Fully Understanding the Hashing Trick

Casper Benjamin Freksen* Lior Kamma* Kasper Green Larsen †

May 22, 2018

Abstract

Feature hashing, also known as *the hashing trick*, introduced by Weinberger *et al.* (2009), is one of the key techniques used in scaling-up machine learning algorithms. Loosely speaking, feature hashing uses a random sparse projection matrix $A : \mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) in order to reduce the dimension of the data from n to m while approximately preserving the Euclidean norm. Every column of A contains exactly one non-zero entry, equals to either -1 or 1.

Weinberger *et al.* showed tail bounds on $\|Ax\|_2^2$. Specifically they showed that for every ε, δ, if $\|x\|_\infty / \|x\|_2$ is sufficiently small, and m is sufficiently large, then

$$\Pr[\ |\ \|Ax\|_2^2 - \|x\|_2^2\ | < \varepsilon \|x\|_2^2\ ] \geq 1 - \delta\ .$$

These bounds were later extended by Dasgupta *et al.* (2010) and most recently refined by Dahlgaard *et al.* (2017), however, the true nature of the performance of this key technique, and specifically the correct tradeoff between the pivotal parameters $\|x\|_\infty / \|x\|_2, m, \varepsilon, \delta$ remained an open question.

We settle this question by giving tight asymptotic bounds on the exact tradeoff between the central parameters, thus providing a complete understanding of the performance of feature hashing. We complement the asymptotic bound with empirical data, which shows that the constants "hiding" in the asymptotic notation are, in fact, very close to 1, thus further illustrating the tightness of the presented bounds in practice.

1 Introduction

Dimensionality reduction that approximately preserves Euclidean distances is a key tool used as a preprocessing step in many geometric, algebraic and classification algorithms, whose performance heavily depends on the dimension of the input. Loosely speaking, a distance-preserving dimensionality reduction is an (often random) embedding of a high-dimensional Euclidean space into a space of low dimension, such that the distance between every two points is approximately preserved (with high probability). Its applications range upon nearest neighbor search [AC09, HIM12], classification and regression [RR08, MM09, PBMID14], manifold learning [HWB08] sparse recovery [CT06] and numerical linear algebra [CW09, MM13, Sár06]. For more applications see, e.g. [Vem05].

One of the most fundamental results in the field was presented in the seminal paper by Johnson and Lindenstrauss [JL84].

*Computer Science Department. Aarhus University. Supported by a Villum Young Investigator Grant. {cfreksen, lior.kamma}@cs.au.dk.

†Computer Science Department. Aarhus University. Supported by a Villum Young Investigator Grant and an AUFF Starting Grant. larsen@cs.au.dk.

Lemma 1 (Distributional JL Lemma)**.** *For every $n \in \mathbb{N}$ and $\varepsilon, \delta \in (0,1)$, there exists a random $m \times n$ projection matrix A, where $m = \Theta(\varepsilon^{-2} \lg \frac{1}{\delta})$ such that for every $x \in \mathbb{R}^n$*

$$\Pr[\ |\ \|Ax\|_2^2 - \|x\|_2^2\ | < \varepsilon \|x\|_2^2\ ] \geq 1 - \delta \tag{1}$$

The target dimension m in the lemma is known to be optimal [JW13, LN17].

Running Time Performances. Perhaps the most common proof of the lemma (see, e.g. [DG03, Mat08]) samples a projection matrix by independently sampling each entry from a standard Gaussian (or Rademacher) distribution. Such matrices are by nature very dense, and thus a naïve embedding runs in $O(m\|x\|_0)$ time, where $\|x\|_0$ is the number of non-zero entries of x.

Due to the algorithmic significance of the lemma, much effort was invested in finding techniques to accelerate the embedding time. One fruitful approach for accomplishing this goal is to consider a distribution over *sparse* projection matrices. This line of work was initiated by Achlioptas [Ach03], who constructed a distribution over matrices, in which the *expected* fraction of non-zero entries is at most one third, while maintaining the target dimension. The best result to date in constructing a sparse Johnson-Lindenstrauss matrix is due to Kane and Nelson [KN14], who presented a distribution over matrices satisfying (1) in which every column has at most $s = O(\varepsilon^{-1} \lg(1/\delta))$ non-zero entries. Conversely Nelson and Nguyễn [NN13] showed that this is almost asymptotically optimal. That is, every distribution over $n \times m$ matrices satisfying (1) with $m = \Theta(\varepsilon^{-2} \lg(1/\delta))$, and such that every column has at most s non-zero entries must satisfy $s = \Omega((\varepsilon \lg(1/\varepsilon))^{-1} \lg(1/\delta))$.

While the bound presented by Nelson and Nguyễn is theoretically tight, we can provably still do much better in practice. Specifically, the lower bound is attained on vectors $x \in \mathbb{R}^n$ for which, loosely speaking, the "mass" of x is concentrated in few entries. Formally, the ratio $\|x\|_\infty / \|x\|_2$ is large. However, in practical scenarios, such as the term frequency - inverse document frequency representation of a document, we may often assume that the mass of x is "well-distributed" over many entries (That is, $\|x\|_\infty / \|x\|_2$ is small). In these common scenarios projection matrices which are significantly sparser turn out to be very effective.

Feature Hashing. In the pursuit for sparse projection matrices, Weinberger *et al.* [WDL+09] introduced dimensionality reduction via *Feature Hashing*, in which the projection matrix A is, in a sense, as sparse as possible. That is, every column of A contains exactly one non-zero entry, randomly chosen from $\{-1, 1\}$. This technique is one of the most influencial mathematical tools in the study of scaling-up machine learning algorithms, mainly due to its simplicity and good performance in practice [Dal13, Sut15]. More formally, for $n, m \in \mathbb{N}^+$, the projection matrix A is sampled as follows. Sample $h \in_R [n] \to [m]$, and $\sigma = \langle \sigma_j \rangle_{j \in [n]} \in_R \{-1, 1\}^n$ independently. For every $i \in [m], j \in [n]$, let $a_{ij} = a_{ij}(h, \sigma) := \sigma_j \cdot \mathbb{1}_{h(j)=i}$ (that is, $a_{ij} = \sigma_j$ iff $h(j) = i$ and 0 otherwise). Weinberger *et al.* additionally showed exponential tail bounds on $\|Ax\|_2^2$ when the ratio $\|x\|_\infty / \|x\|_2$ is sufficiently small, and m is sufficiently large. These bounds were later improved by Dasgupta *et al.* [DKS10] and most recently by Dahlgaard, Knudsen and Thorup [DKT17] improved these concentration bounds. Conversely, a result by Kane and Nelson [KN14] implies that if we allow $\|x\|_\infty / \|x\|_2$ to be too large, then there exist vectors for which (1) does not holds.

Finding the correct tradeoffs between $\|x\|_\infty / \|x\|_2$, and m, ε, δ in which feature hashing performs well remained an open problem. Our main contribution is settling this problem, and providing a complete and comprehensive understanding of the performance of feature hashing.

1.1 Main results

The main result of this paper is a tight tradeoff between the target dimension m, the approximation ratio ε, the error probability δ and $\|x\|_\infty/\|x\|_2$. More formally, let $\varepsilon, \delta > 0$ and $m \in \mathbb{N}^+$. Let $\nu(m, \varepsilon, \delta)$ be the maximum $\nu \in [0, 1]$ such that for every $x \in \mathbb{R}^n$, if $\|x\|_\infty \leq \nu\|x\|_2$ then (1) holds. Our main result is the following theorem, which gives tight asymptotic bounds for the performance of feature hashing, thus closing the long-standing gap.

Theorem 2. *There exist constants $C \geq D > 0$ such that for every $\varepsilon, \delta \in (0, 1)$ and $m \in \mathbb{N}^+$ the following holds. If $\frac{C \lg \frac{1}{\delta}}{\varepsilon^2} \leq m < \frac{2}{\varepsilon^2 \delta}$ then*

$$\nu(m, \varepsilon, \delta) = \Theta\left(\sqrt{\varepsilon} \cdot \min\left\{\frac{\lg \frac{\varepsilon m}{\lg \frac{1}{\delta}}}{\lg \frac{1}{\delta}}, \sqrt{\frac{\lg \frac{\varepsilon^2 m}{\lg \frac{1}{\delta}}}{\lg \frac{1}{\delta}}}\right\}\right).$$

Otherwise, if $m \geq \frac{2}{\varepsilon^2 \delta}$ then $\nu(m, \varepsilon, \delta) = 1$. Moreover if $m < \frac{D \lg \frac{1}{\delta}}{\varepsilon^2}$ then $\nu(m, \varepsilon, \delta) = 0$.

While the bound presented in the theorem may strike as surprising, due to the intricacy of the expressions involved, the tightness of the result shows that this is, in fact, the correct and "true" bound. Moreover, the proof of the theorem demonstrates how both branches in the min expression are required in order to give a tight bound.

Experimental Results. Our theoretical bounds are accompanied by empirical results that shed light on the nature of the constants in Theorem 2. Our empirical results show that in practice the constants inside the Theta-notation are significantly tighter than the theoretical proof might suggest, and in fact feature hashing performs well for a larger scope of vectors. Specifically, our main result implies that whenever $\frac{4 \lg \frac{1}{\delta}}{\varepsilon^2} \leq m < \frac{2}{\varepsilon^2 \delta}$,

$$\nu(m, \varepsilon, \delta) \geq 0.725\sqrt{\varepsilon} \cdot \min\left\{\frac{\lg \frac{\varepsilon m}{\lg \frac{1}{\delta}}}{\lg \frac{1}{\delta}}, \sqrt{\frac{\lg \frac{\varepsilon^2 m}{\lg \frac{1}{\delta}}}{\lg \frac{1}{\delta}}}\right\},$$

(except for very sparse vectors, i.e. $\|x\|_0 \leq 7$) whereas the theoretical proof provides a smaller constant 2^{-6} in front of $\sqrt{\varepsilon}$. Since feature hashing satisfies (1) whenever $\|x\|_\infty \leq \nu(m, \varepsilon, \delta)\|x\|_2$, this implies that feature hashing works well on even a larger range of vectors than the theory suggests.

Proof Technique As a fundamental step in the proof of Theorem 2 we prove tight asymptotic bounds for high-order norms of the approximation factor.[1] More formally, for every $x \in \mathbb{R}^n \setminus \{0\}$ let $X(x) = |\|Ax\|_2^2 - \|x\|_2^2|$. The technical crux of our results is tight bounds on high-order moments of $X(x)$. Note that by rescaling we may restrict our focus without loss of generality to unit vectors.

[1] Given a random variable X and $r > 0$, the rth norm of X (if exists) is defined as $\|X\|_r := \sqrt[r]{\mathbb{E}(|X|^r)}$.

Notation 1. *For every $m, r, k > 0$ denote*

$$\Lambda(m,r,k) = \begin{cases} \sqrt{\frac{r}{m}}, & k \ge mr \\ \max\left\{\sqrt{\frac{r}{m}}, \frac{r^2}{k\ln^2\left(\frac{emr}{k}\right)}\right\}, & mr > k \ge \sqrt{mr} \\ \max\left\{\sqrt{\frac{r}{m}}, \frac{r^2}{k\ln^2\left(\frac{emr}{k}\right)}, \frac{r}{k\ln\left(\frac{emr}{k^2}\right)}\right\}, & \sqrt{mr} > k \end{cases}.$$

In these notations our main technical lemmas are the following.

