The Performance of Johnson-Lindenstrauss Transforms: Beyond the Classical Framework

Meena Jagadeesan

Undergraduate thesis submitted in partial fulfillment of the degree of Bachelor of Arts in Computer Science/Mathematics at Harvard College.

Advisor: Prof. Jelani Nelson

April 2020

Abstract

Euclidean dimensionality reduction is a commonly used technique to scale up algorithms in machine learning and data science. The goal of Euclidean dimensionality reduction is to reduce the dimensionality of a dataset, while preserving Euclidean distances between data points. Given the high-dimensionality of modern datasets, Euclidean dimensionality reduction serves as an effective pre-processing technique: it enables a significant speedup of computational tasks (such as clustering and nearest neighbors) while preserving their accuracy. Beginning with a seminal work by Johnson and Lindenstrauss in the 1980s, Euclidean dimensionality reduction has been studied for decades in the theoretical computer science and mathematics literatures. Recently, the performance of Euclidean dimensionality reduction has been studied in settings that depart from the classical framework, motivated by machine learning and neuroscience considerations.

In this undergraduate thesis, we continue the study of how Euclidean dimensionality reduction performs in settings beyond the classical framework. Our first result is a characterization of the performance of the standard dimensionality reduction schemes (called sparse Johnson-Lindenstrauss transforms) on "well-behaved" datasets; we generalize a line of work in the machine learning literature initiated by Weinberger et al. (ICML '09). Our second result is an analysis of neuroscience-motivated dimensionality reduction schemes (called sparse, sign-consistent Johnson-Lindenstrauss transforms); we build on work by Allen-Zhu et al. (PNAS '15). A shared technical theme between our results is a new perspective on analyzing Johnson-Lindenstrauss transforms.

Comments on Published Work

Our results in this undergraduate thesis have been published at NeurIPS 2019 [23] and at RANDOM 2019 [22].

Contents

Acknowledgments

First and foremost, I would like to thank Prof. Jelani Nelson, for being an incredible advisor during my undergraduate years. I first became interested in computer science because of Jelani's algorithms and complexity course during my freshman year, and my first research experience in computer science was advised by Jelani that summer. Since then, I really can't quantify how much he has taught me about algorithms, research, and mathematical thinking. For that, and for his constant encouragement, I'm very grateful.

I am also grateful to many other faculty members, including Prof. Shuchi Chawla, Prof. Cynthia Dwork, Prof. Stratos Idreos, Prof. Scott Kominers, and Prof. James Mickens, for advising me on research and helping shape my perspective of computer science. I would also like to thank Prof. Eddie Kohler and Prof. David Parkes for their incredible advice and support, and Prof. Boaz Barak for being a thesis reader. Additionally, I would like to gratefully acknowledge travel and research funding grants from the Harvard College Research Program, the Herchel Smith-Harvard Fellowship, the Harvard Program for Research in Science and Engineering (PRISE), and the Harvard Computer Science department.

Lastly, I would like to thank all of my friends for making college so amazing; my Kali Praxi family and the Dunster community for so many fun times; Katherine for being there for me through it all; Alex for endlessly encouraging me and for always making me smile; and my family, for always believing in me.

In memory of my grandfather:

Krishnaiyer Jagadeesan.

Chapter 1

Introduction

The *high-dimensionality* of modern datasets has posed significant computational challenges in machine learning and data science. For example, machine learning tasks such as text-based classification and image clustering often must handle data with hundreds of thousands of features (i.e. dimensions). The fundamental issue is that algorithms for computational tasks often suffer from the *curse of dimensionality*: that is, the runtime and memory usage grow quickly with the dimension of the data. As a result, these algorithms incur a prohibitively high computational cost on high-dimensional data.

Dimensionality reduction is a powerful pre-processing technique that enables a drastic speedup of these algorithms without compromising their accuracy. The key idea of dimensionality reduction is to produce a low-dimensional projection of a dataset that preserves of the "geometry" of the original dataset. The algorithm is then given this projection of the dataset, and since the projection has low dimension, the algorithm can achieve a fast runtime and low memory usage. Moreover, the algorithm's accuracy is often preserved since the projected dataset retains aspects of the "geometry" of the original dataset. For example, in settings such as clustering and nearest neighbors, the algorithm only needs to understand the *distances* between data points. These distances are preserved by the projected dataset, thus enabling the algorithm to achieve a high accuracy.

In this undergraduate thesis, we investigate the design and analysis of dimensionality reduction schemes that preserve the geometry of datasets. We focus on a fundamental goal, called *Euclidean dimensionality reduction*: reduce the dimensionality of a dataset with real-valued entries, while preserving the Euclidean distances between data points. In particular, we consider Euclidean dimensionality reduction in settings motivated by machine learning and neuroscience considerations, and prove new results for each of these settings. (Our results have been published in the following papers [23, 22].)

1.1 History of Euclidean dimensionality reduction

Dimensionality reduction for Euclidean distances has been studied since the 1980s. Given a set of n-dimensional points with real-valued entries, the classical goal is to project to a set of m-dimensional points such that distances are preserved. More specifically, the goal is as follows: given a set $S = \{x_1, \ldots, x_N\}$ of data points in $\mathbb{R}^n$, construct a function $f : S \to \mathbb{R}^m$

such that for any pairs of points $x_i, x_j \in S$, the distance between x_i and x_j is close to the distance between $f(x_i)$ and $f(x_j)$.

The primary objective is to make the projected dimension m as small possible. Surprisingly, a seminal result in the mathematics literature by Johnson and Lindenstrauss [25] showed that it is possible to achieve a projected dimension m that is independent of the original dimension n. (In particular, it's possible to achieve m that grows logarithmically with the number of points N and grows inverse-polynomially with the permitted error in distance preservation.) The standard construction of these dimensionality reduction schemes are based on *random projections*. The idea is to project the original dataset using a random matrix A drawn from a probability distribution over $m \times n$ dimensional matrices. In particular, the (random) matrix A is used to project each data point $x \in S$ to a m-dimensional point $Ax \in \mathbb{R}^m$. The random projection is constructed so that with high probability, the projected data points preserve distances. Random projections that achieve this distance-preserving condition are called *Johnson-Lindenstrauss transforms*.

In the past two decades, computer scientists have become particularly interested in dimensionality reduction for Euclidean distances as a pre-processing step for algorithmic tasks. The algorithmic setting creates a computational constraint on the Johnson-Lindenstrauss transforms: the projection time of applying the (random) matrix A to a data point $x \in \mathbb{R}^n$ needs to be low. In particular, computing Ax needs to be fast, so that pre-processing doesn't create too much of an overhead in computational cost. Unfortunately, the Johnson-Lindenstrauss transforms considered in [25] do not lend themselves well to efficient projection.

A line of work [2, 46, 3, 33, 15, 28] in the theoretical CS literature designed Johnson-Lindenstrauss transforms that achieve faster projection. An elegant way to achieve a fast projection time is to make the (random) matrix A *sparse* (i.e make A have few nonzero entries per column). In this context, it is useful to consider random projections over sparse matrices. The state-of-the-art Johnson-Lindenstrauss transforms over sparse matrices, called *sparse Johnson-Lindenstrauss transforms*, were constructed and analyzed by Kane and Nelson [28]. Sparse Johnson-Lindenstrauss transforms simultaneously allow for substantial dimensionality reduction and a fast projection time, achieving a near-optimal sparsity in some contexts [37].

1.2 Beyond the classical framework

In this undergraduate thesis, we consider how Johnson-Lindenstrauss transforms perform in two settings that depart from the classical framework. The first setting explores whether better dimensionality reduction is possible when the dataset is known to be well-behaved; the second setting explores whether dimensionality reduction is still possible under neuroscience-motivated restrictions. Our main contribution is new results for each of these settings. Moreover, as a shared technical theme between these two settings, we offer a new perspective on analyzing random projections.

1.2.1 Setting 1: "Well-behaved" datasets

The classical results in the CS theory literature consider dimensionality reduction schemes that need to preserve distances even on "badly behaved" datasets. Nonetheless, these "badly behaved" datasets may not arise in many machine learning tasks. In the spirit of "beyond worst-case analysis"[1], a natural question to ask is whether better and faster dimensionality reduction can be achieved, when the dataset is known to be "well-behaved". A line of work [46, 15, 27, 13, 28, 16] in the machine learning literature considers this question in the context of a restricted family of dimensionality reduction schemes (called "feature hashing"). This line of work shows that feature hashing can achieve a much lower projected dimension when the dataset is known to have a certain structure.

Our contribution. We generalize the previous line of work [46, 15, 27, 13, 28, 16] to a much larger family of dimensionality reduction schemes: the family of sparse Johnson-Lindenstrauss transforms. We demonstrate that sparse Johnson-Lindenstrauss transforms can achieve much better and faster dimensionality reduction on well-behaved datasets. In particular, we prove tight bounds on the tradeoff between projected dimension, sparsity of the random matrix, accuracy, and ℓ_∞-to-ℓ_2 norm ratio of the data points (i.e. the formal measure of the extent to which data is "well-behaved").

We show the following:

Theorem 1.1 (Informal) Consider a sparse Johnson-Lindenstrauss transform with projected dimension m and sparsity s. For accuracy parameters ϵ and δ, let $v(m, \epsilon, \delta, s)$ be the maximal value in $[0, 1]$ such that the sparse Johnson-Lindenstrauss transform achieves Euclidean norm preservation on vectors in $S_v = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq v \|x\|_2\}$. Then, $v(m, \epsilon, \delta, s)$ is of the following form:

$$v(m, \epsilon, \delta, s) = \begin{cases} 1 & \text{High } m \\ \sqrt{s}B_1 & \text{Medium-high } m \\ \sqrt{s}\min(B_1, B_2) & \text{Medium-low } m \\ 0 & \text{Small } m, \end{cases}$$

where $p = \ln(1/\delta)$, $B_1 = \sqrt{\ln(m\epsilon^2/p)}/\sqrt{p}$ and $B_2 = \ln(m\epsilon/p)/p$.

(See Theorem 3.2 for a formal statement of the result.)

Theorem 1.1 takes a step towards explaining why dimensionality reduction schemes in practice can outperform the classical results. We also show empirical results that confirm the findings in Theorem 1.1 on synthetic and real-world data.

1.2.2 Setting 2: Neuroscience-motivated restrictions

Dimensionality reduction provides a promising framework to model information compression in the brain, where a large number of input neurons are connected to a small number of

[1] As discussed in a survey by Roughgarden [41], worst-case analysis captures the worst-possible performance of an algorithm on any given input (i.e. in our setting, any set S of vectors in $\mathbb{R}^n$). "Beyond worst-case analysis" is about identifying properties of real-world inputs and proving stronger guarantees for inputs with these properties.

target neurons, but somehow, information is sufficiently preserved. In particular, the data at the input neurons can be modeled as a high-dimensional vector, and the data at the output neurons can be modeled as a low-dimensional vector [18]. The challenge is that the dimensionality reduction schemes need to satisfy constraints imposed by the biological restrictions of these convergent pathways (i.e. neurons are usually either purely excitatory or purely inhibitory). Previous work [4] demonstrates that a low projected dimension is indeed possible, even with these restrictions, using a *sparse, sign-consistent Johnson-Lindenstrauss transform.*

Our contribution. We generalize the analysis of sparse, sign-consistent Johnson-Lindenstrauss transforms to a larger class of dimensionality reduction parameters, thus capturing a greater spectrum of potential biological restrictions. In particular, we prove dimension-sparsity tradeoffs, showing that a lower sparsity is possible, with an appropriate gain in projected dimension.

We show the following:

Theorem 1.2 (Informal) Consider a sparse, sign-consistent Johnson-Lindenstrauss transform with projected dimension m and sparsity s. For accuracy parameters ϵ and δ and for any constant $e \leq B \leq \delta^{-1}$, Euclidean norm preservation is achieved if $s = \Theta(\epsilon^{-1} \log_B(1/\delta))$ and $m = \Theta(B\epsilon^{-2} \log_B^2(1/\delta))$.

(See Theorem 4.4 for a formal statement.)

Theorem 1.2 generalizes the result in [4], which is limited to the case of $B = e$. In particular, Theorem 1.2 demonstrates that a reduction in sparsity is possible with an exponential gain in dimension.

1.2.3 Methods for analyzing random projections

Standard analyses of random projections utilize complicated combinatorial arguments [28, 36, 16, 4]. Moreover, these arguments are fragile in that new combinatorial techniques are needed to prove each new result. (All of the aformentioned papers involve different combinatorial techniques.)

Our contribution. Our work offers a new conceptual perspective on analyzing random projections. In particular, we give a "unified" method for analyzing a class of random projections that obviates the need for constructing new techniques for each setting. We propose a non-combinatorial approach[2] based on results from the probability theory literature; we believe our approach is cleaner than the standard combinatorial approaches.

1.3 Outline for the rest of the undergraduate thesis

In Chapter 2, we present the mathematical framework and classical results for dimensionality reduction. In Chapter 3, we describe our results on the behavior of dimensionality reduction schemes on well-behaved datasets (Setting 1). In Chapter 4, we describe our results on

[2]Our methods borrow some intuition from previous non-combinatorial methods [12, 11]. However, previous approaches turn out not be sufficiently precise for our settings; our bounds have the advantage of being sufficiently precise to recover tight bounds.

dimensionality reduction schemes with neuroscience-motivated constraints (Setting 2). In Chapter 5, we present our new perspective on analyzing Johnson-Lindenstrauss transforms. In Chapter 6, we describe proof sketches of our main results in Chapter 3 and Chapter 4. In the Appendix, we describe the details of the proofs for Chapter 3.

Chapter 2

Background: The Classical Setting

The mathematical goal of Euclidean dimensionality reduction is, for any given set S of N points living in $\mathbb{R}^n$, to design a function $f : S \to \mathbb{R}^m$ (for some $m \ll n$) that preserves Euclidean distances between points.[1] The distance-preserving requirement is that for every $x_i, x_j \in S$, it must hold that $\|f(x_i) - f(x_j)\|_2 \approx \|x_i - x_j\|_2$. The exact form of the distance-preserving requirement is parameterized by a constant $\epsilon > 0$ as follows:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|f(x_i) - f(x_j)\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2 . \tag{2.1}$$

That is, the distance $\|f(x_i) - f(x_j)\|_2$ needs to approximate $\|x_i - x_j\|_2$ up to a muliplicative error of ϵ.

The Johnson-Lindenstrauss lemma [25], a cornerstone result in the dimensionality reduction literature, shows that there exists a function f that projects into $m = \Theta\left(\frac{\log N}{\epsilon^2}\right)$ dimensions. More specifically:

Lemma 2.1 ([25]) For any positive integers $n, N \geq 2$, any parameter $0 < \epsilon < 1$, and any set S of N points in $\mathbb{R}^n$, there exists a function $f : S \to \mathbb{R}^m$ with $m = \Theta\left(\frac{\log N}{\epsilon^2}\right)$ that achieves (2.1) with error ϵ.

In particular, the projected dimension m is independent of n, the original number of dimensions! Moreover, it is logarithmic in the number of points N in the dataset and grows inverse-polynomially with the error ϵ.

The proof of Lemma 2.1 relies on constructing a random projection (i.e. a probability distribution over $m \times n$ matrices) to avoid having to explicitly construct a function f for each set S. More specifically, in the proof of Lemma 2.1 in [25], a probability distribution $\mathcal{A}$ over $m \times n$ real matrices is constructed such that for any given pair of points $y_1, y_2 \in \mathbb{R}^n$, the distance $\|Ay_1 - Ay_2\|_2$ is close to $\|y_1 - y_2\|_2$ with high probability over the choice of A. Since A is linear, this requirement can be expressed as follows.

Requirement 2.2 For error $\epsilon \in (0,1)$ and failure probability $\delta \in (0,1)$, the requirement is that for each $x \in \mathbb{R}^n$:

$$\mathbb{P}_{A \in \mathcal{A}}[(1-\epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1+\epsilon)\|x\|_2] > 1 - \delta.$$

[1] For a vector $y = [y_1, \ldots, y_n] \in \mathbb{R}^n$, the Euclidean norm, also called the ℓ_2-norm, is defined so $\|y\|_2 = \sqrt{\sum_{i=1}^n y_i^2}$. The Euclidean distance between vectors $x, y \in \mathbb{R}^n$ is $\|x - y\|_2$.

We will often use the shorthand $\mathbb{P}_{A\in\mathcal{A}}[\|Ax\|_2 \in (1\pm\epsilon)\|x\|_2]$ to denote $\mathbb{P}_{A\in\mathcal{A}}[(1-\epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1+\epsilon)\|x\|_2]$.

Observe that random projections achieving Requirement 2.2 can yield the N-point version of the dimensionality reduction requirement for a set S as follows. Consider a probability distribution $\mathcal{A}$ that achieves Requirement 2.2 with failure probability $\delta = 0.01/N^2$ and error ϵ. Now, we can union bound over all N^2 differences $\{x_i - x_j\}_{x_i,x_j\in S}$ to see that with probability at least 0.99, it will hold that $(1-\epsilon)\|x_i - x_j\|_2 \leq \|Ax_i - Ax_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$ for all $x_i, x_j \in S$, so (2.1) is achieved. (In fact, when S has infinitely many points, random projections can sometimes still achieve (2.1), and thus achieve distance-preservation for S, although the proofs become significantly more involved [10, 11, 36, 6].)

Constructing dimensionality reduction schemes using random projections has a number of nice consequences.

1. The dimensionality reduction scheme is linear and thus enjoys a simple structure (i.e. Euclidean distance preservation boils down to Euclidean norm preservation).

2. The dimensionality reduction scheme is *data-oblivious* and thus can be constructed without seeing the dataset S ahead of time.

3. The projection time can be directly tuned via the sparsity of matrices in the support of the random projection.

Random projections that are successful in achieving Requirement 2.2 are sometimes referred to as *Johnson-Lindenstrauss transforms*.

2.1 The gaussian construction

The proof of Lemma 2.1 boils down to designing an appropriate random projection achieving Requirement 2.2. In particular, it turns out that there is a random projection with projected dimension $m = \Theta(\epsilon^{-2}\log(1/\delta))$ that achieves Requirement 2.2.

Lemma 2.3 (Distributional JL lemma [25]) For any positive integer n and parameters $0 < \epsilon, \delta < 1$, there exists a random projection over $m \times n$ matrices with $m = \Theta(\epsilon^{-2}\log(1/\delta))$ that satisfies Requirement 2.2.

After Johnson and Lindenstrauss's seminal paper [25], the proof of Lemma 2.3 has been distilled down to a random projection that lends itself to a simple analysis [45]. In particular, the random projection can be taken to be a probability distribution over $m \times n$ matrices with i.i.d. gaussian entries.

Proof. For $1 \leq i \leq m$ and $1 \leq j \leq n$, let $g_{i,j}$ be i.i.d. standard gaussians with mean 0 and variance 1. Now, the (i,j)th entry of A is defined to be the random entry $A_{i,j} = \frac{1}{\sqrt{m}} g_{i,j}$. We claim that for some $m = \Theta(\epsilon^{-2}\log(1/\delta))$, Requirement 2.2 is satisfied. Since A is linear, it suffices to consider each x in the unit ℓ_2 ball, i.e. such that $\|x\|_2 = 1$.

Since it is easier to reason about the ℓ_2^2 than ℓ_2, we consider $\|Ax\|_2^2 - 1$. It is straightforward to see that if $\mathbb{P}[|\,\|Ax\|_2^2 - 1| \leq \epsilon] > 1 - \delta$, then we know that $\mathbb{P}_{A\in\mathcal{A}}[\|Ax\|_2 \in$

$(1 \pm \epsilon') \|x\|_2] > 1 - \delta]$ is satisfied[2] for some $\epsilon' = \Theta(\epsilon)$. Thus, it suffices to consider the error term $\|Ax\|_2^2 - 1$. Notice that

$$\|Ax\|_2^2 - 1 = \frac{1}{m} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} g_{i,j} x_i \right)^2 - 1.$$

Let's consider $G_j = \sum_{i=1}^{n} g_{i,j} x_i$, Since $\|x\|_2 = 1$, each G_j is itself distributed as a gaussian with mean 0 and variance 1. Moreover, the independence of the $g_{i,j}$ tells us that $G_1, \ldots, G_m$ are independent. Thus, we can just consider

$$\frac{1}{m} \left(\sum_{j=1}^{m} G_j^2 \right) - 1 \sim \frac{1}{m} \chi_m^2 - 1,$$

where χ_m^2 is a chi-squared random variable with parameter m. Now, the result follows from tail bounds on the chi-squared random variables [45]. In particular, it follows that:

$$\mathbb{P}\left[\left| \frac{1}{m} \chi_m^2 - 1 \right| \geq \epsilon \right] < 2e^{-m\epsilon^2/8}.$$

We bound $2e^{-m\epsilon^2/8}$ by δ by taking $m = \Theta(\epsilon^{-2} \log(1/\delta))$, as desired. □

The dimensionality reduction achieved by Lemma 2.3 is surprising and substantial. Nonetheless, a natural question is whether a different Johnson-Lindenstrauss transform (i.e. a different random projection achieving Requirement 2.2) might enable an even better projected dimension. The answer turns out to be no: it turns out that dimension $m = \Omega(\epsilon^{-2} \log(1/\delta))$ is necessary for random projections achieving Requirement 2.2 [26, 24]. Hence, the dimensionality reduction achieved by Lemma 2.3 is optimal! In fact, an even stronger result holds [30]: there exists a set S of N points in $\mathbb{R}^n$ such that any function $f : S \to \mathbb{R}^m$ that achieves (2.1) requires $m = \Omega(\epsilon^{-2} \log(N))$. This result has a surprising consequence: we can restrict to considering linear maps without incurring any cost in the projected dimension m.[3]

2.2 Sparse Johnson-Lindenstrauss transforms

In this section, we consider dimensionality reduction as a pre-processing step for algorithmic tasks, where it is useful from a computational perspective to achieve a fast projection time. In particular, we describe *sparse Johnson-Lindenstrauss transforms* [28]: state-of-the-art Johnson-Lindenstrauss transforms over sparse matrices that enable fast projection.

From a computational perspective, the issue with the i.i.d. gaussian construction is that the matrix can be very dense. As a result, for a vector $x \in \mathbb{R}^n$, the time required to compute

[2] Recall that this is shorthand for $\mathbb{P}_{A \in \mathcal{A}}[(1 - \epsilon) \|x\|_2 \leq \|Ax\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

[3] Allowing non-linear maps does enable a much stronger guarantee on points outside of the set of datapoints. In particular, terminal embeddings [35] exploit the fact that non-linear maps can have no kernel in order to provide guarantees on distances to datapoints outside of the dataset.

the projection Ax is $O(m \|x\|_0)$ (where $\|x\|_0$ is the number of nonzero entries in x). Indeed, if the error is very small or the number of points N is large, this will unfortunately result in a high projection time. In this context, over the past two decades, the CS theory literature has developed dimensionality reduction schemes that enable a fast projection time.

An elegant way to reduce the projection time is to restrict the support of $\mathcal{A}$ to sparse matrices, and thus design *sparse random projections*.[4] Here, sparsity of a matrix is defined to the maximum number of nonzero entries in any column. If A has sparsity s, then the projection time on a vector x goes down to $O(s \|x\|_0)$. (This can be seen by expressing Ax as $\sum_{i \in \text{supp}(x)} A^i x_i$, where A^i is the ith column of A.) Thus, the objective is to set s to be as low as possible while still satisfying Requirement 2.2. To get some intuition for why the distance-preserving properties become harder to satisfy when s is small, let's consider the case of $s = 1$. In this case, the nonzero entries in a matrix with sparsity 1 can represented as a hash function $h : \{1, \ldots, n\} \to \{1, \ldots, m\}$. Now, unless m is very large, with nontrivial probability, there will be a collision (i.e. $h(i) = h(j)$) between two entries where x_i and x_j are large. In this case, the ℓ_2 norm of the projected data point will intuitively be far from the ℓ_2 norm of the original data point.

The idea is that there is a sweet spot where the sparsity is large enough to avoid having to increase the dimension m. Using this intuition, Kane and Nelson [28] constructed a family of sparse random projections, improving upon a line of previous work [46, 2, 33, 15]. These sparse random projections are called sparse Johnson-Lindenstrauss transforms (sparse JL transforms). Roughly speaking, a sparse JL transform, as constructed in [28], boils down to drawing a random $m \times n$ matrix where each column contains exactly s nonzero entries, each equal to $-1/\sqrt{s}$ or $1/\sqrt{s}$.

The sparse JL transform is easy to visualize in the case of $s = 1$, as we show in Figure 2.1. The distribution over $m \times n$ matrices can be viewed as follows. A uniformly chosen hash function $h : \{1, \ldots, n\} \to \{1, \ldots, m\}$ is used to select the nonzero entries in each column, so $(h(i), i)$ is the position of the nonzero entry in column i. Each column $1 \leq i \leq n$ is also given its own random sign σ_i, in order to handle collisions. With these components in place, the entry y_j of the projected vector can be written as the weighted sum $\sum_{i \in h^{-1}(j)} \sigma_j x_j$. In fact, this construction (i.e. sparse JL for $s = 1$) is also known as "feature hashing" in the machine learning literature.

The sparse JL transform for $s > 1$ can be viewed as selecting s hash functions $h_1, \ldots, h_s$, rather than just 1 hash function. These hash functions are anti-correlated in order to avoid an entry colliding with itself (i.e. in order to avoid $h_{j_1}(i) = h_{j_2}(i)$ for two different hash functions h_{j_1} and h_{j_2}).[5] Moreover, collisions are handled by s sets of independent column signs. (We formally define sparse JL transforms in Definition 2.6.)

Kane and Nelson show that sparse JL transforms satisfy Requirement 2.2 with the same (optimal) dimension as Lemma 2.3, while also achieving a sparsity property (in particular,

[4] Ailon and Chazelle [3] proposed a dimensionality reduction scheme (called a Fast Johnson–Lindenstrauss Transform) that achieves projection time $O(n \log n)$. This construction can beat the sparse JL construction of Kane and Nelson [28] on dense vectors. However, it is much slower on sparse vectors, which are common in many machine learning applications (e.g. a bag-of-words model).

[5] If this anti-correlation is not introduced and the hash functions are instead chosen independently, the resulting distribution is called "multiple hashing" [46, 15]. However, it turns out that a greater sparsity is actually needed for this construction than for sparse JL transforms [28].

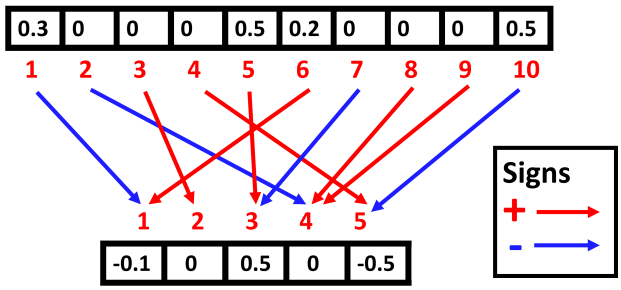

Figure 2.1: A hash function depiction of sparse JL with $s = 1$, $n = 10$, and $m = 5$.

only an ϵ-fraction of the matrix entries are nonzero).

Theorem 2.4 (Sparse JL [28]) For any $n \in \mathbb{N}$ and $\epsilon, \delta \in (0, 1)$, a sparse JL transform $\mathcal{A}_{s,m,n}$ (defined formally in Definition 2.6) over $m \times n$ matrices, with dimension $m = \Theta(\epsilon^{-2} \ln(1/\delta))$ and sparsity $s = \Theta(\epsilon^{-1} \ln(1/\delta))$, satisfies Requirement 2.2.

Sparse JL transforms are state-of-the-art sparse random projections, and achieve a sparsity that is nearly optimal when the dimension m is $\Theta(\epsilon^{-2} \ln(1/\delta))$. In particular, Nelson and Nguyen [37] showed that *any* distribution over matrices satisfying Requirement 2.2 requires sparsity $\Omega(\epsilon^{-1} \ln(1/\delta)/\ln(1/\epsilon))$ when the dimension m is $\Theta(\epsilon^{-2} \ln(1/\delta))$.[6] As a result, we cannot hope for much better sparse random projections with dimension $m = \Theta(\epsilon^{-2} \ln(1/\delta))$ that satisfy Requirement 2.2.

Despite this barrier, it can be necessary in practice to utilize a much lower sparsity s, since the projection time is linear in s. For example, if ϵ is very small, the sparsity $s = \log(1/\delta)/\epsilon$ may still be very high, even though it is much lower than the dimension m. Resolving this issue, Cohen [11] extended the upper bound in Theorem 2.4 to show that sparse JL transforms can achieve a lower sparsity with an appropriate gain in dimension. He proved the following dimension-sparsity tradeoffs:

Theorem 2.5 (Dimension-Sparsity Tradeoffs [11]) For any $n \in \mathbb{N}$ and $\epsilon, \delta \in (0, 1)$, a uniform sparse JL transform $\mathcal{A}_{s,m,n}$ (defined formally in Definition 2.6), with $s \leq \Theta(\epsilon^{-1} \ln(1/\delta))$ and $m \geq \min\left(2\epsilon^{-2}/\delta, \epsilon^{-2} \ln(1/\delta) e^{\Theta\left(\epsilon^{-1} \ln(1/\delta)/s\right)}\right)$, satisfies Requirement 2.2.

This result can be viewed as a tradeoff between projection time (tuned via the sparsity s) and projected dimension (described by m). Theorem 2.5 enables an application to select its own tradeoff between projection time and projected dimension.

2.2.1 Formal definition of Sparse JL transforms

Sparse JL transforms, as constructed by Kane and Nelson [28], are defined as follows.

[6] Kane and Nelson [28] also showed that the analysis of sparse JL transforms in Theorem 2.4 is tight at $m = \Theta(\epsilon^{-2} \ln(1/\delta))$. Thus, closing this (small) gap would either require moving to a different family of sparse random projections or improving the lower bound.

Definition 2.6 Let $\mathcal{A}_{s,m,n}$ be a **sparse JL transform** if the entries of a matrix $A \in \mathcal{A}_{s,m,n}$ are generated as follows. Let $A_{r,i} = \eta_{r,i}\sigma_{r,i}/\sqrt{s}$ where $\{\sigma_{r,i}\}_{r\in[m],i\in[n]}$ and $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ satisfy the following conditions:

- The families $\{\sigma_{r,i}\}_{r\in[m],i\in[n]}$ and $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ are independent from each other.
- The variables $\{\sigma_{r,i}\}_{r\in[m],i\in[n]}$ are i.i.d Rademachers (± 1 coin flips). (These random variables assign signs to entries in the matrix.)
- The variables $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ are identically distributed Bernoullis ($\{0,1\}$ random variables) with expectation s/m. (These random variables determine the nonzero entries in the matrix.)
- The $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ are independent across columns but not independent within each column. For every column $1 \le i \le n$, it holds that $\sum_{r=1}^{m} \eta_{r,i} = s$. Moreover, the random variables are *negatively correlated*: for every subset $S \subseteq [m]$ and every column $1 \le i \le n$, it holds that $\mathbb{E}\left[\prod_{r\in S} \eta_{r,i}\right] \le \prod_{r\in S} \mathbb{E}[\eta_{r,i}]$.

We describe two common constructions of sparse JL transforms satisfying Definition 2.6. The first construction is a **uniform sparse JL transform**, generated as follows: for every $1 \le i \le n$, we *uniformly* choose exactly s of these variables in $\{\eta_{r,i}\}_{r\in[m]}$ to be 1. When $s = 1$, every sparse JL transform is a uniform sparse JL transform, but for $s > 1$, this is not the case. The second common construction is a **block sparse JL transform**, which produces a different construction for $s > 1$. In this construction, each column $1 \le i \le n$ is partitioned into s blocks of $\left\lfloor \frac{m}{s} \right\rfloor$ consecutive rows. In each block in each column, the distribution of the variables $\{\eta_{r,i}\}$ is defined by uniformly choosing *exactly one* of these variables to be 1.

2.2.2 Overview of proofs of Theorem 2.4

We describe, at a high-level, the standard proofs of Theorem 2.4. Since A is linear, it suffices to consider x in the unit ℓ_2 ball, i.e. such that $\|x\|_2 = 1$. Like in the proof of Theorem 2.3, it is easier to reason about ℓ_2^2 than ℓ_2, so we consider $\|Ax\|_2^2 - 1$. For this setting, notice that:

$$\begin{aligned}
\|Ax\|_2^2 - 1 &= \frac{1}{s}\sum_{r=1}^{m}\left(\sum_{i=1}^{n}\sigma_{r,i}\eta_{r,i}x_i\right)^2 - 1 \\
&= \frac{1}{s}\sum_{r=1}^{m}\left(\sum_{1\le i\ne j\le n}\sigma_{r,i}\sigma_{r,j}\eta_{r,i}\eta_{r,j}x_ix_j + \sum_{i=1}^{n}\sigma_{r,i}^2\eta_{r,i}^2x_i^2\right) - 1 \\
&= \frac{1}{s}\sum_{r=1}^{m}\sum_{1\le i\ne j\le n}(\sigma_{r,i}\sigma_{r,j}\eta_{r,i}\eta_{r,j}x_ix_j) + \frac{1}{s}\left(\sum_{i=1}^{n}\sum_{r=1}^{m}\eta_{r,i}x_i^2\right) - 1 \\
&= \frac{1}{s}\sum_{r=1}^{m}\sum_{1\le i\ne j\le n}(\sigma_{r,i}\sigma_{r,j}\eta_{r,i}\eta_{r,j}x_ix_j) + \frac{1}{s}\left(\sum_{i=1}^{n}sx_i^2\right) - 1 \\
&= \frac{1}{s}\sum_{r=1}^{m}\sum_{1\le i\ne j\le n}\sigma_{r,i}\sigma_{r,j}\eta_{r,i}\eta_{r,j}x_ix_j.
\end{aligned}$$

We define the error term $R(x_1, \ldots, x_n)$ to be

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \sum_{1 \leq i \neq j \leq n} \sigma_{r,i} \sigma_{r,j} \eta_{r,i} \eta_{r,j} x_i x_j.$$

Thus, we need to show tail bounds for $R(x_1, \ldots, x_n)$. The challenge is that unlike the proof of Theorem 2.3, we can't relate the distribution of $R(x_1, \ldots, x_n)$ to that of a nice random variable with known tail bounds.

