Entropy Matters: Understanding Performance of Sparse Random Embeddings

Maciej Skorski
University of Luxembourg

Abstract

This work shows how the performance of sparse random embedding depends on the Renyi entropy of the dataset, improving upon recent prior works which looked into less fine-grained data statistics (NIPS'18, 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 celebrated 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

Digital Object Identifier 10.4230/LIPIcs.Submitted.2020.23

1 Introduction

1.1 Sparse Random Projections

The celebrated result due to Johnson and Lindenstrauss [38] states that random linear mappings are perfect embedding: they *almost preserve distances* even when mapping into a *much lower dimension*. More precisely, for any distortion parameter $\epsilon > 0$ if the entries of the $m \times n$ matrix A are sampled independently from the standard gaussian distribution $\mathcal{N}(0,1)$ and $m = \Theta(\log(1/\delta)\epsilon^{-2})$ then for every vector $x \in \mathbb{R}^n$ we have

$$(1-\epsilon)\|x\|_2 \leqslant \|Ax\|_2 \leqslant (1+\epsilon)\|x\|_2 \quad \text{with probability } \delta \tag{1}$$

In applications we may want the above to hold simultaneously for a number of vectors of form $x = x' - x''$ (pairwise differences); then the confidence δ needs to be set up accordingly (by means of the union bound or covering arguments [46]). The optimality of the dimension m has been proven in [40, 37] and the gaussian distribution can be replaced by the Rademacher distribution (± 1 randomly sampled) [1] or more generally by the sub-gaussian condition [11].

The result can be seen as a *dimension-distortion tradeoff*: for an acceptable value of ϵ (which doesn't have to be extremely small in practice) we may obtain $m \ll n$, that is the embedding dimension much smaller than the dimension of the input data x. Reducing the dimension allows for savings in time and memory when processing big data, while the small distortion guarantees that tasks can be done with a similar effect on the embedded data (for example the *cosine similarity* used in data mining [65]). Over years, variants of the above *Johsnon-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 [44, 49], approximation algorithms for clustering high dimensional data [6, 12, 56], speeding up linear algebraic computations [59, 63, 16], analyzing combinatorial properties of graphs [28, 54] and even to privacy [9, 43]. On the pure theory side, it is worth mentioning the importance for understanding Hilbert spaces in functional

42nd Conference on Very Important Topics (CVIT 2016).
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

analysis [39]. Finally, we note that while Equation (1) gives high-probability guarantees, it is possible to find the explicit matrix in randomized polynomial time [23] or by means of derandomization [41].

The focus of this paper is on the *sparse variant* of the Johnson and Lindenstrauss lemma. More precisely, we want A in Equation (1) to have at most s entries in each column. This allows for speeding up projection time, particularly when$x$ itself is sparse. This variant has been covered by a long line of research [1, 21, 53, 3, 57, 42, 18]. The state-of-art result show that keeping the optimal dimension of $m = \Theta(\log(1/\delta)\epsilon^{-2})$ one can take $s = \Theta(\log(1/\delta)\epsilon^{-1})$; in other words one gains at least a factor of $m/s = \epsilon^{-1}$ in the computation time[1]. These results still do not explain the empirically observed performance (much better!), particularly the remarkably powerful technique of *feature hashing* [68] which uses $s = 1$. It turns out, that what explains this phenomena is the underlying *data structure*. The relevant research in [68, 21, 42, 30, 36] has finally established that the certain *data characteristic* which captures sparsity, more precisely the ratio $v = \|x\|_\infty/\|x\|_2$, allows for setting

$$s = \Theta(v^2\epsilon^{-1}) \cdot \max(\log(1/\delta), \log(1/\delta)^2/\log(1/\epsilon)) \qquad (2)$$

as shown in [36]. This offers an additional improvement by a factor of $1/v^2$.

The motivation for this work is the following criticism of prior works

1. The idea of looking at the ratio $v = \|x\|_\infty/\|x\|_2$ does not cope well with datasets that occur in practice; as explained in [36] the implied bounds are asymptotically tight when x is uniformly distributed while real datasets are usually skewed or quite dispersed. For example this is the case in text-mining when data x arise from vectorizing documents followed by the TF-IDF transform [4]. One should also note that the JL Lemma is, in practice, to be applied to *pairwise differences* of the form $x := x' - x''$ where $x', x'' \in \mathcal{X}$, and it is very unlikely for such data to be nearly uniform; in fact datasets such as images [50] tend to produce vectors with entries distributed with "spikes". This motivates looking at parameters other than $v = \|x\|_\infty/\|x\|_2$ in the context of random projections.
2. The proofs are quite complicated, ocasionally sketchy with some numerical mistakes[2] and do not seem to utilize some relevant techniques for simplifications. Their approach is based on seeing Equation (1) as the concentration of the *quadratic form* $x \to \|Ax\|_2^2$, which is estimated via multinomial expansions coupled with some combinatorial arguments and technical bounds. Regarding relevant techniques, we make the following key points a) the standard way of handling quadratic forms is via the *Hanson-Wright inequality*; here prior works does recognize the limitation of the original result [31], but did not consider its modern variants [62, 70] and the useful techniques thereof, such as decoupling of quadratic forms [66, 24] which effectively bridge quadratic and bilinear forms b) when estimating moments of sums of random variables, variants of the (sharp) state-of-art result [51] are re-developed; however parts of calculations could have been carried out using basic facts from *high-dimensional probability* which consider moment conditions when speaking of sub-gaussian, sub-gamma and other distributions [11, 10].