Lemma 3. *For every even $r \le m/4$ and unit vector $x \in \mathbb{R}^n$, $\|X(x)\|_r = O(\Lambda(m, r, \|x\|_\infty^{-2}))$.*

Lemma 4. *For every $k \le n$ and even $r \le \min\{m/4, k\}$, $\|X(x^{(k)})\|_r = \Omega\left(\Lambda(m, r, k)\right)$, where $x^{(k)} \in \mathbb{R}^n$ is the unit vector whose first k entries equal $\frac{1}{\sqrt{k}}$.*

While it might seem at a glance that bounding the high-order moments of $X(x)$ is merely a technical issue, known tools and techniques could not be used to prove Lemmas 3, 4. Particularly, earlier work by Kane and Nelson [KN14, CJN18] and Freksen and Larsen [FL17] used high-order moments bounds as a step in proving probability tail bounds of random variables. The existing techniques, however, can not be adopted to bound high-order moments of $X(x)$ (see also Section 1.2), and novel approaches were needed. Specifically, our proof incorporates a novel combinatorial scheme for counting edge-labeled Eulerian graphs.

Previous Results. Weinberger *et al.* [WDL+09] showed that if $m = \Omega(\varepsilon^{-2}\lg(1/\delta))$, then $\nu(m, \varepsilon, \delta) = \Omega(\varepsilon \cdot (\lg(1/\delta)\lg(m/\delta))^{-1/2})$. Dasgupta *et al.* [DKS10] showed that under similar conditions $\nu(m, \varepsilon, \delta) = \Omega(\sqrt{\varepsilon} \cdot (\lg(1/\delta)\lg^2(m/\delta))^{-1/2})$. These bounds were recently improved by Dahlgaard *et al.* [DKT17] who showed that $\nu(m, \varepsilon, \delta) = \Omega\left(\sqrt{\varepsilon} \cdot \sqrt{\frac{\lg(1/\varepsilon)}{\lg(1/\delta)\lg(m/\delta)}}\right)$. Conversely, Kane and Nelson [KN14] showed that for the restricted case of $m = \Theta(\varepsilon^{-2}\lg(1/\delta))$, $\nu(m, \varepsilon, \delta) = O\left(\sqrt{\varepsilon} \cdot \frac{\lg(1/\varepsilon)}{\lg(1/\delta)}\right)$, which matches the bound in Theorem 2 if, in addition, $\lg(1/\varepsilon) \le \sqrt{\lg(1/\delta)}$.

Key Tool : Counting Labeled Eulerian Graphs. Our proof presents a new combinatorial result concerning Eulerian graphs. Loosely speaking, we give asymptotic bounds for the number of labeled Eulerian graphs containing a predetermined number of nodes and edges. Formally, let α, β, r be integers such that $1 \le \beta \le \alpha/2 \le \min\{n/2, r/2\}$. Let $\mathcal{G}_{\alpha,\beta,r}$ denote the family of all edge-labeled Eulerian multigraphs $G = ([\alpha], E_G, \pi_G)$, such that

1. G has no isolated vertices;
2. $|E_G| = r$, and $\pi_G : E_G \to [r]$ is a bijection, which assigns a label in $[r]$ to each edge; and
3. the number of connected components in G is β.

Notation 2. *Denote $\Delta = \Delta(\alpha, \beta) := \alpha^{2\alpha}\beta^{-\beta}\left[(\alpha - 2\beta)^2 + 4(\alpha - \beta)\right]^{r-\alpha}$.*

Theorem 5. $2^{-O(r)} \cdot \Delta(\alpha, \beta) \le |\mathcal{G}_{\alpha,\beta,r}| \le 2^{O(r)} \cdot \Delta(\alpha, \beta)$.

1.2 Related Work

The CountSketch scheme, presented by Charikar *et al.* [CCF04], was shown to satisfy (1) by Thorup and Zhang [TZ12]. The scheme essentially samples $O(\lg(1/\delta))$ independent copies of a feature hashing matrix with $m = O(\varepsilon^{-2})$ rows, and applies them all to x. The estimator for $\|x\|_2^2$ is then given by computing the median norm over all projected vectors. The CountSketch scheme thus constructs a sketching matrix A such that every column has $O(\lg(1/\delta))$ non-zero entries. However, this construction does not provide a norm-preserving embedding into a Euclidean space (that is, the estimator of $\|x\|_2^2$ cannot be represented as a norm of Ax), which is essential for some applications such as nearest-neighbor search [HIM12].

Kane and Nelson [KN14] presented a simple construction for the so-called sparse Johnson Lindenstrauss transform. This is a distribution of $m \times n$ matrices, for $m = \Theta(\varepsilon^{-2}\lg(1/\delta))$, where every column has s non-zero entries, randomly chosen from $\{-1, 1\}$. Note that if $s = 1$, this distribution yields the feature hashing one. Kane and Nelson showed that for $s = \Theta(\varepsilon m)$ this construction satisfies (1). Recently, Cohen *et al.* [CJN18] presented two simple proofs for this result. While their proof methods give (simple) bounds for high-order moments similar to those in Lemmas 3 and 4, they rely heavily on the fact that s is relatively large. Specifically, for $s = 1$ the bounds their method or an extension thereof give are trivial.

2 Counting Labeled Eulerian Graphs

In this section we prove Theorem 5. In order to upper bound $|\mathcal{G}_{\alpha,\beta,r}|$, we give an encoding scheme and show that every graph $G \in \mathcal{G}_{\alpha,\beta,r}$ can be encoded in a succinct manner, thus bounding $|\mathcal{G}_{\alpha,\beta,r}|$.

Encoding Argument. Fix a graph $G \in \mathcal{G}_{\alpha,\beta,r}$, and let $\langle\{j_p, \ell_p\}\rangle_{p\in[r]}$ be its ordered sequence of edges. In what follows, we give an encoding algorithm that, given G, produces a "short" bit-string $\mathcal{E}$ that encodes G. The string $\mathcal{E}$ is a concatenation of three strings $\mathcal{E}_T, \mathcal{E}_{Eu}, \mathcal{E}_R$, encoded as follows.

Let $\mathcal{CC}(G) = \{C_1, \ldots, C_\beta\}$ be the set of connected components of G ordered by the smallest labeled node in each component, and for every $j \in [\beta]$, denote the graph induced by C_j in G by $G[C_j] = (C_j, E_j)$. For every $j \in [\beta]$ the encoding algorithm chooses a set $T_j \subseteq E_j$ of edges of a spanning tree in C_j. Denote by E_T the union of all trees in G.

Proposition 6. $|E_T| = \alpha - \beta$.

Proof. For every $j \in [\beta]$, $|E_j| = |C_j| - 1$. Therefore $|E_T| = \sum_{j\in[\beta]} |E_j| = \alpha - \beta$. □

Let $e_1, \ldots, e_{\alpha-\beta}$ be the ordering of E_T induced by π_G. The algorithm encodes $\mathcal{E}_T$ to be the list of $\alpha-\beta$ edges in $\binom{V}{2}$, followed by an encoding of $\pi_G(E_T)$ as a set in $\binom{[r]}{\alpha-\beta}$. Next, since every connected component is Eulerian, for every $j \in [\beta]$, there is an edge $e_j \in E_j \setminus E_T$. Let E_{Eu} denote the set of all β such edges, and let $e_{j_1}, \ldots, e_{j_\beta}$ be the ordering of E_{Eu} induced by π_G. For every $i \in [\beta]$, the algorithm encodes a pair $(j_i, (x_i, y_i)) \in [\beta] \times \binom{C_{j_i}}{2}$, and appends them in order together with $\pi_G(E_{Eu})$ to encode $\mathcal{E}_{Eu}$. Finally, the algorithm encodes $E_G \setminus (E_T \cup E_{Eu})$ in the ordering induced by π_G as a list of length $r - \alpha$ in $\bigcup_{j\in[\beta]} \binom{C_j}{2}$. Denote this list of the rest of the edges by $\mathcal{E}_R$.

Lemma 7. *$\mathcal{E}$ can be encoded using at most* $\lg \Delta(\alpha, \beta) + O(r)$ *bits.*

Proof. In order to bound the length of $\mathcal{E}$ we shall bound each of the three strings separately. One can encode an ordered list of $\alpha-\beta$ distinct unordered pairs in V using at most $(\alpha-\beta)\lg\binom{\alpha}{2}$ bits. Therefore $\mathcal{E}_T$ can be encoded using at most

$$(\alpha-\beta)\lg\binom{\alpha}{2}+\lg\binom{r}{\alpha-\beta}\leq 2(\alpha-\beta)\lg\alpha+r \tag{2}$$

bits.

Next, for every $i\in[\beta]$, $(j_i,(x_i,y_i))$ can be encoded using $\lg\beta\binom{|C_{j_i}|}{2}$ bits. Therefore $\mathcal{E}_{Eu}$ can be encoded using at most

$$\sum_{i\in[\beta]}\lg\beta\binom{|C_{j_i}|}{2}+\lg\binom{r-\alpha}{\beta}\leq\beta\lg\beta+2\lg\prod_{i\in[\beta]}|C_{j_i}|+r\leq\beta\lg\beta+2\beta\lg\frac{\alpha}{\beta}+r \tag{3}$$

bits, where the last inequality follows from the AM-GM inequality, since $\sum_{i\in\beta}|C_{j_i}|=\alpha$.

Finally, note that $\mathcal{E}_R$ can be encoded using $(r-\alpha)\lg\left(\sum_{j\in\beta}\binom{C_j}{2}\right)\leq(r-\alpha)\lg\left(\sum_{j\in\beta}|C_j|^2\right)$ bits. Since

$$\max\left\{\sum_{j\in[\beta]}x_j^2:\sum_{j\in[\beta]}x_j=\alpha\geq 2\beta\ \mathit{and}\ \forall j\in[\beta].\ x_j\geq 2\right\}=(\alpha-2(\beta-1))^2+4(\beta-1)\ ,$$

we get that $\mathcal{E}_R$ can be encoded using

$$(r-\alpha)\lg\left[(\alpha-2(\beta-1))^2+4(\beta-1)\right]=(r-\alpha)\lg\left[(\alpha-2\beta)^2+4(\alpha-\beta)\right] \tag{4}$$

bits. Summing over (2), (3) and (4) implies the lemma. □

Lemma 8. *Given $\mathcal{E}$, one can reconstruct G.*

Proof. In order to prove the lemma, we give a decoding algorithm that receives $\mathcal{E}$ and constructs G. The algorithm first reads the first list of $\alpha-\beta$ elements of $\binom{V}{2}$ from $\mathcal{E}_T$, followed by $\pi_G(E_T)$, to decode E_T, and the restriction $\pi_G|_{E_T}$ of π_G to E_T. Given the set of spanning trees, the algorithm constructs $\mathcal{CC}(G)=\{C_1,\ldots,C_\beta\}$ (note that the ordering on $\mathcal{CC}(G)$ is inherent in the components themselves, and does not depend on π_G). Next, the algorithm reads $\mathcal{E}_{Eu}$ and recovers the set E_{Eu} of edges, along with the restriction $\pi_G|_{E_{Eu}}$ of π_G to E_{Eu}. Finally, the algorithm reads $\mathcal{E}_R$ and reconstructs the remaining $r-\alpha$ edges, with their induced ordering. Since $\pi_G(E_G\setminus(E_T\cup E_{Eu}))=[r]\setminus\pi_G(E_T\cup E_{Eu})$, the algorithm can reconstruct the restriction $\pi_G|_{E_R}$ of π_G to E_R, thus reconstructing π_G.

□

Corollary 9. $|\mathcal{G}_{\alpha,\beta,r}|\leq 2^{O(r)}\Delta(\alpha,\beta)$.