The standard approach is to obtain a bound on the $\Theta(\log(1/\delta))$th moment of this random variable, and use this to obtain a tail bound. Markov's inequality, a standard tool, enables us to move from a moment bound to a tail bound. The driving idea of Markov's inequality is the probability that a nonnegative random variable is far from its mean must be small.

Lemma 2.7 (Markov's inequality) If X is a nonnegative random variable, then for any $b > 0$, it holds that:

$$\mathbb{P}[X > b] \leq \frac{\mathbb{E}[X]}{b}.$$

We can use Markov's inequality to relate a high moment of $R(x_1, \ldots, x_n)$ to a tail bound as follows. Let $p = \Theta(\log(1/\delta)$ be an even integer. We can apply Markov's inequality to $R(x_1, \ldots, x_n)^p$ to obtain that:

$$\mathbb{P}[R(x_1, \ldots, x_n) > \epsilon] = \mathbb{P}[R(x_1, \ldots, x_n)^p > \epsilon^p] \leq \epsilon^{-p} \mathbb{E}[R(x_1, \ldots, x_n)^p].$$

The idea is that if $\mathbb{E}[R(x_1, \ldots, x_n)^p] \leq 0.5\epsilon^p$, then we know that $\epsilon^{-p}\mathbb{E}[R(x_1, \ldots, x_n)^p] \leq \delta$.

As a result, the analysis boils down to bounding $\mathbb{E}[R(x_1, \ldots, x_n)^p]$. This is technically challenging since the $\eta_{r,i}$ random variables are not independent. Due to the interaction of these correlations with the potentially differing weights x_i, the behavior of this random variable is difficult to reason about. The original analysis of sparse JL [28] handled the moment bounds via expanding out $R(x_1, \ldots, x_n)^p$ as a sum of a large collection of monomials. They then carefully rearranged and bounded sums of these monomials, and wrote this expression as a sum of certain graphs. Then, they counted the number of graphs and used this to bound the overall expression. The resulting analysis was quite intricate and delicate.

Cohen et al. [12] later showed a much simpler and cleaner proof of Theorem 2.4. They give an elegant approach for bounding the moments of $R(x_1, \ldots, x_n)$ that utilizes the structure of $R(x_1, \ldots, x_n)$. The random variables can be peeled off in layers, as follows: $\mathbb{E}[R(x_1, \ldots, x_n)^p] = \mathbb{E}_\eta \left[\mathbb{E}_\sigma \left[R(x_1, \ldots, x_n)^p\right]\right]$. First, they write $R(x_1, \ldots, x_n)$ as a quadratic form of the $m \cdot n$ Rademachers (± 1 random variables), i.e. it's of the form $\frac{1}{s}\sum_{k,l} a_{k,l}\sigma_k\sigma_l$ for a $m \cdot n$-dimensional block-diagonal matrix with a zero-diagonal, where the (i,j)th entry of the rth block is $\eta_{r,i}\eta_{r,j}x_i x_j$. While $(a_{k,l})_{k,l}$ is a random matrix, for a fixed instantation of the $\eta_{r,i}$, it becomes a scalar matrix.

The quantity $\mathbb{E}_\sigma \left[R(x_1, \ldots, x_n)^p\right]$ can then be handled using that Rademachers are subgaussian random variables. In particular, they consider:

$$R_g(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \sum_{1 \leq i \neq j \leq n} g_{r,i} g_{r,j} \eta_{r,i} \eta_{r,j} x_i x_j,$$

and use that $\mathbb{E}_\sigma[R(x_1,\dots,x_n)^p] \leq \mathbb{E}_g[R_g(x_1,\dots,x_n)^p]$ since the σ are subgaussian. Then, they use gaussian bounds[7] on $\mathbb{E}_g[R_g(x_1,\dots,x_n)^p]$ to obtain an upper bound $B(\eta)$. As a result, they deduce that $\mathbb{E}[R_g(x_1,\dots,x_n)^p] \leq \mathbb{E}_\eta B(\eta)$, and then they handle $\mathbb{E}_\eta B(\eta)$ using binomial random variable moment bounds.

[7]See Lemma 5.3.

Chapter 3

Machine Learning: JL on Feature Vectors

In this chapter, we consider how sparse Johnson-Lindenstrauss transforms perform on well-behaved datasets, in the context of a model inspired by machine learning applications [46]. In Section 3.1, we review previous results. In Section 3.2, we present our tight bounds on the tradeoff between projected dimension, sparsity of the matrix, accuracy, and ℓ_∞-to-ℓ_2 norm ratio of the data points.

3.1 Framework and previous results

In machine learning settings, feature hashing and other random projection schemes are influential in helping manage large data [14], since dimensionality reduction enables a classifier to process vectors in $\mathbb{R}^m$, instead of vectors in $\mathbb{R}^n$. In this context, feature hashing was first introduced by Weinberger et. al [46]. Feature hashing has close ties to sparse JL transforms, and the feature hashing scheme in [46] can be viewed as a sparse JL transform with $s = 1$.

Feature hashing was initially proposed in [46] for document-based classification tasks such as email spam filtering. For such tasks, feature hashing yields a lower dimensional embedding of a high-dimensional feature vector derived from a bag-of-words model (i.e. a vector representation of a text document consisting of the number of occurrences of each word). Since then, feature hashing has become a mainstream approach, applied to numerous domains including ranking text documents [5], compressing neural networks [9], and protein sequence classification [7]. Indeed, feature hashing (also called the "hashing trick") is considered one of the key techniques in scaling up machine learning algorithms [16, 43]!

Interestingly, in practice, feature hashing can do much better than theoretical results, such as Theorem 2.4 and Theorem 2.5, would indicate [16]. A line of work in the machine learning literature [46, 15, 27, 13, 28] considers whether feature hashing can provide better distance-preservation guarantees on "well-behaved" datasets. This line of work fits nicely into the "beyond worst-case analysis" framework [41] that has recently become a main theme in the computer science literature. In particular, the condition in Requirement 2.2 can be viewed as a worst-case guarantee since it requires good norm-preservation guarantees on any given vector $x \in \mathbb{R}^n$. In the N-point setting, Requirement 2.2 translates to requiring good

distance-preservation guarantees on *any* set of N points. Looking at distance-preservation guarantees on "well-behaved" datasets can be viewed as "beyond worst-case analysis" (which is about identifying properties of real-world inputs and proving stronger guarantees for inputs with these properties).

The model for well-behaved datasets considered in this line of work [46, 15, 27, 13, 28] is based on the fact that mass of real-world feature vectors is likely to be spread out between many coordinates. For example, consider a feature vector for a text document obtained from a bag-of-words model with the standard `tf-idf` pre-processing.[1] It's unlikely that a single coordinate of a feature vector has all of the mass, and the mass is likely distributed between a set of document-specific terms. While the highest error terms in sparse JL often stem from vectors with mass concentrated on two entries (i.e. $[1, 1, 0, \ldots, 0]$), it is unlikely in practice for such a vector to occur. This motivates studying the performance of sparse JL on vectors with mass spread out between a set of coordinates, i.e. vectors with low ℓ_∞-to-ℓ_2 ratio.[2]

More formally, take S_v to be $\left\{x \in \mathbb{R}^n \mid \frac{\|x\|_\infty}{\|x\|_2} \leq v\right\}$, so that $S_1 = \mathbb{R}^n$ and $S_v \subsetneq S_w$ for $0 \leq v < w \leq 1$. Let $v(m, \epsilon, \delta, s)$ be the supremum over all $0 \leq v \leq 1$ such that a sparse JL transform with sparsity s and dimension m satisfies $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$ for each $x \in S_v$.[3] (That is, $v(m, \epsilon, \delta, s)$ is the maximum $v \in [0, 1]$ such that for every $x \in \mathbb{R}^n$, if $\|x\|_\infty \leq v\|x\|_2$ then $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$ holds.)

When $v(m, \epsilon, \delta, s) = 0$, this means that the performance in distance preservation is poor: there exist vectors with arbitrarily small ℓ_∞-to-ℓ_2 norm ratio where $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2]$ is smaller than $1-\delta$. When $v(m, \epsilon, \delta, s) = 1$, this means that there is full performance in distance preservation: all vectors in $\mathbb{R}^n$ satisfy $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$, so Requirement 2.2 is satisfied. When $v(m, \epsilon, \delta, s) \in (0, 1)$, this means that there is good performance in distance preservation for $x \in S_{v(m,\epsilon,\delta,s)}$, but there can be poor performance for $x \notin S_{v(m,\epsilon,\delta,s)}$.

Technically, the quantity $v(m, \epsilon, \delta, s)$, as defined here, also depends on n. In particular, every vector $x \in \mathbb{R}^n$ satisfies $\|x\|_\infty \geq \|x\|_2 / \sqrt{n}$, so ℓ_∞-to-ℓ_2 norm ratios below $1/\sqrt{n}$ are not possible in $\mathbb{R}^n$. To avoid this dependence on n and thus make the bounds cleaner, the quantity $v(m, \epsilon, \delta, s)$ is actually defined to be the infimum over all $n \in \mathbb{N}$ of the supremum over all $0 \leq v \leq 1$ such that a sparse JL transform with sparsity s and dimension m satisfies $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$ for each $x \in S_v$. (That is, $v(m, \epsilon, \delta, s)$ is the maximum $v \in [0, 1]$ such that for every $n \in \mathbb{N}$ and every $x \in \mathbb{R}^n$, if $\|x\|_\infty \leq v\|x\|_2$ then $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$ holds.)

For $s = 1$, a line of work [46, 15, 27, 13, 28] improved bounds on $v(m, \epsilon, \delta, 1)$, and was recently closed by Freksen et al. [16].

Theorem 3.1 ([16]) For any $m \in \mathbb{N}$ and $\epsilon, \delta \in (0, 1)$, the function $v(m, \epsilon, \delta, 1)$ is equal to

[1] This is the term frequency-inverse document frequency, which adjusts for the fact that some words occur frequently in general, in order to highlight document-specific terms.

[2] One might imagine other models for vectors with mass spread out between several coordinates, such as the ℓ_p-to-ℓ_2 norm ratio for some other values of p. This would be an interesting direction for future work. Selecting $p = \infty$ enabled the analysis of sparse JL to be tractable, since ℓ_∞-to-ℓ_2 ratio is largely a local property of the vector.

[3] Recall that this is shorthand for $\mathbb{P}_{A \in \mathcal{A}}[(1 - \epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \epsilon)\|x\|_2] > 1 - \delta$.

$f(m, \epsilon, \ln(1/\delta))$ where:

$$f(m, \epsilon, p) = \begin{cases} 1 & \text{if } m \geq 2\epsilon^{-2}e^p \\ \Theta\left(\sqrt{\epsilon}\min\left(\frac{\ln(\frac{m\epsilon}{p})}{p}, \frac{\sqrt{\ln(\frac{m\epsilon^2}{p})}}{\sqrt{p}}\right)\right) & \text{if } \Theta(\epsilon^{-2}p) \leq m < 2\epsilon^{-2}e^p \\ 0 & \text{if } m \leq \Theta(\epsilon^{-2}p). \end{cases}$$

While Theorem 3.1 is restricted to the case of $s = 1$, dimensionality reduction schemes constructed using sparse random projections with sparsity $s > 1$ have been used in practice for projecting feature vectors. For example, sparse JL-like methods (with $s > 1$) have been used to project feature vectors in machine learning domains including visual tracking [42], face recognition [34], and recently in ELM [8]. Now, a variant of sparse JL is included in the Python sklearn library.[4]

In this context, it is natural to explore how constructions with $s > 1$ perform on feature vectors, by studying $v(m, \epsilon, \delta, s)$ for sparse JL with $s > 1$. In fact, a related question was considered by Weinberger et al. [46] for "multiple hashing," an alternate distribution over sparse matrices constructed by adding s draws from $\mathcal{A}_{1,m,n}$ and scaling by $1/\sqrt{s}$. More specifically, they show that $v(m, \epsilon, \delta, s) \geq \min(1, \sqrt{s} \cdot v(m, \epsilon, \delta, 1))$ for multiple hashing. However, Kane and Nelson [28] later showed that multiple hashing has worse geometry-preserving properties than sparse JL: that is, multiple hashing requires a larger sparsity than sparse JL to satisfy Requirement 2.2.

3.1.1 Discussion of combinatorial analysis in [16]

An upper bound on the tail probability of $R(x_1, \ldots, x_n)$ is needed to prove the lower bound on $v(m, \epsilon, \delta, s)$ in Theorem 3.2, and a lower bound is needed to prove the upper bound on $v(m, \epsilon, \delta, s)$ in Theorem 3.2. It turns out that it suffices to tightly analyze the random variable moments $\mathbb{E}[(R(x_1, \ldots, x_n))^q]$. For the upper bound on the tail probability, they use Markov's inequality, like in Section 2.2.2. For the lower bound on the tail probability, they use the Paley-Zygmund inequality, which gives a lower bound on the tail probability from upper and lower bounds on moments.

Thus, the key ingredient of their analysis is a *tight bound* on the moments of $R(x_1, \ldots, x_n)$ on $S_v = \left\{x \in \mathbb{R}^n \mid \frac{\|x\|_\infty}{\|x\|_2} \leq v\right\}$ for the case for $s = 1$. Unfortunately, the combinatorial approach in [28] cannot be directly generalized to obtain Theorem 3.1. In [16], they require a novel graph-counting argument, along with more precise monomial bounds. The analysis is quite intricate and delicate, and it is not clear how to generalize this analysis to $s > 1$.

3.2 Our results

Characterizing $v(m, \epsilon, \delta, s)$ for sparse JL transforms, which are state-of-the-art, remained an open problem. We settle how $v(m, \epsilon, \delta, s)$ behaves for sparse JL with a general sparsity $s > 1$, giving tight bounds. Our theoretical result shows that sparse JL with $s > 1$, even if s

[4] See `https://scikit-learn.org/stable/modules/random_projection.html`.

is a small constant, can achieve significantly better norm-preservation properties for feature vectors than sparse JL with $s = 1$. Moreover, we empirically demonstrate this finding.

3.2.1 Mathematical result

We show the following tight bounds on $v(m, \epsilon, \delta, s)$ for a general sparsity s:

Theorem 3.2 For any $s, m \in \mathbb{N}$ such that $s \leq m/e$, consider a uniform sparse JL transform (defined in Definition 2.6) with sparsity s and dimension m.[5] If ϵ and δ are small enough[6], the function $v(m, \epsilon, \delta, s)$ is equal to $f'(m, \epsilon, \ln(1/\delta), s)$, where $f'(m, \epsilon, p, s)$ is[7]:

$$\begin{cases} 1 & \text{if } m \geq \min\left(2\epsilon^{-2}e^p, \epsilon^{-2}pe^{\Theta\left(\max\left(1, \frac{p\epsilon^{-1}}{s}\right)\right)}\right) \\ \Theta\left(\sqrt{\epsilon s}\frac{\sqrt{\ln(\frac{m\epsilon^2}{p})}}{\sqrt{p}}\right) & \text{else, if } \max\left(\Theta(\epsilon^{-2}p), s \cdot e^{\Theta\left(\max\left(1, \frac{p\epsilon^{-1}}{s}\right)\right)}\right) \leq m \leq \epsilon^{-2}e^{\Theta(p)} \\ \Theta\left(\sqrt{\epsilon s}\min\left(\frac{\ln(\frac{m\epsilon}{p})}{p}, \frac{\sqrt{\ln(\frac{m\epsilon^2}{p})}}{\sqrt{p}}\right)\right) & \text{else, if } \Theta(\epsilon^{-2}p) \leq m \leq \min\left(\epsilon^{-2}e^{\Theta(p)}, s \cdot e^{\Theta\left(\max\left(1, \frac{p\epsilon^{-1}}{s}\right)\right)}\right) \\ 0 & \text{if } m \leq \Theta(\epsilon^{-2}p). \end{cases}$$

Our main result, Theorem 3.2, significantly generalizes Theorem 2.4, Theorem 2.5, and Theorem 3.1. Notice our bound in Theorem 3.2 has up to four regimes. In the first regime, which occurs when $m \geq \min(2\epsilon^{-2}/\delta, \epsilon^{-2}\ln(1/\delta)e^{\Theta(\max(1, \ln(1/\delta)\epsilon^{-1}/s))})$, Theorem 3.2 shows $v(m, \epsilon, \delta, s) = 1$, so $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$ holds on the full space $\mathbb{R}^n$. Notice this boundary on m occurs at the dimensionality-sparsity tradeoff in Theorem 2.5. In the last regime, which occurs when $m \leq \Theta(\epsilon^{-2}\ln(1/\delta))$, Theorem 3.2 shows that $v(m, \epsilon, \delta, s) = 0$, so there are vectors with arbitrarily small ℓ_∞-to-ℓ_2 norm ratio where $\mathbb{P}_{A \in \mathcal{A}}[\|Ax\|_2 \in (1 \pm \epsilon)\|x\|_2] > 1 - \delta$ does not hold. When $s \leq \Theta(\epsilon^{-1}\ln(1/\delta))$, Theorem 3.2 shows that up to two intermediate regimes exist.

One of the regimes, $\Theta(\sqrt{\epsilon s}\min(\ln(\frac{m\epsilon}{p})/p, \sqrt{\ln(\frac{m\epsilon^2}{p})}/\sqrt{p}))$, matches the middle regime of $v(m, \epsilon, \delta, 1)$ in Theorem 3.1 with an extra factor of $\sqrt{s}$, much like the bound for multiple hashing in [46] that we mentioned previously. However, unlike the multiple hashing bound, Theorem 3.2 sometimes has another regime, $\Theta(\sqrt{\epsilon s}\sqrt{\ln(\frac{m\epsilon^2}{p})}/\sqrt{p})$, which does not arise for $s = 1$ (i.e. in Theorem 3.1).[8] Intuitively, we expect this additional regime for sparse JL with s close to $\Theta(\epsilon^{-1}\ln(1/\delta))$: at $s = \Theta(\epsilon^{-1}\ln(1/\delta))$ and $m = \Theta(\epsilon^{-2}\ln(1/\delta))$, Theorem 2.4 tells us $v(m, \epsilon, \delta, s) = 1$, but if ϵ is a constant, then the branch $\Theta(\sqrt{\epsilon s}\ln\left(\frac{m\epsilon}{p}\right)/p)$ yields $\Theta(1/\sqrt{\ln(1/\delta)})$, while the branch $\Theta(\sqrt{\epsilon s}\sqrt{\ln(\frac{m\epsilon^2}{p})}/\sqrt{p})$ yields $\Theta(1)$. Thus, it is natural that the first branch disappears for large m.

Our result elucidates that $v(m, \epsilon, \delta, s)$ increases approximately as $\sqrt{s}$, thus providing insight into how even small constant increases in sparsity can be useful in practice. Another

[5] We prove the lower bound on $v(m, \epsilon, \delta, s)$ in Theorem 3.2 for *any* sparse JL transform.

[6] By "small enough", we mean the condition that $\epsilon, \delta \in (0, C')$ for some positive constant C'.

[7] Notice that the function $f'(m, \epsilon, p, s)$ is not defined for certain "constant-factor" intervals between the boundaries of regimes (e.g. $C_1\epsilon^{-2}p \leq m \leq C_2\epsilon^{-2}p$). See Appendix A for a discussion.

[8] This regime does not arise for $s = 1$, since $e^{\Theta\left(p\epsilon^{-1}\right)} \geq \epsilon^{-2}e^{\Theta(p)}$ for sufficiently small ϵ.

consequence of our result is a lower bound on dimension-sparsity tradeoffs (Corollary A.1 in Appendix A) that essentially matches the upper bound in Theorem 2.5. Moreover, we require new techniques to prove Theorem 3.2. (The combinatorial approaches in [28, 16] do not directly generalize to this setting, and we show that the approach in [12] is also not sufficiently precise, even for $s = 1$.)

To prove Theorem 3.2, we give a new perspective on bounding moments of these random variables. We believe our style of analysis is less brittle than combinatorial approaches [16, 28, 4, 36]: in this setting, once the sparsity $s = 1$ case is recovered, it becomes straightforward to generalize to other s values. Moreover, our approach can yield greater precision than the previous non-combinatorial approach [12], which is necessary for this setting. For this reason, we believe that our *structural* approach could be of more general use. Our approach also shares many common technical ingredients with our proof of Theorem 4.4, and thus takes a step towards a unified analysis of JL transforms. We discuss the analysis methods in greater detail in Chapter 5.

3.2.2 Empirical evaluation

We also empirically support our theoretical findings in Theorem 3.2. First, we illustrate with real-world datasets the potential benefits of using small constants $s > 1$ for sparse JL on feature vectors. We specifically show that $s = \{4, 8, 16\}$ consistently outperforms $s = 1$ in preserving the ℓ_2 norm of each vector, and that there can be up to a *factor of ten* decrease in failure probability for $s = 8, 16$ in comparison to $s = 1$. Second, we use synthetic data to illustrate phase transitions and other trends in Theorem 3.2. More specifically, we empirically show that $v(m, \epsilon, \delta, s)$ is not smooth, and that the middle regime(s) of $v(m, \epsilon, \delta, s)$ increases with s.

Recall that for sparse JL transforms with sparsity s, the projection time for an input vector x is $O(s \|x\|_0)$, where $\|x\|_0$ is the number of nonzero entries in x. Since this grows linearly in s, in order to minimize the impact on projection time, we restrict to small constant s values (i.e. $1 \leq s \leq 16$). In Section 3.2.2, we demonstrate on real-world data the benefits of using $s > 1$. In Section 3.2.2, we illustrate trends in our theoretical bounds on synthetic data. Additional graphs can be found in Appendix I. For all experiments, we use a block sparse JL transform to demonstrate that our theoretical upper bounds also empirically generalize to non-uniform sparse JL transforms.

Real-world datasets

We considered two bag-of-words datasets: the News20 dataset [1] (based on newsgroup documents), and the Enron email dataset [38] (based on e-mails from the senior management of Enron).[9] Both datasets were pre-processed with the standard `tf-idf` preprocessing. In this experiment, we evaluated how well sparse JL preserves the ℓ_2 norms of the vectors in the dataset. An interesting direction for future work would be to empirically evaluate how well sparse JL preserves other aspects of the geometry of real-world data sets, such as the ℓ_2 distances between pairs of vectors.

[9]Note that the News20 dataset is used in [13], and the Enron dataset is from the same collection as the dataset used in [16], but contains a larger number of documents.

In our experiment, we estimated the failure probability $\hat{\delta}(s, m, \epsilon)$ for each dataset as follows. Let D be the number of vectors in the dataset, and let n be the dimension ($n = 101631$, $D = 11314$ for News20; $n = 28102$, $D = 39861$ for Enron). We drew a matrix $M \sim \mathcal{A}_{s,m,n}$ from a block sparse JL transform. Then, we computed $\frac{\|Mx\|_2}{\|x\|_2}$ for each vector x in the dataset, and used these values to compute an estimate $\hat{\delta}(s, m, \epsilon) = \frac{\text{number of vectors } x \text{ such that } \frac{\|Mx\|_2}{\|x\|_2} \notin 1 \pm \epsilon}{D}$. (Here $\frac{\|Mx\|_2}{\|x\|_2} \notin 1 \pm \epsilon$ is short-hand notation for denoting that either $\frac{\|Mx\|_2}{\|x\|_2} > 1 + \epsilon$ or $\frac{\|Mx\|_2}{\|x\|_2} < 1 - \epsilon$.) We ran 100 trials to produce 100 estimates $\hat{\delta}(s, m, \epsilon)$.

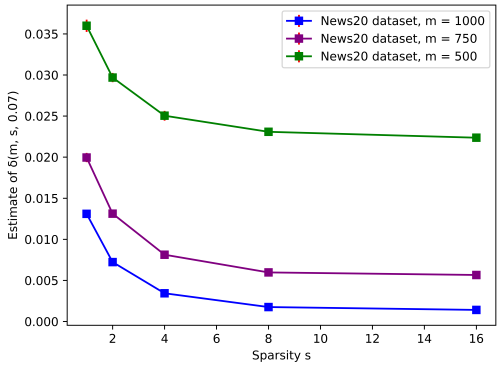

Figure 3.1: News20: $\hat{\delta}(m, s, 0.07)$ v. s

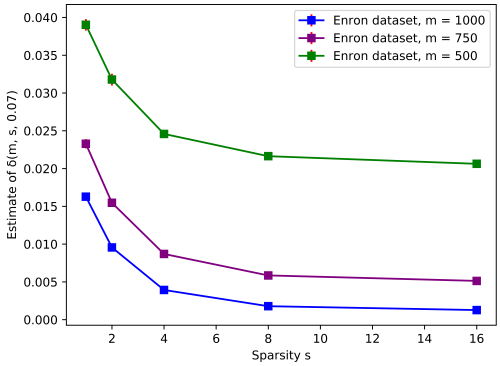

Figure 3.2: Enron: $\hat{\delta}(m, s, 0.07)$ vs. s

Figure 3.1 and Figure 3.2 show the mean and error bars (3 standard errors of the mean) of $\hat{\delta}(s, m, \epsilon)$ at $\epsilon = 0.07$. We consider $s \in \{1, 2, 4, 8, 16\}$, and choose m values so that $0.01 \leq \hat{\delta}(1, m, \epsilon) \leq 0.04$.

All of the plots show that $s \in \{2, 4, 8, 16\}$ achieves a lower failure probability than $s = 1$, with the differences most pronounced when m is larger. In fact, at $m = 1000$, there is a *factor of four* decrease in δ between $s = 1$ and $s = 4$, and a *factor of ten* decrease between $s = 1$ and $s = 8, 16$. We note that in plots in the Appendix, there is a slight increase between $s = 8$ and $s = 16$ at some ϵ, δ, m values (see Appendix I for a discussion of this non-monotonicity in s); however $s > 1$ still consistently beats $s = 1$. Thus, these findings demonstrate the potential benefits of using small constants $s > 1$ in sparse JL in practice, which aligns with our theoretical results.

Synthetic datasets

We used synthetic data to illustrate the phase transitions in our bounds on $v(m, \epsilon, \delta, s)$ in Theorem 3.2 for a block sparse JL transform. For several choices of s, m, ϵ, δ, we computed an estimate $\hat{v}(m, \epsilon, \delta, s)$ of $v(m, \epsilon, \delta, s)$ as follows. Our experiment borrowed aspects of the experimental design in [16]. Our synthetic data consisted of binary vectors (i.e. vectors whose entries are in $\{0, 1\}$). The binary vectors were defined by a set W of values exponentially spread between 0.03 and 1[10]: for each $w \in W$, we constructed a binary vector x^w where the first $1/w^2$ entries are nonzero, and computed an estimate $\hat{\delta}(s, m, \epsilon, w)$ of the failure probability of the block sparse JL transform on the specific vector x^w (i.e. $\mathbb{P}_{A \in \mathcal{A}_{s,m,1/w^2}}[\|Ax^w\|_2 \notin (1 \pm \epsilon) \|x^w\|_2]$).

[10]We took $W = \left\{w \mid w^{-2} \in \{986, 657, 438, 292, 195, 130, 87, 58, 39, 26, 18, 12, 9, 8, 7, 6, 5, 4, 3, 2, 1\}\right\}$.

We computed each $\hat{\delta}(s, m, \epsilon, w)$ using 100,000 samples from a block sparse JL transform, as follows. In each sample, we independently drew a matrix $M \sim A_{s,m,1/w^2}$ and computed the ratio $\frac{\|Mx^w\|_2}{\|x^w\|_2}$. Then, we took $\hat{\delta}(s, m, \epsilon, w) :=$ (number of samples where $\frac{\|Mx^w\|_2}{\|x^w\|_2} \notin 1 \pm \epsilon)/T$. (Here $\frac{\|Mx^w\|_2}{\|x^w\|_2} \notin 1 \pm \epsilon$ is short-hand notation for denoting that either $\frac{\|Mx^w\|_2}{\|x^w\|_2} > 1 + \epsilon$ or $\frac{\|Mx^w\|_2}{\|x^w\|_2} < 1 - \epsilon$.) Finally, we used the estimates $\hat{\delta}(s, m, \epsilon, w)$ to obtain the estimate

$$\hat{v}(m, \epsilon, \delta, s) = \max\left\{v \in W \mid \hat{\delta}(s, m, \epsilon, w) < \delta \text{ for all } w \in W \text{ where } w \leq v\right\}.$$

Why does this procedure estimate $v(m, \epsilon, \delta, s)$? With enough samples, $\hat{\delta}(s, m, \epsilon, w) \to \mathbb{P}_{A \in \mathcal{A}_{s,m,1/w^2}}[\|Ax^w\|_2 \notin (1 \pm \epsilon)\|x^w\|_2]$.[11] As a result, if x^w is a "violating" vector, i.e. $\hat{\delta}(s, m, \epsilon, w) \geq \delta$, then likely $\mathbb{P}_{A \in \mathcal{A}_{s,m,n}}[\|Ax^w\|_2 \notin (1 \pm \epsilon)\|x^w\|_2] \geq \delta$, and so $\hat{v}(m, \epsilon, \delta, s) \geq v(m, \epsilon, \delta, s)$. For the other direction, we use that in the proof of Theorem 1.5, we show that asymptotically, if a "violating" vector (i.e. x s.t. $\mathbb{P}_{A \in \mathcal{A}_{s,m,n}}[\|Ax\|_2 \notin (1 \pm \epsilon)\|x\|_2] \geq \delta$) exists in S_v, then there's a "violating" vector of the form x^w for some $w \leq \Theta(v)$. Thus, the estimate $\hat{v}(m, \epsilon, \delta, s) = \Theta(v(m, \epsilon, \delta, s))$ as $T \to \infty$ and as precision in W goes to ∞.

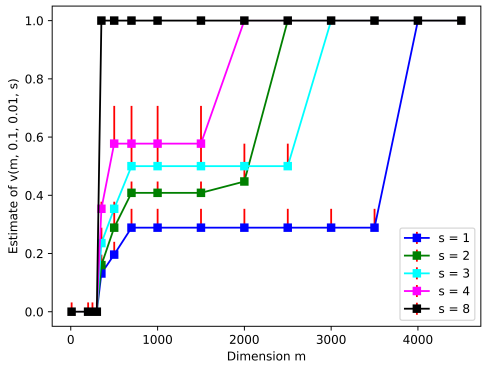

Figure 3.3: $\hat{v}(m, 0.1, 0.01, s)$

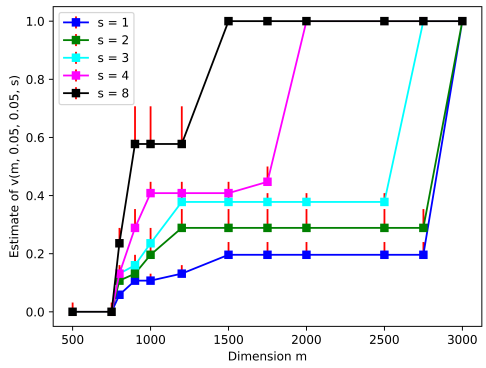

Figure 3.4: $\hat{v}(m, 0.05, 0.05, s)$

Figure 3.3 and Figure 3.4 show $\hat{v}(m, \epsilon, \delta, s)$ as a function of dimension m for $s \in \{1, 2, 3, 4, 8\}$ for two settings of ϵ and δ. The error-bars are based on the distance to the next highest v value in W.

Our first observation is that for each set of s, ϵ, δ values considered, the curve $\hat{v}(m, \epsilon, \delta, s)$ has "sharp" changes as a function of m. More specifically, $\hat{v}(m, \epsilon, \delta, s)$ is 0 at small m, then there is a phase transition to a nonzero value, then an increase to a higher value, then an interval where the value appears "flat", and lastly a second phase transition to 1. The first phase transition is shared between s values, but the second phase transition occurs at different dimensions m (but is within a factor of 3 between s values). Here, the first phase transition likely corresponds to $\Theta(\epsilon^{-2}\ln(1/\delta))$ and the second phase transition likely corresponds to $\min\left(\epsilon^{-2}e^{\Theta(\ln(1/\delta))}, \epsilon^{-2}\ln(1/\delta)e^{\Theta\left(\ln(1/\delta)\epsilon^{-1}/s\right)}\right)$.