   Historically, variants of the JL Lemma have been difficult to prove (the original result used sophisticated geometric approximations, while the sparse variant [21] relied on correlation inequalities [27]). Given the relevance of the problem, there has been always demand for simplifying proofs and developing novel techniques; this actually emerged

---
[1] As shown by [18] one can reduce further sparsity s by $B > 1$ at the price of increasing the dimension m by a factor of $2^{\Theta(B)}$ (exponentially). But this seems to be of little use

[2] See Appendix A.

into an established and independent line of research [28, 29, 23, 19]. Thus further effort in revisiting and modernizing the toolkit used in recent state-of-art works [30, 36] is well-motivated.

1.2 Our Contribution

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

1.2.1 Performance of Sparse Random Projections

We introduce the following parameter, which captures the *data dispersion*

$$v_d(x) \triangleq \sup_{|I|<d/2} \left(\frac{\sum_{i\not\in I} |x|_i^d}{\sum_{i\not\in I} x_i^2} \right)^{\frac{1}{d-2}} / \|x\|_2, \quad d > 2. \tag{3}$$

where I are taken as strict subsets of the support of x. Sample the matrix A as follows

Sampling Distribution for $A \in \mathbb{R}^{m\times n}$
- for every column i, select s positions at random (sampling without replacement)
- on the selected positions put randomly $\pm$
- scale the matrix by $1/\sqrt{s}$

For the matrix as above we prove the following result

▶ **Theorem 1.** *Let $d = \log(1/\delta)$, then the JL Lemma, that is (1), holds for the dimension*

$$m = \Theta(d\epsilon^{-2})$$

and any sparsity s such that

$$v_d(x) \leqslant \Theta(s\epsilon)^{1/2} \min(\log(m\epsilon/d)/d, 1/d^{1/2}). \tag{4}$$

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

▶ 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 of order d, particularly bounding $\|x\|_\infty$ seems to be overshooting.

▶ Remark 3 (Comparison with previous bounds). Since $v_d(x) \leqslant \|x\|_\infty/\|x\|_2$, we immediately obtain the previous state-of-art bounds from [36]. This approximation is however rather crude, as it merely replaces the d-th norm $\|\cdot\|_d$ by $\|\cdot\|_\infty$, 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 $v_d(x) = \Theta(n^{-\frac{2}{d-2}})$ while $\|x\|_\infty/\|x\|_2 = \Theta(n^{-\frac{1}{d}}))$. 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 $n^{\frac{4}{d-2}-\frac{2}{d}}$ which is huge for moderate values of d and large n (that is in the typical application regime).

▶ Remark 4 (Relation to Renyi Entropy). Let's introduce the probability measure $w_i \sim x_i^2$, then $(\sum_i |x_i|^d / \sum_i x_i^2)^{\frac{1}{d-2}}/\|x\|_2 = (\sum_i w_i^{\frac{d}{2}})^{\frac{1}{d-2}} = 2^{H_{d/2}((w_i))/2}$ where the Renyi entropy [60] 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 $d=\infty$. Under the mild assumption that x such that $\sum_{i\not\in I} x_i^2 = \Theta(\|x\|_2^2)$ for all $|I| \leqslant d$ 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 $\frac{1}{d/2-1}H_{d/2}((w_i))$ (which unbounded without some restrictions on x).

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

1.2.2 Techniques of Independent Interest

1.2.2.1 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 past [51, 48], the generic bounds were found intractable to analyze by the authors of prior works [30, 36] and other workarounds have been proposed. While they are interesting (for example [36] develops a moment bound in spirit of Latala's result for linear forms [51]), it has remained an open problem whether we need them at all. In fact, we answer negatively, due to the following result

▶ **Lemma 6.** *Let X_i be independent zero-mean, with possibly different distributions. Then for even $d \geqslant 2$ we have*

$$\|\sum_{i\neq j} X_i X_j\|_d \leqslant 32\|\sum_i X_i\|_d^2.$$

▶ Remark 7. The result is fairly general, not requiring symmetry or identical distributions. In fact, the constant reduces to 4 if X_i 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 [52]. The proof uses *decoupling* [66] which allows for upper-bounding the moments of the quadratic form $\sum_{i\neq j} X_i X_j$ by the moments of bilienar form $\sum_{i\neq j} X_i X_j'$, and *symmetrization* [67] which allows for replacing X_i by their symmetrized versions $X_i - X_i'$ at the expense of a constant factor.

1.2.2.2 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 σ_i are Rademachers, for a given sequence of weights (x_i); it took a while until the original bounds [45] 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 $X_i \sim^{IID} \eta_i\sigma_i$ where η_i are iid Bernoulli and σ_i are iid Rademacher r.vs. Then for $v = v_d(x)$ where $v_d(x)$ is as in Equation (3), and even $d > 0$*

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

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[3] . 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 $S_k = \sum_{i_1\neq\ldots i_k} x_{i_1}^2\cdots x_{i_k}^2$ 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.

