_i x_i^{2d_i},$$

where the summation is over non-negative sequences $(d_i)$ for $i = 1, \ldots, n$ such that $\sum_i d_i = d/2$, and we denote $\|[(d_i)]\|_0 = \#\{i : d_i > 0\}$. Considering possible values of $k = \|[(d_i)]\|_0$, we find that the above expression is a non-negative combination of

$$S_k^d(x) = \sum_{i_1 \neq \cdots \neq i_k} x_{i_1}^{2d_1} \cdots x_{i_k}^{2d_k},$$

where possible values of $k$ are $1 \leq k \leq \min(d/2, n_0)$ where $n_0 = \|[(x_i)]\|_0$. We now apply our assumption on $x$ iteratively to $x_{i_k}, x_{i_{k-1}}, \ldots$, obtaining

$$S_k^d(x) \leq v^2 \sum_{i_1 \neq \cdots \neq i_k} (d_1-1) \sum_{i_1 \neq \cdots \neq i_k} x_{i_1}^2 \cdots x_{i_k}^2.$$

Here we have used the fact that $v_d(x)$ is increasing in $d$, so $v_k(x) \leq v$ when $k \leq d$; this follows from seeing $v_d(x)$ as the power mean of order $d - 2$ and weights $x_i^2 / \sum_{i \neq x} x_i^2$ [32, 66].

We make the following important observation: the equality holds whenever $x_i$ is flat with the value $v$, e.g. all non-zero entries are equal to $v$. Observe that the sums $S_k(x) = \sum_{i_1 \neq \cdots \neq i_k} x_{i_1}^2 \cdots x_{i_k}^2$ are elementary symmetric polynomials in variables $y_i = x_i^2$ where $\sum y_i = \sum x_i^2 = 1$, hence over the probability simplex. The elementary symmetric functions are Schur-concave [17], and thus they are maximized at the uniform distribution, in our case when $x_i = n^{-1/2}$. In fact, $S_k(x)$ is the probability that $k$ independent samples from the distribution $p_i = x_i^2$ do not collide. For any sequence $(x_i^2)$ which has $N$ non-zero equal entries and $\sum_i x_i^2 = 1$ we have that:

$$S_k(x) = N \cdot (N - 1) \cdots (N - k + 1)/N^k.$$

Since $N \geq k$ and since $k \leq d$, using Stirling’s approximation [59] we obtain

$$S_k(x) = \prod_{i=0}^{k-1} (1 - i/N) \geq k! / k^k = \Theta(1)^k \geq \Theta(1)^d.$$

Clearly $S_k(x) \leq 1$ for any $x$. If we replace $(x_i)$ by a sequence such that $x_i = v$ for $K = v^{-2}$ values of $i$ (e.g., flat), we lose at most a factor of $\Theta(1)^k \leq \Theta(1)^d$ in every term $S_k^d(x)$. ▶
4.3 Moments of IID Sums

Proof of Lemma 9. We have the following chain of estimates

\[
E(\sum_i X_i)^d = \sum_{d_1, d_2, \ldots, d_n \geq d, d_i \geq 2} \left(\frac{d}{d_1 \ldots d_n}\right) \prod_i E[X_i^{d_i}]
\]
\[
\leq \sum_{d_1, d_2, \ldots, d_n \geq d, d_i \geq 2} \prod_i \left(\frac{d}{d_i}\right) E|X_i|^{d_i}
\]
\[
\leq \sum_{d_i \geq 2} \prod_i \left(\frac{d}{d_i}\right) E|X_i|^{d_i}
\]
\[
\leq \left(\sum_{k=2}^{d} \left(\frac{d}{k}\right) \|X\|_k^k\right)^n.
\]

Applying this for \(X_i := X_i/t\) we have for any \(t > 0\)

\[
E\left(t^{-1} \sum_i X_i\right)^d \leq \left(\sum_{k=2}^{d} \left(\frac{d}{k}\right) \|X\|_k^k/t^k\right)^n.
\]

Thus \(\|\sum_i X_i\|_d \leq et\) for any \(t\) such that the right-hand side is at most \(e\), equivalently

\[
\sum_{k=2}^{d} \left(\frac{d}{k}\right) \|X\|_k^k/t^k \leq \exp(d/n) - 1,
\]

which is satisfied for

\[
t = 2 \max_{k=2 \ldots d} \frac{\left(\frac{d}{k}\right)^{1/k}}{(\exp(d/n) - 1)^{-1/k}} \|X\|_k.
\]