Our second observation is that as s increases, the "flat" part occurs at a higher y-coordinate. Here, the increase in the "flat" y-coordinate as a function of s corresponds to

[11] With 100,000 samples, running our procedure twice yielded the same $\hat{v}(m, \epsilon, \delta, s)$ values both times.

the $\sqrt{s}$ term in $v(m, \epsilon, \delta, s)$. Technically, according to Theorem 3.2, the "flat" parts should be increasing in m at a slow rate: the empirical "flatness" likely arises since W is a finite set in the experiments.

Our third observation is that $s > 1$ generally outperforms $s = 1$ as Theorem 3.2 suggests: that is, $s > 1$ generally attains a higher $\hat{v}(m, \epsilon, \delta, s)$ value than $s = 1$. We note at large m values (where $\hat{v}(m, \epsilon, \delta, s)$ is close to 1), lower s settings sometimes attain a higher $\hat{v}(m, \epsilon, \delta, s)$ than higher s settings (e.g. the second phase transition doesn't quite occur in decreasing order of s in Figure 3.3): see Appendix I for a discussion of this non-monotonicity in s.[12] Nonetheless, in practice, it's unlikely to select such a large dimension m, since the ℓ_∞-to-ℓ_2 guarantees of smaller m are likely sufficient. Hence, a greater sparsity generally leads to a better $\hat{v}(m, \epsilon, \delta, s)$ value, thus aligning with our theoretical findings.

[12] In Appendix I, we also show more examples where at large m values, lower s settings attain a higher $\hat{v}(m, \epsilon, \delta, s)$ than higher s settings.

Chapter 4

Neuroscience: JL for Sign-Consistent Matrices

In this chapter, we consider dimensionality reduction as a model for information compression in the brain. In Section 4.1, we review the motivation for and construction of random projections, called sparse, sign-consistent JL transforms [4], that satisfy certain biological restrictions. In Section 4.2, we present our novel dimension-sparsity tradeoffs that result from our simplified and generalized analysis of these JL transforms.

4.1 Framework and previous results

Neuroscience-based constraints give rise to the additional condition of sign-consistency on the matrices in the probability distribution. *Sign-consistency* refers to the constraint that the nonzero entries of each column are either all positive or all negative. The relevance of dimensionality reduction schemes in neuroscience is described in a survey by Ganguli and Sompolinsky [18]. In convergent pathways in the brain, information stored in a massive number of neurons is compressed into a small number of neurons, and nonetheless the ability to perform the relevant computations is preserved. Modeling this information compression scheme requires a hypothesis regarding what properties of the original information must be accurately transmitted to the receiving neurons.

A plausible minimum requirement is that convergent pathways preserve the similarity structure of neuronal representations at the source area. This requirement is based on the experimental evidence that semantically similar objects in higher perceptual or association areas in the brain elicit similar neural activity patterns [29] and on the hypothesis that the similarity structure of the neural code is the basis of our ability to categorize objects and generalize responses to new objects [40]. It remains to select the appropriate mathematical measure of similarity. The candidate similarity measure considered in [18] is vector inner product, which conveniently gives rise to a model based on the JL transform. It is not difficult to see that for vectors x and y in the ℓ_2 unit ball, a $(1+\epsilon)$-approximation of $\|x\|_2$, $\|y\|_2$, and $\|x-y\|_2$ implies an additive error $\Theta(\epsilon)$ approximation of the inner product $\langle x, y\rangle$.

Suppose there are n "input" neurons at a source area and m "output" neurons at a target area. As shown in Figure 4.1, in this framework, the information at the input neurons is

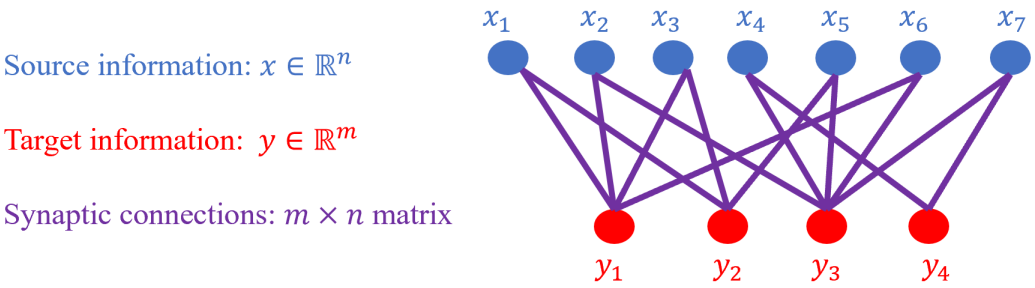

Figure 4.1: A model for information compression in the brain

represented as a vector in $\mathbb{R}^n$, the synaptic connections to output neurons are represented as a $m \times n$ matrix (with (i, j)th entry corresponding to the strength of the connection between input neuron j and output neuron i), and the information received by the output neurons is represented as a vector in $\mathbb{R}^m$. The similarity measure between two vectors v, w of neural information being taken to be $\langle v, w \rangle$ motivates modeling a synaptic connectivity matrix as a random $m \times n$ matrix drawn from a probability distribution that satisfies Requirement 2.2. Certain constraints on synaptic connectivity matrices arise from the biological limitations of neurons: the matrices must be *sparse* since a neuron is only connected to a small number (e.g. a few thousand) of postsynaptic neurons and *sign-consistent* since a neuron is usually purely excitatory or purely inhibitory.

This biological setting motivates the mathematical question: what is the optimal dimension and sparsity that can be achieved by a probability distribution over sparse, sign-consistent matrices that satisfies Requirement 2.2? Notice that the sparse JL transforms given in Definition 2.6 are far from sign-consistent: each nonzero entry is a column is given its own random sign. Nonetheless, as observed by Allen-Zhu, Gelashvili, Micali, and Shavit [4], a sparse, sign-consistent JL transform can be constructed with a modification: draw a single random sign for each column.[1] They proved that this distribution surprisingly permits efficient dimensionality reduction.

Theorem 4.1 (Sparse, sign-consistent JL [4]) For every $\epsilon > 0$, and $0 < \delta < 1/e$, a sparse, sign-consistent JL transform $\mathcal{A}'_{s,m,n}$ (defined formally in Definition 4.2) over $m \times n$ matrices with dimension $m = \Theta(\epsilon^{-2} \log^2(1/\delta))$ and sparsity $s = \Theta(\epsilon^{-1} \log(1/\delta))$ satisfies Requirement 2.2.

In [4], it was also proven that the additional $\log(1/\delta)$ factor on m is essentially necessary: namely, any distribution over sign-consistent matrices satisfying Requirement 2.2 requires $m = \tilde{\Omega}(\epsilon^{-2} \log(1/\delta) \min(\log(1/\delta), \log n))$. Thus, the dimension in Theorem 4.1 is essentially optimal. However, in order to achieve these upper bounds on m and s, the proof presented in [4] involved complicated combinatorics even more delicate than in the analysis of sparse JL in [28].

In Section 4.1.1, we describe how to construct the probability distribution of sparse, sign-consistent matrices analyzed in Theorem 4.1. In Section 4.1.2, we briefly describe the combinatorial proof of Theorem 4.1 presented in [4].

[1] Related mathematical work includes a construction of a dense, sign-consistent JL transform [39, 19].

4.1.1 Construction of Sparse, Sign-Consistent JL

Sparse, sign-consistent JL transforms, as constructed by Allen-Zhu et al. [4], are defined as follows.

Definition 4.2 Let $\mathcal{A}_{s,m,n}$ be a **sparse, sign-consistent JL transform** if the entries of a matrix $A \in \mathcal{A}_{s,m,n}$ are generated as follows. Let $A_{r,i} = \eta_{r,i}\sigma_i/\sqrt{s}$ where $\{\sigma_i\}_{i\in[n]}$ and $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ satisfy the following conditions:

- The families $\{\sigma_i\}_{i\in[n]}$ and $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ are independent from each other.
- The variables $\{\sigma_i\}_{i\in[n]}$ are i.i.d Rademachers (± 1 coin flips).
- The variables $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ are identically distributed Bernoullis ($\{0,1\}$ random variables) with expectation s/m.
- The $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ are independent across columns but not independent within each column. For every column $1 \le i \le n$, it holds that $\sum_{r=1}^{m} \eta_{r,i} = s$. Moreover, the random variables are *negatively correlated*: for every subset $S \subseteq [m]$ and every column $1 \le i \le n$, it holds that $\mathbb{E}\left[\prod_{r\in S} \eta_{r,i}\right] \le \prod_{r\in S} \mathbb{E}[\eta_{r,i}]$.

The key difference between Definition 4.2 and Definition 2.6 is that there is one random sign per column, rather than one random sign per entry in each column.

4.1.2 Discussion of the combinatorial analysis of [4]

Since A is linear, it suffices to consider $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$. Like in Section 2.2.2, it is easier to reason about ℓ_2^2 than ℓ_2, so we consider $\|Ax\|_2^2 - 1$. By a similar calculation to in Section 2.2.2, for sparse, sign-consistent JL, this becomes

$$T(x_1, \ldots, x_n) := \|Ax\|_2^2 - 1 = \frac{1}{s} \sum_{i\neq j} \sum_{r=1}^{m} \eta_{r,i}\eta_{r,j}\sigma_i\sigma_j x_i x_j.$$

Like in Section 2.2.2, it suffices to bound $\mathbb{E}[T(x_1, \ldots, x_n)^p]$ for an even integer $p = \Theta(\log(1/\delta))$.

In the analysis in [4], a complicated combinatorial argument was used to prove the following lemma, from which Theorem 4.1 follows:

Lemma 4.3 ([4]) If $s^2 \le m$ and $p < s$, then $\|T(x_1, \ldots, x_n)\|_p \lesssim \frac{p}{s}$ for all vectors $x \in \mathbb{R}^n$.

The argument in [4] to prove Lemma 4.3 was based on expanding $\mathbb{E}[Z^p]$ into a polynomial with $\approx n^{2p}$ terms, establishing a correspondence between the monomials and the multigraphs, and then doing combinatorics to analyze the resulting sum. The approach of mapping monomials to graphs is commonly used in analyzing the eigenvalue spectrum of random matrices [47, 17] and was also used in [28] to analyze sparse JL. The analysis in [4] borrowed some methods from the analysis in [28]; however, the additional correlations between the Rademachers imposed by sign-consistency forced the analysis in [4] to require more delicate manipulations at several stages of the computation.

The expression to be analyzed was $s^p\mathbb{E}[T(x_1,\ldots,x_n)^p]$, which was written as:

$$\sum_{i_1,\ldots,i_p,j_1,\ldots,j_p\in[n],i_1\neq j_1,\ldots,i_p\neq j_p}\left(\prod_{u=1}^{p}x_{i_u}x_{j_u}\right)\left(\mathbb{E}_\sigma\prod_{u=1}^{p}\sigma_{i_u}\sigma_{j_u}\right)\left(\mathbb{E}_\eta\prod_{u=1}^{t}\sum_{r=1}^{m}\eta_{r,i_u}\eta_{r,j_u}\right).$$

After layers of computation, it was shown that

$$s^p\mathbb{E}[Z^p]\leq e^p\sum_{v=2}^{p}\sum_{G\in\mathcal{G}_{v,p}}\left((1/p^p)\prod_{q=1}^{v}\sqrt{d_q}^{d_q}\right)\sum_{r_1,\ldots,r_p\in[m]}\prod_{i=1}^{w}(s/m)^{v_i}$$

where $\mathcal{G}_{v,p}$ is a set of directed multigraphs with v labeled vertices and t labeled edges, where d_q is the total degree of vertex $q\in[v]$ in a graph $\mathcal{G}_{v,p}$, and where w and $v_1,\ldots,v_w$ are defined by G and the edge colorings $r_1,\ldots,r_t$. The problem then boiled down to carefully enumerating the graphs in $\mathcal{G}_{v,p}$ in six stages and analyzing the resulting expression.

4.2 Our results

We present a simpler, combinatorics-free proof of Theorem 4.1. This analysis shares common technical ingredients with our proof of Theorem 3.2, and thus takes a step towards a unified analysis of JL transforms. We discuss the analysis methods in greater detail in Chapter 5.

Moreover, our analysis also yields dimension/sparsity tradeoffs, which were not previously known.[2] We prove the following:

Theorem 4.4 For every $\epsilon>0$, $0<\delta<1$, and $e\leq B\leq\frac{1}{\delta}$, there exists a probability distribution $\mathcal{A}$ over $m\times n$ real, sign-consistent matrices with $m=\Theta(B\epsilon^{-2}\log_B^2(1/\delta))$ and sparsity $s=\Theta(\epsilon^{-1}\log_B(1/\delta))$ that satisfies Requirement 2.2.

Notice Theorem 4.1 is recovered if $B=e$. For larger B values, Theorem 4.4 enables a $\log B$ factor reduction in sparsity at the cost of a $B/\log^2 B$ factor gain in dimension. These dimension-sparsity tradeoffs bear a strong resemblance to the dimension-sparsity tradeoffs (Theorem 2.5) for the standard sparse JL transform.

[2] In Appendix A of [22], we point out the limiting lemma in the combinatorial analysis in [4], which prevents dimension-sparsity tradeoffs from being attainable through this approach, due to an assumption that is implicitly used in the analysis. For sparse JL, it is similarly not known how to obtain these tradeoffs via the combinatorial approach of [28].

Chapter 5

A Perspective on Analyzing Johnson-Lindenstrauss transforms

The core ingredient in analyzing Johnson-Lindenstrauss transforms is understanding the error term $\|Ax\|_2^2 - 1$. Since tail bounds are related to moment bounds by Markov's inequality (Lemma 2.7), the analysis boils down to analyzing moments of this error term.

For sparse JL, as discussed in Section 2.2.2, the error term becomes:

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{i \neq j} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j = \frac{1}{s} \sum_{i \neq j} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j.$$

For sparse, sign-consistent JL, as discussed in Section 4.1.2, the error term becomes:

$$T(x_1, \ldots, x_n) = \|Ax\|_2^2 - 1 = \frac{1}{s} \sum_{i \neq j} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_i \sigma_j x_i x_j.$$

Standard proofs [28, 4, 16, 36] bound moments via expanding out the pth power of the error term as a sum of a large collection of monomials. Then, the sum of monomials is written as expressed as a sum of a certain collection of multigraphs, which are carefully enumerated and handled. (See Section 4.1.2 for an overview of this combinatorial analysis for sparse sign-consistent JL.) Each of the aforementioned papers uses a different, delicate method to enumerate the graphs and bound each term, due to the specific constraints of each setting.

In this chapter, we describe a new conceptual perspective on analyzing Johnson-Lindenstrauss transforms that we believe provides a greater degree of simplicity and unification. In particular, we believe our unified approach obviates the need to build entirely distinct analyses for each setting.

High-level approach

Our approach builds on the framework for Cohen et al. [12] for giving a cleaner proof of Theorem 2.4 for sparse JL. As described in Section 2.2.2, the idea in [12] is to view the error term $R(x_1, \ldots, x_n)$ as a quadratic form of Rademacher random variables. That is, it can

be expressed as $\sum_{k,l} a_{k,l}\sigma_k\sigma_l$ for an appropriate (random) matrix $(a_{k,l})_{k,l}$. Now, the random variables are peeled off in layers by expressing

$$\mathbb{E}[R(x_1,\ldots,x_n)^p] = \mathbb{E}_\eta[\mathbb{E}_\sigma[R(x_1,\ldots,x_n)^p]].$$

For each instantiation of the η random variables, the inner expression $\mathbb{E}_\sigma[R(x_1,\ldots,x_n)^p]$ is a quadratic form of Rademachers with a non-random matrix. In [12], this quadratic form is handled by using that that Rademachers are sub-gaussian variables, and applying gaussian quadratic form bounds (Lemma 5.3).

Our approach diverges from [12] in how we handle the expression $\mathbb{E}_\sigma[R(x_1,\ldots,x_n)^p]$. As we will describe in Section 5.2, it turns out that using gaussian quadratic forms is too weak to result in the desired results in our settings in Chapter 3 and Chapter 4. While the analysis in [12] achieves the optimal dimension for Theorem 2.4 using a tight bound on gaussian quadratic form moments, we give counterexamples that show that this bound is too loose in our settings to result in the desired results. We thus require a separate treatment of quadratic forms of Rademachers. As we will discuss in Section 5.1, Rademacher quadratic forms can have much smaller moments than gaussian quadratic forms, and we need to take advantage of that distinction in our proofs to recover the desired results.

There has been a thorough study of tight bounds on Rademacher linear and quadratic forms in the probability theory literature [21, 31, 32]. However, it is not clear how to directly apply these bounds in our settings since they suffer from tractability issues. The crux is that while these bounds are focused on obtaining tight estimates for quadratic forms with scalar coefficients, our analysis needs to be tractable for quadratic forms with random variable coefficients. In particular, to bound $\mathbb{E}_\sigma[\mathbb{E}_\eta[||Ax||_2^2 - 1]]$, we need to be able take an expectation over the η random variables of our bound on the Rademacher quadratic form $\mathbb{E}_\eta[||Ax||_2^2 - 1]$.

With these restrictions in mind, our analysis boils down to constructing bounds on Rademacher quadratic form moments that borrow intuition from previous bounds [21, 31, 32], while being sufficiently tractable for our setting. For the remainder of the thesis, we use the notation $\|X\|_T$ for a random variable X to denote the T-norm of X, i.e. $\mathbb{E}[|X|^T]^{1/T}$. Our main technical ingredient is bounds on the T-norm of Rademacher quadratic forms that are tractable to analyze when the coefficients are random variables.

In Section 5.1, we give intuition for why gaussian bounds can be very loose in the Rademacher setting, and review known Rademacher bounds. In Section 5.2, we give counterexamples that demonstrate that gaussian bounds are too weak for our settings in Chapter 3 and Chapter 4. In Section 5.3, we describe our cleaner bounds on Rademacher quadratic forms and other related sums of random variables.

5.1 Rademachers versus gaussians: the distinction

We give some intuition for the distinction between Rademacher and gaussian moment bounds. In Section 5.1.1, we describe the structure of these bounds for the linear form setting. In Section 5.1.2, we draw upon the intuition from the linear form setting to describe the structure of these bounds for the quadratic form setting.

5.1.1 Linear forms

At first glance, it may seem surprising that (sub)-gaussian moment bounds can become very loose for Rademachers. The concept that drives this difference can be illustrated in the linear form setting. Gaussian moment bounds yield the following bound on linear forms of Rademachers:

Lemma 5.1 (Khintchine) Suppose $\sigma_1, \sigma_2 \ldots, \sigma_n$ are i.i.d Rademachers, $x = [x_1, \ldots, x_n]$ is a vector in $\mathbb{R}^n$ such that $|x_1| \geq |x_2| \geq \ldots \geq |x_n|$. Suppose $g_1, \ldots, g_n$ are i.i.d gaussians. Then, for any even integer $T \geq 2$, it holds that:

$$\left\| \sum_{i=1}^n \sigma_i x_i \right\|_T \lesssim \left\| \sum_{i=1}^n g_i x_i \right\|_T \simeq \sqrt{T} \left\| x \right\|_2 .$$

Proof. Since the Rademachers σ_i are subgaussian, we can easily see (from expanding) that $\|\sum_{i=1}^n \sigma_i x_i\|_T \lesssim \|\sum_{i=1}^n g_i x_i\|_T$. Notice that $\sum_{i=1}^n g_i x_i$ is itself a gaussian with mean 0 and variance $\|x\|_2^2$. As a result, it follows from gaussian moment bounds that $\|\sum_{i=1}^n g_i x_i\|_T \simeq \sqrt{T} \|x\|_2$. □

While $\sqrt{T} \|x\|_2$ is tight for gaussian linear form moments, this bound *cannot* be a tight bound on $\|\sum_{i=1}^n \sigma_i x_i\|_T$ for the following reason: As $T \to \infty$, the quantity $\sqrt{T} \|x\|_2$ goes to infinity, while for any $T \geq 1$, the quantity $\|\sum_{i=1}^n \sigma_i x_i\|_T$ is bounded by $\|x\|_1$. The fundamental issue is that the gaussian bound does not take advantage of the fact that Rademachers are *bounded*, while gaussians have tails that extend to infinitely. For sufficiently high moments, which are determined by tail behavior, this distinction causes gaussian bounds to fail to capture Rademacher moments.

A result due to Hitczenko [21] indicates that the tight bound for Rademacher linear forms is actually the following combination of the ℓ_2 and ℓ_1 norm bounds:.

Lemma 5.2 (Hitczenko [21]) Suppose $\sigma_1, \sigma_2 \ldots, \sigma_n$ are i.i.d Rademachers, $x = [x_1, \ldots, x_n]$ is a vector in $\mathbb{R}^n$ such that $|x_1| \geq |x_2| \geq \ldots \geq |x_n|$. Then, for any even integer $2 \leq T \leq n$, it holds that:

$$\left\| \sum_{i=1}^n \sigma_i x_i \right\|_T \simeq \sum_{i=1}^T |x_i| + \sqrt{T} \sqrt{\sum_{i>T} x_i^2}.$$

In this bound, the "big" terms (i.e. terms involving x_1, x_2, …, x_T) are handled with an ℓ_1-norm bound, while the remaining terms are approximated as gaussians and bounded with an ℓ_2-norm bound. The upper bound can be seen easily by splitting into "big" and "small" terms. Indeed, we can express:

$$\left\| \sum_{i=1}^n \sigma_i x_i \right\|_T \lesssim \left\| \sum_{i=1}^T \sigma_i x_i \right\|_T + \left\| \sum_{i=T+1}^n \sigma_i x_i \right\|_T \lesssim \sum_{i=1}^T |x_i| + \sqrt{T} \sqrt{\sum_{i>T} x_i^2}.$$

The fact that this simple bound is tight is surprising, and we refer the reader to the proofs of the lower bound in [31, 21].

5.1.2 Quadratic forms

The situation is similar for quadratic form bounds, though the resulting bounds become much more complex. First, we recall the standard bound for gaussian quadratic forms, that yields the following bound for Rademachers:

Lemma 5.3 (Hanson-Wright bound [20]) Let σ be a d-dimensional vector of independent Rademachers, and let g be a d-dimensional vector of gaussians. Let $A = (a_{k,l})$ be a symmetric $d \times d$ matrix with zero diagonal. Then, for any even integer $T \geq 2$, it holds that:

$$\left\|\sigma^T A \sigma\right\|_T \lesssim \left\|g^T A g\right\|_T \sim \sqrt{T}\sqrt{\sum_{k=1}^{d}\sum_{l=1}^{d} a_{k,l}^2} + T\left(\sup_{\|y\|_2=1} |y^T A y|\right).$$

Notice that the gaussian quadratic form bound in Lemma 5.3 is already far more complicated than the gaussian linear form bound in Lemma 5.1.

A complication similar to the linear form setting arises when the Hanson-Wright bound is applied to Rademachers. While the Hanson-Wright bound is tight for gaussians, this bound *can't* be a tight bound on $\left\|\sigma^T A \sigma\right\|_T$ for the following reason: As $T \to \infty$, the quantity $\sqrt{T}\sqrt{\sum_{k=1}^{d}\sum_{l=1}^{d} a_{k,l}^2}$ goes to ∞, while for any $T \geq 1$, the quantity $\left\|\sigma^T A \sigma\right\|_T$ is bounded by the entrywise ℓ_1-norm $\sum_{k=1}^{d}\sum_{l=1}^{d} |a_{k,l}|$.

There is analogously a closed-form tight bound for the Rademacher quadratic forms, though the expression is not nearly as intuitive as in the linear form case. Latała showed the following moment bounds on Rademacher quadratic forms [32].[1]

Lemma 5.4 ([32]) Let T be an even natural number. Let $\sigma_1, \ldots, \sigma_n$ be independent Rademachers and let $(a_{i,j})$ a symmetric matrix with zero diagonal. Then:

$$\left\|\sum_{1\leq i,j\leq n} a_{i,j}\sigma_i\sigma_j\right\|_T \simeq \left(\sup_{\|b\|_2,\|c\|_2\leq\sqrt{T},\|b\|_\infty,\|c\|_\infty\leq 1}\sum_{1\leq i,j\leq n} a_{i,j}b_i c_j\right) + \sum_{1\leq i\leq T} A_{(i)} + \sqrt{T}\sqrt{\sum_{T<i\leq n}(A_{(i)})^2}$$

where $A_i = \sqrt{\sum_{1\leq j\leq n} a_{i,j}^2}$ and $A_{(1)} \geq A_{(2)} \ldots \geq \ldots A_{(n)}$ is a permutation of $A_1, \ldots, A_n$.

The challenge in applying Lemma 5.4 in our setting is that the terms can become intractable when $(a_{i,j})$ is a random matrix.

In the case of random variable coefficients, we observe that Latała's bound takes the following form.

Lemma 5.5 Let T be an even integer, $\{\sigma_i\}_{1\leq i\leq n}$ be independent Rademachers, and $(Y_{i,j})_{1\leq i,j\leq n}$ be a $n \times n$ symmetric, nonnegative random matrix with zero diagonal (i.e. $Y_{i,i} = 0$) such that $\{Y_{i,j}\}_{1\leq i,j\leq n}$ is independent from $\{\sigma_i\}_{1\leq i\leq n}$. If $W_i = \sqrt{\sum_{1\leq j\leq n} Y_{i,j}^2}$, then:

$$\left\|\sum_{1\leq i,j\leq n} Y_{i,j}\sigma_i\sigma_j\right\|_T \simeq \left\|\sup_{\|b\|_2,\|c\|_2\leq\sqrt{T},\|b\|_\infty,\|c\|_\infty\leq 1}\sum_{1\leq i,j\leq n} Y_{i,j}b_i c_j\right\|_T + \left\|\sum_{1\leq i\leq T} W_{(i)} + \sqrt{T}\sqrt{\sum_{T<i\leq n} W_{(i)}^2}\right\|_T$$

[1] In fact, Latała shows moment bounds for much more general quadratic forms, but for the application to JL, we only need the following bound in the special case of Rademachers.

where $W_{(1)} \geq W_{(2)} \geq \ldots \geq \ldots W_{(n)}$ is a permutation of $W_1, \ldots, W_n$.

Proof of Lemma 5.5. To prove Lemma 5.5, we apply Lemma 5.4 to the case where the $a_{i,j}$ are themselves random variables. Let $Q = \sum_{1 \leq i \neq j \leq n} Y_{i,j}\sigma_i\sigma_j$. Applying Lemma 5.4, we have that:

$$\begin{aligned}
\left(\mathbb{E}_{Y,\sigma}[Q^T]\right)^{1/T} &= \left(\mathbb{E}_Y \mathbb{E}_\sigma[Q^T]\right)^{1/T} \\
&= \left(\mathbb{E}_Y \left[\left\| \sum_{1 \leq i \neq j \leq n} Y_{i,j}\sigma_i\sigma_j \right\|_T^T\right]\right)^{1/T} \\
&\simeq \left\| \sup_{\|b\|_2, \|c\|_2 \leq \sqrt{T}, \|b\|_\infty, \|c\|_\infty \leq 1} \sum_{1 \leq i \neq j \leq n} Y_{i,j} b_i c_j + \sum_{1 \leq i \leq T} W_{(i)} + \sqrt{T}\sqrt{\sum_{T < i \leq n} W_{(i)}^2} \right\|_T \\
&\simeq \left\| \sup_{\|b\|_2, \|c\|_2 \leq \sqrt{T}, \|b\|_\infty, \|c\|_\infty \leq 1} \sum_{1 \leq i \neq j \leq n} Y_{i,j} b_i c_j \right\|_T + \left\| \sum_{1 \leq i \leq T} W_{(i)} + \sqrt{T}\sqrt{\sum_{T < i \leq n} W_{(i)}^2} \right\|_T
\end{aligned}$$

where the last line follows from the fact that the the $Y_{i,j}$ are nonnegative, so each term is nonnegative, so the triangle inequality results in at most a factor of 2 of gain. □

In Lemma 5.5, notice that term $\left\|\sup_{\|b\|_2, \|c\|_2 \leq \sqrt{T}, \|b\|_\infty, \|c\|_\infty \leq 1} \sum_{1 \leq i,j \leq n} Y_{i,j} b_i c_j \right\|_T$ can be viewed as a generalization of the operator norm to the ℓ_2-ball truncated by ℓ_∞ planes. Due to the asymmetrical geometry of the ℓ_2 ball truncated by ℓ_∞ planes, this term becomes especially messy in our setting where the coefficients are random variables. In particular, this operator-norm-like term can become intractable to directly handle when taking an expectation over the η random variables.

5.2 Failure of Hanson-Wright bound for our settings

We show that the Hanson-Wright bound (Lemma 5.3) is too loose for our settings in Chapter 3 and Chapter 4.

5.2.1 Hanson-Wright is too loose for Chapter 3

We show that applying the Hanson-Wright bound (Lemma 5.3) to analyze $\mathbb{E}_\sigma[R(x_1, \ldots, x_n)^p]$ is not sufficiently precise for the setting in Chapter 3, even for the simplest case where $s = 1$. Notice that applying the Hanson-Wright bound to analyze $\mathbb{E}_\sigma[R(x_1, \ldots, x_n)^p]$ effectively approximates $R(x_1, \ldots, x_n)$ as

$$R_g(x_1, \ldots, x_n) = \frac{1}{s}\sum_{r=1}^{m} \sum_{i \neq j} \eta_{r,i}\eta_{r,j} g_{r,i} g_{r,j} x_i x_j$$

where the $g_{r,i}$ are i.i.d standard gaussians. However, we need different technical tools for two reasons.

1. First, in order to upper bound $v(m, \epsilon, \delta, s)$, we need to *lower bound* $\|\|R(x_1, \ldots, x_n)\|_q$, and thus cannot simply consider $\|R_g(x_1, \ldots, x_n)\|_q$.

2. Second, even to lower bound $v(m, \epsilon, \delta, s)$, using $\|R_g(x_1, \ldots, x_n)\|_q$ as a upper bound for $\|R(x_1, \ldots, x_n)\|_q$ is not sufficiently strong. Below, we give a counter-example, i.e. a vector x, where $\|R_g(x_1, \ldots, x_n)\|_q$ is too large to recover a tight lower bound.

Thus, we cannot use the Hanson-Wright bound in this setting, and need to come up with a better bound on $\|R(x_1, \ldots, x_n)\|_q$ that does not implicitly replace Rademachers by gaussians.

We now show point 2: that the Hanson-Wright bound is not sufficiently strong to obtain a tight lower bound $v(m, \epsilon, \delta, s)$. We consider $R_g(x_1, \ldots, x_n) = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\eta_{r,i}\eta_{r,j}g_{r,i}g_{r,j}x_i x_j$, as above, where the $g_{r,i}$ are i.i.d standard gaussians. We consider p equal to $\ln(1/\delta)$ rounded up to the nearest even integer, and we consider a vector of the form $[v, \ldots, v, 0, \ldots, 0]$ where $\frac{1}{v^2}$ is an integer and $v \geq 0$. We show $\|R_g(v, \ldots, v, 0, \ldots, 0)\|_p \gtrsim \omega(\epsilon)$ for a certain v value, where we know it to be true that $\|R(v, \ldots, v, 0, \ldots, 0)\|_p \lesssim \epsilon$.