Next we turn to lower bound $|\mathcal{G}_{\alpha,\beta,r}|$. To this end, we construct a subset of $|\mathcal{G}_{\alpha,\beta,r}|$ of size at least $2^{-O(r)}\Delta(\alpha,\beta)$, thus lower bounding $|\mathcal{G}_{\alpha,\beta,r}|$.

Consider the following family $\mathcal{H}_{\alpha,\beta,r}$ of labeled multigraphs over the vertex set $[\alpha]$. For every $H=([\alpha],E_H,\pi_H)\in\mathcal{H}_{\alpha,\beta,r}$, H contains β connected components, where $\beta-1$ components, referred to as *small* are composed of 2 vertices each, and one *large* component contains the remaining $\alpha-2(\beta-1)$ nodes. The first α edges (according to π_H) are a union of β simple cycles, where each cycle contains the entire set of nodes of one connected component.

Claim 10. $|\mathcal{H}_{\alpha,\beta,r}| \geq 2^{-O(r)}\Delta(\alpha,\beta)$.

Proof. The number of possible ways to choose the partition of $[\alpha]$ into β connected components such that all but one contain exactly 2 vertices is $\frac{1}{\beta!}\binom{\alpha}{2,2,\ldots,2,\alpha-2(\beta-1)}$. Each small component has exactly one spanning cycle, while the large component has $(\alpha-2(\beta-1)-1)!$ spanning cycles. Once the cycles are chosen, there are at least $2^{-\beta}\alpha!$ ways to order the edges. The number of possible edges in H is $(\beta-1)\cdot\binom{2}{2}+\binom{\alpha-2(\beta-1)}{2}$. Therefore there are $\left[\binom{\alpha-2(\beta-1)}{2}+(\beta-1)\right]^{r-\alpha}$ ways to choose the ordered sequence of $r-\alpha$ remaining edges. We conclude that

$$\begin{aligned}|\mathcal{H}_{\alpha,\beta,r}| &\geq \frac{1}{\beta!}\binom{\alpha}{2,\ldots,2,\alpha-2(\beta-1)}(\alpha-2\beta+1)!2^{-\beta}\alpha!\left[\binom{\alpha-2(\beta-1)}{2}+2(\beta-1)\right]^{r-\alpha}\\ &\geq 2^{-O(r)}\beta^{-\beta}\cdot(\alpha!)^2\cdot\left[(\alpha-2\beta+2)^2+4(\beta-1)\right]^{r-\alpha}\geq 2^{-O(r)}\Delta(\alpha,\beta)\ .\end{aligned}$$

□

Lemma 11. $\Pr_{H\in_R\mathcal{H}_{\alpha,\beta,r}}[H\in_R\mathcal{G}_{\alpha,\beta,r}]\geq 2^{-O(r)}$.

Proof. Every $H\in\mathcal{H}_{\alpha,\beta,r}$ contains r labeled edges, no isolated vertices and β connected components. Therefore, $H\in\mathcal{G}_{\alpha,\beta,r}$ if and only if the degree of every node in H is even. Let E_λ be the set the last $r-\alpha$ edges in H, then since $E\setminus E_\lambda$ is a union of disjoint cycles spanning all vertices, for every $v\in V$, $deg_{E_H}(v)=deg_{E_H\setminus E_\lambda}(v)+deg_{E_\lambda}(v)=2+deg_{E_\lambda}(v)$. Hence $H\in\mathcal{G}_{\alpha,\beta,r}$ if and only if for every $v\in V$, $deg_{E_\lambda}(v)$ is even. Consider the set of all possible $\left[(\alpha-2\beta+2)^2+4(\beta-1)\right]^{(r-\alpha)/2}$ sequences of $(r-\alpha)/2$ edges in H. For every such sequence s, let the signature of s be the indicator vector $\sigma(s)\in\{0,1\}^V$, where for every $v\in V$, $\sigma(s)_v=1$ if and only if $deg_s(v)$ is odd. Let s_1,s_2 be of the same signature, and let E_λ be the edge sequence of length $r-\alpha$, which is the concatenation of s_1 and s_2. Then $deg_{E_\lambda}(v)$ is even. Since the number of possible signatures is 2^α, there exists a set S of edge sequences of length $(r-\alpha)/2$ that all have the same signature such that $|S|\geq 2^{-\alpha}\left[(\alpha-2\beta+2)^2+4(\beta-1)\right]^{(r-\alpha)/2}$. Therefore

$$\Pr_{H\in_R\mathcal{H}_{\alpha,\beta,r}}[H\in_R\mathcal{G}_{\alpha,\beta,r}]\geq\Pr[E_\lambda\in S\times S]\geq 2^{-O(r)}\ .$$

□

We therefore conclude the following, which finishes the proof of Theorem 5.

Corollary 12. $|\mathcal{G}_{\alpha,\beta,r}|\geq 2^{-O(r)}\Delta(\alpha,\beta)$.

3 Bounding $\nu(m,\varepsilon,\delta)$

In this section we prove Theorem 2, assuming Lemmas 3 and 4, whose proof is deferred to section 4. Fix $\varepsilon,\delta\in(0,1)$ and an integer m. We first address the case where $m\geq\frac{2}{\varepsilon^2\delta}$. Let $x\in\mathbb{R}^n$ be a unit vector. Then

$$\begin{aligned}\mathbb{E}\left[\left|\|Ax\|_2^2-1\right|^2\right]&=\mathbb{E}\left[\left(\sum_{j\neq\ell\in[n]}\mathbb{1}_{h(j)=h(\ell)}\cdot\sigma_j\sigma_\ell\cdot x_jx_\ell\right)^2\right]\\ &=\mathbb{E}\left[2\sum_{j\neq\ell\in[n]}\mathbb{1}_{h(j)=h(\ell)}x_j^2x_\ell^2\right]\leq\frac{2}{m}\end{aligned}$$

Therefore by Chebyshev's inequality $\Pr\left[\ \left|\|Ax\|_2^2 - 1\right| \geq \varepsilon\ \right] \leq \delta$.

We therefore continue assuming $m < \frac{2}{\varepsilon^2\delta}$. From Lemmas 3 and 4 there exist $C_1, C_2 > 0$ such that for every r, k, if $r \leq m/4$ then for every unit vector x, $\|X(x)\|_r \leq 2^{C_2}\Lambda(m,r,k)$. Moreover, if $r \leq k$ then

$$2^{-C_1}\Lambda(m,r,k) \leq \|X(x^{(k)})\|_r \leq 2^{C_2}\Lambda(m,r,k)\ .$$

Note that in addition $\Lambda(m,2r,k) \leq 4\Lambda(m,r,k)$. Denote $\hat{C} = 2^{C_2+2}$, and $C = 2C_1 + 2C_2 + 5$.

Lemma 13. *If* $m < \frac{\log\frac{1}{\delta}}{4C\varepsilon^2}$ *then* $\nu(m,\varepsilon,\delta) = 0$.

Proof. Let $r = \frac{1}{C}\lg\frac{1}{\delta}$, and let $k \geq 2r$ be some integer. Then

$$\mathbb{E}\left[\left(X(x^{(k)})\right)\right] = \|X(x^{(k)})\|_r^r \geq \left(\frac{r}{m}\right)^{r/2} \geq 2\varepsilon^r\ .$$

Applying the Paley-Zygmund inequality

$$\begin{aligned}
\Pr\left[\ \left|\|Ax^{(k)}\|_2^2 - 1\right| > \varepsilon\ \right] &= \Pr\left[\ \left|Ax^{(k)} - 1\right|^r > \varepsilon^r\ \right] \\
&\geq \Pr\left[\left(X(x^{(k)})\right)^r > 2^{-1}\mathbb{E}[X(x^{(k)})^r]\ \right] \\
&\geq \frac{\mathbb{E}^2[X(x^{(k)})^r]}{4\mathbb{E}[X(x^{(k)})^{2r}]} = \frac{\|X(x^{(k)})\|_r^{2r}}{4\|X(x^{(k)})\|_{2r}^{2r}} \geq \frac{1}{4}\left(\frac{2^{-C_1}\Lambda(m,r,k)}{2^{C_2}\Lambda(m,2r,k)}\right)^{2r} \qquad (5)\\
&\geq \frac{1}{4}\left(\frac{2^{-C_1}}{2^{C_2+2}}\right)^{2r} = 2^{-(2C_1-2C_2-4)r-2} \geq \delta
\end{aligned}$$

Therefore $\nu(m,\varepsilon,\delta) \leq \|x^{(k)}\|_\infty = \frac{1}{\sqrt{k}}$ for every $k \geq 2r$, which implies $\nu(m,\varepsilon,\delta) = 0$. □

For the rest of the proof we assume that $\frac{\hat{C}\log\frac{1}{\delta}}{\varepsilon^2} \leq m < \frac{2}{\varepsilon^2\delta}$, and we start by proving a lower bound on ν.

Lemma 14. $\nu(m,\varepsilon,\delta) = \Omega\left(\min\left\{\frac{\sqrt{\varepsilon}}{\lg\frac{1}{\delta}}\lg\frac{\varepsilon m}{\lg\frac{1}{\delta}}, \sqrt{\frac{\varepsilon\lg\frac{\varepsilon^2 m}{\lg\frac{1}{\delta}}}{\lg\frac{1}{\delta}}}\right\}\right)$.

Proof. Let $r = \lg\frac{1}{\delta}$, let $x \in \mathbb{R}^n$ be a unit vector such that $\|x\|_\infty \leq \min\left\{\frac{\sqrt{\varepsilon}\ln\frac{e\varepsilon m}{r}}{\sqrt{2^{C_2}er}}, \sqrt{\frac{\varepsilon\lg\frac{e\varepsilon^2 m}{r}}{2^{C_2}er}}\right\}$, and let $k := \frac{1}{\|x\|_\infty^2} \geq \max\left\{\frac{2^{C_2}er^2}{\varepsilon\ln^2\frac{e\varepsilon m}{r}}, \frac{2^{C_2}er}{\varepsilon\lg\frac{e\varepsilon^2 m}{r}}\right\}$. If $k \leq mr$, then since $\frac{r^2}{k\ln^2\frac{emr}{k}}$ is convex as a function of $k \in \left[\frac{2^{C_2}er^2}{\varepsilon\ln^2\frac{e\varepsilon m}{r}}, mr\right]$ then

$$2^{C_2}\frac{r^2}{k\ln^2\frac{emr}{k}} \leq \max\left\{\frac{r}{m}, \left(\frac{\varepsilon\ln^2\frac{e\varepsilon m}{r}}{e\ln^2\frac{e\varepsilon m\ln^2\frac{e\varepsilon m}{r}}{r}}\right)\right\} < \varepsilon/2\ .$$

Moreover, if $k \leq \sqrt{mr}$ then since $\frac{r}{k\ln\frac{emr}{k^2}}$ is convex as a function of $k \in \left[\frac{2^{C_2}er}{\varepsilon\ln\frac{e\varepsilon^2 m}{r}}, \sqrt{mr}\right]$, then

$$2^{C_2}\frac{r}{k\ln\frac{emr}{k^2}} \leq \max\left\{\sqrt{\frac{r}{m}}, \left(\frac{\varepsilon\ln\frac{e\varepsilon^2 m}{r}}{e\ln\frac{\varepsilon^2 m\ln^2\frac{e\varepsilon^2 m}{r}}{r}}\right)\right\} \leq \varepsilon/2\ .$$

Since clearly, $\sqrt{\frac{2^{2C_2 r}}{m}} \le \varepsilon/2$, then by Lemma 3 we have $\|X(x)\|_r^r \le (\varepsilon/2)^r$, and thus

$$\begin{aligned}\Pr\left[\ \left|\|Ax\|_2^2 - 1\right| > \varepsilon\ \right] &= \Pr\left[\ \left|\|Ax\|_2^2 - 1\right|^r > \varepsilon^r\ \right] \\ &\le \Pr\left[(X(x))^r > 2^r \mathbb{E}[X(x)^r]\ \right] \le 2^{-r} = \delta\ .\end{aligned} \tag{6}$$

Hence $\nu(m,\varepsilon,\delta) \ge \min\left\{\frac{\sqrt{\varepsilon}\ln\frac{e\varepsilon m}{r}}{\sqrt{2^{C_2}er}}, \sqrt{\frac{\varepsilon\lg\frac{e\varepsilon^2 m}{r}}{2^{C_2}er}}\right\} = \Omega\left(\min\left\{\frac{\sqrt{\varepsilon}}{\lg\frac{1}{\delta}}\lg\frac{\varepsilon m}{\lg\frac{1}{\delta}}, \sqrt{\frac{\varepsilon\lg\frac{\varepsilon^2 m}{\lg\frac{1}{\delta}}}{\lg\frac{1}{\delta}}}\right\}\right)$. □

Lemma 15. $\nu(m,\varepsilon,\delta) = O\left(\min\left\{\frac{\sqrt{\varepsilon}}{\lg\frac{1}{\delta}}\lg\frac{\varepsilon m}{\lg\frac{1}{\delta}}, \sqrt{\frac{\varepsilon\lg\frac{\varepsilon^2 m}{\lg\frac{1}{\delta}}}{\lg\frac{1}{\delta}}}\right\}\right)$.