This proves the first part. Observe that for \(k \geq 2\) we have

\[
\left(\frac{d}{k}\right)^{1/k} (\exp(d/n) - 1)^{-1/k} \leq \frac{ed}{k \exp(d/kn)} \cdot \frac{1}{(1 - \exp(-1))^{1/2}},
\]

where we use the elementary inequalities \(\left(\frac{d}{k}\right) \leq (de/k)^k\) and \(\exp(u) - 1 \geq \exp(u) \cdot (1 - e^{-1})\) for \(u \geq 1\). The function \(u \to u/\exp(u)\) decreases for \(u \geq 1\); applying this to \(u = d/kn\) gives

\[
\left(\frac{d}{k}\right)^{1/k} (\exp(d/n) - 1)^{-1/k} \leq \frac{en}{(1 - e^{-1})^{1/2}}, \quad k \leq d/n.
\]

Since \(\|X\|_k\) increases in \(k\) we have

\[
\max_{k=2 \ldots d, k \leq d/n} \left(\frac{d}{k}\right)^{1/k} (\exp(d/n) - 1)^{-1/k} \|X\|_k \leq \frac{en\|X\|_{d/n}}{(1 - e^{-1})^{1/2}}.
\]

We have \((\exp(d/n) - 1)^{-1/k} \leq (d/n)^{-1/k}\) due to the elementary inequality \(\exp(u) - 1 \geq u\), and \((\frac{d}{k}) \leq (de/k)^k\) for any \(k\). This gives

\[
\max_{k=2 \ldots d} \left(\frac{d}{k}\right)^{1/k} (\exp(d/n) - 1)^{-1/k} \|X\|_k \leq e \max_{k=2 \ldots d} d/k \cdot (n/d)^{1/k} \cdot \|X\|_k
\]

When \(d/n \geq 2\) we have that \(d/k \cdot (n/d)^{1/k} \cdot \|X\|_k = n \|X\|_{d/n} \cdot 2^{-1/2}\) for \(k = d/n\). Comparing the last two equations, we obtain

\[
\max_{k=2 \ldots d, k \leq d/n} \left(\frac{d}{k}\right)^{1/k} (\exp(d/n) - 1)^{-1/k} \|X\|_k \leq C \max_{k=2 \ldots d, k \geq d/n} d/k \cdot (n/d)^{1/k} \cdot \|X\|_k,
\]

with \(C = \frac{e}{(1 - e^{-1})^{1/2}}\). This completes the proof.
4.4 Binomial Moments

Proof of Lemma 11. Applying Lemma 9 we obtain

\[ \|S - ES\|_d \leq O(1) \cdot \max \left\{ \frac{d}{k} \cdot \frac{(Kp/d)^{1/k}}{\max(2, d/K)} : \max(2, d/K) \leq k \leq d \right\}. \]

because \( S \sim \sum_i X_i \) where \( X_i \sim \text{Bern}(p) \) and \( \|X_i - EX_i\|_d = (p(1-p)d^{-1} + (1-p)p^{d-1})^{1/d} \) so that \( \|X_i - EX_i\|_d = \Theta(p)^{1/d} \) for \( p \leq 1/2 \).

The expression under the maximum is proportional to \( k^{-1} \cdot a^{1/k} \) where \( a = Kp/d \). The claim follows by applying Proposition 15, namely a) when \( \max(2, d/K) \leq \log(1/a) \leq d \) (that is, inside the interval) we have necessarily \( a \leq e^{-2} < 1 \) our maximum is at \( k = \log(1/a) \), b) when \( \log(1/a) > d \) we must have \( a < 1 \) and our maximum is at \( k = d \) and c) when \( \log(1/a) < \max(2, d/K) \) then the maximum is at \( k = \max(2, d/K) \).

5 Conclusion

We have proven novel bounds for sparse random projections, showing that the performance depends on the data statistic closed to Renyi entropy. Some intriguing problems we leave for future work are

- How do results extend to non-Rademacher matrices?
- Can we use majorization theory to fully characterize worst case for the linear chaos?