Let's consider a vector $[v, \ldots, v, 0, \ldots, 0]$ where $\frac{1}{v^2}$ is an integer and $v \geq 0$. We apply the Hanson-Wright bound (which is tight for gaussians) to obtain:

$$\begin{aligned}\|R_g(v, \ldots, v, 0, \ldots, 0)\|_p &\gtrsim \frac{pv^2}{s}\left\|\sup_{\|x\|_2,\|y\|_2\leq 1}\sum_{r=1}^{m}\sum_{1\leq i\neq j\leq N}\eta_{r,i}\eta_{r,j}x_{r,i}y_{r,j}\right\|_p \\ &\geq \frac{pv^2}{s}\left\|\sup_{\|x\|_2,\|y\|_2\leq 1}\sum_{1\leq i\neq j\leq N}\eta_{1,i}\eta_{1,j}x_i y_j\right\|_p.\end{aligned}$$

Let $M = \sum_{i=1}^{N}\eta_{1,i}$. Let $S \subseteq [N]$ be the set of indices where $\eta_{1,i} = 1$. We can set the vector to $x_i = y_i = \frac{1}{\sqrt{M}}$ for all $i \in S$ and 0 elsewhere. This gives us:

$$\left\|\sup_{\|x\|_2,\|y\|_2\leq 1}\sum_{1\leq i\neq j\leq N}\eta_{1,i}\eta_{1,j}x_i y_j\right\|_p \geq \|M-1\|_p = \left\|\sum_{i=1}^{N}\eta_{1,i} - 1\right\|_p \gtrsim \left\|I_{\sum_{i=1}^{N}\eta_{1,i}\geq 2}\sum_{i=1}^{N}\eta_{1,i}\right\|_p.$$

We can expand out this moment to obtain:

$$\begin{aligned}\mathbb{E}\left[\left(I_{\sum_{i=1}^{N}\eta_{1,i}\geq 2}\sum_{i=1}^{N}\eta_{1,i}\right)^p\right] &\geq C^p\sum_{M=2}^{p}\binom{N}{M}M^p\left(\frac{s}{m}\right)^M\left(1-\frac{s}{m}\right)^M \\ &\geq C^p\sum_{M=2}^{p}\left(\frac{N}{p}\right)^M\left(\frac{p}{M}\right)^M M^p\left(\frac{s}{m}\right)^M\left(1-\frac{s}{m}\right)^M \\ &= C^p\sum_{M=2}^{p}\left(\frac{s}{pmv^2}\right)^M M^p\left(\frac{p}{M}\right)^M\left(1-\frac{s}{m}\right)^M.\end{aligned}$$

Since $M \leq p$, we know that $\left(\frac{p}{M}\right)^M \geq 1$. Moreover, as long as $p \geq \frac{se}{mv^2}$, we know that $\left(1-\frac{s}{m}\right)^{M/p} \geq \left(1-\frac{s}{m}\right)^{N/p} \geq \left(1-\frac{s}{m}\right)^{\frac{m}{s}} \geq 0.3$. Thus we obtain a bound of

$$D^p\sum_{M=2}^{p}M^p\left(\frac{s}{pmv^2}\right)^M.$$

If $2 \leq \frac{p}{\ln(pmv^2/s)} \leq p$ (which can be written as $1 \leq \ln(pmv^2/s) \leq \frac{p}{2}$), then we know that:

$$pv^2 \left\| \sup_{\|x\|_2, \|y\|_2 \leq 1} \sum_{1 \leq i \neq j \leq N} \eta_{1,i}\eta_{1,j} x_i y_j \right\|_p \gtrsim \frac{p^2 v^2}{\ln(pmv^2/s)}.$$

Now, we show that when $s = 1$, the value $v = \sqrt{\epsilon}\frac{\ln(\frac{m\epsilon}{p})}{p}$ results in $\|R_g(v, \ldots, v, 0, \ldots, 0)\|_p \gtrsim \omega(\epsilon)$, when $\sqrt{\epsilon}\min\left(\frac{\ln(\frac{m\epsilon}{p})}{p}, \frac{\sqrt{\ln(\frac{m\epsilon^2}{p})}}{\sqrt{p}}\right) = \sqrt{\epsilon}\frac{\ln(\frac{m\epsilon}{p})}{p}$, $m \geq \Theta(\epsilon^{-2}p)$, and $\ln(\frac{m\epsilon}{p}) \leq \sqrt{p}$. This is problematic because under these conditions, it holds that $\|R(v, \ldots, v, 0, \ldots, 0)\|_p \lesssim \epsilon$ (as shown in the proof of Theorem 3.1); moreover, this bound is required to deduce the tight bounds in Theorem 3.1 and Theorem 3.2.

We know that $v \leq \frac{\sqrt{\epsilon}}{\sqrt{p}}$, so $\sqrt{\epsilon}\min\left(\frac{\ln(\frac{m\epsilon}{p})}{p}, \frac{\sqrt{\ln(\frac{m\epsilon^2}{p})}}{\sqrt{p}}\right) = v$. Moreover, we show that $1 \leq \ln(pmv^2/s) \leq \frac{p}{2}$. In particular, notice that:

$$\frac{pmv^2}{e} = \ln^2\left(\frac{m\epsilon}{p}\right)\frac{m\epsilon}{pe} \geq 1,$$

$$pmv^2 = \ln^2\left(\frac{m\epsilon}{p}\right)\frac{m\epsilon}{p} \leq m\epsilon \leq p \cdot e^{\sqrt{p}} \leq e^{p/2}.$$

As a result, it holds that:

$$\|R_g(v, \ldots, v, 0, \ldots, 0)\|_p \gtrsim \frac{p^2 v^2}{\ln(pmv^2)} \gtrsim \frac{\epsilon \ln^2(m\epsilon/p)}{\ln(\frac{m\epsilon}{p})} \geq \epsilon \ln\left(\frac{m\epsilon}{p}\right) = \omega(\epsilon).$$

5.2.2 Hanson-Wright is too loose for Chapter 4

We show that applying the Hanson-Wright bound (Lemma 5.3) to analyze $\mathbb{E}_\sigma[T(x_1, \ldots, x_n)^p]$ is not sufficiently precise for the setting in Chapter 4. We view the random variable $T(x_1, \ldots, x_n)$ as a quadratic form $\frac{1}{s}\sigma^T A\sigma$, where σ an n-dimensional vector of independent Rademachers and A is a symmetric, zero-diagonal $n \times n$ matrix where the (i, j)th entry (for $i \neq j$) is $x_i x_j \sum_{r=1}^m \eta_{r,i}\eta_{r,j}$. We let $Q_{i,j} = \sum_{r=1}^m \eta_{r,i}\eta_{r,j}$, so that we can write the (i, j)th entry as $Q_{i,j}x_i x_j$. Applying the Hanson-Wright bound followed by an expectation over the η values yields

$$\left\|\sigma^T A\sigma\right\|_p \lesssim \left\| \sqrt{p}\sqrt{\sum_{i=1}^n \sum_{j \leq n, j \neq i} Q_{i,j}^2 x_i^2 x_j^2} + p \sup_{\|y\|_2 = 1} \left| \sum_{i=1}^n \sum_{j \leq n, j \neq i} Q_{i,j} x_i x_j y_i y_j \right| \right\|_p =: U_p. \quad (5.1)$$

We show that the vector $x = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0] \in \mathbb{R}^n$ forces U_p to be too large to yield the optimal m value, thus proving that the Hanson-Wright bound does not provide a sufficiently tight bound on $\left\|T(1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)\right\|_p$ to achieve Theorem 4.1. The main ingredient in our proof is the following lemma, which we prove in subsection C.1:

Lemma 5.6 For every column $1 \leq i \leq n$, suppose that the random variables $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ have the distribution defined by uniformly choosing exactly s of the variables per column. If $x = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0\right]$, $p < s$ and $B = m/s^2 \leq \frac{e^p}{p}$, then

$$U_p \simeq \begin{cases} \frac{p^2}{\log Bp} & \text{if } B \geq \frac{e}{p} \\ \frac{p}{B} & \text{if } B < \frac{e}{p}. \end{cases}$$

We can obtain bounds on s and m from Lemma 5.6 via Markov's inequality. We disregard the case where $B \geq \frac{e^p}{p}$, since this case would yield a value for m that is not polynomial in $\log(1/\delta)$. If $B < e/p$, then it follows that $s = \Theta(\epsilon^{-1}B^{-1}\log(1/\delta)) = \Omega(\epsilon^{-1}\log^2(1/\delta))$ and $m = \Theta(\epsilon^{-2}B^{-1}\log^2(1/\delta)) = \Omega(\epsilon^{-2}\log^3(1/\delta))$. If $B \geq e/p$, then it follows that $s = \Theta(\epsilon^{-1}p^2/\log(Bp)) = \Omega(\epsilon^{-1}\log(1/\delta))$ and $m = \Theta(\epsilon^{-2}p^4B/\log^2(Bp)) = \Omega(\epsilon^{-2}\log^3(1/\delta))$. These bounds on m incur an extra $\log(1/\delta)$ factor, and thus the Hanson-Wright bound is too weak for this setting. Now, it suffices to prove Lemma 5.6.

We assume that $x = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0\right]$ and that the random variables $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ have the distribution defined by uniformly choosing exactly s of the variables per column. We first show the following computation of $||Q_{i,j}||_p$.

Proposition 5.7 Assume that the random variables $\{\eta_{r,i}\}_{r\in[m],i\in[n]}$ have the distribution defined by uniformly choosing exactly s of the variables per column. Then, if $p < s$ and $X \sim \text{Bin}(s, s/m)$, we have that $||Q_{i,j}||_p \simeq ||X||_p$.

Proof. We condition on the event that the nonzero locations in column i are at $r_1, r_2, \ldots, r_s$. Notice that the random variable $(Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1)$ is distributed as $Z_{r_1} + Z_{r_2} + \ldots + Z_{r_s}$ where Z_{r_k} is an indicator for the kth entry in the jth column being nonzero. Let Z'_{r_k} for $1 \leq k \leq s$ be i.i.d random variables distributed as $\text{Bern}(s/m)$. Now, observe that

$$\mathbb{E}[(Z_{r_1} + Z_{r_2} + \ldots + Z_{r_s})^p] = \sum_{\substack{0\leq t_1,t_2,\ldots,t_s\leq p \\ t_1+t_2+\ldots+t_s=p}} \mathbb{E}[\prod_{i=1}^{s} Z_{r_i}^{t_i}] = \sum_{\substack{0\leq t_1,t_2,\ldots,t_s\leq p \\ t_1+t_2+\ldots+t_s=p}} \mathbb{E}[\prod_{i|t_i>0} Z_{r_i}].$$

Notice that $\mathbb{E}[(Z'_{r_1} + Z'_{r_2} + \ldots + Z'_{r_s})^p] = \sum_{0\leq t_1,t_2,\ldots,t_s\leq p, t_1+t_2+\ldots+t_s=p} \mathbb{E}[\prod_{i|t_i>0} Z'_{r_i}]$. Thus, it suffices to compare $\mathbb{E}[\prod_{i|t_i>0} Z_{r_i}]$ and $\mathbb{E}[\prod_{i|t_i>0} Z'_{r_i}]$. We see that $\mathbb{E}[\prod_{i|t_i>0} Z'_{r_i}] = \left(\frac{s}{m}\right)^{|\{i|t_i>0\}|}$. Since $p < s$, we see that $\mathbb{E}[\prod_{i|t_i>0} Z_{r_i}] = \prod_{j=0}^{|\{i|t_i>0\}|-1} \frac{s-j}{m-j}$. It is not difficult to verify that this ratio is bounded by $2^{O(p)}$ as desired, so

$$\frac{\mathbb{E}[(Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1)^p]}{\mathbb{E}[X^p]} = \frac{\mathbb{E}[(Z_{r_1} + Z_{r_2} + \ldots + Z_{r_s})^p]}{\mathbb{E}[X^p]} \geq 2^{-O(p)}.$$

Now, by the law of total expectation, we know that

$$\frac{\mathbb{E}[Q_{i,j}^p]}{\mathbb{E}[X^p]} \geq 2^{-O(p)}$$

as desired. □

We now prove the following relation between U_p and $\|Q_{1,2}\|_p$:

Lemma 5.8 Assume the notation and restrictions above. Then $U_p \simeq p \|Q_{1,2}\|_p$.

Proof of Lemma 5.8. For ease of notation, we define

$$S_1 := p \sup_{\|y\|_2=1} \left| \sum_{i=1}^{n} \sum_{j \le n, j \neq i} Q_{i,j} x_i x_j y_i y_j \right|$$

$$S_2 := \sqrt{p} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} Q_{i,j}^2 x_i^2 x_j^2}.$$

Our goal is to calculate $U_p = \|S_1 + S_2\|_p$. We make use of the following upper and lower bounds on $\|S_1 + S_2\|_p$:

$$\left| \|S_1\|_p - \|S_2\|_p \right| \le \|S_1 - S_2\|_p \le \|S_1 + S_2\|_p \le \|S_1\|_p + \|S_2\|_p . \tag{5.2}$$

In order to compute $\left| \|S_1\|_p - \|S_2\|_p \right|$ and $\|S_1\|_p + \|S_2\|_p$, we first compute $\|S_1\|_p$ and $\|S_2\|_p$. For our choice of x, notice

$$\|S_1\|_p \simeq p \left\| \sup_{\|y\|_2=1} |Q_{1,2} y_1 y_2| \right\|_p \simeq p \|Q_{1,2}\|_p$$

$$\|S_2\|_p \simeq \sqrt{p} \left\| \sqrt{Q_{1,2}^2} \right\|_p = \sqrt{p} \|Q_{1,2}\|_p .$$

From these bounds, coupled with (5.2), it follows that $\|U\|_p \simeq p \|Q_{1,2}\|_p$ as desired. □

We now show Lemma 5.6 follows from Lemma 5.8 and Proposition 5.7.

Proof of Lemma 5.6. After applying Lemma 5.8, it suffices to calculate $\|Q_{1,2}\|_p$. It follows from Proposition 5.7 that $\|Q_{1,2}\|_p \simeq \|X\|_p$ where X is distributed as $\mathrm{Bin}(s, s/m)$. Now, the following calculation $\|X\|_p$ for $p < s$ and $B = m/s^2 \le \frac{e^p}{p}$ follows from the lower and upper bounds of Lemma 5.11 (Latała's bound on moments of sums of i.i.d nonnegative random variables):

$$\|X\|_p \simeq \begin{cases} \frac{p}{\log Bp} & \text{if } B \ge \frac{e}{p} \\ \frac{1}{B} & \text{if } B < \frac{e}{p} \end{cases} .$$

From this, Lemma 5.6 follows. □

5.3 Rademacher moment bounds

The main ingredient in our proofs is to design moment bounds for Rademacher quadratic forms that are tractable when the coefficients are random variables. As mentioned in Section 5.1, the tight bound on quadratic forms of Rademachers in Lemma 5.4 becomes particularly messy when the coefficients are random variables. We give cleaner (though weaker) bounds that take further advantage of structure of the error terms $R(x_1, \ldots, x_n)$ and $T(x_1, \ldots, x_n)$. Due to the ubiquity of moment bounds in theoretical computer science, we hope that these bounds can be of more general use.

5.3.1 Revisiting the linear form setting

As a starting point, we revisit the linear form setting, considering linear forms of symmetric random variables. Linear forms naturally arise in the sparse JL error term $R(x_1, \ldots, x_n)$ since

$$\sum_{i \neq j} \eta_{r,i}\sigma_{r,i}\eta_{r,j}\sigma_{r,j}x_j x_i = \left(\sum_{1 \leq i \leq n} \eta_{r,i}\sigma_{r,i}x_i\right)^2 - \sum_{1 \leq i \leq n} \eta_{r,i}x_i^2 \leq \left(\sum_{1 \leq i \leq n} \eta_{r,i}\sigma_{r,i}x_i\right)^2.$$

Notice that $\mathbb{E}_\sigma\left[\sum_{1 \leq i \leq n} \eta_{r,i}\sigma_{r,i}x_i\right]$ falls within scope of Lemma 5.2. Let's imagine though that we use Lemma 5.2 to handle $\mathbb{E}_\sigma\left[\sum_{1 \leq i \leq n} \eta_{r,i}\sigma_{r,i}x_i\right]$. The issue is that the bound involves sorting the weighted terms $\eta_{r,i}x_i$ in ascending order; however, this becomes somewhat messy since it requires reasoning about the order statistics of these random variables. We provide an alternate (though potentially weaker) bound for linear forms of symmetric random variables that avoids these complications. Moreover, this bound also turns out to be of use for analyzing the sparse, sign-consistent error term.

Proposition 5.9 Suppose that $T \geq 1$ is an integer. Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d symmetric random variables and suppose that $x = [x_1, \ldots, x_n]$ satisfies $\|x\|_2 \leq 1$ and $\|x\|_\infty \leq v$. Then, we have that

$$\left\|\sum_i Y_i x_i\right\|_{2T} \lesssim v\left(\sup_{1 \leq t \leq T} \frac{T}{t}\left(\frac{1}{Tv^2}\right)^{\frac{1}{2t}} \|Y_i\|_{2t}\right).$$

Theoretically, moments of the form in Proposition 5.9 can be bounded using Lemma 5.2 (or from Theorem 2 in [31], a tight bound on moments of weighted sums of symmetric random variables). However, reducing the tight bound to the form that we want would require some simplifications. Instead, we give a direct proof of our weaker bound that is sufficiently tight for our setting.

Proof of Proposition 5.9. Let $k = 2v\left(\sup_{1 \leq t \leq T} \frac{T}{t}\left(\frac{1}{Tv^2}\right)^{1/(2t)} \|Y_i\|_{2t}\right)$. Observe that

$$\begin{aligned}
\mathbb{E}[\left(\frac{\sum_i Y_i x_i}{k^2}\right)^{2T}] &= \sum_{d_1+d_2+\ldots+d_n=T, d_i \leq T} \frac{2T!}{2d_1! \ldots 2d_n!} \prod_{i=1}^n \mathbb{E}\left[\left(\frac{Y_i x_i}{k}\right)^{2d_i}\right] \\
&\leq C^T \sum_{d_1+d_2+\ldots+d_n=T, d_i \leq T} \frac{(2T)^{2T}}{(2d_1)^{2d_1} \ldots (2d_n)^{2d_n}} \prod_{i=1}^n \mathbb{E}\left[\left(\frac{Y_i x_i}{k}\right)^{2d_i}\right] \\
&\leq C^T \prod_{i=1}^n \sum_{0 \leq d_i \leq T} \frac{(2T)^{2d_i}}{(2d_i)^{2d_i}} \mathbb{E}\left[\left(\frac{Y_i x_i}{k}\right)^{2d_i}\right] \\
&= C^T \prod_{i=1}^n \left(1 + \sum_{1 \leq d_i \leq T} \left(\frac{Tx_i \|Y_i\|_{2d_i} v}{v d_i k}\right)^{2d_i}\right)
\end{aligned}$$

Now, we use the fact that $|x_i| \leq v$ and the condition on k to obtain that this is bounded by

$$C^T \prod_{i=1}^n \left(1 + \frac{x_i^2}{v^2} \sum_{1 \leq d_i \leq T} \left(\frac{Tv \|Y_i\|_{2d_i}}{d_i k}\right)^{2d_i}\right) \leq C^T \prod_{i=1}^n \left(1 + Tx_i^2\right) \leq C^T \prod_{i=1}^n e^{Tx_i^2} \leq C^T e^T.$$

□

Proposition 5.9 provides a bound on linear forms of symmetric random variables that can be used to obtain a bound on $\left(\sum_{1\leq i\leq n}\eta_{r,i}\sigma_{r,i}x_i\right)^2$, and thus on $\left(\sum_{1\leq i\leq n}\eta_{r,i}\sigma_{r,i}x_i\right)^2 - \sum_{1\leq i\leq n}\eta_{r,i}x_i^2$. However, observe that $\left(\sum_{1\leq i\leq n}\eta_{r,i}\sigma_{r,i}x_i\right)^2 - \sum_{1\leq i\leq n}\eta_{r,i}x_i^2$ is actually a square of a linear form with a zero diagonal. It turns out that bounding the moments of this random variable using Proposition 5.9 is weak in some parts of the analysis in Chapter 3.

In this context, we give a bound on moments of squares of linear forms with a zero diagonal, i.e. $\sum_{i\neq j}Y_iY_jx_ix_j$. Since random variables with a zero diagonal are common in the JL literature [28, 4, 36], we believe this moment bound could be of broader use.

Lemma 5.10 Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d symmetric random variables and suppose that $x = [x_1, \ldots, x_n]$ satisfies $\|x\|_2 = 1$ and $\|x\|_\infty \leq v$. Let T be an even natural number. Then, we have that

$$\left\|\sum_{i\neq j}Y_iY_jx_ix_j\right\|_T \lesssim v^2\left(\sup_{1\leq t\leq T/2}\frac{T^2}{t^2}\left(\frac{1}{Tv^2}\right)^{1/t}\|Y_i\|_{2t}^2\right).$$

The structure of random variable in Lemma 5.10 theoretically falls under the scope of Lemma 5.5. However, the first term of the bound in Lemma 5.5, which is an operator-norm-like term for an asymmetric random matrix in this setting, becomes intractable to manage. We give an alternate (weaker) upper bound that is both tractable to analyze and sufficiently tight for our setting. Our proof of this bound is similar to our proof of Proposition 5.9 presented above.

Proof of Lemma 5.10. Let $k = 2v\left(\sup_{1\leq t\leq T/2}\frac{T}{t}\left(\frac{1}{Tv^2}\right)^{1/(2t)}\|Y_i\|_{2t}\right)$. Observe that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\sum_{i\neq j}Y_iY_jx_ix_j}{k^2}\right)^T\right] &\leq \sum_{d_1+d_2+\ldots+d_n=T, d_i\leq T/2}\frac{2T!}{2d_1!\ldots 2d_n!}\prod_{i=1}^n\mathbb{E}\left[\left(\frac{Y_ix_i}{k}\right)^{2d_i}\right] \\
&\leq C^T\sum_{d_1+d_2+\ldots+d_n=T, d_i\leq T/2}\frac{(2T)^{2T}}{(2d_1)^{2d_1}\ldots(2d_n)^{2d_n}}\prod_{i=1}^n\mathbb{E}\left[\left(\frac{Y_ix_i}{k}\right)^{2d_i}\right] \\
&\leq C^T\prod_{i=1}^n\sum_{0\leq d_i\leq T/2}\frac{(2T)^{2d_i}}{(2d_i)^{2d_i}}\mathbb{E}\left[\left(\frac{Y_ix_i}{k}\right)^{2d_i}\right] \\
&= C^T\prod_{i=1}^n\left(1+\sum_{1\leq d_i\leq T/2}\left(\frac{Tx_i\|Y_i\|_{2d_i}v}{vd_ik}\right)^{2d_i}\right)
\end{aligned}$$

Now, we use the fact that $|x_i| \leq v$ and the condition on k to obtain that this is bounded by

$$C^T\prod_{i=1}^n\left(1+\frac{x_i^2}{v^2}\sum_{1\leq d_i\leq T/2}\left(\frac{Tv\|Y_i\|_{2d_i}}{d_ik}\right)^{2d_i}\right) \leq C^T\prod_{i=1}^n(1+Tx_i^2) \leq C^T\prod_{i=1}^n e^{Tx_i^2} \leq C^Te^T.$$

□

Latała [31] gives the following nice bound on sums of i.i.d symmetric random variables that turns to be generally useful in our proofs.

Lemma 5.11 ([31]) Suppose that q is an even natural number. Suppose that $Y_1, \ldots, Y_n$ are i.i.d symmetric random variables. Then:

$$\left\| \sum_{i=1}^{n} Y_i \right\|_q \lesssim \sup_{2 \leq T \leq q} \frac{q}{T} \left(\frac{n}{q} \right)^{1/T} \|Y_i\|_T .$$

We give a proof of this result using Proposition 5.9.

Proof of Lemma 5.11. Consider $[x_1, \ldots, x_n] = [1/\sqrt{n}, \ldots, 1/\sqrt{n}]$. Notice that:

$$\left\| \sum_{i=1}^{n} Y_i \right\|_q = \sqrt{n} \left\| \sum_{i=1}^{n} Y_i x_i \right\|_q \lesssim \sup_{2 \leq T \leq q} \frac{q}{T} \left(\frac{n}{q} \right)^{1/T} \|Y_i\|_T$$

by Proposition 5.9. □

We give a general lower bound on moments of certain (potentially correlated) sums of identically distributed random variables, that is useful in our analysis of sparse JL.

Proposition 5.12 Let $Y_1, \ldots, Y_n$ be identically distributed (but not necessarily independent) random variables, such that the joint distribution is a symmetric function of $Y_1, \ldots, Y_n$ and for any integers $d_1, \ldots d_n \geq 0$, it is true that $\mathbb{E}[\prod_{1 \leq i \leq n} Y_i^{d_i}] \geq 0$. For any natural number q and natural number T that divides q, it is true that

$$\left\| \sum_{i=1}^{n} Y_i \right\|_q \geq T \left(\frac{n}{q} \right)^{T/q} \|Y_1 Y_2 \ldots Y_T\|_{q/T}^{1/T}$$

Proof of Proposition 5.12. The proof follows from expanding $\mathbb{E}[(\sum_{i=1}^{n} Y_i)^q]$ and using the fact that $\mathbb{E}[\prod_{1 \leq i \leq n} Y_i^{d_i}] \geq 0$ so that we can restrict to a subset of the terms. By the symmetry of the joint distribution, we know that for $1 \leq r_1 \neq r_2 \neq r_T \leq n$, we know that $\mathbb{E}[Y_{r_1}^{q/T} \ldots Y_{r_T}^{q/T}] = \mathbb{E}[Y_1^{q/T} \ldots Y_T^{q/T}]$. The number of terms of the form $\mathbb{E}[Y_{r_1}^{q/T} \ldots Y_{r_T}^{q/T}]$ in $\mathbb{E}[(\sum_{i=1}^{n} Y_i)^q]$ is:

$$\begin{aligned}
\binom{n}{T} \binom{q}{q/T, q/T, \ldots, q/T} &\geq C^q \left(\frac{n}{T} \right)^T \frac{q!}{((q/T)!)^T} \\
&\geq C^q \left(\frac{n}{T} \right)^T T^q \\
&\geq C_2^q \left(\frac{n}{q} \right)^T \left(\frac{q}{T} \right)^T T^q \\
&\geq C'^q \left(\frac{n}{q} \right)^T T^q.
\end{aligned}$$

This implies that

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i \right)^q \right] \geq C'^q \left(\frac{n}{q} \right)^T T^q \mathbb{E}\left[Y_1^{q/T} \ldots Y_T^{q/T} \right]$$

and the statement follows from taking $1/q$th powers. □

5.3.2 A simpler quadratic form bound

For the sparse, sign-consistent JL, the error term $T(x_1, \ldots, x_n)$ can't be decomposed into linear forms, so the results from the previous subsection do not suffice in handling the error term. For this setting, we give an alternate quadratic form that enjoys a greater degree of simplicity than Lemma 5.4. Using this bound, combined with the bounds in the previous subsection, gives us the necessary technology to analyze sparse, sign-consistent JL.

Our bound is based on a degree-2 analog of Hitczenko's observation in Lemma 5.2. We analogously handle the "big" terms with an ℓ_1-norm bound and bound the remaining terms by approximating some of the Rademachers by gaussians. From this, we obtain a combination of ℓ_2 and ℓ_1 norm bounds, similar to the linear form setting. Our simple bound has the surprising feature that it yields tighter guarantees than the Hanson-Wright bound (Lemma 5.3) yields for sparse, sign-consistent JL. While our bound is weaker than Latała's tight bound on the moments of Rademacher quadratic forms (Lemma 5.4) in the general case, it provides a greater degree of simplicity: our bound avoids the operator-norm-like term in Lemma 5.4 that is especially difficult to analyze when A is a random matrix, as is the case in this setting. Moreover, our bound still retains the necessary precision to recover the optimal dimension for sparse, sign-consistent JL.

We derive the following moment bound on quadratic forms of Rademachers[2] that yields tighter guarantees than the Hanson-Wright bound yields for $\|T(x_1, \ldots, x_n)\|_p$:

Lemma 5.13 If $A = (a_{i,j})$ is a symmetric square $n \times n$ matrix with zero diagonal, $\{\sigma_i\}_{i\in[n]}$ is a set of independent Rademachers, and even $q \geq 1$, then

$$\left\| \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}\sigma_i\sigma_j \right\|_q \lesssim \left(\sum_{i=1}^{\min(q,n)} \sum_{j=1}^{\min(q,n)} |a_{i,j}| \right) + \sqrt{q} \sqrt{\sum_{i=1}^{n} \left\| \sum_{j>q} a_{i,j}\sigma_j \right\|_q^2}.$$

Observe that our bound avoids the weakness of the Hanson-Wright bound in the limit as $q \to \infty$. As discussed in Section 5.1, we know that $\left\| \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}\sigma_i\sigma_j \right\|_q$ can be bounded by the entrywise ℓ_1-norm bound $\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}|$ for any $q \geq 1$. While the Hanson-Wright bound goes to ∞ as $q \to \infty$, the bound in Lemma 5.13 approaches the entrywise ℓ_1 bound in the limit: for $q > n$, the second term in Lemma 5.13 vanishes since the summand $\sum_{j>q}$ is empty. As a result, the bound becomes the first-term, which becomes $\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}|$ as desired. Our simplified bound, though weaker in the general case, consists of much easier-to-analyze terms and has a cleaner proof, while still being sufficiently tight for this setting. For $1 \leq q < n$, our bound becomes an interpolation of ℓ_1 and ℓ_2 norm bounds that bears resemblance to Hitczenko's Rademacher linear form bound in Lemma 5.2.

Although our bound is weaker than Lemma 5.4 in the general case, it is much simpler to analyze, especially when A is a random matrix. While the bound in Lemma 5.4 is focused on obtaining tight estimates for quadratic forms where A is a scalar matrix, our bound is

[2] As mentioned before, Latała [32] provides a tight bound on the moments of $\sigma^T A\sigma$ (and on the moments of more general quadratic forms). However, his bound consists of terms that are difficult to analyze when the quadratic form coefficients are random variables. Moreover, his proof is quite complicated, though the bound can be used in a black box to generate a much messier solution (by unravelling some of his proof to avoid the operator-norm-like term).

much more tractable when A is a random matrix. Recall that the main complication in the bound in Lemma 5.4 arises from the operator-norm-like term. Observe that our bound in Lemma 5.13 manages to avoid this term altogether. Moreover, our ℓ_1 norm term is straightforward to calculate, and our ℓ_2 norm term can be handled cleanly even when the $a_{i,j}$ are themselves random variables, through a bound that we describe later in this section

The following lemma allows us to decouple the two sets of Rademachers in our quadratic form so that we can reduce analyzing the moments of the quadratic form to analyzing the moments of a linear form.

Lemma 5.14 (Decoupling, Theorem 6.1.1 of [44]) If $A = (a_{i,j})$ is a symmetric, zero-diagonal $n \times n$ matrix and $\{\sigma_i\}_{i\in[n]} \cup \{\sigma_i'\}_{i\in[n]}$ are independent Rademachers, then

$$\left\| \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\sigma_i\sigma_j \right\|_q \lesssim \left\| \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\sigma_i'\sigma_j \right\|_q .$$

Now, we are ready to prove Lemma 5.13.

Proof of Lemma 5.13. By Lemma 5.14 and the triangle inequality, we know

$$\left\| \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\sigma_i\sigma_j \right\|_q \lesssim \underbrace{\left\| \sum_{i=1}^{\min(q,n)}\sum_{j=1}^{\min(q,n)} a_{i,j}\sigma_i'\sigma_j \right\|_q}_{\alpha} + \underbrace{\left\| \sum_{i=1}^{n}\sum_{j>q} a_{i,j}\sigma_i'\sigma_j \right\|_q}_{\beta} + \underbrace{\left\| \sum_{i>q}\sum_{j=1}^{q} a_{i,j}\sigma_i'\sigma_j \right\|_q}_{\gamma} .$$

We first bound α. Since a Rademacher σ satisfies $|\sigma| = 1$, it follows that α can be upper bounded by the entrywise ℓ_1-norm bound $\sum_{i=1}^{\min(q,n)}\sum_{j=1}^{\min(q,n)} |a_{i,j}|$ as desired. Using Lemma 5.1, we know that β can be upper bounded by:

$$\sqrt{q}\left\| \sqrt{\sum_{i=1}^{n}\left(\sum_{j>q} a_{i,j}\sigma_j\right)^2} \right\|_q = \sqrt{q}\sqrt{\left\| \sum_{i=1}^{n}\left(\sum_{j>q} a_{i,j}\sigma_j\right)^2 \right\|_{q/2}} \le \sqrt{q}\sqrt{\sum_{i=1}^{n}\left\| \sum_{j>q} a_{i,j}\sigma_j \right\|_q^2}.$$

We now bound γ. An analogous argument shows $\gamma \le \sqrt{q}\sqrt{\sum_{j=1}^{q}\left\|\sum_{i>q} a_{i,j}\sigma_i\right\|_q^2}$. Thus:

$$\gamma \le \sqrt{q}\sqrt{\sum_{j=1}^{q}\left\| \sum_{i>q} a_{i,j}\sigma_i \right\|_q^2} \le \sqrt{q}\sqrt{\sum_{j=1}^{n}\left\| \sum_{i>q} a_{i,j}\sigma_i \right\|_q^2} = \sqrt{q}\sqrt{\sum_{i=1}^{n}\left\| \sum_{j>q} a_{i,j}\sigma_j \right\|_q^2}.$$

□

Chapter 6

Proof of Main Results

In Section 6.1, we provide a proof sketch of our main result in Chapter 3. In Section 6.2, we provide a proof of our main result in Chapter 4.

6.1 Proof sketches for Chapter 3

We sketch a proof of Theorem 3.2. (The full proofs of the supporting lemmas can be found in the Appendix.) For every $[x_1, \ldots, x_n] \in \mathbb{R}^n$ such that $\|x\|_2 = 1$, we need to analyze tail bounds of

$$R(x_1, \ldots, x_n) = \|Ax\|_2^2 - 1 = \frac{1}{s} \sum_{i \neq j} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j.$$

An upper bound on the tail probability of $R(x_1, \ldots, x_n)$ is needed to prove the lower bound on $v(m, \epsilon, \delta, s)$ in Theorem 3.2, and a lower bound is needed to prove the upper bound on $v(m, \epsilon, \delta, s)$ in Theorem 3.2. It turns out that it suffices to tightly analyze the random variable moments $\mathbb{E}[(R(x_1, \ldots, x_n))^q]$. For the upper bound, we use Markov's inequality (Lemma 2.7) like in [16, 28, 4, 36], and for the lower bound, we use the Paley-Zygmund inequality (Lemma 6.1) like in [16]: Markov's inequality gives a tail upper bound from upper bounds on moments, and the Paley-Zygmund inequality gives a tail lower bound from upper and lower bounds on moments.

More specifically, the Paley-Zygmund inequality says the following:

Lemma 6.1 (Paley-Zygmund) Suppose that Z is a nonnegative random variable with finite variance. If $0 < \theta < 1$, then:

$$\mathbb{P}[Z > \theta \mathbb{E}[Z]] \geq (1 - \theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

The Paley-Zygmund inequality essentially says that when the 1-norm $\|Z\|_1$ is close to the 2-norm $\|Z\|_2$, the random variable can't have significant mass below the mean.