To this end, let $r = \frac{1}{C}\lg\frac{1}{\delta}$, and denote

$$t = \min\left\{\frac{\sqrt{e\varepsilon}}{r}\ln\frac{e\varepsilon m}{r}, \sqrt{\frac{e\varepsilon\ln\frac{e\varepsilon^2 m}{r}}{r}}\right\} = O\left(\min\left\{\frac{\sqrt{\varepsilon}}{\lg\frac{1}{\delta}}\lg\frac{\varepsilon m}{\lg\frac{1}{\delta}}, \sqrt{\frac{\varepsilon\lg\frac{\varepsilon^2 m}{\lg\frac{1}{\delta}}}{\lg\frac{1}{\delta}}}\right\}\right)\ .$$

Assume first that $t \le \frac{1}{\sqrt{r}}$, and let $k = \frac{1}{t^2}$. We will show that $\mathbb{E}\left[\left(X(x^{(k)})\right)^r\right] \ge 2\varepsilon^r$. Since $t \le \frac{1}{\sqrt{r}}$, then $k \ge r$. If $\frac{\sqrt{e\varepsilon}}{r}\ln\frac{e\varepsilon m}{r} \le \sqrt{\frac{e\varepsilon\ln\frac{e\varepsilon^2 m}{r}}{r}}$, then $k = \frac{r^2}{e\ln^2\frac{e\varepsilon m}{r}}$. Since $\frac{e\varepsilon m}{r} > e$, then $k \le mr$. Therefore

$$\mathbb{E}\left[\left(X(x^{(k)})\right)^r\right] = \|X(x^{(k)})\|_r^r \ge \left(\frac{r^2}{k\ln^2\frac{emr}{k}}\right)^r = \left(\frac{e\varepsilon\ln^2\frac{e\varepsilon m}{r}}{\ln^2\frac{e^2\varepsilon m\ln^2\frac{e\varepsilon m}{r}}{r}}\right)^r \ge 2\varepsilon^r\ .$$

Otherwise, $k = \frac{r}{e\varepsilon\ln\frac{e\varepsilon^2 m}{r}}$. Moreover, since $\frac{\varepsilon^2 m}{r} > 1$, then $k \le r/\varepsilon \le \sqrt{mr}$. Therefore

$$\mathbb{E}\left[\left(X(x^{(k)})\right)^r\right] = \|X(x^{(k)})\|_r^r \ge \left(\frac{r}{k\ln\frac{emr}{k^2}}\right)^r = \left(\frac{e\varepsilon\ln\frac{e\varepsilon^2 m}{r}}{\ln\frac{e^3\varepsilon^2 m\ln^2\frac{e\varepsilon^2 m}{r}}{r}}\right)^r \ge 2\varepsilon^r\ .$$

Applying the Paley-Zygmund inequality we get that similarly to (5)

$$\Pr\left[\ \left|\|Ax^{(k)}\|_2^2 - 1\right| > \varepsilon\ \right] \ge \Pr\left[\left(X(x^{(k)})\right)^r > 2^{-1}\mathbb{E}[X(x^{(k)})^r]\ \right] \ge \delta$$

Therefore $\nu(m,\varepsilon,\delta) \le \|x^{(k)}\|_\infty = t$.

Assume next that $\frac{1}{\sqrt{r}} < t < \sqrt{\frac{\varepsilon}{4}}$, and note that since $\frac{\sqrt{e\varepsilon}}{r}\ln\frac{e\varepsilon m}{r} \ge \frac{1}{\sqrt{r}}$, then $m > e^{\sqrt{r/(e\varepsilon)}}$, and since $\sqrt{\frac{e\varepsilon\ln\frac{e\varepsilon^2 m}{r}}{r}} \ge \frac{1}{\sqrt{r}}$ then $m > e^{1/(e\varepsilon)}$. Let $k = \frac{1}{t^2}$, and consider independent $h \in_R [n] \to [m]$, and $\sigma = (\sigma_1, \ldots, \sigma_m) \in_R \{-1,1\}^m$. Let $y \in \mathbb{R}^n$ be defined as follows. For every $j \in [n]$, $y_j = x_j^{(k)}$ if and only if $h(j) = 1$, and $y_j = 0$ otherwise. Denote $z = x^{(k)} - y$. Then $\|x^{(k)}\|_2^2 = \|y\|_2^2 + \|z\|_2^2$, and moreover, $\|Ax^{(k)}\|_2^2 = \|Ay\|_2^2 + \|Az\|_2^2$, where $A = A(h,\sigma)$. Let $\mathcal{E}_{first}$ denote the event that $|h^{-1}(\{1\})| = 2\sqrt{\varepsilon k}$, and that for all $j \in [n]$, if $h(j) = 1$

then $\sigma_j = 1$, and let $\mathcal{E}_{rest}$ denote the event that $\left|\|Az\|_2^2 - \|z\|_2^2\right| < \varepsilon\|z\|_2^2$. By Chebyshev's inequality, $\Pr[\mathcal{E}_{rest} \mid \mathcal{E}_{first}] = \Omega(1)$. Note that if $k = \frac{r^2}{e\varepsilon \ln^2 \frac{e\varepsilon m}{r}}$, then

$$2\sqrt{\varepsilon k} = \frac{r}{\sqrt{e}\ln\frac{e\varepsilon m}{r}} \le \frac{\lg\frac{1}{\delta}}{C\sqrt{e}\ln\frac{e\varepsilon m}{r}} \le \frac{\lg\frac{1}{\delta}}{C\sqrt{e}(\ln em - \ln\frac{1}{\varepsilon} - \ln r)}$$
$$\le \frac{\lg\frac{1}{\delta}}{C\sqrt{e}(\ln m - 3\ln\ln m)} \le \frac{\lg\frac{1}{\delta}}{2\ln m} ,$$

where the inequality before last is due to the fact that $m > \max\{e^{1/e\varepsilon}, e^{\sqrt{r}}\}$, and otherwise, $k = \frac{r}{e\varepsilon \ln\frac{e\varepsilon^2 m}{r}}$, and

$$2\sqrt{\varepsilon k} \le \varepsilon k = \frac{\lg\frac{1}{\delta}}{eC\ln\frac{e\varepsilon^2 m}{r}} = \frac{\lg\frac{1}{\delta}}{eC(\ln em - 2\ln\frac{1}{\varepsilon} - \ln r)} \le \frac{\lg\frac{1}{\delta}}{eC(\ln em - 4\ln\ln m)} \le \frac{\lg\frac{1}{\delta}}{2\ln m} .$$

Therefore for small enough ε,

$$\Pr[\mathcal{E}_{first}] = \binom{k}{2\sqrt{\varepsilon k}} \cdot \left(\frac{1}{m}\right)^{2\sqrt{\varepsilon k}} \cdot \left(1 - \frac{1}{m}\right)^{k-2\sqrt{\varepsilon k}} \cdot 2^{-2\sqrt{\varepsilon k}}$$
$$\ge \left(\frac{1}{m}\right)^{2\sqrt{\varepsilon k}} \cdot \left(1 - \frac{1}{m}\right)^{r} \cdot 2^{-r} \ge \left(\frac{1}{m}\right)^{2\sqrt{\varepsilon k}} \cdot 2^{-\frac{2}{C}\lg\frac{1}{\delta}} \ge \delta^{3/4} .$$

We conclude that for small enough δ, $\Pr[\mathcal{E}_{first} \wedge \mathcal{E}_{rest}] \ge \delta$. Conditioned on $\mathcal{E}_{first} \wedge \mathcal{E}_{rest}$ we get that

$$\|Ax^{(k)}\|_2^2 = \|Ay\|_2^2 + \|Az\|_2 \ge \frac{4\varepsilon k}{k} + (1-\varepsilon)\|z\|_2^2$$
$$= 4\varepsilon + (1-\varepsilon) \cdot \frac{k - 2\sqrt{\varepsilon k}}{k} \ge 4\varepsilon + (1-\varepsilon)^2 > 1 + \varepsilon ,$$

where the inequality before last is due to the fact that $k \ge \frac{4}{\varepsilon}$. Therefore $\nu(m, \varepsilon, \delta) \le \|x^{(k)}\|_\infty = t$.

Finally, assume $t > \sqrt{\frac{\varepsilon}{4}}$. Since $\sqrt{\frac{e\varepsilon\ln\frac{e\varepsilon^2 m}{r}}{r}} \ge t > \sqrt{\frac{\varepsilon}{4}}$, we get that $m \ge \frac{r}{e\varepsilon^2} e^{r/(4e)} \ge \frac{r}{e\varepsilon^2\delta^{1/(4eC)}}$. Let $k = \frac{2}{\varepsilon}$. Consider independent $h \in_R [n] \to [m]$, and $\sigma = (\sigma_1, \ldots, \sigma_m) \in_R \{-1, 1\}^m$, and let $A = A(h, \sigma)$. Let $\mathcal{E}_{col}$ denote the event that there are $j \ne \ell \in [k]$ such that for every $p \ne q \in [k]$, $h(p) = h(q)$ if and only if $\{p, q\} = \{j, \ell\}$. Then for small enough ε, δ,

$$\Pr[\mathcal{E}_{col}] = \binom{k}{2} \cdot \frac{1}{m} \cdot \prod_{j \in [k-1]} \left(1 - \frac{j}{m}\right) \ge \frac{k^2}{2m} \cdot (1 - \varepsilon/2) \cdot \left(1 - \frac{k}{m}\right)^k$$
$$\ge \frac{k^2}{2m} \cdot (1 - \varepsilon/2) \cdot \left(1 - \frac{k^2}{m}\right) \ge 2\delta \cdot (1 - \varepsilon/2) \cdot \left(1 - 4e\delta^{1/(4Ce)}\right) \ge \delta .$$

Conditioned on $\mathcal{E}_{col}$ we get that $\left|\|Ax^{(k)}\|_2^2 - 1\right| = \frac{2}{k} = \varepsilon$. Therefore $\nu(m, \varepsilon, \delta) \le \sqrt{\frac{\varepsilon}{2}} \le O(t)$.