1.2.2.3 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 [52], 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 [52].

▶ **Lemma 9.** *For zero-mean r.vs. $X_i \sim^{IID} X$ and even $d > 0$*

$$\|\sum_{i=1}^{n} X_i\|_d \leqslant 2\mathrm{e}\cdot\max_k\left[\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k}\|X\|_k : \max(2,d/n)\leqslant k\leqslant d\right] \tag{5}$$

which implies the following simpler bound

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

▶ 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 the original proof. We also note that there is the same matching, up to a constant, lower bound [51], so that in the result above we have the equality up to a constant.

[3] The map $(x_i) \to \|\sum_i x_i\sigma_i\|_d$ is Schur-concave in variables x_i^2 [26].

1.2.2.4 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 $\|S\|_d$ where S is binomial. Somewhat surprisingly, the literature does not offer good bounds for binomial moments. What we know are combinatorial formulas [47] not in a closed asymptotic form, and nearly perfect estimates (up to $o(1)$ relative error) for binomial probabilities [64] as well as the tails [20, 55, 58] (see also the survey in [2]); these could be in principle used to recover moments but this leads to intractable integrals with Kullback-Leibler terms in exponents.

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

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

$$\|S-\mathbf{E}S\|_d = \Theta(1)\begin{cases} (dKp)^{1/2} & \log(d/Kp) < d/K \leqslant 2 \\ Kp^{K/d} & \log(d/Kp) < 2 \leqslant d/K \\ \frac{d}{\log(d/Kp)} & \max(2,d/K) \leqslant \log(d/Kp) \leqslant d \\ (Kp)^{1/d} & d < \log(d/Kp) \end{cases}. \tag{7}$$

▶ 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 [51].

1.3 Proof Outline

We actually prove that

$$(1-\epsilon)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+\epsilon)\|x\|_2^2 \quad \text{with probability } \delta \tag{8}$$

from which Equation (8) follows by taking the square roots and using the elementary inequalities $\sqrt{1+\epsilon} \leqslant 1+\epsilon$, $1-\epsilon \leqslant \sqrt{1-\epsilon}$. Denoting $Z = \|Ax\|_2^2$ we find that (see also [36])

$$Z = \frac{1}{s}\sum_{r=1}^{m} Z_r, \quad Z_r \triangleq \sum_{i\neq j} x_i x_j \eta_i \eta_j \sigma_i \sigma_j. \tag{9}$$

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

$$\|Z\|_d \leqslant \frac{1}{s}\|\sum_{r=1}^{m} Z_r\|_d, \quad Z_r \sim^{IID} \sum_{i\neq j} x_i x_j \eta_i \eta_j \sigma_i \sigma_j. \tag{10}$$

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

$$\|Z_r\|_d \leqslant O(K^{-1}\|S-S'\|_d^2), \quad S, S' \sim^{IID} \mathrm{Binom}(K,p). \tag{11}$$

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

▶ **Corollary 13.** *For any even $d > 0$ we have*

$$\|Z_r\|_d \leqslant O(1)\begin{cases} dp & \log(d/Kp) < d/K \leqslant 2 \\ Kp^{2K/d} & \log(d/Kp) < 2 \leqslant d/K \\ \frac{K^{-1}d^2}{\log^2(d/Kp)} & \max(2,d/K) \leqslant \log(d/Kp) \leqslant d \\ K^{-1}(Kp)^{2/d} & d < \log(d/Kp) \end{cases}. \tag{12}$$

It now suffices to plug this bound in Lemma 8 (it applies for negatively dependent r.vs.) and analyze the 4 different regimes, to obtain moment bounds for Z defined in Equation (9); then Theorem 1 is a simple consequence of Markov's inequality. We stress that the most of work has been already done up to this point, due to our modular approach; the details of application of Lemma 8 are deferred to the appendix , we note that they also simplify over an analogous analysis in [36].

▶ 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]).

1.4 Organization

The rest of the paper is organized as follows: in Section 2 we introduce basic notation and some simple auxiliary facts that will be used throughout the discussion, in Section 3 we present proofs of the key ingredients of our proof. Details omitted in the proof outline are provided in Appendix B and In Section 4 we conclude the work.

2 Preliminaries

2.1 Basic Notation

For a random variable X we define its d-th moment as $\mathbf{E}|X|^d$ and its d-th norm as $\|X\|_d = (\mathbf{E}|X|^d)^{1/d}$ (this is indeed a norm when $d \geqslant 1$). 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 $\|x_i\|_0 = \#\{i : x_i \neq 0\}$.

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

2.2 Auxiliary Functions

During our analysis we will often see two particular functions. Their properties follow by a standard application of the derivative test and are summarized below.

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

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

2.3 Probabilistic Techniques

The following fact (follows by a clever use of the triangle inequality) which shows that, roughly, we can replace zero-mean random variables by their symmetrization when calculating norms and moments.

▶ **Proposition 17** (Symmetrization trick [67]). *We have*

$$\frac{1}{2}\|\sum_i X_i\sigma_i\| \leqslant \|\sum_i X_i\sigma_i\| \leqslant 2\|\sum_i X_i\sigma_i\|$$

for any zero-mean independent X_i and independent Rademacher random variables σ_i; this is valid for any norm $\|\cdot\|$.