Analogously to how we applied Markov's inequality to $\Theta(\log(1/\delta))$th moments in Section 2.2.2 to upper bound the failure probability by δ, we wish to apply the Paley-Zygmund inequality to $\Theta(\log(1/\delta))$th moments to lower bound the failure probability by δ. In particular, we use the following corollary:

Lemma 6.2 Suppose that $K > 0$ and Z is a nonnegative random variable, such that $\|Z\|_q \geq 2K$ and $\|Z\|_{2q}$ is finite. Then,

$$\mathbb{P}[Z > K] \geq 0.25 \left(\frac{\|Z\|_q}{\|Z\|_{2q}} \right)^{2q}.$$

Proof of Lemma 6.2. We apply Lemma 6.1 with $\theta = 1/2$ to Z^p to obtain that:

$$\mathbb{P}[Z^q > 2^{-1}\mathbb{E}[Z^q]] \geq 0.25 \frac{\mathbb{E}[Z^q]^2}{\mathbb{E}[Z^{2q}]} = 0.25 \left(\frac{\|Z\|_p}{\|Z\|_{2q}} \right)^{2q}.$$

If $\|Z\|_q \geq 2K$, then we know that

$$\mathbb{P}[Z > K] = \mathbb{P}[Z^q > K^q] \geq \mathbb{P}[Z^q > 2^{-q}\mathbb{E}[Z^q]] \geq \mathbb{P}[Z^q > 2^{-1}\mathbb{E}[Z^q]]$$

and then we can apply the above result. □

For both the upper and lower tail bounds, the key ingredient of our analysis is thus a *tight bound* for $\|R(x_1, \ldots, x_n)\|_q$ on $S_v = \left\{ x \in \mathbb{R}^n \mid \frac{\|x\|_\infty}{\|x\|_2} \leq v \right\}$ at *each* threshold v value. For the upper bound on moments, we need to analyze $\|R(x_1, \ldots, x_n)\|_q$ for general vectors $[x_1, \ldots, x_n]$. We analyze $\|R(x_1, \ldots, x_n)\|_q$ using the bounds in Chapter 5. For the lower bound on moments, we only need to show $\|R(x_1, \ldots, x_n)\|_q$ is large for single vector in each S_v, and we show we can select the vector in the ℓ_2-unit ball with $1/v^2$ nonzero entries, all equal to v. For ease of notation, we denote this vector by $[v, \ldots, v, 0, \ldots, 0]$ for the remainder of the paper.

Our strategy for bounding $\|R(x_1, \ldots, x_n)\|_q$ is to break down into rows. We define

$$Z_r(x_1, \ldots, x_n) := \sum_{1 \leq i \neq j \leq n} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j$$

so that $R(x_1, \ldots, x_n) = \frac{1}{s}\sum_{r=1}^m Z_r(x_1, \ldots, x_n)$. We analyze the moments of $Z_r(x_1, \ldots, x_n)$, and then combine these bounds to obtain moment bounds for $R(x_1, \ldots, x_n)$. In our bounds, we use the notation $f \lesssim g$ (resp. $f \gtrsim g$) to denote $f \leq Cg$ (resp. $f \geq Cg$) for some constant $C > 0$.

6.1.1 Bounding $\|Z_r(x_1, \ldots, x_n)\|_q$

We show the following bounds on $\|Z_r(x_1, \ldots, x_n)\|_q$. For the lower bound, as we discussed before, it suffices to bound $\|Z_r(v, \ldots, v, 0, \ldots, 0)\|_q$. For the upper bound, we need to bound $\|Z_r(x_1, \ldots, x_n)\|_q$ for general vectors as a function of the ℓ_∞-to-ℓ_2 norm ratio.

Lemma 6.3 Let $\mathcal{A}_{s,m,n}$ be a sparse JL transform such that $s \leq m/e$. Suppose that $x = [x_1, \ldots, x_n]$ satisfies $\|x\|_\infty \leq v$ and $\|x\|_2 = 1$. If T is even, then:

$$\|Z_r(x_1, \ldots, x_n)\|_T \lesssim \begin{cases} \frac{Ts}{m}, & \text{for } T = 2, 3 \leq T \leq \frac{se}{mv^2} \\ \min\left(\frac{T^2v^2}{\ln(mTv^2/s)^2}, \frac{T}{\ln(m/s)} \right) & \text{for } T \geq 3, T \geq \frac{se}{mv^2}, \ln(Tmv^2/s) \leq T \\ v^2\left(\frac{s}{mTv^2}\right)^{2/T}, & \text{for } T \geq 3, T \geq \frac{se}{mv^2}, \ln(Tmv^2/s) > T. \end{cases}$$

Lemma 6.4 Let $\mathcal{A}_{s,m,n}$ be a sparse JL transform. Suppose $\frac{1}{v^2}$ and T are even integers. Then, $\|Z_r(v,\dots,v,0,\dots,0)\|_2 \gtrsim \frac{s}{m}$. Moreover, if $s \le m/e$ and $T \ge \frac{se}{mv^2}$, then

$$\left\| Z_r(v,\dots,v,0,\dots,0) I_{\sum_{i=1}^{1/v^2} \eta_{1,i}=2} \right\|_T \gtrsim v^2 \left(\frac{s}{mTv^2}\right)^{2/T}$$

and

$$\|Z_r(v,\dots,v,0,\dots,0)\|_T \gtrsim \begin{cases} \frac{T^2v^2}{\ln^2(mv^2T/s)} & \text{for } 1 \le \ln(mv^2T/s) \le T, v \le \frac{\sqrt{\ln(m/s)}}{\sqrt{T}} \\ v^2\left(\frac{s}{mTv^2}\right)^{2/T} & \text{for } \ln(mv^2T/s) > T. \end{cases}$$

We now sketch our methods to prove Lemma 6.3 and Lemma 6.4. For the lower bound (Lemma 6.4), we can view $Z_r(v,\dots,v,0,\dots,0)$ as a quadratic form $\sum_{t_1,t_2} a_{t_1,t_2}\sigma_{t_1}\sigma_{t_2}$ where $(a_{t_1,t_2})_{t_1,t_2\in[mn]}$ is an appropriately defined block-diagonal mn dimensional matrix. We can write $\mathbb{E}_{\sigma,\eta}[(Z_r(v,\dots,v,0,\dots,0))^q]$ as $\mathbb{E}_\eta[\mathbb{E}_\sigma[(Z_r(v,\dots,v,0,\dots,0))^q]]$: for *fixed* $\eta_{r,i}$ values, the coefficients are scalars. We make use of Lemma 5.4 to analyze $\mathbb{E}_\sigma[(Z_r(v,\dots,v,0,\dots,0))^q]$ as a function of the $\eta_{r,i}$. Then, we handle the randomness of the $\eta_{r,i}$ by taking an expectation of the resulting bound on $\mathbb{E}_\sigma[(Z_r(v,\dots,v,0,\dots,0))^q]$ over the $\eta_{r,i}$ values to obtain a bound on $\|Z_r(v,\dots,v,0,\dots,0)\|_q$.

For the upper bound (Lemma 6.3), since Lemma 5.4 is tight for scalar quadratic forms, the natural approach would be to use it to upper bound $\mathbb{E}_\sigma[(Z_r(x_1,\dots,x_n))^q]$ for general vectors. However, when the vector is not of the form $[v,\dots,v,0,\dots,0]$, the asymmetry makes the resulting bound intractable to simplify. Specifically, the first term, which can be viewed as a generalization of an operator norm to an ℓ_2 ball cut out by ℓ_∞ hyperplanes, becomes problematic when taking an expectation over the $\eta_{r,i}$ to obtain a bound on $\mathbb{E}_{\sigma,\eta}[(Z_r(x_1,\dots,x_n))^q]$. Thus, we utilize our simpler estimates (Proposition 5.9 and Lemma 5.10). These estimates take advantage of the structure of $Z_r(x_1,\dots,x_n)$ and enable us to show Lemma 6.3.

6.1.2 Obtaining bounds on $\|R(x_1,\dots,x_n)\|_q$

Now, we use Lemma 6.3 and Lemma 6.4 to show the following bounds on $\|R(x_1,\dots,x_n)\|_q$:

Lemma 6.5 Suppose $\mathcal{A}_{s,m,n}$ is a sparse JL transform such that $s \le m/e$, and let $x = [x_1,\dots,x_n]$ be such that $\|x\|_2 = 1$. Then, $\|R(x_1,\dots,x_n)\|_2 \le \frac{\sqrt{2}}{\sqrt{m}}$. Now, suppose that $2 < q \le m$ is an even integer and $\|x\|_\infty \le v$. If $\frac{se}{mv^2} \ge q$, then $\|R(x_1,\dots,x_n)\|_q \lesssim \frac{\sqrt{q}}{\sqrt{m}}$. If $\frac{se}{mv^2} < q$ and if there exists a constant $C_2 \ge 1$ such that $C_2q^3mv^4 \ge s^2$, then $\|R(x_1,\dots,x_n)\|_q \lesssim g$ where g is:

$$\begin{cases} \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}\right) & \text{if } \ln(\frac{qmv^4}{s^2}) \le 2, \ln(\frac{qmv^2}{s}) \le q \\ \frac{\sqrt{q}}{\sqrt{m}} & \text{if } \ln(\frac{qmv^4}{s^2}) \le 2, \ln(\frac{qmv^2}{s}) > q \\ \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \min\left(\frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}, \frac{q}{s\ln(m/s)}\right)\right) & \text{if } \ln(\frac{qmv^4}{s^2}) > 2, \ln(\frac{qmv^2}{s}) \le q \\ \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}\right) & \text{if } \ln(\frac{qmv^4}{s^2}) > 2, \ln(\frac{qmv^2}{s}) > q. \end{cases}$$

Lemma 6.6 Suppose $\mathcal{A}_{s,m,n}$ is a uniform sparse JL transform. Let q be a power of 2, and suppose that $0 < v \leq 0.5$ and $\frac{1}{v^2}$ is an even integer. If $qv^2 \leq s$, then $\|R(v,\ldots,v,0,\ldots,0)\|_q \gtrsim \frac{\sqrt{q}}{\sqrt{m}}$. If $m \geq q$, $2 \leq \ln(qmv^4/s^2) \leq q$, $2qv^2 \leq 0.5s\ln(qmv^4/s^2)$, and $s \leq m/e$, then $\|R(v,\ldots,v,0,\ldots,0)\|_q \gtrsim \frac{qv^2}{s\ln(qmv^4/s^2)}$. If $s \leq m/e$, $v \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{q}}$, and $1 \leq \ln(qmv^2/s) \leq q$, then $\|R(v,\ldots,v,0\ldots,0)\|_q \gtrsim \frac{q^2v^2}{s\ln^2(qmv^2/s)}$.

We now sketch how to prove bounds on $\|R(x_1,\ldots,x_n)\|_q$ using bounds on $\|Z_r(x_1,\ldots,x_n)\|_T$. To show Lemma 6.5, we show that making the row terms $Z_r(x_1,\ldots,x_n)$ independent does not decrease $\|R(x_1,\ldots,x_n)\|_q$, and then we apply Lemma 5.11 for moments of sums of i.i.d symmetric random variables. For Lemma 6.6, handling the correlations between the row terms $Z_r(x_1,\ldots,x_n)$ requires more care. We show that the negative correlations induced by having exactly s nonzero entries per column do not lead to significant loss, and then stitch together $\|R(v,\ldots,v,0,\ldots,0)\|_q$ using Proposition 5.12.

6.1.3 Proof of main result from moment bounds

We now sketch how to prove Theorem 3.2, using Lemma 6.5 and Lemma 6.6. First, we simplify these bounds at the target parameters to obtain the following:

Lemma 6.7 Let $\mathcal{A}_{s,m,n}$ be a sparse JL transform, and suppose ϵ and δ are small enough, $s \leq m/e$, $\Theta(\epsilon^{-2}\ln(1/\delta)) \leq m < 2\epsilon^{-2}/\delta$, $v \leq f'(m,\epsilon,\ln(1/\delta),s)$, and $p = \Theta(\ln(1/\delta))$ is even. If $x = [x_1,\ldots,x_n]$ satisfies $\|x\|_\infty \leq v$ and $\|x\|_2 = 1$, then $\|R(x_1,\ldots,x_n)\|_p \leq \frac{\epsilon}{2}$.

Lemma 6.8 There is a universal constant D satisfying the following property. Let $\mathcal{A}_{s,m,n}$ be a uniform sparse JL transform, and suppose ϵ,δ are small enough, $s \leq m/e$, $f'(m,\epsilon,\ln(1/\delta),s) \leq 0.5$, and q is an even integer such that $q = \min(m/2, \Theta(\ln(1/\delta))$. For each $\psi > 0$, there exists $v \leq f'(m,\epsilon,\ln(1/\delta),s) + \psi$, such that $\|R(v,\ldots,v,0,\ldots,0)\|_q \geq 2\epsilon$ and $\frac{\|R(v,\ldots,v,0,\ldots,0)\|_q}{\|R(v,\ldots,v,0,\ldots,0)\|_{2q}} \geq D$.

Now, we use Lemma 6.7 and Lemma 6.8 to prove Theorem 3.2.

Proof of Theorem 3.2. Since the maps in $\mathcal{A}_{s,m,n}$ are linear, it suffices to consider unit vectors x. First, we prove the lower bound on $v(m,\epsilon,\delta,s)$. To handle $m \geq 2\epsilon^{-2}/\delta$, we take $q = 2$ in Lemma 6.7 and apply Chebyshev's inequality. Otherwise, we take $p = \ln(1/\delta)$ (approximately) and apply Lemma 6.7 and Markov's inequality (Lemma 2.7). We see that $\mathbb{P}[|\,\|Ax\|_2^2 - 1| \geq \epsilon]$ can be expressed as:

$$\mathbb{P}[|R(x_1,\ldots,x_n)| \geq \epsilon] = \mathbb{P}[R(x_1,\ldots,x_n)^p \geq \epsilon^p] \leq \epsilon^{-p}\mathbb{E}[R(x_1,\ldots,x_n)]^p \leq \delta.$$

Thus, $\mathbb{P}[|\,\|Ax\|_2^2 - 1| \geq \epsilon]$ is satisfied for unit vectors $x \in S_v$ when $v \leq f'(m,\epsilon,\ln(1/\delta),s)$ as desired.

Now, we prove the upper bound on $v(m,\epsilon,\delta,s)$. We need to lower bound the tail probability of $R(v,\ldots,v,0,\ldots,0)$, and to do this, we use Lemma 6.2 (the Paley-Zygmund inequality applied to qth moments). Let D be defined as in Lemma 6.8, and take $q = \min(m/2, \max(2, \frac{\ln(1/\delta)-2}{-2\ln(D)}))$. By Lemma 6.2 and Lemma 6.8, there exists $v \leq f'(m,\epsilon,\ln(1/\delta),s)+$

ψ such that:

$$\mathbb{P}[|R(v,\ldots,v,0,\ldots,0)|>\epsilon] \geq 0.25\left(\frac{\|R(v,v,\ldots,v,0,\ldots,0)\|_q}{\|R(v,v,\ldots,v,0,\ldots,0)\|_{2q}}\right)^{2q} \geq 0.25D^{2q} > \delta.$$

Thus, it follows that $\sup_{x\in S_{f'(m,\epsilon,\ln(1/\delta),s)+\psi},\|x\|_2=1}\mathbb{P}[|\,\|Ax\|_2^2-1|>\epsilon]>\delta$ as desired. □

6.2 Proofs for Chapter 4

We prove Theorem 4.4. Our main lemma is the following bound on the moments of $T(x_1,\ldots,x_n)$.

Lemma 6.9 Let $B=m/s^2$. If $p\geq 2$, then

$$\|T(x_1,\ldots,x_n)\|_p \lesssim \begin{cases}\frac{p}{s\log B}, & \text{if } B\geq e\\ \frac{p}{sB} & \text{if } B<e.\end{cases}$$

In order to analyze $\|T(x_1,\ldots,x_n)\|_p$, we view $T(x_1,\ldots,x_n)$ as a quadratic form $\frac{1}{s}\sigma^T A\sigma$, where the vector σ is an n-dimensional vector of independent Rademachers, and $A=(a_{i,j})$ is a symmetric, zero-diagonal $n\times n$ matrix where the (i,j)th entry (for $i\neq j$) is $x_ix_j\sum_{r=1}^m \eta_{r,i}\eta_{r,j}$. Since Z is symmetric in $x_1,\ldots,x_n$, we can assume WLOG that $|x_1|\geq|x_2|\geq\ldots\geq|x_n|$. For convenience, we define (like in [12]),

$$Q_{i,j} := \sum_{r=1}^m \eta_{r,i}\eta_{r,j} \tag{6.1}$$

to be the number of collisions between the nonzero entries of the ith column and the nonzero entries of the jth column. Now, the (i,j)th entry of A (for $i\neq j$) can be written as $Q_{i,j}x_ix_j$.

We now use Lemma 5.13 and the triangle inequality to obtain the following bound on $||T(x_1,\ldots,x_n)||_p$:

$$\begin{aligned}||T(x_1,\ldots,x_n)||_p &= \frac{1}{s}\left(\mathbb{E}_\eta\mathbb{E}_\sigma\left[\sum_{i=1}^n\sum_{j\leq n,j\neq i}Q_{i,j}x_ix_j\sigma_i\sigma_j\right]^p\right)^{1/p}\\ &\lesssim \frac{1}{s}\left(\mathbb{E}_\eta\left[\sum_{i=1}^p\sum_{\substack{j\leq p\\ j\neq i}}|Q_{i,j}x_ix_j| + \sqrt{p}\sqrt{\sum_{i=1}^n\left(\mathbb{E}_\sigma\left[\sum_{\substack{j>p\\ j\neq i}}Q_{i,j}x_ix_j\sigma_j\right]^p\right)^{2/p}}\right]^p\right)^{1/p}\\ &\leq \frac{1}{s}\left(\underbrace{\left\|\sum_{i=1}^p\sum_{\substack{j\leq p\\ j\neq i}}|Q_{i,j}x_ix_j|\right\|_p}_{(*)} + \underbrace{\sqrt{p}\sqrt{\sum_{i=1}^n\left\|\sum_{\substack{j>p\\ j\neq i}}Q_{i,j}x_ix_j\sigma_j\right\|_p^2}}_{(**)}\right).\end{aligned}$$

We first discuss some intuition for why using this bound to analyze $\|T(x_1, \ldots, x_n)\|_p$ avoids the loss incurred by the Hanson-Wright bound here. In the Hanson-Wright bound, all of the Rademachers are essentially approximated by gaussians. In our bound, we make use of Rademachers in the appropriate places to avoid loss. For $1 \leq i \leq p$ and $1 \leq j \leq p$ (the upper left $p \times p$ minor where the $|x_i|$ and $|x_j|$ values are the largest), our approach utilizes an ℓ_1-norm bound rather than $\sqrt{p}$ times an ℓ_2 bound, which turns out to allow us to save a factor of $\sqrt{p}$ in the resulting bound on $\|T(x_1, \ldots, x_n)\|_p$. Now, since the original matrix is symmetric, it only remains to consider $1 \leq i \leq n$ and $p+1 \leq j \leq n$. In this range, we approximate the σ_i Rademachers by gaussians and use an ℓ_2-norm bound. It turns out that approximating the σ_j Rademachers by gaussians as well would yield too loose of a bound for our application, so we preserve the σ_j Rademachers. For the remaining Rademacher linear forms, the interaction between the x_j values (all of which are upper bounded in magnitude by $\frac{1}{\sqrt{p}}$) and the σ_j Rademachers yields the desired bound.

Now, it remains to bound $(*)$ and $(**)$. Bounding these quantities requiring understanding the moments of $Q_{i,j}$. We use the binomial-like properties of the $Q_{i,j}$s coupled with standard moment bounds involving the binomial distribution to analyze the moments.

First, we analyze the independence structure of the $Q_{i,j}$ random variables.

Proposition 6.10 Let X be a random variable distributed as $\text{Bin}(s, s/m)$. For any $1 \leq i \leq n$, given any choice of s nonzero rows $r_1 \neq r_2 \neq \ldots \neq r_s$ in the ith column, the set of $n-1$ random variables[1] $\{(Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1)\}_{1 \leq j \leq n, j \neq i}$ are independent. Moreover, for any even $q \geq 1$ and any $j \neq i$:

$$\|Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1\|_q \leq \|X\|_q .$$

The independence properties use that the nonzero entries in different columns are independent. Moreover, the binomial bound on the moments of $Q_{i,j}$ follows from the decomposition of $Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1$ into a sum of Bernoulli random variables.

Proof of Proposition 6.10. Let A be a matrix drawn from $\mathcal{A}$, and pick any $1 \leq i \leq n$. We condition on the event that the s nonzero entries in column i of A occur at rows $r_1, \ldots, r_s$. For $1 \leq j \leq n$, $j \neq i$ and $1 \leq k \leq s$, let $Y_{k,j} = \eta_{r_k,j}$, so that $(Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i}) = \sum_{k=1}^{s} Y_{k,j}$. Notice that the sets $\{Y_{k,j}\}_{k \in [s]}$ for $1 \leq j \leq n, j \neq i$ are independent from each other, which means random variables in the set $\{Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1\}_{1 \leq j \leq n, j \neq i}$ are independent. For $1 \leq j \leq n, j \neq i$, and $1 \leq k \leq s$, let $Z_{k,j}$ be distributed as i.i.d Bernoulli random variables with expectation s/m. Notice that for a fixed j, each $Y_{k,j}$ is distributed as $Z_{k,j}$ and the random variables $\{Y_{k,j}\}_{1 \leq k \leq s}$ are negatively correlated (and nonnegative), which means

$$\|Q_{i,j} \mid \eta_{r_1,i} = \eta_{r_2,i} = \ldots = \eta_{r_s,i} = 1\|_q = \left\| \sum_{k=1}^{s} Y_{k,j} \right\|_q \leq \left\| \sum_{k=1}^{s} Z_{k,j} \right\|_q = \|X\|_q .$$

□

[1] See Appendix A of [22] for a formal discussion of viewing these quantities as random variables over a different probability space.

Proposition 6.10 relates the moments of $Q_{i,j}$ to the moments of binomial random variables. In order to bound the moments, we thus bound the moments of binomial random variables. This is an application of Lemma 5.11.

Proposition 6.11 Suppose that X is a random variable distributed as $\text{Bin}(N, \alpha)$ for any $\alpha \in (0,1)$ and any integer $N \geq 1$. If $q \geq 1$ and $C = \frac{q}{\alpha \max(N,q)}$, then

$$\|X\|_q \lesssim \begin{cases} \frac{q}{\log C} & \text{if } B \geq e \\ \frac{q}{C} & \text{if } B < e \end{cases}.$$

Proof. The main tool that we use in this proof is Lemma 5.11 (Latała's bound on moments of sums of i.i.d nonnegative random variables). Notice that it suffices to obtain an upper bound on $\|X\|_q$ for all $N \geq q$. (Since $\|X\|_q$ is an increasing function of N, an upper bound on $\|X\|_q$ at $N = q$ is also an upper bound on $\|X\|_q$ for all $N < q$). For the rest of the proof, we assume $N \geq q$.

Notice X has the same distribution as $\sum_{j=1}^N Z_j$ where $Z, Z_1, \ldots, Z_N$ are i.i.d Bernoulli random variables with expectation α. Since $\|Z\|_t = \alpha^{1/t}$, we know by Lemma 5.11,

$$\begin{aligned} \|X\|_q &\simeq \sup_{1 \leq t \leq q} \frac{q}{t} \left(\frac{N}{q}\right)^{1/t} \alpha^{1/t} \\ &= \sup_{1 \leq t \leq q} \frac{q}{t} \left(\frac{1}{B}\right)^{1/t} \end{aligned}$$

At $t = 1$, this quantity is equal to $\frac{q}{C}$, and at $t = q$, this quantity is equal to $\left(\frac{1}{C}\right)^{1/q} = e^{\log(1/C)/q}$. The only $t \in \mathbb{R}$ for which this quantity has derivative 0 is $t = \log C$. Notice that $1 \leq \log C \leq q$ if and only if $e \leq C \leq e^q$. Thus

$$\|X\|_q \simeq \begin{cases} \max(\frac{q}{C}, \frac{q}{\log C}, e^{\log(1/C)/q}) & \text{if } e \leq C \leq e^q \\ \max(\frac{q}{C}, e^{\log(1/C)/q}) & \text{if } C < e \text{ or if } C > e^q. \end{cases}.$$

For $C \geq e$, we want to show $\|X\|_q \lesssim q/\log C$. Since $\log C > 0$, we see $e^{\log(1/C)/q} = e^{-\log C/q} \leq q/\log C$ and $q/C \leq q/\log C$.

For $C < e$, we want to show $\|X\|_q \lesssim q/C$. Since $\frac{1}{C} > \frac{1}{e}$, we see $e^{\log(1/C)/q} = \left(\frac{1}{C}\right)^{1/q} \leq \frac{e}{C} \lesssim \frac{q}{C}$. □

Now, we are ready to bound the quantities $(*)$ and $(**)$. We prove the following sublemmas, which assume the notation used throughout the paper:

Lemma 6.12 If $m/s^2 = B$, then

$$\left\| \sum_{i=1}^p \sum_{j \leq p, j \neq i} |Q_{i,j} x_j x_i| \right\|_p \lesssim \begin{cases} \frac{p}{\log B} & \text{if } B \geq e \\ \frac{p}{B} & \text{if } B < e \end{cases}.$$

Lemma 6.13 If $m/s^2 = B$, then

$$\sqrt{p} \sqrt{\sum_{i=1}^n \left\| \sum_{j > p, j \neq i} Q_{i,j} x_i x_j \sigma_j \right\|_p^2} \lesssim \begin{cases} \frac{p}{\log B} & \text{if } B \geq e \\ \frac{p}{B} & \text{if } B < e \end{cases}.$$

We now use Proposition 6.10 as well as the moment bound on binomial random variables from Proposition 6.11 to prove Lemma 6.12 and thus bound $(*)$.

Proof of Lemma 6.12. We carefully use the triangle inequality to see[2]:

$$\left\| \sum_{i=1}^{p} \sum_{\substack{j \le p \\ j \neq i}} |Q_{i,j} x_j x_i| \right\|_p \le 2 \left\| \sum_{i=1}^{p} \sum_{\substack{j \le p \\ j > i}} Q_{i,j} |x_j| |x_i| \right\|_p \lesssim \left\| \sum_{i=1}^{p} x_i^2 \sum_{\substack{j \le p \\ j > i}} Q_{i,j} \right\|_p \lesssim \sum_{i=1}^{p} x_i^2 \left\| \sum_{\substack{j \le p \\ j > i}} Q_{i,j} \right\|_p .$$

Let $X \sim \text{Bin}(s, s/m)$ and $Y \sim \text{Bin}(sp, s/m)$. By Proposition 6.10, for any i and any $r_1 \neq r_2 \neq \ldots \neq r_s$, the random variables $\{Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1\}_{j \neq i}$ are independent and $\|Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1\|_p \le \|X\|_p$. It follows from taking pth powers of both sides that

$$\left\| \left(\sum_{\substack{j \le p \\ j > i}} Q_{i,j} \right) \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1 \right\|_p = \left\| \sum_{\substack{j \le p \\ j > i}} (Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1) \right\|_p \le \|Y\|_p .$$

Now, Proposition 6.11 gives us a bound on $\|Y\|_p$, and the result follows from the law of total expectation.[3] □

We now use Proposition 6.10 as well as the moment bound on weighted sums of binomial random variables from Proposition 5.9 to prove Lemma 6.13 and thus bound $(**)$.

Proof of Lemma 6.13. Observe that

$$\sqrt{p} \sqrt{\sum_{i=1}^{n} \left\| \sum_{\substack{j > p \\ j \neq i}} Q_{i,j} x_i x_j \sigma_j \right\|_p^2} = \sqrt{p} \sqrt{\sum_{i=1}^{n} x_i^2 \left\| \sum_{\substack{j > p \\ j \neq i}} Q_{i,j} x_j \sigma_j \right\|_p^2} \le \sqrt{p} \max_{1 \le i \le n} \left\| \sum_{\substack{j > p \\ j \neq i}} Q_{i,j} x_j \sigma_j \right\|_p .$$

Let $X \sim \text{Bin}(s, s/m)$. By Proposition 6.10, for any i and any $r_1 \neq r_2 \neq \ldots \neq r_s$, the random variables $\{Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1\}_{j \neq i}$ are independent and

$$\|Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1\|_p \le \|X\|_p .$$

Moreover, $|x_j| \le \frac{1}{\sqrt{p}}$ for $j > p$. Now, we consider $\left\| \sum_{j>p, j \neq i} Q_{i,j} x_j \sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1 \right\|_p$ which is equal to

$$\left\| \sum_{j>p, j \neq i} (Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1)(\sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1) x_j \right\|_p .$$

[2]Naively applying the triangle inequality yields a suboptimal bound, so we require this more careful treatment.

[3]See Appendix A of [22] for a formal discussion of why a uniform bound on the conditional p-norm implies a bound on the p-norm here.

Now, we use the fact that $(\sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1)$ is distributed as a Rademacher and that the set of $n-1$ random variables $\{\sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1\}_{j \neq i}$ are independent and also independent of $\{Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1\}_{j \neq i}$.

We apply Proposition 5.9.[4] Let $Y_i = (Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1)(\sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1)$. Using that $\|y\|_\infty \leq \frac{1}{\sqrt{q}}$ and setting $T = q/2$, we have that:

$$\frac{1}{\sqrt{q}} \left(\sup_{1 \leq t \leq q/2} \frac{q}{2t} \left(\frac{2q}{q} \right)^{\frac{1}{2t}} \|Y_i\|_{2t} \right) = \frac{1}{\sqrt{q}} \left(\sup_{1 \leq t \leq q/2} \frac{q}{2t} 2^{\frac{1}{2t}} \|Y_i\|_{2t} \right) \leq \frac{1}{\sqrt{q}} \left(\sup_{1 \leq z \leq q, \text{ z even}} \frac{q}{z} \|Y_i\|_z \right).$$

Let's bound $\|Y_i\|_z$. Using the above properties, we know that

$$\|Y_i\|_z = \|(Q_{i,j} \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1)(\sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1)\| \leq \|X\|_z .$$

We can apply Proposition 6.11 to bound $\|X\|_z$. Let's consider $\frac{z}{(s/m)\max(s,z)}$. If $s \leq z$, then this is equal to $\frac{z}{\alpha z} = \frac{1}{(s/m)} = \frac{m}{s}$. If $s \geq z$, then this is equal to $\frac{z}{(s/m)s} \geq \frac{1}{(s/m)s} = \frac{m}{s^2}$. Thus, we know that $\frac{z}{(s/m)\max(s,z)} \geq B$. Since the bound in Proposition 6.11 is a decreasing function of C, we can upper bound by the case where $C = B$.

When $B \geq e$, the expression is thus bounded by:

$$\frac{1}{\sqrt{q}} \left(\sup_{1 \leq t \leq q/2} \frac{q}{2t} \frac{2t}{\log B} \right) = \frac{\sqrt{q}}{\log B}.$$

When $B < e$, the expression is thus bounded by:

$$\frac{1}{\sqrt{q}} \left(\sup_{1 \leq t \leq q/2} \frac{q}{2t} \frac{2t}{B} \right) = \frac{\sqrt{q}}{B}.$$

Thus, we obtain a bound on the conditional p-norm $\left\| \sum_{j>p, j \neq i} Q_{i,j} x_j \sigma_j \mid \eta_{r_1,i} = \ldots = \eta_{r_s,i} = 1 \right\|_p$. Now, the result follows from the law of total expectation. □

We now show the bound on $||T(x_1, \ldots, x_n)||_p$ follows from the bounds on $(*)$ and $(**)$ in Lemmas 6.12, 6.13.

Proof of Lemma 6.9. Applying Lemmas 6.12, 6.13 after the following simplification proves the lemma:

$$\|T(x_1, \ldots, x_n)\|_p \lesssim \frac{1}{s} \left\| \sum_{i=1}^{p} \sum_{j \leq p, j \neq i} |Q_{i,j} x_i x_j| \right\|_p + \frac{\sqrt{p}}{s} \sqrt{\sum_{i=1}^{n} \left\| \sum_{j>p, j \neq i} Q_{i,j} x_i x_j \sigma_j \right\|_p^2}.$$

□

We show Lemma 6.9 implies Theorem 4.4, completing the proof.

[4] Approximating the σ_j by gaussians yields a suboptimal bound, so we require the bound given in Proposition 5.9.