References


Some remarks on prior works

A.1 Some issues with numeric constants

Lemma 2.1 in [36] gives the following bound (expressed in our notation)

$$\|Z_r\|_d \lesssim \begin{cases} dp & d = 2 \text{ or } d \leq pe/u^2 \\ \min\left(\frac{d^2 v^2}{\ln(dv^2/p)}, \frac{d^2}{\ln(1/p)}\right) & 1 \leq \log(dv^2/p) \leq d \\ v^2(p/dv^2)^{2/d} & d < \log(dv^2/p) \end{cases}$$

There is a minor mistake in splitting the branches: they emerge from taking the derivative test of the function $d^2 v^2 u^{-2} (p/dv^2)^{1/u}$ where $1 \leq u \leq d/2$ (Lemma D.1). Here the local maxima occurs at $u = \log(dv^2/p)/2$ and when comparing this with edges $u = 1$ and $u = d/2$ we obtain the conditions $2 \leq \log(dv^2/p)$ and $\log(dv^2/p) \leq d$. Thus, the splitting conditions should be a bit different; this particular issue doesn’t affect the bounds expressed in the asymptotic notation; we report it with intent to motivate our effort in giving a simple and clear proof.

A.2 Gaps in symmetrization

Section 2.2 of [36], when explaining the proof strategy, proposes to apply the bounds on $Z_r$ defined in Equation (10) assuming they are symmetric. But $Z_r$ are not symmetric (it is easy to see they have positive higher-order moments), thus extra work is needed to push this argument forward.
Concluding Main Theorem

Without losing generality, we assume that \( d = \log(1/\delta) \) is even. Recall that we denote \( v = v_d(x) \), also without losing generality we assume that \( v^{-2} \) is an integer. For \( K = v^{-2} \) define the following quantities

\[
I_1 \triangleq \max_q \left\{ \frac{d}{q} \cdot \left( \frac{m}{d} \right)^{1/q} : \log(q/Kp) \leq q/K \leq 2, 2 \leq q \leq d \right\}
\]

\[
I_2 \triangleq \max_q \left\{ \frac{d}{q} \cdot \left( \frac{m}{d} \right)^{1/q} : K(p^{2/k})^2 : \log(q/Kp) \leq 2 \leq q/K, 2 \leq q \leq d \right\}
\]

\[
I_3 \triangleq \max_q \left\{ \frac{d}{q} \cdot \left( \frac{m}{d} \right)^{1/q} : K^{-1} q^2 / \log^2(q/Kp) : \max(2, q/K) \leq \log(q/Kp) \leq q, 2 \leq q \leq d \right\}
\]

\[
I_4 \triangleq \max_q \left\{ \frac{d}{q} \cdot \left( \frac{m}{d} \right)^{1/q} : K^{-1} (Kp)^{2/q} : q \leq \log(q/Kp), 2 \leq q \leq d \right\}
\]

Following the proof outline we arrive at Corollary 13. Taking into account Lemma 11 and Lemma 9, implies that:

\[
\| \sum_{r=1}^m Z_r \|_d \leq O(\max(I_1, I_2, I_3, I_4)).
\]

The goal is to prove that for \( t = se \) we have

\[
\| \sum_{r=1}^m Z_r \|_d \leq t/e,
\]

and then the result follows from Markov’s inequality. We give first bounds for \( I_1, I_2, I_4 \) as they are fairly easy to obtain. The case of \( I_3 \) is analyzed as the last one.

B.1 First Branch

We will show the following bound

\begin{itemize}
  \item \textbf{Lemma 20.} We have  
  \[ I_1 \leq O(dmp^2)^{1/2}. \]
\end{itemize}

Proof of Lemma 20. We have

\[
I_1 = \max_q \left\{ pd(m/d)^{1/q} : \log(q/Kp) \leq q/K \leq 2, 2 \leq q \leq d \right\}
\]

\[
\leq (dmp^2)^{1/2}
\]

where the inequality follows because \( m \geq d \) and \( 1/q \leq 1/2 \) (for \( q \) satisfying the constraints). This completes the proof.