We will also need the following decoupling inequality has been proven very useful in attacking quadratic forms

▶ **Proposition 18** (Decoupling inequality [66]). *Let X_i be zero-mean independent r.vs. and X_i' be their independent copies. Then for any weights $a_{i,j}$*

$$\mathbf{E} f(\sum_{i\neq j} a_{i,j} X_i X_j) \leqslant \mathbf{E} f(4\sum_{i\neq j} a_{i,j} X_i X_j')$$

for any convex function f.

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

3 Proofs

3.1 Quadratic vs Linear Chaos

Proof of Lemma 6. Let X_i' be independent copies of X_i. The decoupling inequality gives

$$\|\sum_{i\neq j} X_i X_j\|_d \leqslant 4\|\sum_{i\neq j} X_i X_j'\|_d. \tag{13}$$

We apply the symmetrization trick to the d-th norm twice: first for random variables X_i with any fixed choice of X_j' which gives $\|\sum_{i\neq j} X_i X_j'\|_d \leqslant 2\|\sum_{i\neq j} X_i \sigma_i X_j'\|_d$ (here we use the independence of X_i and X_j') and second for random variables X_j' under the fixed values of $X_i\sigma_i$) which gives $\|\sum_{i\neq j} X_i X_j'\|_d \leqslant 4\|\sum_{i\neq j} X_i\sigma_i X_j'\sigma_j'\|_d$ (σ_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

$$\mathbf{E}[(\sum_{i\neq j} X_i X_j')^d | (X_j')] = \sum_{(d_i)} \binom{d}{2d_1 \ldots 2d_n} \prod_i (\sum_{j\neq i} X_j')^{2d_i} \mathbf{E} X_i^{2d_i}.$$

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

$$\mathbf{E}(\sum_{j\neq i} X_j')^d \leqslant \mathbf{E}(\sum_j X_j')^d.$$

Combining the last two bounds gives

$$\begin{aligned}
\mathbf{E}(\sum_{i\neq j} X_i X_j')^d &\leqslant \mathbf{E}_{(X_j')}[\mathbf{E}[(\sum_{i\neq j} X_i X_j')^d | (X_j')]] \\
&\leqslant \sum_{(d_i)} \binom{d}{2d_1 \ldots 2d_n} \mathbf{E}[\prod_i (\sum_j X_j')^{2d_i} X_i^{2d_i}] \\
&\leqslant \mathbf{E}(\sum_i (\sum_j X_j') X_i)^d \\
&= \mathbf{E}(\sum_i X_i)^d (\sum_j X_j')^d \\
&= \mathbf{E}(\sum_i X_i)^{2d}
\end{aligned}$$

which can be stated as

$$\| \sum_{i \neq j} X_i X'_j \|_d \leqslant \| \sum_i X_i \|_d^2. \tag{14}$$

By combining Equation (13) and Equation (14), and keeping in mind that X_i above are the symmetrized versions of original random variables, we obtain that for original (only centered) random variables X_i

$$\mathbf{E}\| \sum_{j \neq i} X_i X_j \|_d \leqslant 16 \mathbf{E}\| \sum_{j \neq i} X_i \sigma_i \|_d$$

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

3.2 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

$$\mathbf{E}(\sum_i x_i X_i)^d = \sum_{(d_i)} \binom{d}{2d_1 \ldots 2d_n} p^{\|(d_i)\|_0} \prod_i x_i^{2d_i}$$

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

$$S_k^{[d]}(x) = \sum_{i_1 \neq \ldots \neq i_k} x_{i_1}^{2d_1} \ldots x_{i_k}^{2d_k}$$

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

$$S_k^{[d]}(x) \leqslant v^{2\sum_{i:d_i>1}(d_i-1)} \sum_{i_1 \neq \ldots \neq i_k} x_{i_1}^2 \ldots x_{i_k}^2.$$

Here we have used the fact that $v_d(x)$ is increasing in d, so $v_k(x) \leqslant v$ when $k \leqslant d$; this follows from seeing $v_d(x)$ as the power mean of order $d-2$ and weights $x_i^2 / \sum_{i \notin I} x_i^2$ [32, 69].

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 $S_k(x) = \sum_{i_1 \neq \ldots \neq i_k} x_{i_1}^2 \ldots x_{i_k}^2$ are elementary symmetric polynomials in variables $y_i = x_i^2$ where $\sum_i y_i = \sum_i x_i^2 = 1$, 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 $N \geqslant k$ and since $k \leqslant d$, using Stirling's approximation [61] we obtain

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

Clearly we also have $S_k(x) \leqslant 1$ for any x. Thus if we replace (x_i) by a sequence such that $x_i = v$ for $K = v^{-2}$ values of i (e.g., flat) we loose at most a factor of $\Theta(1)^k \leqslant \Theta(1)^d$ in the upper bound. ◀

3.3 Moments of IID Sums

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

$$
\begin{aligned}
\mathbf{E}(\sum_i X_i)^d &= \sum_{d_i:d_1+\ldots+d_n=d, d_i\geqslant 2} \binom{d}{d_1\ldots d_n}\prod_i \mathbf{E}X_i^{d_i} \\
&\leqslant \sum_{d_i:d_1+\ldots+d_n=d, d_i\geqslant 2} \prod_i \binom{d}{d_i}\mathbf{E}X_i^{d_i} \\
&\leqslant \sum_{d_i\geqslant 2}\prod_i \binom{d}{d_i}\mathbf{E}X_i^{d_i} \\
&\leqslant \left(\sum_{k=2}^{d}\binom{d}{k}\|X\|_k^k\right)^n .
\end{aligned}
$$