Proof of Theorem 4.4. It suffices to show $\mathbb{P}_{\eta,\sigma}[|T(x_1,\ldots,x_n)| > \epsilon] < \delta$. By Markov's inequality, we know

$$\mathbb{P}_{\eta,\sigma}[|T(x_1,\ldots,x_n)| > \epsilon] < \epsilon^{-p}\mathbb{E}[|T(x_1,\ldots,x_n)|^p] = \left(\frac{\|T(x_1,\ldots,x_n)\|_p}{\epsilon}\right)^p.$$

Suppose that $B \geq e$. Then by Lemma 6.9, we know

$$\left(\frac{\|T(x_1,\ldots,x_n)\|_p}{\epsilon}\right)^p \leq \left(\frac{Cp}{(\log B)s\epsilon}\right)^p.$$

Thus, to upper bound this quantity by δ, we can set $s = \Theta(\epsilon^{-1}p/\log B) = \Theta(\epsilon^{-1}\log_B(1/\delta))$ and $m = \Theta(Bs^2)$. We impose the additional constraint that $B \leq \frac{1}{\delta}$ to guarantee that $s \geq 1$. This proves the desired result.[5] □

[5]If we set $B < e$, if we use Lemma 6.9, we know that in order to obtain an upper bound of δ, we would have to set $s = \Theta(\epsilon^{-1}p/B) = \Theta(\epsilon^{-1}\log(1/\delta)/B)$ and $m = \Theta(\epsilon^{-1}\log^2(1/\delta)/B)$. This yields no better s or m values than those achieved when $B = e$.

Bibliography

[1] The 20 newsgroups text dataset. `https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html`.

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

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

[4] Z. Allen-Zhu, R. Gelashvili, S. Micali, and N. Shavit. Sparse sign-consistent Johnson–Lindenstrauss matrices: Compression with neuroscience-based constraints. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 111, pages 16872–16876, 2014.

[5] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger. Learning to rank with (a lot of) word features. *Information Retrieval*, 13(3):291–314, Jun 2010.

[6] J. Bourgain, S. Dirksen, and J. Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pages 499–508, 2015.

[7] C. Caragea, A. Silvescu, and P. Mitra. Protein sequence classification using feature hashing. *Proteome Science*, 10(1), 2012.

[8] C. Chen, C. Vong, C. Wong, W. Wang, and P. Wong. Efficient extreme learning machine via very sparse random projection. *Soft Computing*, 22, 03 2018.

[9] W. Chen, J. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen. Compressing neural networks with the hashing trick. *Proceedings of the 32nd Annual International Conference on Machine Learning (ICML)*, pages 2285–2294, 2015.

[10] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Symposium on Theory of Computing Conference (STOC)*, pages 81–90, 2013.

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

[12] M. B. Cohen, T. S. Jayram, and J. Nelson. Simple analyses of the sparse Johnson-Lindenstrauss transform. In *Proceedings of the 1st Symposium on Simplicity in Algorithms (SOSA)*, pages 1–9, 2018.

[13] S. Dahlgaard, M. Knudsen, and M. Thorup. Practical hash functions for similarity estimation and dimensionality reduction. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 6618–6628, 2017.

[14] B. Dalessandro. Bring the noise: Embracing randomness is the key to scaling up machine learning algorithms. *Big Data*, 1(2):110–112, 2013.

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

[16] C. Freksen, L. Kamma, and K. G. Larsen. Fully understanding the hashing trick. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5394–5404, 2018.

[17] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 62:233–241, 1981.

[18] S. Ganguli and H. Sompolinsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual Review of Neuroscience*, 35:485–508, 2012.

[19] R.T. Gray and P.A. Robinson. Stability and structural constraints of random brain networks with excitatory and inhibitory neural populations. *Journal of Computational Neuroscience*, 27(1):81–101, 2009.

[20] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[21] P. Hitczenko. Domination inequality for martingale transforms of Rademacher sequence. *Israel Journal of Mathematics*, 84:161–178, 1993.

[22] M. Jagadeesan. Simple analysis of sparse, sign-consistent JL. In *Proceedings of the 23rd International Conference and 24th International Conference on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (RANDOM)*, pages 61:1–61:20, 2019.

[23] M. Jagadeesan. Understanding sparse JL for feature hashing. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 15177–15187, 2019.

[24] T.S. Jayram and D. P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and steaming problems with subconstant error. In *ACM Transactions on Algorithms (TALG) - Special Issue on SODA'11*, volume 9, pages 1–26, 2013.

[25] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[26] D. M. Kane, R. Meka, and J. Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Proceedings of the 14th International Workshop and 15th International Conference on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (RANDOM)*, pages 628–639, 2011.

[27] D. M. Kane and J. Nelson. A derandomized sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1006.3585, 2010.

[28] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 16872–16876. ACM Press, 2012.

[29] R. Kiani, H. Esteky, K. Mirpour, and K. Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97:4296–4309, 2007.

[30] K. G. Larsen and J. Nelson. Optimality of the johnson-lindenstrauss lemma. In *Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638, 2017.

[31] R. Latała. Estimation of moments of sums of independent real random variables. *Annals of Probability*, 25(3):1502–1513, 1997.

[32] R. Latała. Tail and moment estimates for some types of chaos. *Studia Mathematica*, 135(1):39–53, 1999.

[33] P. Li, T. Hastie, and K. Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 287–296, 2006.

[34] C. Ma, J. Jung, S. Kim, and S. Ko. Random projection-based partial feature extraction for robust face recognition. *Neurocomputing*, 149:1232 – 1244, 2015.

[35] S. Narayanan and J. Nelson. Optimal terminal dimensionality reduction in Euclidean space. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1064–1069, 2019.

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

[37] J. Nelson and H.L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 101–110, 2013.

[38] D. Newman. Bag of words data set. `https://archive.ics.uci.edu/ml/datasets/Bag+of+Words`, 2008.

[39] K. Rajan and L.F. Abbot. Eigenvalue spectra of random matrices for neural networks. *Physical Review Letters*, 97:188104, 2006.

[40] T. Rogers and J. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press, 2004.

[41] T. Roughgarden. Beyond worst-case analysis. In *Communications of the ACM*, volume 62, pages 88–96, 2019.

[42] H. Song. Robust visual tracking via online informative feature selection. *Electronics Letters*, 50(25):1931–1932, 2014.

[43] S. Suthaharan. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, volume 36 of *Integrated Series in Information Systems*. Springer US, Boston, MA, 2016.

[44] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

[45] M. Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge series on statistical and probabilistic mathematics*. Cambridge University Press, New York, 2019.

[46] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1113–1120, 2009.

[47] E.P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62:548–564, 1955.

Appendix A

Proof Details for Chapter 3

In Appendix A.1, we prove our corollary regarding dimension-sparsity tradeoffs and discuss some of the subtleties of Theorem 3.2. In Appendix A.2, we prove our moment bounds for $Z_r(x_1, \ldots, x_n)$ in Lemma 6.3 and Lemma 6.4. In Appendix A.3, we prove our moment bounds for $R(x_1, \ldots, x_n)$ in Lemma 6.5 and Lemma 6.6. In Appendix A.4, we prove auxiliary lemmas needed in the proof of Lemma 6.5. In Appendix A.5, we prove auxiliary lemmas needed in the proof of Lemma 6.6. In Appendix A.6, we prove our simplified moment bounds for $R(x_1, \ldots, x_n)$ in Lemma 6.7 and Lemma 6.8. In Appendix A.7, we provide additional experimental results on real-world and synthetic datasets as well as additional discussion.

A.1 Discussion of theoretical results

We discuss some of the subtleties of Theorem 3.2. When $m \geq \min(2\epsilon^{-2}e^p, \epsilon^{-2}pe^{\Theta(\max(1,p\epsilon^{-1}/s))})$, where $p = \ln(1/\delta)$, we show that $v(m, \epsilon, \delta, s) = 1$, which means that the norm-preserving condition holds on the full space. This generalizes Cohen's bound [11] to a slightly more general family of sparse JL transforms, as we discuss below. When $m \leq \Theta(\epsilon^{-2}\ln(1/\delta))$, we show that $v(m, \epsilon, \delta, s) = 0$. For the remaining regimes, $\sqrt{\epsilon s}\sqrt{\ln(\frac{m\epsilon^2}{p})}/\sqrt{p}$ and $\sqrt{\epsilon s}\min\left(\ln(\frac{m\epsilon}{p})/p, \sqrt{\ln(\frac{m\epsilon^2}{p})}/\sqrt{p}\right)$, our upper and lower bounds on $v(m, \epsilon, \delta, s)$ match up to constant factors.

In terms of the boundaries between regimes, we emphasize that in Theorem 3.2, the function $f'(m, \epsilon, \delta, s)$ may not be defined for certain intervals between the boundaries of regimes, since there may be different absolute constants in different boundaries. More specifically, these intervals are $C_1\epsilon^{-2}p \leq m \leq C_2\epsilon^{-2}p$, $\epsilon^{-2}e^{C_1 p} \leq m \leq 2\epsilon^{-2}e^p$, and $s \cdot e^{C_1 \max(1,p\epsilon^{-1}/s)} \leq m \leq s \cdot e^{C_2 \max(1,p\epsilon^{-1}/s)}$. These gaps arise because the boundaries between the regimes on our upper and lower bounds on $v(m, \epsilon, \delta, s)$ can have different absolute constants, so we don't have precise control on $v(m, \epsilon, \delta, s)$ in these gaps. Nonetheless, the gaps only span a constant factor range on the exponent in the dimension m.

We now state the dimension-sparsity tradeoffs that follow from our bounds:

Corollary A.1 Suppose that ϵ and δ are sufficiently small and $s \leq m/e$. If $\mathcal{A}_{s,m,n}$ is any sparse JL transform, then $v(m, \epsilon, \delta, s) = 1$ when $m \geq \min\left(2\epsilon^{-2}/\delta, \epsilon^{-2}\ln(1/\delta)e^{\Theta(\max(1,\ln(1/\delta)\epsilon^{-1}/s))}\right)$.

If $\mathcal{A}_{s,m,n}$ is a uniform sparse JL transform, then $v(m,\epsilon,\delta,s) \le 1/2$ when $m \le \min\left(\epsilon^{-2}e^{\Theta(\ln(1/\delta))}, \epsilon^{-2}\ln(1/\delta)e^{\Theta(\max(1,\ln(1/\delta)\epsilon^{-1}/s))}\right)$, apart from a constant-factor interval $C_1\epsilon^{-2}\ln(1/\delta) \le m \le C_2\epsilon^{-2}\ln(1/\delta)$ where we do not have a bound on the behavior of sparse JL.

Proof of Corollary A.1. The first statement follows from the fact the lower bound in Theorem 3.2 holds for any sparse JL transform. For the upper bound, we also use Theorem 3.2. Let's set $C_v\sqrt{\epsilon s}\frac{\sqrt{\ln(m\epsilon^2}}{\sqrt{p}} = \frac{1}{2}$, where C_v is the implicit constant in the upper bound. This solves to $m = \epsilon^{-2}pe^{\frac{C_L p\epsilon^{-1}}{s}}$ for some constant C_L as desired. We also have the condition that $m \le \epsilon^{-2}e^{\Theta(\ln(1/\delta))}$ for this regime to be reached. We can obtain the max with 1 on the exponent, by using that $v(m,\epsilon,\delta,s) = 0$ when $m \le \Theta(\epsilon^{-2}\ln(1/\delta))$. To avoid having a gap when $m = s\cdot e^{\Theta(\max(1,\ln(1/\delta)\epsilon^{-1}/s))}$, we implicitly use that our lower bound actually doesn't have a gap between these regimes (though there may be a gap in the boundary between the lower bound and upper bound). Thus, we only have to keep the gap $C_1\epsilon^{-2}\ln(1/\delta) \le m \le C_2\epsilon^{-2}\ln(1/\delta)$ where we do not have a lower bound. □

Notice that the upper and lower bounds in Corollary $A.1$ also match up to constant factors on the exponent in the dimension m.

A.2 Proofs of Lemma 6.3 and Lemma 6.4

We analyze the moments of $Z_r(x_1,\ldots,x_n)$, proving Lemma 6.4 and Lemma 6.3. Our lower bound in Lemma 6.4 holds for $\|Z_r(v,\ldots,v,0,\ldots,0\|_q$ as well as $\left\|Z_r(v,\ldots,v,0,\ldots,0)I_{\sum_{i=1}^{1/v^2}\eta_{r,i}=2}\right\|_T$ (for technical reasons discussed in Appendix A.3). Our upper bound in Lemma 6.3 holds for $\|Z_r(x_1,\ldots,x_n)\|_q$. In Appendix A.2.1, we prove Lemma 6.4. In Appendix A.2.2, we prove Lemma 6.3.

A.2.1 Proof of Lemma 6.4

The key ingredient of the proof is Lemma 5.5 (for Rademacher quadratic forms). We can view $Z_r(v,\ldots,v,0,\ldots,0)$ as the following quadratic form:

$$Z_r(v,\ldots,v,0,\ldots,0) = v^2\sum_{1\le i\ne j\le N}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j},$$

where $N = \frac{1}{v^2}$. Since the support of $\eta_{r,i}$ is $\{0,1\}$ and due to symmetry of this random variable, it is tractable to analyze the expressions in Lemma 5.5.

Proof of Lemma 6.4. First, we handle the case of $T = 2$:

$$\mathbb{E}[Z_r(v,\ldots,v,0,\ldots,0)]^2 = v^4\mathbb{E}\left[\left(\sum_{i\ne j}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}\right)^2\right]$$

$$= 2v^4 \mathbb{E}\left[\sum_{i\neq j} \eta_{r,i}\eta_{r,j}\right] = 2v^4 \left(\frac{s}{m}\right)^2 N(N-1) \geq \frac{v^4 N^2 s^2}{m^2} = \frac{s^2}{m^2}$$

as desired.

Now we consider $T > 2$, and we prove a bound on $\|Z_1(v,\ldots,v,0,\ldots,0)\|_T$. We see that $\|Z_1(v,\ldots,v,0,\ldots,0)\|_T = v^2 \left\|\sum_{i\neq j} \eta_{1,i}\eta_{1,j}\sigma_{1,i}\sigma_{1,j}\right\|_T$. Fix $1 \leq M \leq \min(N,T)$. We use Lemma 5.5 with $Y_{i,j} = \eta_{1,i}\eta_{1,j} I_{M=\sum_{k=1}^N \eta_{1,k}}$ to compute $\left\|\sum_{i\neq j} \eta_{1,i}\eta_{1,j}\sigma_{1,i}\sigma_{1,j} I_{M=\sum_{k=1}^N \eta_{1,k}}\right\|_T$. We will then aggregate over $2 \leq M \leq T$ and not even count $M = 1$ or $T < M \leq N$. We only use the operator-norm-like term in Lemma 5.5. Observe that

$$I_{M=\sum_{k=1}^N \eta_{1,k}} \sup_{\|b\|_2,\|c\|_2\leq\sqrt{T},\|b\|_\infty,\|c\|_\infty\leq 1} \sum_{i\neq j} \eta_{1,i}\eta_{1,j} b_i c_j$$

is equal to

$$I_{M=\sum_{k=1}^N \eta_{1,k}} \sup_{\|b\|_2,\|c\|_2\leq\sqrt{T},\|b\|_\infty,\|c\|_\infty\leq 1} \sum_{i,j|\eta_{1,i}=1,\eta_{1,j}=1} b_i c_j \geq I_{M=\sum_{k=1}^N \eta_{1,k}} M(M-1),$$

where we set $b_i = 1$ on all i such that $\eta_{1,i} = 1$ and $c_j = 1$ on all j such that $\eta_{1,j} = 1$.

Since the events $M = \sum_{k=1}^N \eta_{1,k}$ are disjoint across different M values, we know that:

$$\begin{aligned}
\left\|\sum_{i\neq j} \eta_{1,i}\eta_{1,j}\sigma_{1,i}\sigma_{1,j}\right\|_T &\gtrsim \left(\sum_{M=2}^{\min(T,N)} \left\|\sum_{i\neq j} \eta_{1,i}\eta_{1,j}\sigma_{1,i}\sigma_{1,j} I_{M=\sum_{k=1}^N \eta_{1,k}}\right\|_T^T\right)^{1/T} \\
&\gtrsim \left(\sum_{M=2}^{\min(T,N)} \left\|I_{M=\sum_{k=1}^N \eta_{1,k}} \sup_{\|b\|_2,\|c\|_2\leq\sqrt{T},\|b\|_\infty,\|c\|_\infty\leq 1} \sum_{i\neq j} \eta_{1,i}\eta_{1,j} b_i c_j\right\|_T^T\right)^{1/T} \\
&\gtrsim \left(\sum_{M=2}^{\min(T,N)} \left\|I_{M=\sum_{k=1}^N \eta_{1,k}} M^2\right\|_T^T\right)^{1/T} \\
&= \left(\sum_{M=2}^{\min(T,N)} \mathbb{P}[M = \sum_{i=1}^N \eta_{1,i}] M^{2T}\right)^{1/T} \\
&= \left(\sum_{M=2}^{\min(T,N)} \binom{N}{M} \left(\frac{s}{m}\right)^M \left(1-\frac{s}{m}\right)^{N-M} M^{2T}\right)^{1/T} \\
&\gtrsim \left(\sum_{M=2}^{\min(T,N)} \left(\frac{Ns}{mT}\right)^M \left(\frac{T}{M}\right)^M \left(1-\frac{s}{m}\right)^{N-M} M^{2T}\right)^{1/T} \\
&\gtrsim \left(\sum_{M=2}^{\min(T,N)} \left(\frac{s}{mTv^2}\right)^M \left(1-\frac{s}{m}\right)^{N-M} M^{2T}\right)^{1/T}
\end{aligned}$$

$$\gtrsim \left(\sum_{M=2}^{\min(T,N)} \left(\frac{s}{mTv^2}\right)^M M^{2T} \right)^{1/T},$$

where the last line follows from the fact that since $T \geq \frac{se}{mv^2}$ and $s \leq m/e$, we know that:

$$\left(1-\frac{s}{m}\right)^{\frac{N-M}{T}} \geq \left(1-\frac{s}{m}\right)^{\frac{N}{T}} \geq \left(1-\frac{s}{m}\right)^{\frac{Nmv^2}{se}} \geq \left(1-\frac{s}{m}\right)^{\frac{m}{s}} \geq 0.25.$$

Setting $t = T/M$, we obtain, up to constants:

$$\sup_{2\leq M\leq \min(T,N)} \left(\frac{s}{mTv^2}\right)^{M/T} M^2 = \sup_{\max(1,T/N)\leq t\leq T/2} \left(\frac{T^2}{t^2}\right) \left(\frac{s}{mTv^2}\right)^{1/t}.$$

We can take a derivative to obtain the two expressions in the lemma statement at the following regimes of parameters: $\max(1, Tv^2) \leq \ln(Tmv^2/s) \leq T$ and $\ln(Tmv^2/s) > T$. The second regime aligns with the lemma statement. Thus it suffices to show that when $v \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$, it is true that $Tv^2 \leq \ln(Tmv^2/s)$. This is a straightforward calculation[1].

Now, let's consider the case where we want to bound $\left\| Z_1(v,\ldots,v,0,\ldots,0) I_{\sum_{k=1}^N \eta_{1,k}=2} \right\|_T$. It follows from the above calculations, without taking the sum that we obtain a lower bound of

$$\left(\binom{N}{2} \left(\frac{s}{m}\right)^2 \left(1-\frac{s}{m}\right)^{N-2} \right)^{1/T} \gtrsim \left(\frac{s}{mTv^2}\right)^{2/T}.$$

□

A.2.2 Proof of Lemma 6.3

In the paper, we discussed the tractability issues with using the general quadratic form moment bound Lemma 5.5 to upper bound $\|Z_r(x_1,\ldots,x_n)\|_q$. Thus, we require simpler bounds that are easier to analyze. Linear forms naturally arise in the upper bound since $Z_r(x_1,\ldots,x_n) = \left(\sum_{1\leq i\leq n} \eta_{r,i}\sigma_{r,i}x_i\right)^2 - \sum_{1\leq i\leq n} \eta_{r,i}x_i^2 \leq \left(\sum_{1\leq i\leq n} \eta_{r,i}\sigma_{r,i}x_i\right)^2$. However, it turns out that a vanilla linear form bound (e.g. Proposition 5.9) here is weak due to the loss arising from ignoring the $\sum_{1\leq i\leq n} \eta_{r,i}x_i^2$ term. Thus, we use Lemma 5.10 (our generalized bound tailored to squares of linear forms with a zero diagonal) to obtain:

Lemma A.2 If $\|x\|_\infty \leq v$ and $\|x\|_2 \leq 1$, then we have that:

$$\|Z_r(x_1,\ldots,x_n)\|_T = \left\| \sum_{i\neq j} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_ix_j \right\|_T \lesssim v^2 \left(\sup_{1\leq t\leq T/2} \frac{T^2}{t^2} \left(\frac{s}{mTv^2}\right)^{1/t} \right).$$

Proof. This can be seen by simply taking $Y_i = \eta_{r,i}\sigma_{r,i}$ in Lemma 5.10. □

[1]In fact, $v = \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$ is very close to the value where $Tv^2 = \ln(Tmv^2/s)$, so this approximation is essentially tight.

It turns out that using only this bound would lose the $m \geq s \cdot e^{\Theta\left(\max(1, \frac{p\epsilon^{-1}}{s})\right)}$ branch in the lower bound on $v(m, \epsilon, \delta, s)$ in Theorem 3.2. The lower bound on moments of $\|Z_r(v, \ldots, v, 0, \ldots, 0)\|_T$ in Lemma 6.4 sheds light on where this loss may be arising. We see that the problematic case is when $v \geq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}} =: v_1$, and so we require a new bound for this regime. Since the vector $[v_1, \ldots, v_1, 0, \ldots, 0]$ is in S_v when $v_1 \leq v$, we can't hope to beat the bound of $||Z_r(v_1, \ldots, v_1, 0, \ldots, 0)||_T \gtrsim \frac{T^2 v_1^2}{\ln^2(Tmv_1^2/s)} \simeq \frac{T}{\ln(m/s)}$ from Lemma 6.4. We show that we can match this value:

Lemma A.3 Suppose that $x = [x_1, \ldots, x_n]$ satisfies $\|x\|_2 = 1$ and $\|x\|_\infty < v$. If $s \leq m/e$, $T \geq \frac{se}{mv^2}$, $T \geq 3$, $T \geq \ln(mv^2/s)$, then:

$$\|Z_r(x_1, \ldots, x_n)\|_T = \left\| \sum_{i \neq j} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j \right\|_T \leq \left\| \sum_i \eta_{r,i}\sigma_{r,i}x_i \right\|_{2T}^2 \lesssim \frac{T}{\ln(m/s)}.$$

The proof of this bound requires a new technique that handles larger $|x_i|$ entries, while still managing the many smaller $|x_i|$ that are still allowed to be present. We separate out $|x_i| \geq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$ and $|x_i| \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$. In the quadratic form formulation of $Z_r(x_1, \ldots, x_n)$, this separation cannot be carried out, since there would be cross-terms between $|x_i| \geq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$ and $|x_i| \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$. As a result, we require the linear form bound (Proposition 5.9) for $|x_i| \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$, and it turns out to be sufficiently tight in this regime.

Proof of Lemma A.3. WLOG, assume that $|x_1| \geq |x_2| \geq \ldots \geq |x_n|$. Let $P = \left\lceil \frac{T}{\ln(m/s)} \right\rceil$. We know that

$$\left\| \sum_i \eta_{r,i}\sigma_{r,i}x_i \right\|_{2T} \leq \left\| \sum_{1 \leq i \leq P} \eta_{r,i}\sigma_{r,i}x_i \right\|_{2T} + \left\| \sum_{i > P} \eta_{r,i}\sigma_{r,i}x_i \right\|_{2T}.$$

For $1 \leq i \leq P$, we use the bound $|\sum_{i=1}^p \eta_{r,i}\sigma_{r,i}x_i| \leq \sum_{i=1}^p |x_i| \leq \sqrt{\left\lceil \frac{T}{\ln(m/s)} \right\rceil} \leq 2\sqrt{\frac{T}{\ln(m/s)}}$. For the remaining terms, we take $Y_i = \eta_{r,i}\sigma_{r,i}$ in Proposition 5.9 to obtain the following upper bound[2] for $|x_i| \leq v' := \frac{\sqrt{\ln(m/s)}}{\sqrt{T}}$ and $\|x\|_2 \leq 1$:

$$\left\| \sum_i \eta_{r,i}\sigma_{r,i}x_i \right\|_{2T} \lesssim v' \left(\sup_{1 \leq t \leq T} \frac{T}{t} \left(\frac{s}{mTv'^2} \right)^{\frac{1}{2t}} \right).$$

Based on the conditions in this lemma statement, we know that $\frac{mTv'^2}{s} = \frac{mT\ln(m/s)}{sT} = \frac{m}{s\ln(m/s)} \geq e$. Thus taking a derivative, we obtain that this can be upper bounded by taking

[2]Observe that the upper endpoint of T on the sup expression does not match with the upper endpoint of $T/2$ on the sup expression in Lemma A.2, and in fact, it turns out that this bound is not sufficiently strong to recover Theorem 3.2. This is sufficiently tight here, since we are focusing on the case where $\ln(Tmv'^2/s)$ is *small*.

$t = \ln(mTv'^2/s)$ which yields:

$$\frac{Tv'}{\ln(mTv'^2/s)} = \frac{Tv'}{\ln(m/s) - \ln\ln(m/s)} \leq \frac{Tv'}{0.5\ln(m/s)} = 2\frac{\sqrt{T}}{\sqrt{\ln(m/s)}}.$$

□

Finally, combining Lemma $A.2$ and Lemma $A.3$ yields Lemma 6.3:

Proof of Lemma 6.3. We apply Lemma $A.2$ at $T = 2$ to directly obtain $\frac{Ts}{m}$, and for $T \geq 3$, we apply Lemma $A.2$ and take a derivative to obtain:

$$\left\| \sum_{i \neq j} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j \right\|_T \lesssim v^2 \begin{cases} \frac{Ts}{mv^2}, & \text{for } se \geq mTv^2 \\ \frac{T^2}{\ln(mTv^2/s)^2}, & \text{for } se \leq mTv^2, \ln(Tmv^2/s) \leq T \\ \left(\frac{s}{mTv^2}\right)^{2/T}, & \text{for } \ln(Tmv^2/s) \geq T, se \leq mTv^2 \end{cases}.$$

To obtain the desired bound, we also include the bound from Lemma $A.3$ in the middle regime. □

A.3 Combining rows to bound $\|R(x_1, \ldots, x_n)\|_q$

Now, we show to move from bounds on moments of individual rows (i.e. $Z_r(x_1, \ldots, x_n)$) to bounds on moments of $R(x_1, \ldots, x_n)$. In Appendix A.3.1, we obtain an upper bound on $\|R(x_1, \ldots, x_n)\|_q$, thus proving Lemma 6.5. In Appendix A.3.2, we obtain a lower bound on $\|R(x_1, \ldots, x_n)\|_q$, thus proving Lemma 6.6.

A.3.1 Proof of Lemma 6.5

Since the $\eta_{r,i}$ are negatively correlated and q is even, we can always upper bound the moments of $R(x_1, \ldots, x_n)$ by the case of a sum of *independent* random variables[3] $Z'_r(x_1, \ldots, x_n) \sim Z_r(x_1, \ldots, x_n)$. We see that:

$$s \cdot \|R(x_1, \ldots, x_n)\|_q \leq \left\| \sum_{r=1}^{m} Z_r(x_1, \ldots, x_n) \right\|_q \tag{A.1}$$

$$\leq \left\| \sum_{r=1}^{m} Z'_r(x_1, \ldots, x_n) \right\|_q \tag{A.2}$$

$$\lesssim \sup_{2 \leq T \leq q} \frac{q}{T} \left(\frac{m}{q}\right)^{1/T} \|Z_1(x_1, \ldots, x_n)\|_T, \tag{A.3}$$

where the last inequality follows from Lemma 5.11. Thus, it remains to analyze the sup expression. It turns out that each regime of bounds in Lemma 6.3 collapses to one value, so the different regimes in Lemma 6.3 correspond to different parts of the max expressions in Lemma 6.5. Depending on the parameters, some of these regimes may not exist, as is reflected by branches of the max expression sometimes vanishing in Lemma 6.3. We defer the computation to Appendix A.4.

[3]This can easily be seen by expanding.

A.3.2 Proof of Lemma 6.6

Moving from a lower bound on the moments of individual rows given by Lemma 6.4 to moments of
$R(v, \ldots, v, 0, \ldots, 0)$ is more delicate. Unlike in the upper bound, the negative correlations between random variables require some care to handle, even with the simplification that the s nonzero entries in a column are chosen uniformly at random. For example, the conditional distribution of $\eta_{s+1,1} \mid \eta_{1,1} = \eta_{2,1} = \ldots = \eta_{s,1} = 1$ is 0, while the marginal distribution of $\eta_{s+1,1}$ has expectation s/m. One aspect that simplifies our analysis is that we *know* from our proof of Lemma 6.5 which moments of $Z_r(x_1, \ldots, x_n)$ are critical in the sup expression in (A.1). We only need to account for these particular moments in our lower bound approach. It turns out that the three critical values are $q/T = 2$, $q/T = q$, and $q/T = \ln(qmv^4/s^2)$.

For $q/T = q$, where rows are isolated, we can directly obtain a bound from Lemma 5.12 and Lemma 6.4 to obtain.

Lemma A.4 Suppose $\mathcal{A}_{s,m,n}$ is a uniform sparse JL transform. Suppose that q is even, $s \leq m/e$, $q \geq \frac{se}{mv^2}$, $1 \leq \ln(qmv^2/s) \leq q$, $v \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{q}}$ and $\frac{1}{v^2}$ is an even integer. Then it is true that:

$$\|R(v, \ldots, v, 0, \ldots, 0)\|_q \gtrsim \frac{q^2 v^2}{s \ln^2(\frac{qmv^2}{s})}.$$

Proof. By Lemma 5.12 with $T = 1$, we have that:

$$\|R(v, \ldots, v, 0, \ldots, 0)\|_q \geq \frac{m^{1/q}}{s} \|Z_1(v, \ldots, v, 0, \ldots, 0)\|_q \geq \frac{1}{s} \|Z_1(v, \ldots, v, 0, \ldots, 0)\|_q .$$

Now, we apply Lemma 6.4 to obtain the desired expression. □

For $q/T = 2$ and $q/T = \ln(qmv^4/s^2)$, we make use of the Lemma A.5 that relates moments of products of rows to products of moments of rows by taking advantage of either s and $\frac{1}{v^2}$ being sufficiently large. The method essentially uses a counting argument to show that not too many terms vanish as a result of negative correlations, and requires adding in an indicator for the number of nonzero entries in a row being 2 for some cases (which is sufficient to prove Lemma 6.6).

Lemma A.5 Suppose $\mathcal{A}_{s,m,n}$ is a uniform sparse JL transform. If $1 \leq T \leq q/2$ is an integer, q/T is an even integer, $\frac{1}{v^2}$ is an even integer, and $2Tv^2 \leq s$, then:

$$\left\| \prod_{i=1}^{T} Z_i(v, \ldots, v, 0, \ldots, 0) \right\|_{q/T}^{1/T} \gtrsim \begin{cases} \|Z_1(v, \ldots, v, 0, \ldots, 0)\|_2 & \text{if } T = q/2 \\ \left\| Z_1(v, \ldots, v, 0, \ldots, 0) I_{\sum_{i=1}^{N} \eta_{1,i} = 2} \right\|_{q/T} & \text{if } 1 \leq T \leq q/2 \end{cases}.$$

We defer the proof to Appendix A.5.

Now we can use Lemma 5.12 coupled with Lemma A.5 and Lemma 6.4 to handle the cases of $q/T = 2, \ln(qmv^4/s^2)$ and obtain the following bounds. For $q/T = 2$, we obtain:

Lemma A.6 Suppose $\mathcal{A}_{s,m,n}$ is a uniform sparse JL transform. If q is an even integer, $\frac{qv^2}{s} \leq 1$, and $\frac{1}{v^2}$ is an even integer, then it is true that:

$$\|R(v, \ldots, v, 0, \ldots, 0)\|_q \gtrsim \left(\frac{q}{m}\right)^{1/2}.$$

Proof of Lemma A.6. Take $T = \frac{q}{2}$ and $qv^2 \leq s$. By Lemma A.5 and Lemma 5.12, we have that:

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \gtrsim \frac{q}{s}\left(\frac{m}{q}\right)^{1/2}\|Z_1(v,\ldots,v,0,\ldots,0)\|_2 .$$

Now, by Lemma 6.4, we can see that $\|Z_1(v,\ldots,v,0,\ldots,0)\|_2 \gtrsim \frac{s}{m}$. Thus, our bound becomes:

$$\frac{q}{s}\left(\frac{m}{q}\right)^{1/2}\frac{s}{m} = \left(\frac{q}{m}\right)^{1/2}.$$

□

For $q/T = \ln(qmv^4/s^2)$, we similarly obtain the following bound using Lemma 5.12 coupled with Lemma $A.5$.

Lemma A.7 Suppose $\mathcal{A}_{s,m,n}$ is a uniform sparse JL transform. Suppose that q is a power of 2, $s \leq m/e$, $2qv^2 \leq 0.5s\ln(qmv^4/s^2)$, $\frac{1}{v^2}$ is even, $2 \leq \ln(qmv^4/s^2) \leq q$, and $m \geq q$. Then it is true that:

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \gtrsim \frac{qv^2}{s\ln(\frac{qmv^4}{s^2})}.$$

Proof. Let's let $f(x)$ be the function that rounds x to the nearest power of 2. By the conditions, we know that $2 \leq f(\ln(qmv^4/s^2)) \leq q$. Now, we want the condition $2qv^2 \leq sf(\ln(qmv^4/s^2))$ to be satisfied. If $f(\ln(qmv^4/s^2)) \geq \ln(qmv^4/s^2)$, then this is implied by $2qv^2 \leq s\ln(qmv^4/s^2) = s\max(\ln(qmv^4/s^2), 2)$, which is a strictly weaker condition than the one given in the lemma statement. If $f(\ln(qmv^4/s^2)) \leq \ln(qmv^4/s^2)$, then $f(\ln(qmv^4/s^2)) \geq 0.5\ln(qmv^4/s^2)$ and so $2qv^2 \leq 0.5s\ln(qmv^4/s^2) \leq sf(\ln(qmv^4/s^2))$ gives the desired condition.