This completes the proof of Lemma 15, and thus of Theorem 2.

4 Bounding the Moments of $|\|Ax\|_2 - 1|$

In this section we prove Lemmas 3 and 4. Recall that $h \in_R [n] \to [m]$, and $\sigma_1, \ldots, \sigma_m \in_R \{1, -1\}$ are independent. For every $i \in [m], j \in [n]$, $a_{ij} := \sigma_j \cdot \mathbb{1}_{h(j)=i}$. For every unit vector $x \in \mathbb{R}^n \setminus \{0\}$ we let $X = X(x) = |\|Ax\|_2^2 - 1|$. We start with providing a better understanding of X.

$$\|Ax\|_2^2 = 1 + \sum_{j \neq \ell \in [n]} \mathbb{1}_{h(j)=h(\ell)} \cdot \sigma_j \sigma_\ell \cdot x_j x_\ell$$

Denote $I_{[n]} = \{(p,p) : p \in [n]\}$. Then

$$X = \left| \|Ax\|_2^2 - 1 \right| = \left| \sum_{(j,\ell) \in ([n] \times [n] \setminus I_{[n]})} \mathbb{1}_{h(j)=h(\ell)} \cdot \sigma_j \sigma_\ell \cdot x_j x_\ell \right| ,$$

and therefore for every even r

$$\|X\|_r^r = \mathbb{E}[X^r] = \sum_{\langle (j_p, \ell_p) \rangle_{p \in [r]} \in ([n] \times [n] \setminus I_{[n]})^r} \mathbb{E}\left[\prod_{p \in [r]} \mathbb{1}_{h(j_p)=h(\ell_p)} \sigma_{j_p} \sigma_{\ell_p} x_{j_p} x_{\ell_p} \right] \tag{7}$$

Every $S = \langle (j_p, \ell_p) \rangle_{p \in [r]} \in ([n] \times [n] \setminus I_{[n]})^r$ defines a directed multigraph $\overrightarrow{G_S}$ with r ordered directed edges on vertex set $[n]$. Let G_S denote the underlying undirected multigraph.

Definition 1. *Let $q \in [n]$. The* degree of q in S *is the degree of q in G_S. Namely $d_S(q) := |\{p \in [r] : q \in \{j_p, \ell_p\}\}|$.*

Notation 3. *Given $S \in ([n] \times [n] \setminus I_{[n]})^r$, let $\mathcal{CC}(S)$ denote the set of all connected components of G_S that contain at least two nodes. Let $\beta(S) := |\mathcal{CC}(S)|$, $V(S) = \bigcup_{C \in \mathcal{CC}(S)} C$ and $\alpha(S) := |V(S)|$.*

Next, for every integer β and a subset $V \subseteq [n]$, let $\mathcal{S}_{V,\beta} \subseteq ([n] \times [n] \setminus I_{[n]})^r$ be the set of all sequences $S \in ([n] \times [n] \setminus I_{[n]})^r$ such that

1. For every $q \in [n]$, $d_S(q)$ is even; and
2. $V(S) = V$ and $\beta(S) = \beta$.

Lemma 16.

$$\|X\|_r^r \leq \|x\|_\infty^{2r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta}{(\|x\|_\infty^2 m)^\alpha} \sum_{V \in \binom{[n]}{\alpha}} |\mathcal{S}_{V,\beta}| \cdot \prod_{q \in V} x_q^2 , \tag{8}$$

Moreover, if for all $j \in \operatorname{supp}(x)$, $|x_j| = \|x\|_\infty$, then equality holds.

Proof. Fix some $S = \langle (j_p, \ell_p) \rangle_{p\in[r]} \in ([n] \times [n] \setminus I_{[n]})^r$. Then

$$\begin{aligned}\mathbb{E}\left[\prod_{p\in[r]} \mathbb{1}_{h(j_p)=h(\ell_p)} \sigma_{j_p}\sigma_{\ell_p} x_{j_p} x_{\ell_p}\right] &= \mathbb{E}\left[\prod_{p\in[r]} \mathbb{1}_{h(j_p)=h(\ell_p)} \cdot \prod_{q\in V(S)} \sigma_q^{d_S(q)} \cdot \prod_{q\in V(S)} x_q^{d_S(q)}\right] \\ &= \prod_{q\in V(S)} x_q^{d_S(q)} \cdot \mathbb{E}\left[\prod_{p\in[r]} \mathbb{1}_{h(j_p)=h(\ell_p)}\right] \cdot \prod_{q\in V(S)} \mathbb{E}\left[\sigma_q^{d_S(q)}\right]\end{aligned} \tag{9}$$

where the last equality follows from independence. Assume first that for some $q \in V(S)$, $d_S(q)$ is odd. Then $\mathbb{E}\left[\sigma_q^{d_S(q)}\right] = 0$, and therefore (9) equals 0. Otherwise, $\mathbb{E}\left[\sigma_q^{d_S(q)}\right] = 1$ for all $q \in V(S)$. We therefore assume hereafter that $d_S(q)$ is even for all $q \in V(S)$. For every $C \in \mathcal{CC}(S)$, C contains an edge of G_S, thus there exists $p \in [r]$ such that $j_p, \ell_p \in C$. Conversely, for every $p \in [r]$ there exists a unique connected component $C \in \mathcal{CC}(S)$ such that $j_p, \ell_p \in C$. Therefore

$$\mathbb{E}\left[\prod_{p\in[r]} \mathbb{1}_{h(j_p)=h(\ell_p)}\right] = \mathbb{E}\left[\prod_{C\in\mathcal{CC}(S)} \prod_{p\in[r]:j_p\in C} \mathbb{1}_{h(j_p)=h(\ell_p)}\right] = \prod_{C\in\mathcal{CC}(S)} \mathbb{E}\left[\prod_{p\in[r]:j_p\in C} \mathbb{1}_{h(j_p)=h(\ell_p)}\right],$$

where the last equality is due to independence. Next, let $C = \{v_1, \ldots, v_{|C|}\} \in \mathcal{CC}(S)$, then $\mathbb{E}\left[\prod_{p\in[r]:j_p\in C} \mathbb{1}_{h(j_p)=h(\ell_p)}\right] = \mathbb{E}\left[\mathbb{1}_{h(v_1)=\ldots=h(v_{|C|})}\right] = \frac{1}{m^{|C|-1}}$. We thus conclude that

$$\prod_{C\in\mathcal{CC}(S)} \mathbb{E}\left[\prod_{p\in[r]:j_p\in C} \mathbb{1}_{h(j_p)=h(\ell_p)}\right] = \prod_{C\in\mathcal{CC}(S)} \frac{1}{m^{|C|-1}} = \frac{1}{m^{\alpha(S)-\beta(S)}} .$$

For every sequence S that donates a non-zero summand to the sum, since $d_S(q)$ is even for all $q \in V(S)$ every $C \in \mathcal{CC}(S)$ is Eulerian, and therefore contains at least two nodes and two edges. Therefore $1 \le \beta(S) \le r/2$ and $2\beta(S) \le \alpha(S) \le r$. Plugging this into (7) we get that

$$\begin{aligned}\|X\|_r^r &= \sum_{\substack{S\in([n]\times[n]\setminus I_{[n]})^r \\ \forall q\in V(S).\ d_S(q)\in\mathbb{N}_{even}}} \frac{1}{m^{\alpha(S)-\beta(S)}} \prod_{q\in V(S)} x_q^{d_S(q)} \\ &= \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \sum_{V\in\binom{[n]}{\alpha}} \sum_{S\in\mathcal{S}_{V,\beta}} \frac{1}{m^{\alpha-\beta}} \prod_{q\in V} x_q^{d_S(q)}\end{aligned} \tag{10}$$

For every $q \in V(S)$, $d_S(q)$ is a positive even integer, and therefore $d_S(q) - 2 \ge 0$ is also even. Hence for every $q \in V(S)$, $x_q^{d_S(q)-2} = |x_q|^{d_S(q)-2} \le \|x\|_\infty^{d_S(q)-2}$. Since $\sum_{q\in V(S)} d_S(q) = 2r$, then $\prod_{q\in V} x_q^{d_S(q)} \le \|x\|_\infty^{2r-2\alpha} \prod_{q\in V} x_q^2$. Moreover, equality holds if for all $j \in \mathrm{supp}(x)$, $|x_j| = \|x\|_\infty$. Plugging this in (10) we get that

$$\begin{aligned}\|X\|_r^r &\le \|x\|_\infty^{2r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta}{(\|x\|_\infty^2 m)^\alpha} \sum_{V\in\binom{[n]}{\alpha}} \sum_{S\in\mathcal{S}_{V,\beta}} \prod_{q\in V} x_q^2 \\ &= \|x\|_\infty^{2r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta}{(\|x\|_\infty^2 m)^\alpha} \sum_{V\in\binom{[n]}{\alpha}} |\mathcal{S}_{V,\beta}| \cdot \prod_{q\in V} x_q^2\end{aligned}$$

□

4.1 Upper Bounding $\|X\|_r$

We start by proving Lemma 3. To this end, denote $k = \|x\|_\infty^{-2}$, and for every $1 \leq \beta \leq \alpha/2 \leq r/2$, denote

$$M(\alpha,\beta) = \left(m\beta^{-1}\right)^\beta \left(\frac{k\alpha}{m}\right)^\alpha (\alpha - 2\beta)^{2r-2\alpha}$$
$$N(\alpha,\beta) = \left(m\beta^{-1}\right)^\beta \left(\frac{k\alpha}{m}\right)^\alpha (\alpha - \beta)^{r-\alpha} .$$

Applying Theorem 5 to the expression in Lemma 16 we can conclude the following.

Claim 17. $\|X\|_r^r \leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} (M(\alpha,\beta) + N(\alpha,\beta))$.

Proof. Let $1 \leq \beta \leq \alpha/2 \leq r/2$. Then for every $V \in \binom{[n]}{\alpha}$, every sequence in $\mathcal{S}_{V,\beta}$ defines a directed edge-labeled multigraph $\overrightarrow{G_S}$ on V, whose underlying undirected graph G_S is Eulerian and has β connected components. Invoking the notation used in Theorem 5, G_S is isomorphic to some graph in $\mathcal{G}_{\alpha,\beta,r}$, and moreover, it defines at most 2^r sequences in $\mathcal{S}_{V,\beta}$. Thus $|\mathcal{S}_{V,\beta}| \leq 2^r |\mathcal{G}_{\alpha,\beta,r}| \leq 2^{O(r)} \Delta(\alpha,\beta)$. Plugging this in the (8) we get that

$$\|X\|_r^r \leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} m^\beta \left(\frac{k}{m}\right)^\alpha \Delta(\alpha,\beta) \sum_{V \in \binom{[n]}{\alpha}} \prod_{q \in V} x_q^2 . \tag{11}$$

For every $V \in \binom{[n]}{\alpha}$, the coefficient of $\prod_{q \in V} x_q^2$ in the expansion of $\left(\sum_{q\in[n]} x_q^2\right)^\alpha$ is $\alpha!$. Therefore

$$1 = \left(\sum_{q\in[n]} x_q^2\right)^\alpha \geq \alpha! \sum_{V \in \binom{[n]}{\alpha}} \prod_{q \in V} x_q^2 .$$

Plugging this in (11) we get

$$\begin{aligned}
\|X\|_r^r &\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} m^\beta \left(\frac{k}{m}\right)^\alpha \frac{\Delta(\alpha,\beta)}{\alpha!} \\
&\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \left(m\beta^{-1}\right)^\beta \left(\frac{k\alpha}{m}\right)^\alpha \left[(\alpha-2\beta)^2 + 4(\alpha-\beta)\right]^{r-\alpha} \\
&\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} (M(\alpha,\beta) + N(\alpha,\beta)) .
\end{aligned}$$

□

Lemma 18. *For all $1 \le \beta \le \alpha/2 \le r/2$, then if $k \ge mr$, then $M(\alpha,\beta) \le 2^{O(r)}\left(\frac{k^2 r}{m}\right)^{r/2}$. Otherwise, $M(\alpha,\beta) \le 2^{O(r)} \max\left\{\left(\frac{r}{\ln\frac{2emr}{k}}\right)^{2r}, \left(\frac{k^2 r}{m}\right)^{r/2}\right\}$.*

Proof. Let $1 \le \beta \le \alpha/2 \le r/2$. First note that

$$\alpha^\alpha \le 2^{O(r)}\alpha! = 2^{O(r)}(\alpha-2\beta)!\cdot(2\beta)!\cdot\binom{\alpha}{2\beta} \le 2^{O(r)}(\alpha-2\beta)^{\alpha-2\beta}\beta^{2\beta}\ ,$$

and therefore $M(\alpha,\beta) \le 2^{O(r)}(m\beta)^\beta \cdot \left(\frac{k}{m}\right)^\alpha (\alpha-2\beta)^{2r-\alpha-2\beta}$.