Applying this for $X_i := X_i/t$ we have for any $t > 0$

$$
\mathbf{E}(t^{-1}\sum_i X_i)^d \leqslant \left(\sum_{k=2}^{d}\binom{d}{k}\|X\|_k^k/t^k\right)^n .
$$

Thus $\|\sum_i X_i\|_d \leqslant \mathrm{e}t$ for any t such that the right-hand side is at most e, equivalently

$$
\sum_{k=2}^{d}\binom{d}{k}\|X\|_k^k/t^k \leqslant \exp(d/n) - 1
$$

which is satisfied for

$$
t = 2\max_{k=2\ldots d}\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k}\|X\|_k .
$$

This proves the first part. Observe that for $k \geqslant 2$ we have

$$
\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k} \leqslant \frac{\mathrm{e}d}{k\exp(d/kn)}\cdot\frac{1}{(1-\exp(-1))^{1/2}}
$$

where we use the elementary inequalities $\binom{d}{k} \leqslant (de/k)^k$ and $\exp(u) - 1 \geqslant \exp(u)\cdot(1-\mathrm{e}^{-1})$ for $u \geqslant 1$. The function $u \to u/\exp(u)$ decreases for $u \geqslant 1$; applying this to $u = d/kn$ gives

$$
\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k} \leqslant \frac{\mathrm{e}n}{(1-\mathrm{e}^{-1})^{1/2}}, \quad k \leqslant d/n.
$$

Since $\|X\|_k$ increases in k we have

$$
\max_{k=2\ldots d, k\leqslant d/n}\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k}\|X\|_k \leqslant \frac{\mathrm{e}n\|X\|_{d/n}}{(1-\mathrm{e}^{-1})^{1/2}}.
$$

We have $(\exp(d/n)-1)^{-1/k} \leqslant (d/n)^{-1/k}$ due to the elementary inequality $\exp(u) - 1 \geqslant u$, and $\binom{d}{k} \leqslant (d\mathrm{e}/k)^k$ for any k. This gives

$$
\max_{k=2\ldots d}\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k}\|X\|_k \leqslant \mathrm{e}\max_{k=2\ldots d} d/k\cdot(n/d)^{1/k}\cdot\|X\|_k
$$

When $d/n \geqslant 2$ 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

$$
\max_{k=2\ldots d, k\leqslant d/n}\binom{d}{k}^{1/k}(\exp(d/n)-1)^{-1/k}\|X\|_k \leqslant C\max_{k=2\ldots d, k>d/n} d/k\cdot(n/d)^{1/k}\cdot\|X\|_k
$$

with $C = \frac{\mathrm{e}}{(1-\mathrm{e}^{-1})^{1/2}}$, which completes the proof. ◀

3.4 Binomial Moments

Proof of Lemma 11. Applying Lemma 9 we obtain

$$\|S - \mathbf{E}S\|_d \leqslant O(1) \cdot \max\left\{(d/k) \cdot (Kp/d)^{1/k} : \max(2, d/K) \leqslant k \leqslant d\right\}.$$

because $S \sim \sum_i X_i$ where $X_i \sim \mathrm{Bern}(p)$ and $\|X_i - \mathbf{E}X_i\|_d = (p(1-p)^{d-1} + (1-p)p^{d-1})^{1/d}$ so that $\|X_i - \mathbf{E}X_i\|_d = \Theta(p)^{1/d}$ for $p \leqslant 1/2$.

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 $\max(2, d/K) \leqslant \log(1/a) \leqslant d$ (that is, inside of the interval) we have necessarily $a \leqslant \mathrm{e}^{-2} < 1$ 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 $k = \max(2, d/K)$ regardless whether $a < 1$ or $a \geqslant 1$. ◀

4 Conclusion

We have proven novel bounds for sparse random projections, showing that the performance depends on the data statistic closed to *Renyi entropy*. Some intereging 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

1. Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, 2001.
2. Thomas D Ahle. Asymptotic tail bound and applications. 2017.
3. Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, 2006.
4. Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.
5. SN Bernshtein. Probability theory {in Russian}. *Gosizdat, Moscow-Leningrad*, 1927.
6. Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.
7. Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.
8. Henry W Block, Thomas H Savits, Moshe Shaked, et al. Some concepts of negative dependence. *The Annals of Probability*, 10(3):765–772, 1982.
9. Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 410–419. IEEE, 2012.
10. Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, Pascal Massart, et al. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560, 2005.
11. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
12. Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems*, pages 298–306, 2010.

13 V Buldygin and K Moskvichova. The sub-gaussian norm of a binary random variable. *Theory of probability and mathematical statistics*, 86:33–49, 2013.

14 Christian Cachin. Smooth entropy and rényi entropy. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 193–208. Springer, 1997. https://link.springer.com/chapter/10.1007/3-540-69053-0_14.