B.2 Second Branch

We will show the following bound

\begin{itemize}
  \item \textbf{Lemma 21.} For \( p \leq 2v^{-2} \) we have  
  \[ I_2 \leq (dmp^2)^{1/2}. \]
\end{itemize}
18:16  Sparse Random Projections

Proof of Lemma 20. For \( q \) satisfying the constraint we have \( K/q \geq e^{-2} / p \) which, due to \( p \leq 2e^{-2} \), implies \( K/q \geq 1/2 \). Then \( p^{2K/q} \leq p \) (recall that \( p < 1 \)) and thus
\[
I_2 \leq \max_q \left\{ \frac{d}{q} \cdot (m/d)^{1/q} \cdot Kp : \log(q/Kp) \leq 2 \leq q/K, 2 \leq q \leq d \right\}.
\]
For \( q \) within the constraints we have \( K/q \leq \frac{1}{2} \) and therefore
\[
I_2 \leq \frac{p}{2} \max_q \left\{ \frac{d}{q} \cdot (m/d)^{1/q} : \log(q/Kp) \leq 2 \leq q/K, 2 \leq q \leq d \right\}.
\]
Since \( m/d \geq 1 \) the expression under the maximum decreases with \( q \), thus is not bigger than the value at \( q = 2 \). Thus, \( I_2 \leq p(dm)^{1/2}/2 \) and the result follows.

B.3 Fourth Branch

We will prove the following bound

Lemma 22. We have
\[
I_4 \leq \begin{cases} \frac{1}{2} \left( dmp^2 \right)^{1/2} \log(dv^4/mp^2) & \leq 2 \\ dv^2 \log(dv^4/mp^2) & > 2 \end{cases}.
\]

Proof of Lemma 22. We have
\[
I_4 = \max_q \left\{ K^{-1} \cdot d/q \cdot (K^2 p^2 m/d)^{1/q} : q \leq \log(q/Kp), 2 \leq q \leq d \right\}.
\]
Let \( a = K^2 p^2 m/d \), the expression under the maximum is proportional to \( 1/q \cdot a^{1/q} \). We now apply Proposition 15: for \( a \geq 1 \) the maximum is not bigger than the value at \( q = 2 \), so
\[
I_4 \leq \frac{1}{2} \left( dmp^2 \right)^{1/2}.
\]
We now can assume \( a < 1 \), equivalent to \( K^2 p^2 m < d \). The global maximum is at \( q = \log(1/a) \), thus our maximum is still at \( q = 2 \) when \( \log(1/a) \leq 2 \) and otherwise is not bigger than the value at \( q = \log(1/a) \). We then obtain
\[
I_4 \leq K^{-1} d \log(d/mp^2 K^2) \leq K^{-1} d = dv^2.
\]
This complete the proof.

B.4 Third Branch

We will show the following bound

Lemma 23. Suppose that \( v^2 \geq se/d^2 \), then
\[
I_3 \leq O(dmp^2)^{1/2} + O(dv / \log(dv^2/p))^2
\]

Proof of Lemma 23. The proof is based on splitting the maximum into three regimes: \( q \in [2, 3], 3 \leq q \leq \log(m/d) \) and \( \log(m/d) \leq q \leq d \). Define
\[
I^0 = \max_q \left\{ \frac{d}{q} \cdot (m/d)^{1/q} \cdot v^2 q^2 / \log^2(qv^2/p) : 2 \leq \log(qv^2/p) \leq q \leq d, 2 \leq q \leq 3 \right\},
\]
\[
I^- = \max_q \left\{ \frac{d}{q} \cdot (m/d)^{1/q} \cdot v^2 q^2 / \log^2(qv^2/p) : 2 \leq \log(qv^2/p) \leq q \leq d, 3 \leq q \leq \log(m/d) \right\},
\]
\[
I^+ = \max_q \left\{ \frac{d}{q} \cdot (m/d)^{1/q} \cdot v^2 q^2 / \log^2(qv^2/p) : 2 \leq \log(qv^2/p) \leq q \leq d, \log(m/d) \leq q \leq d \right\}.
\]
so that we have $I_3 \leq \max(I_0, I^+, I^-)$ (for convenience, we replace the constraint \[ \max(2, qv^2) \leq \log(qv^2/p) \] in $I_3$ by the weaker one $2 \leq \log(qv^2/p)$). By the assumptions we have $v^2/p \geq me/d^2$. Since $m \geq de^{-2}$ we have $\epsilon \geq (d/m)^{1/2}$, and thus
\[ v^2/p \geq (m/d)^{1/2} \cdot d^{-1}. \]

\textbf{Claim 24.} We have $I_- \leq O(d^2v^2/\log^2(dv^2/p))$ when $\log d \leq \frac{5 \log(m/d)}{12}$.