Next, we fix some $\beta \in [r/2]$. Define $f, \hat{f} : (2\beta, +\infty) \to \mathbb{R}$ by $f(\alpha) = \left(\frac{k}{m}\right)^\alpha \cdot (\alpha-2\beta)^{2r-\alpha-2\beta}$, and $\hat{f}(\alpha) = \ln\frac{k}{m} - \ln(\alpha-2\beta) + \frac{2r-4\beta}{\alpha-2\beta} - 1$ for all $\alpha > 2\beta$. Then $f'(\alpha) = f(\alpha)\cdot\hat{f}(\alpha)$, and $\hat{f}'(\alpha) = -\frac{1}{\alpha-2\beta} - \frac{2(r-2\beta)}{(\alpha-2\beta)^2} < 0$ for all $\alpha > 2\beta$.

Assume first that $\beta \ge \frac{r}{2} - \frac{ek}{2m}$, then $\hat{f}(r) = \ln\frac{k}{m} - \ln(r-2\beta) + 1 \ge 0$. Therefore $\hat{f}(\alpha) \ge 0$, thus $f'(\alpha) \ge 0$ and $f(\alpha) \le f(r)$ for all $\alpha \in (2\beta, r]$. It follows that

$$M(\alpha,\beta) \le 2^{O(r)}(m\beta)^\beta\left(\frac{k}{m}\right)^r (r-2\beta)^{r-2\beta} \le 2^{O(r)}\left(\frac{m}{r/2}\right)^{r/2}\left(\frac{kr}{m}\right)^r \le 2^{O(r)}\left(\frac{k^2 r}{m}\right)^{r/2}\ , \tag{12}$$

where the inequality before last follows from the fact that $m \ge r$ and $\beta \le r/2$.

To prove the first part of the lemma, note that if $k \ge \frac{mr}{e}$, then for all $\beta \in (0, r/2]$, $\beta \ge \frac{r}{2} - \frac{ek}{2m}$.

To prove the second part of the lemma, we next assume that $k < \frac{mr}{e}$ and $\beta < \frac{r}{2} - \frac{ek}{2m}$, and note that this implies a tighter bound on k, namely $k < \frac{1}{e}m(r-2\beta)$. Let

$$\alpha_0 = 2\beta + \frac{2(r-2\beta)}{\ln\frac{2em(r-2\beta)}{k}} \qquad , \qquad \alpha_1 = 2\beta + \frac{2(r-2\beta)}{\ln\frac{2em(r-2\beta)}{k\ln\frac{2em(r-2\beta)}{k}}}\ .$$

Then $2\beta \le \alpha_0 < \alpha_1$, and moreover

$$\hat{f}(\alpha_0) = \ln\frac{k}{m} - \ln\frac{2(r-2\beta)}{\ln\frac{2em(r-2\beta)}{k}} + \frac{2(r-2\beta)}{\frac{2(r-2\beta)}{\ln\frac{2em(r-2\beta)}{k}}} - 1 = \ln\ln\frac{2em(r-2\beta)}{k} > 0\ ,$$

and similarly

$$\hat{f}(\alpha_1) = \ln\ln\frac{2em(r-2\beta)}{k\ln\frac{2em(r-2\beta)}{k}} - \ln\ln\frac{2em(r-2\beta)}{k} < 0\ .$$

Therefore there exists a unique $\alpha^* \in (\alpha_0, \alpha_1)$ such that $\hat{f}(\alpha^*) = 0$, and thus $\frac{k}{em(\alpha^*-2\beta)} = e^{-\frac{2r-4\beta}{\alpha^*-2\beta}} \le 1$. Moreover, for every $\alpha > 2\beta$,

$$f(\alpha) \le f(\alpha^*) = \left(\frac{k}{m}\right)^{\alpha^*}\cdot(\alpha^*-2\beta)^{2r-\alpha^*-2\beta} \le \left(\frac{k}{m}\right)^{2\beta}(\alpha^*-2\beta)^{2r-4\beta}\ ,$$

and we get that

$$
\begin{aligned}
M(\alpha,\beta) &\leq 2^{O(r)}(m\beta)^{\beta}\left(\frac{k}{m}\right)^{2\beta}(\alpha^*-2\beta)^{2r-4\beta} \leq 2^{O(r)}\left(\frac{k^2\beta}{m}\right)^{\beta}(\alpha_1-2\beta)^{2r-4\beta} \\
&\leq 2^{O(r)}\left(\frac{k^2\beta}{m}\right)^{\beta}\left(\frac{r}{\ln\frac{2emr}{k\ln\frac{2emr}{k}}}\right)^{2r-4\beta}
\end{aligned}
\tag{13}
$$

where the last inequality follows from the fact that $y = \frac{x}{\ln\frac{x}{\ln x}}$ is monotonically increasing for $x > 1$.

Finally, define $g, \hat{g} : (0, +\infty) \to \mathbb{R}$ by

$$
g(\beta) = \left(\frac{k^2\beta}{m}\right)^{\beta}\left(\frac{r}{\ln\frac{2emr}{k\ln\frac{2emr}{k}}}\right)^{2r-4\beta} \quad \textit{and} \quad \hat{g}(\beta) = \ln\frac{k^2}{m} + \ln\beta + 1 - 4\ln\left(\frac{r}{\ln\frac{2emr}{k\ln\frac{2emr}{k}}}\right)
$$

for all $\beta > 0$. Then $g'(\beta) = g(\beta)\cdot\hat{g}(\beta)$, and moreover $g''(\beta) = g(\beta)\cdot\hat{g}^2(\beta) + g(\beta)/\beta > 0$. We thus conclude that g is convex as a function of $\beta \in (0, r/2]$, and therefore $g(\beta) \leq \max\{\lim_{\beta\to 0} g(\beta), g(r/2)\}$ for all $\beta \in (0, r/2]$.

$$
\lim_{\beta\to 0} g(\beta) = \left(\frac{r}{\ln\frac{2emr}{k\ln\frac{2emr}{k}}}\right)^{2r}, \quad \textit{and} \quad g(r/2) \leq 2^{O(r)}\left(\frac{k^2 r}{m}\right)^{r/2}.
$$

Since $k < \frac{mr}{e}$, then $\frac{2emr}{k} > e^2$, and therefore $\ln\frac{2emr}{k\ln\frac{2emr}{k}} \geq \frac{1}{2}\ln\frac{2emr}{k}$. Plugging into (13), we thus get that since $\ln\frac{2emr}{k} > 1$,

$$
M(\alpha,\beta) \leq 2^{O(r)}\max\left\{\left(\frac{r}{\ln\frac{2emr}{k}}\right)^{2r}, \left(\frac{k^2 r}{m}\right)^{r/2}\right\},
$$

and the proof of the lemma is now complete. □

Lemma 19. *Let $1 \leq \beta \leq \alpha/2 \leq \ell/2$. Then if $k^2 > mr$, then $N(\alpha,\beta) \leq \left(\frac{k^2 r}{m}\right)^{r/2}$ and otherwise, $N(\alpha,\beta) \leq \max\left\{M(\alpha,\beta), \left(\frac{r}{\ln\frac{emr}{k^2}}\right)^{r}\right\}$.*

Proof. Assume first that $k^2 \geq mr$. If $k > m$, then $\left(\frac{k\alpha}{m}\right)^{\alpha}(\alpha-\beta)^{r-\alpha}$ is increasing as a function of α over $[2\beta, r]$, and therefore,

$$
N(\alpha,\beta) \leq \left(\frac{kr}{m}\right)^{r}\cdot\left(\frac{m}{\beta}\right)^{\beta} \leq 2^{O(r)}\left(\frac{k^2 r}{m}\right)^{r/2}. \tag{14}
$$

Otherwise, since $2\beta \leq \alpha \leq r$,

$$
N(\alpha,\beta) \leq \left(\frac{k^2}{\beta m}\right)^{\beta}\alpha^{\alpha}(\alpha-\beta)^{r-\alpha} \leq r^r\left(\frac{k^2}{\beta m}\right)^{\beta} \leq 2^{O(r)}\left(\frac{k^2 r}{m}\right)^{r/2}. \tag{15}
$$

Next, we assume that $k^2 < rm$, and note that whenever $\alpha > 4\beta$, then $(\alpha - 2\beta)^2 > (\alpha - \beta)$, and therefore $N(\alpha, \beta) \leq M(\alpha, \beta)$. Otherwise, $\alpha^\alpha(\alpha - \beta)^{r-\alpha} \leq 2^{O(r)}\beta^\alpha\beta^{\alpha-\beta} \leq 2^{O(r)}\beta^r$, and therefore

$$N(\alpha, \beta) \leq \left(\frac{k^2}{m\beta}\right)^\beta \cdot \beta^r \tag{16}$$

Define next $g, \hat{g} : (0, +\infty) \to \mathbb{R}$ by $g(\beta) = \beta^r \left(\frac{k^2}{\beta m}\right)^\beta$ and $\hat{g}(\beta) = \frac{r}{\beta} + \ln \frac{k^2}{\beta m} - 1$ for every $\beta > 0$. Then $g'(\beta) = g(\beta)\hat{g}(\beta)$, $g(\beta) > 0$ and $\hat{g}'(\beta) = -\frac{r}{\beta^2} - \frac{1}{\beta} < 0$ for every $\beta > 0$. Let

$$\beta_0 = \frac{r}{\ln \frac{emr}{k^2}} \quad , \quad \beta_1 = \frac{r}{\ln \frac{emr}{k^2 \ln \frac{emr}{k^2}}} .$$

Then $0 < \beta_0 < \beta_1 \leq r$, and moreover

$$\hat{g}(\beta_0) = \frac{r}{\frac{r}{\ln \frac{emr}{k^2}}} + \ln \frac{k^2}{\frac{r}{\ln \frac{emr}{k^2}} m} - 1 = \ln \ln \frac{emr}{k^2} \geq \ln \ln e = 0 ,$$

where the last inequality is due to the fact that $k^2 < mr$. In addition,

$$\hat{g}(\beta_1) = \frac{r}{\frac{r}{\ln \frac{emr}{k^2 \ln \frac{emr}{k^2}}}} + \ln \frac{k^2}{\frac{r}{\ln \frac{emr}{k^2 \ln \frac{emr}{k^2}}} m} - 1 = -\ln \ln \frac{emr}{k^2} + \ln \ln \frac{emr}{k^2 \ln \frac{emr}{k^2}} < 0$$

Therefore there exists a unique $\beta^* \in (\beta_0, \beta_1)$ such that $0 = \hat{g}(\beta^*) = \frac{r}{\beta^*} + \ln \frac{k^2}{\beta^* m} - 1$, which in turn implies $\frac{k^2}{\beta^* m} = e^{1 - r/\beta^*}$. Moreover, and for all $\beta > 0$,

$$g(\beta) \leq g(\beta^*) = (\beta^*)^r \left(\frac{k^2}{\beta^* m}\right)^{\beta^*} \leq \left(\frac{r}{\ln \frac{emr}{k^2 \ln \frac{emr}{k^2}}}\right)^r .$$

Since $\frac{emr}{k^2} > e$, $\ln \ln \frac{emr}{k^2} < \frac{1}{2} \ln \frac{emr}{k^2}$, and we have $N(\alpha, \beta) \leq 2^{O(r)} \left(\frac{r}{\ln \frac{emr}{k^2}}\right)^r$. ☐

We now turn to prove Lemma 3.