15 Herman Chernoff et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

16 Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

17 M Lawrence Clevenson and William Watkins. Majorization and the birthday inequality. *Mathematics Magazine*, 64(3):183–188, 1991.

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

19 Michael B Cohen, TS Jayram, and Jelani Nelson. Simple analyses of the sparse johnson-lindenstrauss transform. In *1st Symposium on Simplicity in Algorithms (SOSA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

20 Harald Cramér. On a new limit theorem of the theory of probability. *Uspekhi Matematicheskikh Nauk*, (10):166–178, 1944.

21 Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350, 2010.

22 Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

23 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *International Computer Science Institute, Technical Report*, 22(1):1–5, 1999.

24 Victor H de la Peña and Stephen J Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, pages 806–816, 1995.

25 Devdatt P Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), 1996.

26 Morris L Eaton. A note on symmetric bernoulli random variables. *The annals of mathematical statistics*, 41(4):1223–1226, 1970.

27 Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

28 Peter Frankl and Hiroshi Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 1988.

29 Peter Frankl and Hiroshi Maehara. Some geometric applications of the beta distribution. *Annals of the Institute of Statistical Mathematics*, 42(3):463–474, 1990.

30 Casper B Freksen, Lior Kamma, and Kasper Green Larsen. Fully understanding the hashing trick. In *Advances in Neural Information Processing Systems*, pages 5389–5399, 2018.

31 David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

32 G.H. Hardy, Karreman Mathematics Research Collection, J.E. Littlewood, G. Pólya, G. Pólya, and D.E. Littlewood. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952. URL: https://books.google.at/books?id=t1RCSP8YKt8C.

33 Paweł Hitczenko. Domination inequality for martingale transforms of a rademacher sequence. *Israel Journal of Mathematics*, 84(1-2):161–178, 1993.

34 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

35 Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

36 Meena Jagadeesan. Understanding sparse jl for feature hashing. In *Advances in Neural Information Processing Systems*, pages 15203–15213, 2019. https://arxiv.org/pdf/1903.03605.pdf.

37 Thathachar S Jayram and David P Woodruff. Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms (TALG)*, 9(3):1–17, 2013.

38 William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

39 William B Johnson and Assaf Naor. The johnson–lindenstrauss lemma almost characterizes hilbert space, but not quite. *Discrete & Computational Geometry*, 43(3):542–553, 2010.

40 Daniel Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit johnson-lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 628–639. Springer, 2011.

41 Daniel M Kane and Jelani Nelson. A derandomized sparse johnson-lindenstrauss transform. *arXiv preprint arXiv:1006.3585*, 2010.

42 Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

43 Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the johnson-lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71, 2013.

44 Michael Kerber and Sharath Raghvendra. Approximation and streaming algorithms for projective clustering via random projections. *arXiv preprint arXiv:1407.2063*, 2014.

45 Aleksandr Khintchine. Über dyadische brüche. *Mathematische Zeitschrift*, 18(1):109–116, 1923.

46 B Klartag and Shahar Mendelson. Empirical processes and random projections. *Journal of Functional Analysis*, 225(1):229–245, 2005.

47 Andreas Knoblauch. Closed-form expressions for the moments of the binomial probability distribution. *SIAM Journal on Applied Mathematics*, 69(1):197–204, 2008.

48 Konrad Kolesko and Rafał Latała. Moment estimates for chaoses generated by symmetric random variables with logarithmically convex tails. *Statistics & Probability Letters*, 107:210–214, 2015.

49 Samory Kpotufe and Bharath Sriperumbudur. Gaussian sketching yields a jl lemma in rkhs. In *International Conference on Artificial Intelligence and Statistics*, pages 3928–3937, 2020.

50 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

51 Rafał Latała. Tail and moment estimates for some types of chaos. *Studia mathematica*, 135(1):39–53, 1999.

52 Rafał Latała et al. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 1997.

53 Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, 2006.

54 Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

55 John E Littlewood. On the probability in the tail of a binomial distribution. *Advances in Applied Probability*, 1(1):43–72, 1969.

56 Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1027–1038, 2019.

57 Jiří Matoušek. On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

58 Brendan D McKay. On littlewood's estimate for the binomial distribution. *Advances in Applied Probability*, 21(2):475–478, 1989.

59 Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pages 117–126. IEEE, 2013.

60 Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

61 Herbert Robbins. A remark on stirling's formula. *The American mathematical monthly*, 62(1):26–29, 1955.

62 Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

63 Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE, 2006.

64 Pantelimon Stanica. Good lower and upper bounds on binomial coefficients. *Journal of Inequalities in Pure and Applied Mathematics*, 2(3):30, 2001.

65 Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

66 Roman Vershynin. A simple decoupling inequality in probability theory. *preprint*, 2011.

67 Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

68 Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120, 2009.

69 Alfred Witkowski. A new proof of the monotonicity of power means. *J. Ineq. Pure and Appl. Math*, 5(1), 2004.

70 Shuheng Zhou. Sparse hanson–wright inequalities for subgaussian quadratic forms. *Bernoulli*, 25(3):1603–1639, 2019. appears in 2015 at https://arxiv.org/pdf/1510.05517.pdf.

A Some remarks on prior works

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

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

There is a minor mistake in splitting the branches: they emerge from taking the derivative test of the function $d^2v^2u^{-2}(p/dv^2)^{1/u}$ where $1 \leqslant u \leqslant d/2$ (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 $2 \leqslant \log(dv^2/p)$ and $\log(dv^2/p) \leqslant d$. Thus the splitting conditions should be 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.