\textbf{Proof of Claim.} For any $q$ satisfying the restrictions it holds that
\[
q \geq \log(v^2/p) \\
\geq \frac{\log(m/d)}{2} - \log d \\
\geq \frac{\log(m/d)}{12}.
\]
We then have $(m/d)^{1/q} \leq O(1)$ and thus
\[ I^- \leq \max_q \{d \cdot qv^2/(2qv^2/p) : 2 \leq \log(qv^2/p) \leq q \leq d, 3 \leq q \leq \log(m/d) \}.
\]
Considering the auxiliary function $u \rightarrow u/\log^2 u$ with $u = qv^2/p \geq v^2$, we see that it decreases in $u$ and hence in $q$ for fixed $v^2$ and $p$. The expression is thus not smaller than its value at $q = d$, which gives
\[ I^- \leq d^2v^2/\log^2(dv^2/p), \]
and completes the proof.

\textbf{Claim 25.} We have $I_- \leq d^2v^2/\log^2(dv^2/p)$ when $\log d > \frac{5 \log(m/d)}{12}$.

\textbf{Proof of Claim.} We have that $dv^2/p \geq me/d \geq (m/d)^{1/2}$ and therefore
\[
I^- \leq dv^2m/(m/d)^{1/3}\log(m/d) \\
\leq dv^2(m/d)^{5/12}/\log^2(m/d) \\
\leq dv^2(m/d)^{5/12}/\log^2(dv^2/p) \\
\leq O(dv^2/\log^2(dv^2/p)),
\]
which completes the proof.

\textbf{Claim 26.} We have $I^+ \leq O(d^2v^2/\log^2(dv^2/p))$

\textbf{Proof of Claim.} We have $(m/d)^{1/q} \leq e$ for $q \geq \log(m/d)$, thus
\[ I^+ \leq d \cdot \max_q \{qv^2/(2qv^2/p) : 2 \leq \max(\log(qv^2/p), \log(m/d)) \leq q \leq d \}.
\]
Considering the auxiliary function $u \rightarrow u/\log^2 u$ with $u = qv^2/p \geq v^2$, we see that it decreases in $u$ and hence in $q$ for fixed $v^2$ and $p$. The expression is thus not smaller than its value at $q = d$, which gives
\[ I^+ \leq O(d^2v^2/\log^2(dv^2/p)) \]
and the claim follows.
Sparse Random Projections

Claim 27. We have $I_0 \leq O((dmp^2)^{1/2})$.

Proof of Claim. We have $I_0 \leq O(v^2(md)^{1/2})$ because $(m/d)^{1/q} \leq (m/d)^{1/2}$ (due to $m/d \geq 1$ and $q \geq 2$). However, for $q \in [2, 3]$ the constraint $\log(qv^2/p) \leq q$ gives $v^2 \leq O(p)$. Thus

$$I_0 \leq O(p(md)^{1/2}),$$

which completes the proof. ◄

The result follows now by combining the above three claims. ◁

B.5 Merging Branch Bounds

To conclude the main result it suffices to satisfy

$$c \cdot \max(I_1, I_2, I_3, I_4) \leq s\epsilon \tag{17}$$

for some absolute constant $c$. The condition in Equation (17) for $I_1, I_2$ is equivalent to $c \cdot (dmp^2)^{1/2} \leq s\epsilon$, which holds when

$$m \geq \Omega(d\epsilon^{-2}). \tag{18}$$

To satisfy Equation (17) for $I_4$ we require, in addition to Equation (18), that $cdv^2 \leq s\epsilon$, equivalent to

$$v \leq O((s\epsilon)^{1/2}/d^{1/2}). \tag{19}$$

Finally, in order to satisfy Equation (17) for $I_3$ we observe that, under the restriction

$$v^2 \geq s\epsilon/d^2, \tag{20}$$

the bound in Lemma 23 gives

$$I_3 \leq O(dmp^2)^{1/2} + O(dv/\log(m\epsilon/d))^2,$$

which follows because $\log(dv^2/p) \geq \log(s\epsilon/dp) = \log(m\epsilon/d)$. Thus, in addition to Equation (18) and Equation (20) it suffices that

$$v \leq O((s\epsilon)^{1/2} \log(m\epsilon/d)/d). \tag{21}$$

Now observe that for

$$v = \Theta(s\epsilon)^{1/2} \min(\log(m\epsilon/d)/d, 1/d^{1/2}) \tag{22}$$

the condition in Equation (20) is automatically satisfied. Thus, the theorem holds for $v$ as above, and clearly for any smaller $v$. 