Proof of Lemma 3. Assume first that $k \geq mr$. Then by Claim 17 and Lemmas 18, 19 we get that

$$\|X\|_r^r \leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} (M(\alpha, \beta) + N(\alpha, \beta)) \leq \frac{2^{O(r)}}{k^r} \left(\frac{k^2 r}{m}\right)^{r/2} ,$$

and therefore $\|X\|_r = O\left(\sqrt{\frac{r}{m}}\right)$.

Next, assume that $mr > k \geq \sqrt{mr}$. Once again by Claim 17 and Lemmas 18, 19 we get that

$$\begin{aligned}\|X\|_r^r &\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} (M(\alpha,\beta)+N(\alpha,\beta)) \\ &\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \max\left\{ \left(\frac{r}{\ln \frac{2emr}{k}}\right)^{2r}, \left(\frac{k^2 r}{m}\right)^{r/2} \right\} + \left(\frac{k^2 r}{m}\right)^{r/2} \qquad (17) \\ &\leq \frac{2^{O(r)}}{k^r} \max\left\{ \left(\frac{r}{\ln \frac{2emr}{k}}\right)^{2r}, \left(\frac{k^2 r}{m}\right)^{r/2} \right\} ,\end{aligned}$$

and therefore $\|X\|_r = O\left(\max\left\{\frac{r^2}{k \ln^2 \frac{2emr}{k}}, \sqrt{\frac{r}{m}}\right\}\right)$.

Finally, assume that $\sqrt{mr} > k$. Once again by Claim 17 and Lemmas 18, 19 we get that

$$\begin{aligned}\|X\|_r^r &\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} (M(\alpha,\beta)+N(\alpha,\beta)) \\ &\leq \frac{2^{O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \max\left\{ \left(\frac{r}{\ln \frac{2emr}{k}}\right)^{2r}, \left(\frac{k^2 r}{m}\right)^{r/2} \right\} + \max\left\{ M(\alpha,\beta), \left(\frac{r}{\ln \frac{emr}{k^2}}\right)^{r} \right\} \qquad (18) \\ &\leq \frac{2^{O(r)}}{k^r} \max\left\{ \left(\frac{r}{\ln \frac{2emr}{k}}\right)^{2r}, \left(\frac{k^2 r}{m}\right)^{r/2} \right\} ,\end{aligned}$$

and therefore $\|X\|_r = O\left(\max\left\{\frac{r}{k \ln \frac{emr}{k^2}}, \frac{r^2}{k \ln^2 \frac{2emr}{k}}, \sqrt{\frac{r}{m}}\right\}\right)$.

□

4.2 Lower Bounding $\|X(x^{(k)})\|_r$

We finish this section by proving Lemma 4. To this end, let $k \leq n$, and recall that by Lemma 16, since for every $j \in \operatorname{supp}(x^{(k)})$, $|x_j^{(k)}| = \|x^{(k)}\|_\infty$, then

$$\|X(x^{(k)})\|_r^r = \|x^{(k)}\|_\infty^{2r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta}{\left(\|x^{(k)}\|_\infty^2 m\right)^\alpha} \sum_{V \in \binom{[n]}{\alpha}} |\mathcal{S}_{V,\beta}| \cdot \prod_{q \in V} (x_q^{(k)})^2 .$$

For every $V \subseteq [n]$, if $V \subseteq [k]$, then $\prod_{q\in V} (x_q^{(k)})^2 = \|x^{(k)}\|_\infty^{2|V|}$, and otherwise $\prod_{q\in V} (x_q^{(k)})^2 = 0$. Substituting $\|x^{(k)}\|_\infty = \frac{1}{\sqrt{k}}$, and applying Theorem 5 we get that since $r \le k$ then

$$\begin{aligned} \|X(x^{(k)})\|_r^r &= \frac{1}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta k^\alpha}{m^\alpha} \sum_{V \in \binom{[k]}{\alpha}} |\mathcal{S}_{V,\beta}| \cdot \frac{1}{k^\alpha} \\ &\ge \frac{1}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta}{m^\alpha} \sum_{V \in \binom{[k]}{\alpha}} 2^{-O(r)} \alpha^{2\alpha} \beta^{-\beta} \left[(\alpha-2\beta)^2 + 4(\alpha-\beta)\right]^{r-\alpha} \\ &= \frac{2^{-O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \frac{m^\beta}{m^\alpha} \cdot \binom{k}{\alpha} \cdot \alpha^{2\alpha} \beta^{-\beta} \left[(\alpha-2\beta)^2 + 4(\alpha-\beta)\right]^{r-\alpha} \\ &\ge \frac{2^{-O(r)}}{k^r} \sum_{\beta=1}^{r/2} \sum_{\alpha=2\beta}^{r} \left(m\beta^{-1}\right)^\beta \left(\frac{k\alpha}{m}\right)^\alpha \left[(\alpha-2\beta)^{2r-2\alpha} + (\alpha-\beta)^{r-\alpha}\right] . \end{aligned} \tag{19}$$

Setting $\alpha = r, \beta = r/2$ we get that

$$\|X(x^{(k)})\|_r \ge \sqrt[r]{\frac{2^{-O(r)}}{k^r} \left(mr^{-1}\right)^{r/2} \left(\frac{kr}{m}\right)^r} = \Omega\left(\sqrt{\frac{r}{m}}\right) .$$

Assume next that $k \le mr$ and let $\alpha = 2 + \frac{r}{\ln\left(\frac{emr}{k}\right)}, \beta = 1$. Then $\left(\frac{k}{m(\alpha-2)}\right)^{\alpha-2} \ge 2^{-O(r)}$, and therefore

$$\begin{aligned} \|X(x^{(k)})\|_r^r &\ge \frac{2^{-O(r)}}{k^r} \cdot m \cdot \left(\frac{k\alpha}{m}\right)^\alpha (\alpha-2)^{2r-2\alpha} \\ &\ge \frac{2^{-O(r)}}{k^r} \cdot (\alpha-2)^{2r} \cdot \frac{k^2}{m} \cdot \left(\frac{k}{m}\right)^{\alpha-2} \cdot (\alpha-2)^\alpha \cdot (\alpha-2)^{-2\alpha} \\ &\ge \frac{2^{-O(r)}}{k^r} \cdot (\alpha-2)^{2r} \cdot \frac{k^2}{m} \cdot \left(\frac{k}{m(\alpha-2)}\right)^{\alpha-2} \ge \left(\frac{2^{-O(1)} r^2}{k \ln^2\left(\frac{emr}{k}\right)}\right)^r . \end{aligned} \tag{20}$$

We conclude that

$$\|X(x^{(k)})\|_r = \Omega\left(\frac{r^2}{k \ln^2\left(\frac{emr}{k}\right)}\right) .$$

Finally, assume that $k \le \sqrt{mr}$ and let $\alpha = \frac{2r}{\ln \frac{emr}{k^2}}, \beta = \frac{r}{\ln \frac{emr}{k^2}}$. Then $\left(\frac{k^2}{m\beta}\right)^\beta \ge 2^{-O(r)}$ and therefore

$$\|X(x^{(k)})\|_r \ge \sqrt[r]{\frac{2^{-O(r)}}{k^r} \left(\frac{k^2}{m\beta}\right)^\beta \beta^r} = \Omega\left(\frac{r}{k \ln\left(\frac{emr}{k^2}\right)}\right) ,$$

and the proof of Lemma 4 is now complete.

5 Empirical Analysis

The goal of the experiments is to give bounds on some of the constants hidden in the main theorem. From our experiments we conclude that for $\frac{4\lg\frac{1}{\delta}}{\varepsilon^2} \leq m < \frac{2}{\varepsilon^2\delta}$ the constant inside the Θ-notation in Theorem 2 is at least 0.725 except for very sparse vectors ($\|x\|_0 \leq 7$), where the constant is at least 0.6. Furthermore, we confirm that $\nu(m, \varepsilon, \delta) = 1$ when $m \geq \frac{2}{\varepsilon^2\delta}$ and that there exists data points where $\nu(m, \varepsilon, \delta) < 1$ while $m = \frac{2-\gamma}{\varepsilon^2\delta}$, for some small γ.

5.1 Experiment Setup and Analysis

To arrive at the results, we ran experiments and analysed the data in several phases. In the first phase we varied the target dimension m over exponentially spaced values in the range $[2^6, 2^{12}]$, and a parameter k which controls the ratio between the ℓ_∞ and the ℓ_2 norm. The values of k varied over exponentially spaced values in the range $[2^1, 2^{13}]$. Then for all m and k, we generated 2^{24} vectors x with entries in $\{0, 1\}$ such that $\|x\|_2 = \sqrt{k}\|x\|_\infty$, and for any given m and k the supports of the vectors were pairwise disjoint. We then hashed the generated vectors using feature hashing, and recorded the ℓ_2 norm of the embedded vectors.

The second phase then calculated the distortion between the original and the embedded vectors, and computed the error probability $\hat{\delta}$. Loosely speaking, $\hat{\delta}(m, k, \varepsilon)$ is the ratio of the 2^{24} vectors for a given m and k that have distortion greater than ε. Formally, $\hat{\delta}$ is calculated using the following formula

$$\hat{\delta}(m, k, \varepsilon) = \frac{\left|\left\{x : \|x\|_2 = \sqrt{k}\|x\|_\infty, \left|\|A_m x\|_2^2 - \|x\|_2^2\right| \geq \varepsilon\|x\|_2^2\right\}\right|}{\left|\left\{x : \|x\|_2 = \sqrt{k}\|x\|_\infty\right\}\right|},$$

where ε was varied over exponentially spaced values in the range $[2^{-10}, 2^{-1}]$. Note that $\hat{\delta}$ tends to the true error probability as the number of vectors tends to infinity. Computing $\hat{\delta}$ yielded a series of 4-tuples $(m, k, \varepsilon, \hat{\delta})$ which can be interpreted as given target dimension m, ℓ_∞/ℓ_2 ratio $1/\sqrt{k}$, distortion ε, we have measured that the failure probability is at most $\hat{\delta}$.