B Concluding Main Theorem

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

$$I_1 \triangleq \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot qp : \log(q/Kp) \leqslant q/K \leqslant 2, 2 \leqslant q \leqslant d \right\}$$
$$I_2 \triangleq \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot K(Kp^{2K/q})^2 : \log(q/Kp) \leqslant 2 \leqslant q/K, 2 \leqslant q \leqslant d \right\}$$
$$I_3 \triangleq \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot K^{-1} q^2/\log^2(q/Kp) : \max(2, q/K) \leqslant \log(q/Kp) \leqslant q, 2 \leqslant q \leqslant d \right\}$$
$$I_4 \triangleq \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot K^{-1}(Kp)^{2/q} : q \leqslant \log(q/Kp), 2 \leqslant q \leqslant d \right\}$$

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

$$\| \sum_{r=1}^{m} Z_r \|_d \leqslant O(\max(I_1, I_2, I_3, I_4))$$

The goal is to prove that for $t = s\epsilon$ we have

$$\| \sum_{r=1}^{m} Z_r \|_d \leqslant t/\mathrm{e} \tag{15}$$

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.0.1 First Branch

We will show the following bound

▶ **Lemma 20.** *We have*

$$I_1 \leqslant O(dmp^2)^{1/2}.$$

Proof of Lemma 20. We have

$$I_1 = \max_q \left\{ pd(m/d)^{1/q} : \log(q/Kp) \leqslant q/K \leqslant 2, 2 \leqslant q \leqslant d \right\}$$
$$\leqslant (dmp^2)^{1/2}$$

where the inequality follows because $m \geqslant d$ and $1/q \leqslant \frac{1}{2}$ (for q satisfying the constraints). This completes the proof. ◀

B.0.2 Second Branch

We will show the following bound

▶ **Lemma 21.** *For $p \leqslant 2\mathrm{e}^{-2}$ we have*

$$I_2 \leqslant (dmp^2)^{1/2}.$$

Proof of Lemma 20. For q satisfying the constraint we have $K/q \geqslant \mathrm{e}^{-2}/p$ which, due to $p \leqslant 2\mathrm{e}^{-2}$, implies $K/q \geqslant 1/2$. Then $p^{2K/q} \leqslant p$ (recall that $p < 1$!) and thus

$$I_2 \leqslant \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot Kp : \log(q/Kp) \leqslant 2 \leqslant q/K, 2 \leqslant q \leqslant d \right\}.$$

For q within the constraints we have $K/q \leqslant \frac{1}{2}$ and therefore

$$I_2 \leqslant \frac{p}{2} \max_q \left\{ d \cdot (m/d)^{1/q} : \log(q/Kp) \leqslant 2 \leqslant q/K, 2 \leqslant q \leqslant d \right\}.$$

Since $m/d \geqslant 1$ the expression under the maximum decreases with q, thus is not bigger than the value at $q = 2$. Thus $I_2 \leqslant p(dm)^{1/2}/2$ and the result follows. ◀

B.0.3 Fourth Branch

We will prove the following bound

▶ **Lemma 22.** *We have*

$$I_4 \leqslant \begin{cases} (dmp^2)^{1/2} & \log(dv^4/mp^2) \leqslant 2 \\ dv^2/\log(dv^4/mp^2) & \log(dv^4/mp^2) > 2 \end{cases}.$$

Proof of Lemma 22. We have

$$I_4 = \max_q \left\{ K^{-1} \cdot d/q \cdot (K^2p^2m/d)^{1/q} : q \leqslant \log(q/Kp), 2 \leqslant q \leqslant d \right\}.$$

Let $a = K^2p^2m/d$, the expression under the maximum is proportional to $1/q \cdot a^{1/q}$. We now apply Proposition 15: for $a \geqslant 1$ the maximum is not bigger than the value at $q = 2$, so

$$I_4 \leqslant (dmp^2)^{1/2}.$$

We now can assume $a < 1$, equivalent to $K^2p^2m < d$. The global maximum is at $q = \log(1/a)$, thus our maximum is still at $q = 2$ when $\log(1/a) \leqslant 2$ and otherwise is not bigger than the value at $q = \log(1/a)$. We then obtain

$$I_4 \leqslant K^{-1}d/\log(d/mp^2K^2) \leqslant K^{-1}d = dv^2.$$

This complete the proof. ◀

B.0.4 Third Branch

We will show the following bound

▶ **Lemma 23.** *Suppose that $v^2 \geqslant s\epsilon/d^2$, then*

$$I_3 \leqslant O(dmp^2)^{1/2} + O(dv/\log(dv^2/p))^2$$

Proof of Lemma 23. The proof is based on splitting the maximum into three regimes: $q \in [2, 3]$,$3 \leqslant q \leqslant \log(m/d)$ and $\log(m/d) \leqslant q \leqslant d$. Define

$$I^0 = \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot v^2q^2/\log^2(qv^2/p) : 2 \leqslant \log(qv^2/p) \leqslant q \leqslant d, 2 \leqslant q \leqslant 3 \right\}$$