We use the fact that $\ln(qmv^4/s^2)/2 \leq f(\ln(qmv^4/s^2)) \leq 2\ln(qmv^4/s^2)$. We apply Lemma A.5 and Lemma 5.12, with $T = \frac{q}{f(\ln(qmv^4/s^2))}$ and Lemma 6.4 to see that if we have the additional condition that $f(\ln(qmv^4/s^2)) \geq \frac{se}{mv^2}$, then we know that:

$$\begin{aligned}
\|R(v,\ldots,v,0,\ldots,0)\|_q &\gtrsim \frac{q}{sf(\ln(\frac{qmv^4}{s^2}))}\left(\frac{m}{q}\right)^{1/f(\ln(\frac{qmv^4}{s^2}))}\|Z_1(v,\ldots,v,0,\ldots,0)I_{M=2}\|_{f(\ln(\frac{qmv^4}{s^2}))} \\
&\gtrsim \frac{qv^2}{2s\ln(\frac{qmv^4}{s^2})}\left(\frac{m}{q}\right)^{1/f(\ln(\frac{qmv^4}{s^2}))}\left(\frac{s}{mf(\ln(\frac{qmv^4}{s^2}))v^2}\right)^{2/f(\ln(\frac{qmv^4}{s^2}))} \\
&= \frac{qv^2}{2s\ln(\frac{qmv^4}{s^2})}\left(\frac{s^2}{qmv^4}\right)^{1/f(\ln(\frac{qmv^4}{s^2}))}\left(\frac{1}{f(\ln(qmv^4/s^2))^2}\right)^{1/f(\ln(qmv^4/s^2))} \\
&\gtrsim \frac{qv^2}{s\ln(\frac{qmv^4}{s^2})}\left(\frac{s^2}{qmv^4}\right)^{\frac{2}{\ln(\frac{qmv^4}{s^2})}}. \\
&\gtrsim \frac{qv^2}{s\ln(\frac{qmv^4}{s^2})}.
\end{aligned}$$

Now, we see that

$$\frac{mv^2}{se} = \sqrt{\frac{qmv^4}{s^2}}\frac{1}{e}\frac{\sqrt{m}}{\sqrt{q}} \geq \frac{\sqrt{m}}{\sqrt{q}} \geq 1.$$

This implies that $\frac{se}{mv^2} \leq 1$, so the condition of $f(\ln(qmv^4/s^2)) \geq 2 \geq \frac{se}{mv^2}$ is automatically satisfied. □

With these bounds, Lemma 6.6 follows.

Proof of Lemma 6.6. We combine Lemma *A*.4, Lemma *A*.6, and Lemma *A*.7. □

A.4 Proofs of auxiliary lemmas for Lemma 6.5

First, we use Lemma 5.11 and Lemma 6.3 to prove a upper bound $\|R(x_1, \ldots, x_q)\|_q$ that is not quite in the desired form for Lemma 6.5.

Lemma A.8 Let $2 \leq q \leq m$ be an even integer and $|x_i| \leq v$ and $\|x\|_2 = 1$. If $\frac{se}{mv^2} \geq q$, then:

$$\|R(x_1, \ldots, x_n)\|_q \lesssim \alpha_1(q, v, s, m).$$

If $\ln(qmv^2/s) > q$ then we have

$$\|R(x_1, \ldots, x_n)\|_q \lesssim \max(\alpha_1(q, v, s, m), \alpha_2(q, v, s, m)).$$

In all other cases, we have that

$$\|R(x_1, \ldots, x_n)\|_q \lesssim \max(\alpha_1(q, v, s, m), \alpha_2(q, v, s, m), \min(\alpha_3(q, v, s, m), \alpha_4(q, v, s, m))).$$

The functions are defined as follows.

$$\alpha_1(q, v, s, m) = \frac{\sqrt{q}}{\sqrt{m}}$$

$$\alpha_2(q, v, s, m) = \begin{cases} \frac{eqv^2}{s\ln(qmv^4/s^2)} & \text{for } \ln(qmv^4/s^2) \geq 2 \\ \frac{\sqrt{q}}{\sqrt{m}} & \text{for } \ln(qmv^4/s^2) \leq 2 \end{cases}$$

$$\alpha_3(q, v, s, m) = \frac{qv^2e}{s} \sup_{T \leq q, T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{T}{\ln^2(mv^2T/s)} \left(\frac{s}{qv^2}\right)^{1/T}$$

$$\alpha_4(q, v, s, m) = \frac{qe^2}{s\ln(m/s)} \begin{cases} 1 & \text{for } \ln(qmv^4/s^2) \geq 2 \\ \left(\frac{s}{qv^2}\right)^{1/\ln(mv^2/s)} & \text{else} \end{cases}$$

Proof of Lemma A.8. As we discussed in Appendix A.3, it suffices to bound

$$\frac{1}{s} \sup_{2 \leq T \leq q} \frac{q}{T} \left(\frac{m}{q}\right)^{1/t} \|Z_1(x_1, \ldots, x_n)\|_t .$$

Our bounds on $\|Z_1(x_1,\ldots,x_n)\|_t$ are based on Lemma 6.3. We split into cases based on the T value, and how it separates into different cases in Lemma 6.3. Let

$$\beta_1(q,v,s,m) = \frac{1}{s}\sup_{T=2,3\leq T\leq \frac{se}{mv^2}} \frac{q}{T}\left(\frac{m}{q}\right)^{1/t}\|Z_1(x_1,\ldots,x_n)\|_t.$$

$$\beta_2(q,v,s,m) = \frac{1}{s}\sup_{\max(3,\frac{se}{mv^2})\leq T\leq \ln(mv^2T/s)} \frac{q}{T}\left(\frac{m}{q}\right)^{1/t}\|Z_1(x_1,\ldots,x_n)\|_t.$$

$$\beta_{34}(q,v,s,m) = \frac{1}{s}\sup_{T\geq \max(3,\frac{se}{mv^2},\ln(mv^2T/s))} \frac{q}{T}\left(\frac{m}{q}\right)^{1/t}\|Z_1(x_1,\ldots,x_n)\|_t.$$

Let β_3 branch arise when we use the $\frac{T^2v^2}{\ln(Tmv^2/s)^2}$ for the $\|Z_1(x_1,\ldots,x_n)\|_t$ bound, and let the β_4 branch arise when we use $\frac{Tv^2}{s\ln(m/s)}$ for the $\|Z_1(x_1,\ldots,x_n)\|_t$ bound. Thus, we know that

$$\beta_{34}(q,v,s,m) \leq \min(\beta_3(q,v,s,m),\beta_4(q,v,s,m)).$$

Let's first consider $\frac{se}{mv^2}\geq q$. In this case, only the β_1 branch arises. Now, suppose that $\frac{se}{mv^2}<q$.

Suppose that $\ln(qmv^2/s)>q$. Then we show that the β_{34} branch does not arise. It suffices to show that $\ln(Tmv^2/s)>T$ for all $T\geq\frac{se}{mv^2}$. Let $x=Tmv^2/s$. It suffices to show that $\frac{s}{mv^2}\frac{x}{\ln x}<1$ for all $e\leq x\leq\frac{qmv^2}{s}$. Since $\frac{s}{mv^2}\frac{x}{\ln x}<1$ at $x=\frac{qmv^2}{s}$ and this is an increasing function of x, we know that the condition is true.

We now produce bounds $\alpha_1(q,v,s,m),\ldots,\alpha_4(q,v,s,m)$ such that $\beta_i(q,v,sm)\lesssim\alpha_i(q,v,s,m)$, which is what we do for the remainder of the analysis.

First, we handle the $\beta_1(q,v,s,m)$ term. We see that

$$\beta_1(q,v,s,m) = \frac{1}{s}\sup_{2\leq T\leq\frac{s}{mv^2}}\frac{q}{T}\left(\frac{m}{q}\right)^{1/T}\frac{Ts}{m} = \frac{1}{s}\frac{qs}{m}\left(\frac{m}{q}\right)^{1/T}\leq\frac{q}{m}\left(\frac{m}{q}\right)^{1/2}=\frac{\sqrt{q}}{\sqrt{m}}.$$

Now, we handle the $\beta_2(q,v,s,m)$ term. We obtain a bound for $\|Z_r\|_T\lesssim v^2\left(\frac{s}{mTv^2}\right)^{2/T}$. The expression becomes:

$$\begin{aligned}\beta_2(q,v,s,m) &= \frac{1}{s}\sup_{T\geq\max(\frac{se}{mv^2},3),T\leq\ln(mv^2T/s)}\frac{qv^2}{T}\left(\frac{m}{q}\right)^{1/T}\left(\frac{s}{mTv^2}\right)^{2/T}\\ &= \frac{1}{s}\sup_{T\geq\max(\frac{se}{mv^2},3),T\leq\ln(mv^2T/s)}\frac{qv^2}{T}\left(\frac{s}{\sqrt{qm}Tv^2}\right)^{2/T}\\ &\leq \frac{1}{s}\sup_{T\geq\max(\frac{se}{mv^2},3),T\leq\ln(mv^2T/s)}\frac{qv^2}{T}\left(\frac{s^2}{qmv^4}\right)^{1/T}.\end{aligned}$$

Suppose that $\ln(qmv^4/s^2)\geq 2$. In this case, we have that this expression is upper bounded by $T=\ln(qmv^4/s^2)$. When we plug this into the expression, we obtain $\frac{qv^2}{s\ln(qmv^4/s^2)}$. Otherwise,

if $\ln(qmv^4/s^2) \leq 2$, then this expression is upper bounded by $T = 3$:

$$\frac{qv^2}{s}\left(\frac{s^2}{qmv^4}\right)^{1/3} = \frac{C_1 C_5 q^{2/3} v^{2/3}}{s^{1/3} m^{1/3}}.$$

We know that that $\frac{q^{2/3}v^{2/3}}{s^{1/3}m^{1/3}} \leq \frac{\sqrt{q}}{\sqrt{m}}$ because this reduces to

$$\frac{q^{1/6}v^{2/3}m^{1/6}}{s^{1/3}} = \left(\frac{qmv^4}{s^2}\right)^{1/6} \leq e^{1/3}.$$

Now, we handle the $\beta_4(q, v, s, m)$ term when $\ln(qmv^2/s) \leq q$.

$$\begin{aligned}
\beta_4(q, v, sm) &= \frac{1}{s} \sup_{T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{q}{T}\left(\frac{m}{q}\right)^{1/T} \frac{T}{\ln(m/s)} \\
&\leq \sup_{T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{q}{s\ln(m/s)}\left(\frac{mv^2}{s}\right)^{1/T}\left(\frac{s}{qv^2}\right)^{1/T} \\
&\leq \frac{qe}{s\ln(m/s)} \sup_{T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \left(\frac{s}{qv^2}\right)^{1/T}
\end{aligned}$$

If $s \leq qv^2$, this is bounded by 1, and if $s \geq qv^2$, this is bounded by $\left(\frac{s}{qv^2}\right)^{1/\ln(mv^2/s)}$. We see that $\frac{s}{qv^2} \leq \frac{mv^2}{s}$, so $\left(\frac{s}{qv^2}\right)^{1/\ln(mv^2/s)} \leq \left(\frac{mv^2}{s}\right)^{1/\ln(mv^2/s)} \leq e$. Thus this is bounded by $\frac{qe^2}{s\ln(m/s)}$.

Now, we handle the $\beta_3(q, v, s, m)$ term. In this case, the expression becomes:

$$\begin{aligned}
\beta_3(q, v, s, m) &= \frac{1}{s} \sup_{T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{qv^2}{T}\left(\frac{m}{q}\right)^{1/T} \frac{T^2}{\ln^2(mv^2T/s)} \\
&\leq \sup_{T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{qv^2T}{s\ln^2(mv^2T/s)}\left(\frac{mv^2}{s}\right)^{1/T}\left(\frac{s}{qv^2}\right)^{1/T} \\
&\leq \frac{qv^2e}{s} \sup_{T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{T}{\ln^2(mv^2T/s)}\left(\frac{s}{qv^2}\right)^{1/T}
\end{aligned}$$

□

We use some function bounding arguments to come with a simpler bound for α_3 for sufficiently large v.

Lemma A.9 Assume that $C_2 q^3 mv^4 \geq s^2$ for some $C_2 \geq 1$. Then it is true that

$$\frac{qv^2e}{s} \sup_{T \leq q, T \geq \frac{se}{mv^2}, 3, \ln(mv^2T/s)} \frac{T}{\ln^2(mv^2T/s)}\left(\frac{s}{qv^2}\right)^{1/T} \leq \frac{C_2^{1/3} q^2 v^2 e^5}{s\ln^2(mv^2q/s)}.$$

Proof of Lemma A.9. With the assumptions that we made we know that $\frac{s}{q^3v^2C_2^2} \leq \frac{mv^2}{s}$. This implies that our expression becomes:

$$\begin{aligned}
E &= \frac{qv^2e}{s} \sup_{T\leq q, T\ \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{T}{\ln^2(mv^2T/s)} \left(\frac{s}{qv^2}\right)^{1/T} \\
&= \frac{qv^2e}{s} \sup_{T\leq q, T\geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} C_2^{1/T} \frac{T}{\ln^2(mv^2T/s)} \left(\frac{s}{C_2q^3v^2}\right)^{1/T} q^{2/T} \\
&\leq \frac{qv^2e^2}{s} C_2^{1/3} \sup_{T\leq q, T\geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{T}{\ln^2(mv^2T/s)} q^{2/T}
\end{aligned}$$

.

It suffices to show that $\sup_{T\leq q, T\geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))} \frac{T}{\ln^2(mv^2T/s)} q^{2/T} \leq \frac{qe^3}{\ln^2(mv^2q/s)}$.

Let T_{min} be the minimum T such that $T \geq \max(\frac{se}{mv^2}, 3, \ln(mv^2T/s))$. We just need to bound

$$\begin{aligned}
\sup_{T_{min}\leq T\leq q} \frac{Tq^{2/T}}{\ln^2(mv^2T/s)} &\leq \max\left(\sup_{T_{min}\leq T\leq \ln q} \frac{Tq^{2/T}}{\ln^2(mv^2T/s)}, \sup_{\max(T_{min}, \ln q)\leq T\leq q} \frac{Tq^{2/T}}{\ln^2(mv^2T/s)}\right) \\
&\leq \max\left(\sup_{T_{min}\leq T\leq \ln q} \frac{Tq^{2/T}}{\ln^2(mv^2T/s)}, e^2 \sup_{\max(T_{min}, \ln q)\leq T\leq q} \frac{T}{\ln^2(mv^2T/s)}\right)
\end{aligned}$$

First, we handle the second term. Let $Q = \frac{mv^2T}{s}$. We use that $T_{min} \geq \frac{se}{mv^2}$, so $\frac{mv^2T_{min}}{s} \geq e$ to conclude $Q \geq e$. We see that

$$e^2 \sup_{\max(T_{min}, \ln q)\leq T\leq q} \frac{T}{\ln^2(mv^2T/s)} \leq e^2 \frac{s}{mv^2} \sup_{e\leq Q\leq \frac{qmv^2}{s}} \frac{Q}{\ln^2(Q)}.$$

We see that setting Q to its maximum value achieves within a factor of e of the maximum value of $\frac{Q}{\ln^2(Q)}$. Thus, we obtain that this is upper bounded by $e^3 \frac{q}{\ln^2(mv^2q/s)}$.

Now, we just need to handle the first term. If $T_{min} \geq \ln q$, then this term doesn't exist. Let's take a log of the expression to obtain:

$$\ln\left(\frac{T}{\ln^2(mv^2T/s)}\right) = \ln T - 2\ln\ln(mv^2T/s) + \frac{2}{T}\ln(q)).$$

The derivative is:

$$\frac{1}{T} - \frac{2}{T\ln(mv^2T/s)} - \frac{2}{T^2}\ln(q).$$

The sign of the derivative is the same as:

$$1 - \frac{2}{\ln(mv^2T/s)} - \frac{2\ln q}{T}.$$

Since $T_{min} \geq \frac{se}{mv^2}$, we know that $\ln(mv^2T/s) \geq 0$. Thus, we know that $1 - \frac{2}{\ln(mv^2T/s)} \leq 1$. Since $T \leq \ln q$, we know that $\frac{\ln q}{T} \geq 1$, so $-\frac{2\ln q}{T} \leq -2$. Thus, the derivative is negative, so the sup is attained at $T_{min} = T$, where the expression is:

$$e^3 \frac{T_{min} q^{2/T_{min}}}{\ln^2(mv^2T_{min}/s)} \leq e^3 \frac{(\ln q) q^{2/3}}{\ln^2(mv^2T_{min}/s)} \leq e^3 \frac{q^{3/4}}{\ln^2(mv^2T_{min}/s)}.$$

Thus, to upper bound by $\frac{q}{\ln^2(mv^2q/s)}$, it suffices to show:

$$\frac{\ln^2(mv^2q/s)}{\ln^2(mv^2T_{min}/s)} \leq q^{0.25}.$$

If $\frac{s}{mv^2} \leq 1$, the ratio is at most

$$\frac{\ln(mv^2q/s)}{\ln(mv^2T_{min}/s)} \leq \frac{\ln(mv^2/s) + \ln q}{\ln(mv^2/s) + \ln T_{min}} \leq \frac{\ln q}{\ln e} = \ln q \leq q^{0.25}.$$

If $\frac{s}{mv^2} \geq 1$, then $qmv^2/s \leq q$. Using this and $\frac{mv^2T_{min}}{s} \geq e$, we know:

$$\frac{\ln(mv^2q/s)}{\ln(mv^2T_{min}/s)} \leq \frac{\ln(q)}{\ln(e)} = \ln q \leq q^{0.25}.$$

□

Now, we combine Lemma $A.8$ and Lemma $A.9$ to prove Lemma 6.5.

Proof of Lemma 6.5. First, we compute the second moment by hand:

$$\begin{aligned}
\mathbb{E}[R(x_1, \ldots, x_n)]^2 &= \frac{1}{s^2}\mathbb{E}\left[\left(\sum_{r=1}^{m}\sum_{i\neq j}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_ix_j\right)^2\right] \\
&= \frac{2}{s^2}\mathbb{E}\left[\sum_{r=1}^{m}\sum_{i\neq j}\eta_{r,i}\eta_{r,j}x_i^2x_j^2\right] \\
&\leq \frac{2}{m}\left(\sum_{i=1}^{n}x_i^2\right)^2 \\
&= \frac{2}{m}.
\end{aligned}$$

For $2 < q \leq m$, we apply Lemma $A.8$ and Lemma $A.9$. We only include α_4 when $\ln(qmv^4/s^2) \geq 2$ to simplify the bound. The bound follows. □

A.5 Proof of auxiliary lemma for Lemma 6.6

We prove Lemma $A.5$.

Proof of Lemma A.5. First, we show the following fact: Suppose that there are T distinguishable buckets and we want to a assign an ordered pair of 2 unequal elements in $[N]$ to each bucket so that the total number of times that any element $i \in [N]$ shows up is $\leq s$. We show that the number of such assignments is at least $C^T N^{2T}$ for some constant C. To prove this, we first consider the case where $N \geq 2T$. In this case, we have that the number of such assignments is at least:

$$N(N-1)(N-2)\dots(N-2T+1) \geq C_1^{2T} N^{2T}.$$

Now, if $N < 2T$, then we define:

$$\beta = \left\lceil \frac{2T}{N} \right\rceil = \lceil 2Tv^2 \rceil \leq s.$$

We partition $2T$ into β blocks, each of size N, until potentially the last block, which may be smaller. We can read off ordered pairs assigned to each bucket from this formulation. Let's assume that each block is a permutation of $1, \dots, N$, and the last block is $2T - (\beta-1)(N)$ non-equal numbers drawn from $1, \dots, N$. (this satisfies the unequal ordered pair condition). Then the number of assignments is $(N!)^{\beta-1} \cdot (N)(N-1)\dots(N-(2T-(\beta-1)(N))+1)$. This is at least as big as $C_1^{2T} N^{2T}$ for some constant C_1.

First, we handle the case where $q/T = 2$. Since we have a uniform sparse JL transform, we know that for $1 \leq x \leq s$:

$$\mathbb{E}[\eta_{1,1} \dots \eta_{x,1}] \geq \frac{s(s-1)\dots(s-x+1)}{(m)(m-1)\dots(m-x+1)} \geq C_2^{-x} \left(\frac{s}{m}\right)^x.$$

We know that

$$Z_r^2 = 2\left(\sum_{i \neq j} \eta_{r,i}\eta_{r,j}\right) + Y_r,$$

where Y_r has expectation 0. In this case we have that

$$Z_1^2 \dots Z_T^2 = 2^T \left(\sum_{i_1 \neq j_1, \dots, i_T \neq j_T} \prod_{k=1}^{T} \eta_{k,i_k}\eta_{k,j_k} \right) + Q.$$

where Q consists of terms that contain a factor of some Y_r. Due to the independence of the Rademachers, the expectation of any term that contains a factor of Y_r has expectation 0, which implies that:

$$\mathbb{E}[Z_1^2 \dots Z_T^2] = v^{2T} 2^T \mathbb{E}\left[\left(\sum_{i_1 \neq j_1, \dots, i_T \neq j_T} \prod_{k=1}^{T} \eta_{k,i_k}\eta_{k,j_k} \right) \right].$$

Let $\eta'_{r,i} \sim \eta_{r,i}$ be *independent* random variables. Suppose that

$$Z'_r := v^{2T} \sum_{i \neq j} \eta'_{r,i}\eta'_{r,j}\sigma_{r,i}\sigma_{r,j}.$$

We know that

$$Z_r'^2 = 2\left(\sum_{i\neq j} \eta'_{r,i}\eta'_{r,j}\right) + Y'_r,$$

where Y'_r has expectation 0. This means that:

$$Z_1'^2 \dots Z_T'^2 = v^{2T}2^T\left(\sum_{i_1\neq j_1,\dots,i_T\neq j_T}\prod_{k=1}^{T}\eta'_{k,i_k}\eta'_{k,j_k}\right) + Q'$$

where Q' consists of terms that contain a factor of some Y'_r. For similar reasons, this implies that

$$\mathbb{E}[Z_1'^2 \dots Z_T'^2] = v^{2T}2^T\mathbb{E}\left[\left(\sum_{i_1\neq j_1,\dots,i_T\neq j_T}\prod_{k=1}^{T}\eta'_{k,i_k}\eta'_{k,j_k}\right)\right].$$

Let's view $\prod_{k=1}^{T}\eta'_{k,i_k}\eta'_{k,j_k}$ and $\prod_{k=1}^{T}\eta_{k,i_k}\eta_{k,j_k}$ as terms in a sum. In the second expression, every term has expectation $\left(\frac{s}{m}\right)^{2T}$, and there are at most N^{2T} terms. In the first expression, if there are $> s$ copies of any i_k value, then the expectation is 0. Otherwise, the expectation varies between $C_2^{-2T}\left(\frac{s}{m}\right)^{2T}$ and $\left(\frac{s}{m}\right)^{2T}$. By the counting argument at the beginning of the proof, we know that there are at least $C_1^{2T}N^{2T}$ terms. This implies that

$$\|Z_1 \dots Z_T\|_2 \gtrsim C^T \|Z'_1 \dots Z'_T\|_2 = C^T \|Z'_1\|_2^T = C^T \|Z_1\|_2^T$$

as desired.

Now, we handle the case of the general q/T. Since we have a uniform sparse JL transform, we know that for $1 \leq x \leq s$:

$$\mathbb{E}[\eta_{1,1} \dots \eta_{x,1}] \geq \frac{s(s-1)\dots(s-x+1)}{(m)(m-1)\dots(m-x+1)} \geq C_2^{-x}\left(\frac{s}{m}\right)^x.$$

We know that

$$(Z_r)^{q/T} = 2^{q/T-1}\sum_{i\neq j}(\eta_{r,i}\eta_{r,j})^{q/T} + Y_r,$$

where Y_r has expectation ≥ 0. In this case we have that

$$(Z_1 \dots Z_T)^{q/T} = 2^{q-T}\left(\sum_{i_1\neq j_1,\dots,i_T\neq j_T}\prod_{k=1}^{T}(\eta_{k,i_k}\eta_{k,j_k})^{q/T}\right) + Q.$$

where Q has expectation ≥ 0. This implies that:

$$\mathbb{E}[Z_1^{q/T} \dots Z_T^{q/T}] \geq v^{2T}2^{q-T}\mathbb{E}\left[\left(\sum_{i_1\neq j_1,\dots,i_T\neq j_T}\prod_{k=1}^{T}(\eta_{k,i_k}\eta_{k,j_k})^{q/T}\right)\right].$$

Let $\eta'_{r,i} \sim \eta_{r,i}$ be *independent* random variables, and let $M'_r = \sum_{i=1}^{N}\eta'_{r,i}$. Suppose:

$$Z'_r := v^{2T}\sum_{i\neq j}\eta'_{r,i}\eta'_{r,j}\sigma_{r,i}\sigma_{r,j}.$$

We know that

$$(Z'_r I_{M'_r=2})^{q/T} = 2^{q/T-1} \sum_{i \neq j} \left(\eta'_{r,i}\eta'_{r,j} I_{M'_r=2}\right)^{q/T} + Y'_r,$$

where Y'_r has expectation 0. In this case we have that

$$(Z'_1 I_{M'_1=2} \dots Z'_T I_{M'_T=2})^{q/T} = 2^{q-T} \left(\sum_{i_1 \neq j_1, \dots, i_T \neq j_T} \prod_{k=1}^{T} \left(\eta'_{k,i_k}\eta'_{k,j_k} I_{M'_k=2}\right)^{q/T} \right) + Q'.$$

where Q' consists of terms that contain a factor of some Y'_r. For similar reasons to the above, we have that:

$$\mathbb{E}[Z_1'^q \dots Z_T'^q] = v^{2T} 2^{q-T} \mathbb{E}\left[\left(\sum_{i_1 \neq j_1, \dots, i_T \neq j_T} \prod_{k=1}^{T} \left(\eta'_{k,i_k}\eta'_{k,j_k} I_{M'_k=2}\right)^{q/T} \right) \right].$$

Let's view $\prod_{k=1}^{T} \left(\eta_{k,i_k}\eta_{k,j_k}\right)^{q/T}$ and $\prod_{k=1}^{T} \left(\eta'_{k,i_k}\eta'_{k,j_k} I_{M'_k=2}\right)^{q/T}$ as terms in a sum. In the second expression, every term has expectation $\leq \left(\frac{s}{m}\right)^{2T}$ (the indicator can only *reduce* the expectation), and there are at most N^{2T} terms. In the first expression, if there are $> s$ copies of any i_k value, then the expectation is 0. Otherwise, the expectation varies between $C_2^{-2T}\left(\frac{s}{m}\right)^{2T}$ and $\left(\frac{s}{m}\right)^{2T}$. By the counting argument, we know that there are at least $C_1^{-2T} N^{2T}$ terms. This implies that

$$\|Z_1 \dots Z_T\|_{q/T} \gtrsim C^T \left\|Z'_1 I_{M'_1=2} \dots Z'_T I_{M'_T=2}\right\|_{q/T} = C^T \left\|Z'_1 I_{M'_1}=2\right\|_{q/T}^{T} = C^T \|Z_1 I_{M_1=2}\|_{q/T}^{T}$$

as desired. □

A.6 Proof of Lemma 6.7 and Lemma 6.8

Recall that our proof of Theorem 3.2 requires cleaner bounds on moments of $\|R(x_1, \dots, x_n)\|_q$ that follow simplifying the bounds in Lemma 6.5 and Lemma 6.6 at the target values of v. The proofs of these lemmas boil down to function bounding and simplification.

A.6.1 Proof of Lemma 6.7

First, we show how Lemma 6.5 implies Lemma 6.7. The proof involves simplifying and bounding the function at the target v value.

Proof of Lemma 6.7. We plug $q = p$ into Lemma 6.5. We use this relaxed version of the bound: If $\frac{se}{mv^2} \geq q$, then $\|R(x_1, \dots, x_n)\|_q \lesssim \frac{\sqrt{q}}{\sqrt{m}}$. Otherwise, if there exists $C_2 q^3 m v^4 \geq s^2$, then

$$\|R(x_1, \dots, x_n)\|_q \lesssim \begin{cases} \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{C_2^{1/3} q^2 v^2}{s \ln^2(qmv^2/s)}\right) & \text{if } \ln(qmv^4/s^2) \leq 2 \\ \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \min\left(\frac{C_2^{1/3} q^2 v^2}{s\ln^2(qmv^2/s)}, \frac{q}{s\ln(m/s)}\right)\right) & \text{if } \ln(qmv^4/s^2) > 2 \end{cases}$$

Suppose that the absolute constant on the upper bounds is $\leq C'$. Let $C = \max(C', 1)$ (we take C to be the constant on the upper bounds). Let's take $C_{v,2} = \frac{0.25}{\sqrt{C}}$, $C_{v,1} = \min\left(\frac{0.1}{C^{3/2}}, C_{v,2}\right)$, $C_S = 4C$, $C_M = \max\left(e^{\frac{1}{C_{v,1}^2}}, 16C^2, e^{\frac{1}{C_{v,2}^2}}, e^2\right)$. For the remainder of the analysis, we assume that $m \geq C_M \epsilon^{-2} p$ and $m < 2\epsilon^{-2}\delta$.

First, observe $m \geq 16C^2\epsilon^{-2}p$ gives us that $C\frac{\sqrt{p}}{\sqrt{m}} \leq 0.25\epsilon$ regardless of v.

Now, let $f_1 = C_{v,1}\sqrt{\epsilon s}\frac{\ln(\frac{m\epsilon}{p})}{p}$, and $f_2 = \frac{C_{v,2}\sqrt{\epsilon s}\sqrt{\ln\frac{m\epsilon^2}{p}}}{\sqrt{p}}$.

First, let's analyze $v = f_2$. We show that $\ln(pmf_2^4/s^2) \geq 2$. Observe that $\ln(pmf_2^4/s^2) = \ln(C_{v,2}^4 \ln^2(m\epsilon^2/p)) + \ln(m\epsilon^2/p)$. Using the fact that $m \geq e^{\frac{1}{C_{v,2}^2}}\epsilon^{-2}p$, we see that

$$C_{v,2}^4 \ln^2(m\epsilon^2/p) \geq C_{v,2}^4 \frac{1}{C_{v,2}^4} = 1.$$

Now, since $m \geq e^2\epsilon^{-2}p$, this implies that

$$\ln(pmf_2^4/s^2) = \ln(C_{v,2}^4 \ln^2(m\epsilon^2/p)) + \ln(m\epsilon^2/p) \geq \ln(m\epsilon^2/p) \geq 2.$$

Moreover, we know that $p^3 m f_2^2 \geq s^2$, since $pmv^4 \geq e^2 s^2$. Now, we show that $C\frac{pf_2^2}{s\ln(pmf_2^4/s^2)} \leq 0.25\epsilon$. Let's observe that

$$\frac{Cpv^2}{s\ln(pmf_2^4/s^2)} \leq \frac{CC_{v,2}^2\epsilon\ln(\frac{m\epsilon^2}{p})}{\ln(\frac{m\epsilon^2}{p})} = CC_{v,2}^2\epsilon$$

Since $C_{v,2} = \frac{0.25}{\sqrt{C}}$, we get a bound of 0.25ϵ.

Now, we handle the case where $m \geq s \cdot e^{\frac{C_S p\epsilon^{-1}}{s}}$. We first show that $C\frac{p}{s\ln(m/s)} \leq 0.25\epsilon$. If $s \geq \Theta(\epsilon^{-1}\ln(1/\delta))$, using that $m \geq se$, this immediately follows from $\frac{p}{s\ln(m/s)} \leq \frac{p}{s} \leq 0.25\epsilon$. Otherwise, we need it to be true that $s\ln(m/s) \geq 4Cp\epsilon^{-1}$. This can be written as $\ln(m/s) \geq \frac{4Cp\epsilon^{-1}}{s}$. Since $C_S = 4C$, this can be written as: $m \geq s \cdot e^{\frac{C_S p\epsilon^{-1}}{s}}$, as desired. This, combined with the above analysis, implies that when $m \geq s \cdot e^{\frac{C_S p\epsilon^{-1}}{s}}$, taking $v = f_2$:

$$\begin{aligned}
\|R(x_1, \ldots, x_n)\|_q &\leq C\max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \min\left(\frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}, \frac{q}{s\ln(m/s)}\right)\right) \\
&\leq C\max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \frac{q}{s\ln(m/s)}\right) \\
&\leq 0.25\epsilon.
\end{aligned}$$

Now, we just need to handle the case where $m \leq s\cdot e^{C_S p\epsilon^{-1}/s}$, $m \geq \Theta(\epsilon^{-1}\ln(1/\delta))$, $m \geq se$. Such values only exist if $s \leq \Theta(\epsilon^{-1}\ln(1/\delta))$. Observe that we can set $C_2 = \frac{1}{C_{v,1}^4}$ and using the fact that $C_{v,1} \leq C_{v,2}$, we obtain that

$$\frac{C_2 p^3 m v^4}{s^2} \geq \frac{C_2 p^3 m}{s^2}\min\left(C_{v,1}, C_{v,2}\right)^4\left(\frac{\sqrt{\epsilon s}}{p}\right)^4 = C_2 C_{v,1}^4 \frac{m\epsilon^2}{p} \geq C_2 C_{v,1}^4.$$

Thus, this is lower bounded by 1 when $C_2 = \frac{1}{C_{v,1}^4}$.