In the third phase, we varied δ over exponentially spaced values in the range $[2^{-20}, 2^0]$, and calculated a value $\hat{\nu}$. Intuitively, $\hat{\nu}(m, \varepsilon, \delta)$ is the largest ℓ_∞/ℓ_2 ratio such that for all vectors having at most this ℓ_∞/ℓ_2 ratio the measured error probability $\hat{\delta}$ is at most δ. Formally,

$$\hat{\nu}(m, \varepsilon, \delta) = \max\left\{\frac{1}{\sqrt{k}} : \forall k' \geq k, \hat{\delta}(m, k', \varepsilon) \leq \delta\right\}.$$

Note once more that $\hat{\nu}$ tends to the true ν value as the number of vectors tends to infinity.

To find a bound on the constant of the Θ-notation in Theorem 2, we truncated data points that did not satisfy $\frac{4\lg\frac{1}{\delta}}{\varepsilon^2} \leq m < \frac{2}{\varepsilon^2\delta}$, and for the remaining points we plotted $\hat{\nu}$ over the theoretical bound in Figure 1:

$$\frac{\hat{\nu}(m, \varepsilon, \delta)}{\min\left\{\frac{\sqrt{\varepsilon}\lg\frac{\varepsilon m}{\lg\frac{1}{\delta}}}{\lg\frac{1}{\delta}}, \sqrt{\frac{\varepsilon\lg\frac{\varepsilon^2 m}{\lg\frac{1}{\delta}}}{\lg\frac{1}{\delta}}}\right\}}.$$

From this plot we conclude that the constant is at least 0.6 on the large range on parameters we tested. However, the smallest values seem to be outliers and come from a combination of

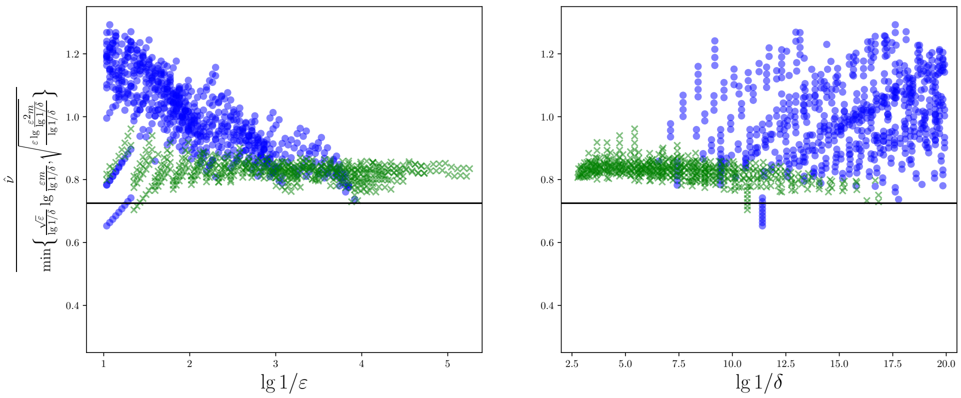

Figure 1: This plot shows the measured $\hat{\nu}$ values over the theoretical bound (abbreviated here): $\min\{\mathsf{left}, \mathsf{right}\}$. This ratio corresponds to the constant in the Θ-notation in Theorem 2. The points are marked with blue circles if $\mathsf{left} < \mathsf{right}$, otherwise they are marked with green $\times$'s. The horizontal line at 0.725 is there to ease comparisons with Figure 2. The data points below the line come from very sparse vectors ($k = 7$) with high target dimension ($m = 2^{14}$).

very sparse vectors ($k = 7$) and high target dimension ($m = 2^{14}$). For the rest of the data points the constant is at least 0.725. While there are data points where the constant is larger (i.e. feature hashing performs better), there are data points close to 0.725 over the entire range of ε and δ.

In Figure 2 we show that we indeed need both terms in the minimum in Theorem 2, by plotting the measured $\hat{\nu}$ values over both terms in the minimum in the theoretical bound separately. For both terms there are points whose value is significantly below 0.725.

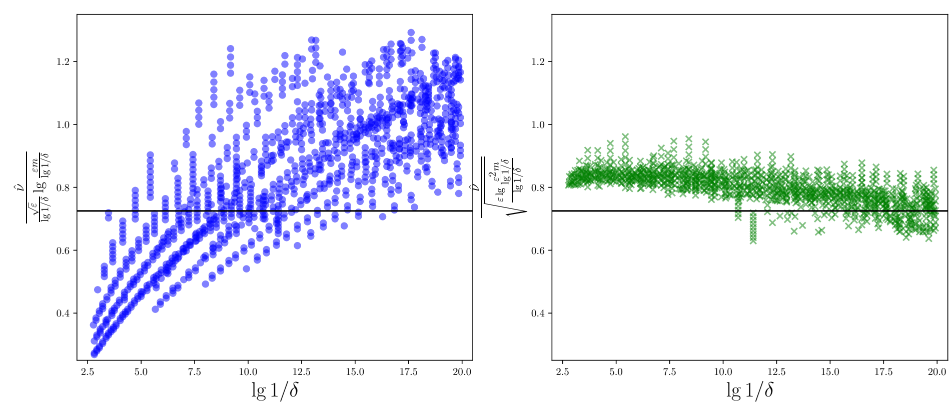

Figure 2: This plot shows the measured $\hat{\nu}$ values over each of the two terms in the minimum in the theoretical bound (abbreviated here): $\min\{\mathsf{left}, \mathsf{right}\}$. In the left subfigure the y-axis of the blue circles is $\frac{\hat{\nu}}{\mathsf{left}}$, while the y-axis of the green $\times$'s in the right subfigure is $\frac{\hat{\nu}}{\mathsf{right}}$. Note that the x-axis (values of $\lg(1/\delta)$) is the same in both subfigures, and the same as in the right subfigure of Figure 1. As in Figure 1, the horizontal line at 0.725 is there to ease comparison between the figures.

To find a bound on m where $\hat{\nu}(m, \varepsilon, \delta) = 1$ we took the untruncated data and recorded

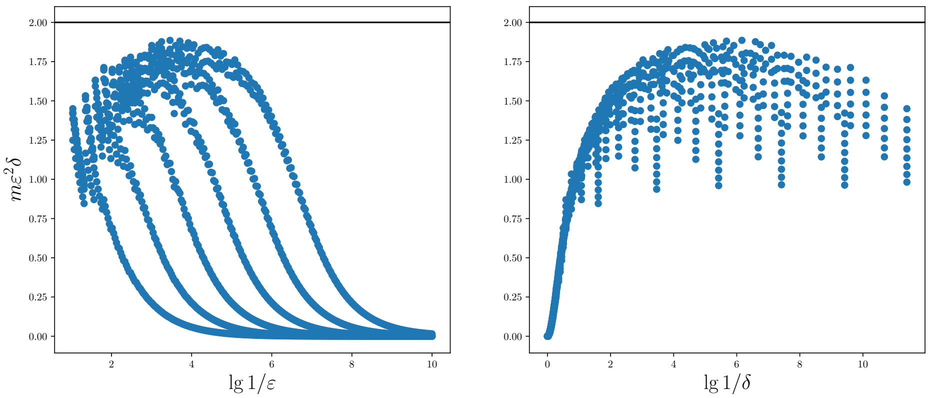

Figure 3: This plot shows the constant border where $\hat{\nu}(m, \varepsilon, \delta)$ becomes 1 for the first time. The theory states that if $2 \leq m\varepsilon^2\delta$ then $\hat{\nu}(m, \varepsilon, \delta) = 1$. The distinct curves in the left plot correspond to distinct values of m.

the maximal $\hat{\delta}$ for each m and ε. We then plotted $m\varepsilon^2\hat{\delta}$ in Figure 3. From Figure 3 it is clear that $\hat{\nu}(m, \varepsilon, \delta) = 1$ when $m \geq \frac{2}{\varepsilon^2\delta}$. Furthermore, the figure also shows that there are data points where $\hat{\nu}(m, \varepsilon, \delta) < 1$ while $m = \frac{2-\gamma}{\varepsilon^2\delta}$, for some small γ. Therefore we conclude the bound $m \geq \frac{2}{\varepsilon^2\delta}$ is tight.

5.2 Implementation Details

As random number generators, we used degree 20 polynomials modulo the Mersenne prime $2^{61} - 1$, where the coefficients were random data from `random.org`. The random data was independent between experiments with diffent values of m, and between the random number generator used for vector generation and hashing.

Feature hashing was done using double tabulation hashing [Tho14] on 64 bit numbers. The tables in our implementation of double tabulation hashing were filled with numbers from the aforementioned random number generator. Double tabulation hashing has been proven to behave fully randomly with high probability [DKRT15].

References

[AC09] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[Ach03] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, June 2003.

[CCF04] M. Charikar, K. C. Chen, and M. Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[CJN18] M. B. Cohen, T. S. Jayram, and J. Nelson. Simple analyses of the sparse Johnson-Lindenstrauss transform. In *1st Symposium on Simplicity in Algorithms, SOSA 2018, January 7-10, 2018, New Orleans, LA, USA*, pages 15:1–15:9, 2018.

[CT06] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[CW09] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, pages 205–214. ACM, 2009.

[Dal13] B. Dalessandro. Bring the noise: Embracing randomness is the key to scaling up machine learning algorithms. *Big Data*, 1(2):110–112, 2013.

[DG03] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[DKRT15] S. Dahlgaard, M. B. T. Knudsen, E. Rotenberg, and M. Thorup. Hashing for statistics over k-partitions. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 1292–1310. IEEE Computer Society, 2015.

[DKS10] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, STOC '10, pages 341–350, 2010.

[DKT17] S. Dahlgaard, M. Knudsen, and M. Thorup. Practical hash functions for similarity estimation and dimensionality reduction. In *Advances in Neural Information Processing Systems 30*, pages 6615–6625. Curran Associates, Inc., 2017.

[FL17] C. B. Freksen and K. G. Larsen. On using toeplitz and circulant matrices for Johnson-Lindenstrauss transforms. In *28th International Symposium on Algorithms and Computation, ISAAC 2017, December 9-12, 2017, Phuket, Thailand*, pages 32:1–32:12, 2017.

[HIM12] S. Har-Peled, P. Indyk, and R. Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.

[HWB08] C. Hegde, M. Wakin, and R. Baraniuk. Random projections for manifold learning. In *Advances in Neural Information Processing Systems 20*, pages 641–648. Curran Associates, Inc., 2008.

[JL84] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[JW13] T. S. Jayram and D. P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):26:1–26:17, 2013.

[KN14] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4:1–4:23, January 2014.

[LN17] K. G. Larsen and J. Nelson. Optimality of the Johnson-Lindenstrauss lemma. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 633–638, 2017.

[Mat08] J. Matoušek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[MM09] O. Maillard and R. Munos. Compressed least-squares regression. In *Advances in Neural Information Processing Systems 22*, pages 1213–1221. Curran Associates, Inc., 2009.

[MM13] X. Meng and M. W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, STOC '13, pages 91–100, 2013.

[NN13] J. Nelson and H. L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 101–110. ACM, 2013.

[PBMID14] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas. Random projections for linear support vector machines. *ACM Trans. Knowl. Discov. Data*, 8(4):22:1–22:25, 2014.

[RR08] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

[Sár06] T. Sárlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152. IEEE Computer Society, 2006.

[Sut15] S. Suthaharan. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning.* Springer Publishing Company, Incorporated, 1st edition, 2015.

[Tho14] M. Thorup. Simple tabulation, fast expanders, double tabulation, and high independence. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science(FOCS)*, pages 90–99. IEEE Computer Society, 2014.

[TZ12] M. Thorup and Y. Zhang. Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.*, 41(2):293–331, April 2012.

[Vem05] S. S. Vempala. *The random projection method.* DIMACS : series in discrete mathematics and theoretical computer science. American Mathematical Society, 2005.

[WDL+09] 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 '09, pages 1113–1120, 2009.