$$I^- = \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot v^2q^2/\log^2(qv^2/p) : 2 \leqslant \log(qv^2/p) \leqslant q \leqslant d, 3 \leqslant q \leqslant \log(m/d) \right\}$$

$$I^+ = \max_q \left\{ d/q \cdot (m/d)^{1/q} \cdot v^2q^2/\log^2(qv^2/p) : 2 \leqslant \log(qv^2/p) \leqslant q \leqslant d, \log(m/d) \leqslant q \leqslant d \right\}$$

so that we have $I_3 \leqslant \max(I^0, I^+, I^-)$ (for convenience we replace the constraint $\max(2, qv^2) \leqslant \log(qv^2/p)$ in I_3 by the weaker one $2 \leqslant \log(qv^2/p)$). By the assumptions we have $v^2/p \geqslant m\epsilon/d^2$. Since $m \geqslant d\epsilon^{-2}$ we have $\epsilon \geqslant (d/m)^{1/2}$, and thus

$$v^2/p \geqslant (m/d)^{1/2} \cdot d^{-1}.$$

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

Proof of Claim. For any q satisfying the restrictions it holds that

$$\begin{aligned} q &\geqslant \log(v^2/p) \\ &\geqslant \frac{\log(m/d)}{2} - \log d \\ &\geqslant \frac{\log(m/d)}{12}. \end{aligned}$$

We then have $(m/d)^{1/q} \leqslant O(1)$ and thus

$$I^- \leqslant \max_q \left\{ d \cdot qv^2/\log^2(qv^2/p) : 2 \leqslant \log(qv^2/p) \leqslant q \leqslant d, 3 \leqslant q \leqslant \log(m/d) \right\}$$

Considering the auxiliary function $u \to u/\log^2 u$ with $u = qv^2/p \geqslant \mathrm{e}^2$, 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

$$I^- \leqslant d^2v^2/\log^2(dv^2/p)$$

and completes the proof. ◀

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

Proof of Claim. We have that $dv^2/p \geqslant m\epsilon/d \geqslant (m/d)^{1/2}$ and therefore

$$\begin{aligned} I^- &\leqslant dv^2 d(m/d)^{1/3}\log(m/d) \\ &\leqslant dv^2(m/d)^{5/12}/\log^2(m/d) \\ &\leqslant dv^2(m/d)^{5/12}/\log^2(dv^2/p) \\ &\leqslant O(d^2v^2/\log^2(dv^2/p)). \end{aligned}$$

which completes the proof. ◀

▷ Claim 26. We have $I^+ \leqslant O(d^2v^2/\log^2(dv^2/p))$

Proof of Claim. We have $(m/d)^{1/q} \leqslant \mathrm{e}$ for $q \geqslant \log(m/d)$, thus

$$I^+ \leqslant d \cdot \max_q\{qv^2/\log^2(qv^2/p) : \ 2 \leqslant \max(\log(qv^2/p), \log(m/d)) \leqslant q \leqslant d\}$$

Considering the auxiliary function $u \to u/\log^2 u$ with $u = qv^2/p \geqslant \mathrm{e}^2$, 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

$$I^+ \leqslant O(d^2v^2/\log^2(dv^2/p))$$

and the claim follows. ◀

▷ Claim 27. We have $I^0 \leqslant O((dmp^2)^{1/2})$.

Proof of Claim. We have $I^0 \leqslant O(v^2(md)^{1/2})$ because $(m/d)^{1/q} \leqslant (m/d)^{1/2}$ (due to $m/d \geqslant 1$ and $q \geqslant 2$). However for $q \in [2,3]$ the constraint $\log(qv^2/p) \leqslant q$ gives $v^2 \leqslant O(p)$. Thus

$$I^0 \leqslant O(p(md)^{1/2})$$

which completes the proof. ◀

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

B.0.5 Merging Branch Bounds

To conclude the main result it suffices to satisfy

$$c \cdot \max(I_1, I_2, I_3, I_4) \leqslant s\epsilon \tag{16}$$

for some absolute constant c. The condition in Equation (16) for I_1, I_2 is equivalent to $c \cdot (dmp^2)^{1/2} \leqslant s\epsilon$, which holds when

$$m \geqslant \Omega(d\epsilon^{-2}). \tag{17}$$

To satisfy Equation (16) for I_4 we require, in addition to Equation (17), that $cdv^2 \leqslant s\epsilon$, equivalent to

$$v \leqslant O((s\epsilon)^{1/2}/d^{1/2}). \tag{18}$$

Finally, in order to satisfy Equation (16) for I_3 we observe that, under the restriction

$$v^2 \geqslant s\epsilon/d^2 \tag{19}$$

the bound in Lemma 23 gives

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

which follows because $\log(dv^2/p) \geqslant \log(s\epsilon/dp) = \log(m\epsilon/d)$. Thus in addition to Equation (17) and it suffices that

$$v \leqslant O((s\epsilon)^{1/2}\log(m\epsilon/d)/d) \tag{20}$$

Now observe that for

$$v = \Theta(s\epsilon)^{1/2}\min(\log(m\epsilon/d)/d, 1/d^{1/2}) \tag{21}$$

the condition in is automatically satisfied. Thus the theorem holds for v as above, and clearly for any smaller v.