First, we analyze the case of $v = f_1$. We show that $\frac{CC_2^{1/3}p^2v^2}{s\ln^2(pmv^2/s)} \leq 0.1\epsilon$. Observe that

$$\frac{CC_2^{1/3}p^2v^2}{s\ln^2(pmv^2/s)} = \frac{\epsilon CC_2^{1/3}C_{v,1}^2\ln^2(\frac{m\epsilon}{p})}{\ln^2(\frac{C_{v,1}^2 m\epsilon\ln^2(\frac{m\epsilon}{p})}{p})} = \frac{\epsilon CC_2^{1/3}C_{v,1}^2\ln^2(\frac{m\epsilon}{p})}{\left(\ln(\frac{m\epsilon}{p}) + \ln(C_{v,1}^2\ln^2(\frac{m\epsilon}{p}))\right)^2}.$$

Now, since $m \geq e^{1/C_{v,1}^2}\epsilon^{-2}p$, we know that $\ln(C_{v,1}^2\ln^2(\frac{m\epsilon}{p})) \geq 0$. Thus we can bound the above expression by:

$$\frac{\epsilon CC_{v,1}^{2/3}\ln^2(\frac{m\epsilon}{p})}{\ln^2(\frac{m\epsilon}{p})} = \epsilon CC_{v,1}^{2/3}\epsilon \leq 0.1\epsilon,$$

where the last inequality uses the fact that $C_{v,1} \leq \frac{0.1}{C^{3/2}}$.

Let's now consider how the term $\frac{pv^2}{s\ln(pmv^4/s^2)}$ how changes as a function of v. This term only arises in the bound if $\ln(pmv^4/s^2) \geq 2$. First, we show this is an increasing function of v. Let $w = pmv^4/s^2$. We see that $\frac{pv^2}{s\ln(pmv^4/s^2)} = \frac{s}{\sqrt{pm}}\frac{\sqrt{w}}{\ln w}$. We observe that this is an increasing function of w as long as $w \geq e^2$, which is exactly our restriction on w. Thus, $\frac{pv^2}{s\ln(pmv^4/s^2)}$ is an increasing function of v in this range.

Now, we consider how the $\frac{C_2^{1/3}p^2v^2}{s\ln^2(pmv^2/s)}$ term changes a function of v. This term only arises in the bound if $\ln(pmv^2/s) \geq 1$. First, we show that $f(v) \leq 2f(v')$ if $v \leq v'$. Let $w = pmv^2/s$. We see that $\frac{p^2v^2}{s\ln(pmv^2/s)} = \frac{p}{m}\frac{w}{\ln^2 w}$. We observe that this is an increasing function of w as long as $w \geq e^2$. When $e \leq w \leq e^2$, observe that this is bounded by at most a factor of 2 above any other w value.

Now, for the remainder of the analysis, let $v = \min(f_1, f_2)$. We show that $\|R(x_1, \ldots, x_n)\|_q \leq 0.25\epsilon$.

If $\ln(pmv^2/s) \leq 1$ (i.e. $\frac{se}{mv^2} \geq p$), then we know that the bound is actually $\frac{\sqrt{p}}{\sqrt{m}}$, and we've already shown that $\|R(x_1, \ldots, x_n)\|_q \leq 0.25\epsilon$.

For the remainder of the analysis, we assume that $\ln(pmv^2/s) > 1$.

First, suppose that $v = f_1$. If $\ln(pmv^4/s^2) \leq 2$, then we know that

$$\|R(x_1, \ldots, x_n)\|_q \leq C\max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}\right) \leq 0.25\epsilon.$$

Otherwise, we know that $\ln(pmv^4/s^2) > 2$. First let's show that that $C\frac{pv^2}{s\ln(pmv^4/s^2)} \leq 0.25\epsilon$. We know that $v \leq f_2$. At $v = f_2$, we know that the expression is upper bounded by 0.25ϵ. Since the $\frac{pv^2}{s\ln(pmv^4/s^2)}$ term is an increasing function of v in this regime, this means that we get a bound of 0.25ϵ in this case too. Thus, we know that:

$$\begin{aligned}\|R(x_1, \ldots, x_n)\|_q &\leq C\max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \min\left(\frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}, \frac{q}{s\ln(m/s)}\right)\right) \\ &\leq C\max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}\right)\end{aligned}$$

$$\leq 0.25\epsilon.$$

Now, suppose that $v = f_2$. We've already shown that $\ln(pmv^4/s^2) \geq 2$ here (near the beginning of the proof). Since $v \leq f_1$, we obtain a bound of $2 \cdot 0.1\epsilon = 0.2\epsilon$. This means:

$$\begin{aligned}\|R(x_1, \ldots, x_n)\|_q &\leq C \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \min\left(\frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}, \frac{q}{s\ln(m/s)}\right)\right) \\ &\leq C \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{qv^2}{s\ln(qmv^4/s^2)}, \frac{C_2^{1/3}q^2v^2}{s\ln^2(qmv^2/s)}\right) \\ &\leq 0.25\epsilon.\end{aligned}$$

□

A.6.2 Proof of Lemma 6.8

Now, we show how Lemma 6.5 and Lemma 6.6 imply Lemma 6.8. The proof simply involves bounding and simplifying the functions in the original lemmas at the target v value.

Proof of Lemma 6.8. We use Lemma 6.5 but put in an absolute constant. Let $D_2 > 0$ be such that: if $\frac{se}{mv^2} \geq q$, then

$$\|R(x_1, \ldots, x_n)\|_q \leq D_2\frac{\sqrt{q}}{\sqrt{m}}.$$

Otherwise, if $q^3mv^4 \geq s^2$, then $\|R(x_1, \ldots, x_n)\|_q$ is upper bounded by:

$$D_2\begin{cases}\max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{q^2v^2}{s\ln^2(qmv^2/s)}\right) & \text{if } \ln(qmv^4/s^2) \leq 2, \ln(qmv^2/s) \leq q \\ \frac{\sqrt{q}}{\sqrt{m}} & \text{if } \ln(qmv^4/s^2) \leq 2, \ln(qmv^2/s) > q \\ \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{4096qv^2}{s\ln(qmv^4/s^2)}, \min\left(\frac{q^2v^2}{s\ln^2(qmv^2/s)}, \frac{q}{s\ln(m/s)}\right)\right) & \text{if } \ln(qmv^4/s^2) > 2, \ln(qmv^2/s) \leq q \\ \max\left(\frac{\sqrt{q}}{\sqrt{m}}, \frac{4096qv^2}{s\ln(qmv^4/s^2)}\right) & \text{if } \ln(qmv^4/s^2) > 2, \ln(qmv^2/s) > q.\end{cases}$$

We use Lemma 6.6 but put in an absolute constant $D_1 > 0$ (which we take to be ≤ 1). Let $2 \leq q \leq m$ be an even integer, and suppose that $0 < v \leq 0.5$ and $\frac{1}{v^2}$ is an even integer. If $qv^2 \leq s$, then

$$\|R(x_1, \ldots, x_n)\|_q \geq D_1\frac{\sqrt{q}}{\sqrt{m}}.$$

If $m \geq q$, $2 \leq \ln(qmv^4/s^2) \leq q$, $2qv^2 \leq 0.5s\ln(qmv^4/s^2)$, and $s \leq m/e$ then:

$$\|R(x_1, \ldots, x_n)\|_q \geq D_1\frac{4096qv^2}{s\ln(qmv^4/s^2)}.$$

If $v \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{q}}$ and $1 \leq \ln(qmv^2/s) \leq q/2$, and $s \leq m/e$, then:

$$\|R(x_1, \ldots, x_n)\|_q \geq D_1\frac{q^2v^2}{s\ln^2(qmv^2/s)}.$$

Let $D = \frac{D_1}{2048 D_2}$. It suffices to show that for v defined in the lemma statement:

$$\|R(v, v, \ldots, v, 0, \ldots, 0)\|_q \geq 2\epsilon$$

and

$$\frac{\|R(v, v, \ldots, v, 0, \ldots, 0)\|_q}{\|R(v, v, \ldots, v, 0, \ldots, 0)\|_{2q}} \geq D.$$

First, we handle the case where $m \leq \Theta(\epsilon^{-2} \ln(1/\delta))$. Let's take $v = \psi$ for any sufficiently small ψ. By sufficiently small, we mean $v^2 \leq \frac{se}{2mq}$ and $0 < v \leq 0.5$. This implies that $\frac{se}{mv^2} \geq 2q$ and $qv^2 \leq s$. Thus we know (using that $q \leq m$) that $\|R(x_1, \ldots, x_n)\|_{2q} \leq D_2 \frac{\sqrt{2q}}{\sqrt{m}}$ and $\|R(x_1, \ldots, x_n)\|_q \geq D_1 \frac{\sqrt{q}}{\sqrt{m}}$. This means that:

$$\frac{\|R(v, v, \ldots, v, 0, \ldots, 0)\|_q}{\|R(v, v, \ldots, v, 0, \ldots, 0)\|_{2q}} \geq D$$

as desired. Suppose that $m \leq \Theta(\epsilon^{-2} \ln(1/\delta))$. Based on the setting q, this means that $\|R(v, \ldots, v, 0 \ldots, 0)\|_q \geq D_1 \frac{\sqrt{q}}{\sqrt{m}} \geq 2\epsilon$ as desired.

Now, we handle the cases where $m \geq \Theta(\epsilon^{-2} \ln(1/\delta))$. Notice that the condition $f'(m, \epsilon, \delta, s) \leq 0.5$ allows us to assume that $s \leq \Theta(\epsilon^{-1} \ln(1/\delta))$ and $m \leq \epsilon^{-2} e^p$. Let $f_1 = 4\sqrt{\epsilon s} \frac{\ln(\frac{m\epsilon}{q})}{q}$ and let $f_2 = \sqrt{\epsilon s} \frac{\sqrt{\ln(\frac{m\epsilon^2}{q})}}{\sqrt{q}}$. We will consider $v = C_{v,1} f_1 =: v_1$ and $v = C_{v,2} f_2 =: v_2$. First, we handle the condition of $q^3 m v^4 \geq s^2$. We enforce the condition $C_{v,1}, C_{v,2} \geq 1$. Assuming that $v \geq \frac{\sqrt{\epsilon s}}{q}$ (which is true at the two values of v that we consider), we know $\frac{q^3 m v^4}{s^2} \geq \frac{m\epsilon^2}{q} \geq 1$. Also, we make $m \geq 2C^2 \epsilon^{-2} q$, so that $\sqrt{\frac{2q}{m}} \leq \sqrt{\frac{2q}{2C^2 \epsilon^{-2} q}} = \frac{\epsilon}{C}$.

Consider $v = v_2$. We first check that the conditions for the upper bound are satisfied. We have that $\frac{qmv^4}{s^2} = C_{v,2}^4 \frac{m\epsilon^2}{q} \ln^2(\frac{m\epsilon^2}{q})$. Observe that when $m \geq e^2 \epsilon^{-2} q$ and $C_{v,2} \geq 1$, this is lower bounded by e^2, so $\ln(qmv^4/s^2) \geq 2$. Also, we have that $\frac{qmv^2}{se} = \sqrt{qm} \sqrt{\frac{qmv^4}{s^2}} \frac{1}{e} \geq 1$. Now, we check the additional conditions needed for the lower bound. Observe that

$$\frac{2qv^2}{s} = 2\epsilon C_{v,2}^2 \ln(\frac{m\epsilon^2}{q}) \leq 0.5 C_{v,2}^4 \frac{m\epsilon^2}{q} \ln^2(\frac{m\epsilon^2}{q}) = 0.5 \ln(qmv^4/s^2)$$

as desired. We check that $\ln(qmv^4/s^2) \leq q$. It suffices to show that

$$\frac{m\epsilon^2}{q} \ln^2\left(\frac{m\epsilon^2}{q}\right) \leq \frac{e^q}{C_{v,2}^4}.$$

Using the condition that $m \leq \epsilon^{-2} \frac{e^q}{q C_{v,2}^4}$ where we obtain that

$$\frac{m\epsilon^2}{q} \ln^2(\frac{m\epsilon^2}{q}) \leq \frac{e^q}{q^2 C_{v,2}^4} \ln^2(\frac{e^q}{q^2 C_{v,2}^4}) \leq \frac{e^q}{q^2 C_{v,2}^4} \ln^2(e^q) \leq \frac{e^q}{C_{v,2}^4}$$

as desired. Now, we compute the value of $\frac{qv^2}{s \ln(qmv^4/s^2)}$ at $v = C_{v,2} f_2$. We obtain:

$$\frac{qv^2}{s \ln(qmv^4/s^2)} = C_{v,2}^2 \epsilon \frac{\ln(m\epsilon^2/q)}{\ln\left(\frac{m\epsilon^2}{q}\right) + \ln\left(C_{v,2}^4 \ln^2\left(\frac{m\epsilon^2}{q}\right)\right)}.$$

Consider $v = v_1$. We first check that the conditions for the upper bound are satisfied. In this case, we have that $\frac{qmv^2}{s} = 16C_{v,1}^2 \frac{m\epsilon}{q} \ln^2(\frac{m\epsilon}{q})$. Observe that when $C_{v,1} \geq 1$ and $m \geq e^2\epsilon^{-2}q \geq e^2\epsilon^{-1}q$, this is lower bounded by e^2, so $\ln(qmv^2/s) \geq 2$. Now, we claim that when $f_1 \leq f_2$, we show that $\ln(qmv^2/s) \leq q/2$. In this case, using that $m \leq \epsilon^{-2}qe^q$, we have: $\frac{4\ln(m\epsilon/q)}{q} \leq \frac{\sqrt{\ln(m\epsilon^2/q)}}{\sqrt{q}}$. This means that $\ln(m\epsilon/q) \leq \sqrt{q}\sqrt{\ln(m\epsilon^2/q)}/4 \leq q/4$. Observe that

$$\begin{aligned}\ln(qmv^2/s) &= \ln(16C_{v,1}^2) + \ln(m\epsilon/q) + 2\ln\ln(m\epsilon/q)\\ &\leq \ln(16C_v^2) + \frac{q}{4} + 2\ln\ln q\\ &\leq \frac{q}{2}.\end{aligned}$$

At this value, observe that:

$$\frac{q^2v^2}{s\ln^2(qmv^2/s)} = 16C_{v,1}^2\epsilon\left(\frac{\ln(m\epsilon/q)}{\ln\left(\frac{m\epsilon}{q}\right) + \ln\left(16C_{v,1}^2\ln^2\left(\frac{m\epsilon}{q}\right)\right)}\right)^2.$$

Let $C = D_1$. Let's set $\sqrt{\frac{1}{C}} \leq C_{v,2} = C_{v,1} = C_v \leq 4\sqrt{\frac{1}{C}}$. Using the fact that $v^2 \leq 0.5$ (so $\frac{1}{v^2} \geq 2$), this means that $\frac{1}{v^2}$ has can take on at least 3 different powers of 2. Let's observe that when $16C_v^2\ln^2(\frac{m\epsilon}{q}) \leq \frac{m\epsilon}{q}$ (we can get this condition by saying that $m \geq C_{M,2}\epsilon^{-2}q$ for a sufficiently large $C_{M,2}$) and $16C_v^2\ln^2(m\epsilon/q) \geq 1$ (we can get this condition by saying that $m \geq C_{M,2}\epsilon^{-2}q$ for a sufficiently large $C_{M,2}$), we know that

$$\frac{4\epsilon}{C} \leq 4C_v^2\epsilon \leq \frac{q^2v_1^2}{s\ln^2(qmv_1^2/s)} \leq 16C_v^2\epsilon \leq \frac{256\epsilon}{C}.$$

Suppose that $C_v^4\ln^2(m\epsilon^2/q) \leq \frac{m\epsilon^2}{q}$ (we can get this condition by saying that $m \geq C_{M,2}\epsilon^{-2}q$ for a sufficiently large $C_{M,2}$) and $C_v^4\ln^2(m\epsilon^2/q) \geq 1$ (we can get this condition by saying that $m \geq C_{M,2}\epsilon^{-2}q$ for a sufficiently large $C_{M,2}$). Let's observe that

$$4096C_v^2\epsilon \geq \frac{4096qv_2^2}{s\ln^2(qmv_2^4/s^2)} \geq 2048C_v^2\epsilon \geq \frac{2048\epsilon}{C}.$$

Let $m' = s \cdot e^{\frac{C\epsilon^{-1}q}{1024s}}$. When $m \geq m'$, we know that $\frac{q}{s\ln(m/s)} \leq \frac{1024\epsilon}{C}$ and when $m \leq m'$, we know that $\frac{q}{s\ln(m/s)} \geq \frac{1024\epsilon}{C}$.

In order to plug in $v = v_1$ and use the $\frac{q^2v^2}{s\ln^2(qmv^2/s)}$ lower bound, we need to show that $v \leq \frac{\sqrt{\ln(m/s)}}{\sqrt{q}}$. At $v = \frac{\sqrt{\ln(m/s)}}{\sqrt{q}}$, we have that $\frac{qmv^2}{s} = \frac{m}{s}\ln\left(\frac{m}{s}\right)$. Observe that when $m \geq e^2s$, this is lower bounded by e^2, so $\ln(qmv^2/s) \geq 2$. At this value, observe that:

$$\frac{q^2v^2}{s\ln^2(qmv^2/s)} = \frac{q\ln(m/s)}{s\ln^2\left(\frac{m}{s}\ln\left(\frac{m}{s}\right)\right)} \geq \frac{q\ln(m/s)}{4s\ln^2\left(\frac{m}{s}\right)} = \frac{q}{4s\ln\left(\frac{m}{s}\right)}.$$

We can write $\frac{q^2v^2}{s\ln^2(qmv^2/s)} = \frac{q}{m}\frac{w}{\ln^2 w}$, where $w = qmv^2/s$. We observe that this is an increasing function of w as long as $w \geq e^2$. Thus, it suffices to show that $\frac{q^2v_1^2}{s\ln^2(qmv_1^2/s)} \leq \frac{q^2v^2}{s\ln^2(qmv^2/s)}$. When $m \leq m'$, we know that

$$\frac{q^2v_1^2}{s\ln^2(qmv_1^2/s)} \leq \frac{256\epsilon}{C} \leq \frac{q}{4s\ln\left(\frac{m}{s}\right)} \leq \frac{q^2v^2}{s\ln^2(qmv^2/s)}.$$

Thus, we have that $v_1 \leq v = \frac{\sqrt{\ln(m/s)}}{\sqrt{q}}$ as desired.

The first case is $m \leq m'$ and $f_2 \leq f_1$. We set $v = C_v f_2$.

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \geq D_1\frac{4096qv^2}{s\ln(qmv^4/s)}.$$

For the upper bound, we see that $\ln(2qmv^4/s^2) > \ln(qmv^4/s^2) \geq 2$ and $\sqrt{\frac{2q}{m}} \leq \frac{\epsilon}{C}$. Here, we have that

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq \begin{cases} D_2\max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{8192qv^2}{s\ln(2qmv^4/s)}, \frac{4q^2v^2}{s\ln^2(2qmv^2/s)}\right) & \text{if } \ln(2qmv^2/s) \leq 2q \\ D_2\max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{8192qv^2}{s\ln(qmv^4/s)}\right) & \text{if } \ln(2qmv^2/s) > 2q \end{cases}.$$

Now, we use the fact that $v \leq C_v f_1 := v_1$ to see that:

$$\frac{4q^2v^2}{s\ln(2qmv^2/s)} \leq \frac{4q^2v^2}{s\ln(qmv^2/s)} \leq \frac{8q^2v_1^2}{s\ln(qmv_1^2/s)} \leq \frac{2048\epsilon}{C}.$$

We also observe that since $2qmv^4/s \leq (qmv^4/s)^2$, we know:

$$\frac{8192qv^2}{s\ln(2qmv^4/s)} \geq \frac{8192qv^2}{2s\ln(qmv^4/s)} \geq \frac{2048\epsilon}{C}.$$

This, coupled with the guarantee on $\frac{\sqrt{2q}}{\sqrt{m}}$, implies we have an upper bound of:

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq D_2\frac{8192qv^2}{s\ln(2qmv^4/s)}.$$

Thus, we have that

$$\frac{\|R(v,\ldots,v,0\ldots,0)\|_q}{\|R(v,\ldots,v,0,\ldots,0)\|_{2q}} \geq \frac{D_1}{2D_2} \geq D.$$

Moreover, we have that

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \geq D_1 \cdot \frac{4096qv^2}{s\ln(qmv^4/s)} \geq D_1\frac{2048\epsilon}{C} = 2048\epsilon$$

The next case is $f_1 \leq f_2$ and $m \leq m'$. We set $v = v_1$. Since $f_1 \leq f_2$, we know that $\ln(qmv^4/s^2) \leq q$. Thus we know:

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \geq \begin{cases} D_1\max\left(\frac{4096qv^2}{s\ln(qmv^4/s)}, \frac{q^2v^2}{s\ln^2(qmv^2/s)}\right) & \text{if } \ln(qmv^4/s^2) \geq 2, qv^2 \leq s \\ D_1\frac{q^2v^2}{s\ln^2(qmv^2/s)} & \text{else} \end{cases}.$$

For the upper bound, we know that:

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq \begin{cases} D_2 \max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{8192qv^2}{s\ln(2qmv^4/s)}, \frac{4q^2v^2}{s\ln^2(2qmv^2/s)}\right) & \text{if } \ln(2qmv^4/s) > 2 \\ D_2 \max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{4q^2v^2}{s\ln^2(2qmv^2/s)}\right) & \text{if } \ln(2qmv^4/s) \leq 2 \end{cases}.$$

To make these bounds compatible, we need to handle the case where $\ln(qmv^4/s) \geq 2$, $qv^2 \geq s$ better. Let $v' = C_v f_2$. Assuming that $\ln(qmv^4/s) \geq 2$, we know that $\frac{8192qv^2}{s\ln(2qmv^4/s)}$ can be upper bounded by:

$$\frac{8192qv^2}{s\ln(qmv^4/s)} \leq \frac{8192qv'^2}{s\ln(qmv'^4/s)} = \frac{8192C_v^2\epsilon\ln(m\epsilon^2/q)}{\ln(m\epsilon^2/q) + \ln(C_v^4\ln^2(m\epsilon^2/q))} \leq 8192C_v^2\epsilon \leq \frac{8192\epsilon}{C}$$

as long as $\ln^2(m\epsilon/q)C_v^4 \geq 1$ (which we can make true by appropriately setting the constants on the bound for m). Observe also that:

$$\frac{4q^2v^2}{s\ln^2(2qmv^2/s)} \geq \frac{q^2v^2}{s\ln^2(qmv^2/s)} \geq \frac{4\epsilon}{C}.$$

Thus:

$$\frac{8192qv^2}{s\ln(qmv^4/s)} \leq \frac{8192q^2v^2}{s\ln^2(2qmv^2/s)}.$$

This, coupled with the guarantee on $\frac{\sqrt{2q}}{\sqrt{m}}$, implies that our upper bound becomes:

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq \begin{cases} D_2\frac{8192q^2v^2}{s\ln^2(2qmv^2/s)} & \text{if } \ln(2qmv^4/s) \leq 2 \\ D_2\frac{8192q^2v^2}{s\ln^2(2qmv^2/s)} & \text{if } \ln(qmv^4/s) \geq 2, qv^2 \geq s \\ D_2\max\left(\frac{8192qv^2}{s\ln(2qmv^4/s)}, \frac{4q^2v^2}{s\ln^2(2qmv^2/s)}\right) & \text{else.} \end{cases}$$

We now show that we can tweak C_v within the factor of $2^{1/4}$ range permitted to show that we can ensure that it is not true that $2 - \ln 2 < \ln(qmv^4/s) \leq 2$. Observe that multiplying by a factor of $2^{1/4}$ in this case yields $\ln(2qmv^4/s) > 2$ and dividing by a factor of $2^{1/4}$ yields $\ln(qmv^4/s) \leq 2 - \ln 2$. Thus, at least one of the C_v values that yields a power of 2 for $\frac{1}{v^2}$ will work. Thus, we have that

$$\frac{\|R(v,\ldots,v,0\ldots,0)\|_q}{\|R(v,\ldots,v,0,\ldots,0)\|_{2q}} \geq \frac{D_1}{8192D_2} = \frac{D}{2048}.$$

Moreover, we have that:

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \geq D_1 \cdot \frac{q^2v^2}{s\ln^2(qmv^2/s)} \geq D_1\frac{4\epsilon}{C} = 4\epsilon$$

The next case is that $m > m'$. We set $v = C_v\sqrt{\epsilon}s\frac{\sqrt{\ln\left(\frac{m\epsilon^2}{q}\right)}}{\sqrt{q}}$. We know:

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \geq D_1\frac{4096qv^2}{s\ln(qmv^4/s)}.$$

For the upper bound, we see that $\ln(2qmv^4/s^2) > \ln(qmv^4/s^2) > 2$. We know:

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq \begin{cases} D_2 \max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{8192qv^2}{s\ln(2qmv^4/s)}, \frac{2q}{s\ln(m/s)}\right) & \text{if } \ln(2qmv^2/s) \leq 2q \\ D_2 \max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{8192qv^2}{s\ln(2qmv^4/s)}\right) & \text{if } \ln(2qmv^2/s) > 2q \end{cases}.$$

This can be relaxed to:

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq D_2 \max\left(\frac{\sqrt{2q}}{\sqrt{m}}, \frac{8192qv^2}{s\ln(2qmv^4/s)}, \frac{2q}{s\ln(m/s)}\right).$$

Now, we know that

$$\frac{2q}{s\ln(m/s)} \leq \frac{2048\epsilon}{C} \leq \frac{4096qv^2}{s\ln(qmv^4/s)} = \frac{8192qv^2}{2s\ln(qmv^4/s)} \leq \frac{8192qv^2}{s\ln(2qmv^4/s)}.$$

This coupled with what we know about $\frac{\sqrt{2q}}{\sqrt{m}}$ means that:

$$\|R(v,\ldots,v,0,\ldots,0)\|_{2q} \leq D_2 \frac{8192qv^2}{s\ln(2qmv^4/s)}.$$

Thus, we have that

$$\frac{\|R(v,\ldots,v,0\ldots,0)\|_q}{\|R(v,\ldots,v,0,\ldots,0)\|_{2q}} \geq \frac{D_1}{2D_2} \geq D.$$

Moreover, we have that

$$\|R(v,\ldots,v,0,\ldots,0)\|_q \geq D_1 \cdot \frac{4096qv^2}{s\ln(qmv^4/s)} \geq D_1 \frac{2048\epsilon}{C} = 2048\epsilon.$$

We use the condition on q not being more than a constant factor away from $p = \ln(1/\delta)$, to conclude that $\epsilon^{-2}q = \Theta(\epsilon^{-2}p)$, $f_2 = \Theta\left(\sqrt{\epsilon s}\frac{\sqrt{\ln\left(\frac{m\epsilon^2}{p}\right)}}{\sqrt{p}}\right)$, and $f_1 = \Theta\left(\sqrt{\epsilon s}\frac{\ln\left(\frac{m\epsilon}{p}\right)}{p}\right)$, and to conclude that the boundaries move within the Θ notation as well. □

A.7 Additional experimental results and discussion

All of the experiments (in Chapter 3 and the Appendix) were run on the default hardware on a Google Colab notebook. The code is available at `https://github.com/mjagadeesan/sparsejl-featurehashing`.

First, we give the results of additional experimental results on real-world and synthetic datasets, using the same experimental setup as Chapter 3.

For the synthetic datasets, the trends in Figure A.1 and Figure A.2 look quite similar to the figures in Chapter 3. We see, though, that Figure A.2 experiences more severe non-monotonic behavior as a function of s in the second phase transition. Consider, for example, in Figure A.2, the behavior at $m = 12000$: we see that $\hat{v}(m,\epsilon,\delta,4) < \hat{v}(m,\epsilon,\delta,3)$. In fact,

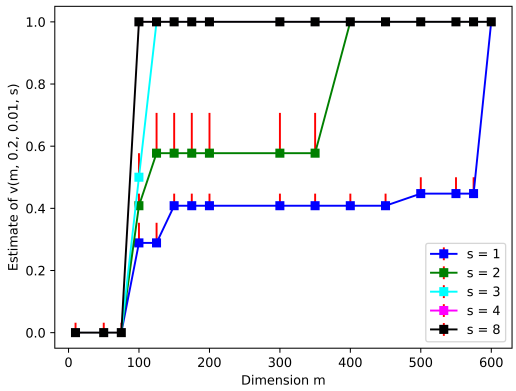

Figure A.1: Phase transitions of $\hat{v}(m, 0.2, 0.01, s)$

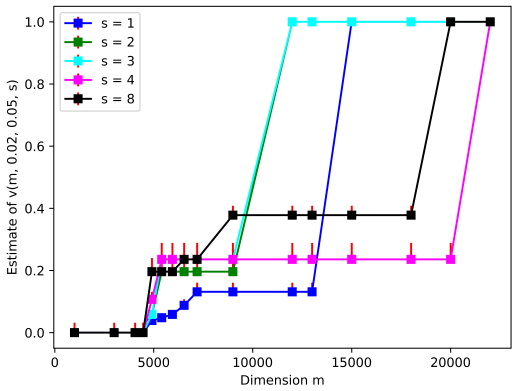

Figure A.2: Phase transitions of $\hat{v}(m, 0.02, 0.05, s)$

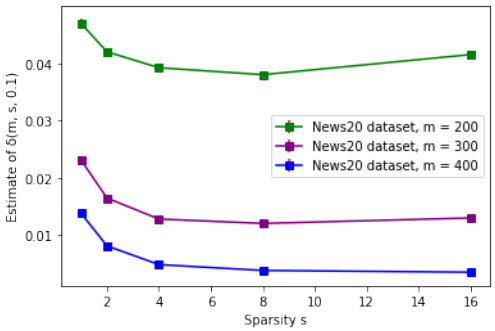

Figure A.3: $\hat{\delta}(m, s, 0.1)$ on News20

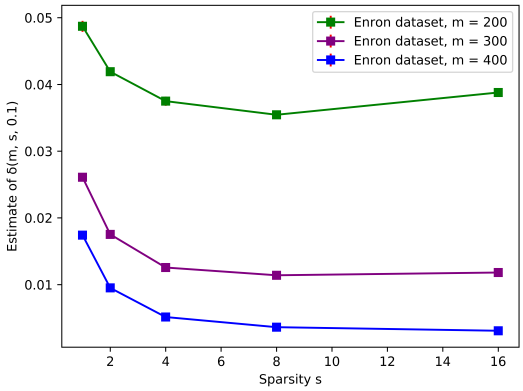

Figure A.4: $\hat{\delta}(m, s, 0.1)$ on Enron

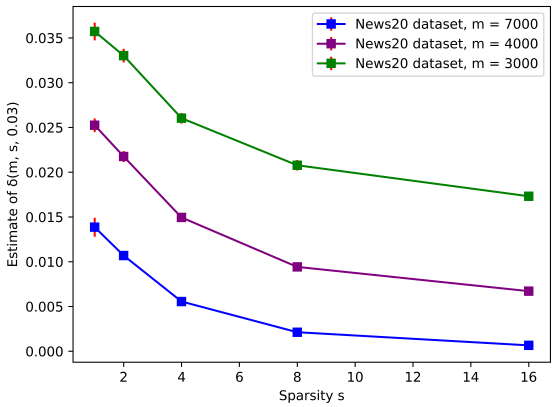

Figure A.5: $\hat{\delta}(m, s, 0.03)$ on News20

the order of the phase transitions in Figure A.2 is far from decreasing. Nonetheless, the general patterns and trends in the theoretical result still hold (e.g. the "flat" part occurs at a lower y-coordinate for lower s values.)

For the real-world datasets, the trends in Figure A.3, Figure A.4, and Figure A.5 look quite similar to the figures in Chapter 3. One slight difference is that the failure probability noticeably increases in Figure A.3 and Figure A.4 between $s = 8$ and $s = 16$. It turns out that the failure probability actually increases to a local maximum somewhere in $12 \leq s \leq 16$, and then decreases when $s \geq 16$, reaching lower than the value at $s = 8$ by the time $s = 20$. There turns out to be a similar local maximum phenomenon when $\epsilon = 0.07$ and $m = 500$, though the local maximum occurs in $24 \leq s \leq 32$ and thus is not as visible in the graph.

As a general comment on non-monotonicity as a function of s, we emphasize that our asymptotic theoretical results characterize the *macroscopic* behavior of $v(m, \epsilon, \delta, s)$, and do not preclude the existence of constant factor fluctuations for small changes in parameters. An interesting direction for future work would be to look further into this non-mononocity and try to characterize when it arises.