Dimensionality Reduction on the Simplex

Rasmus J. Kyng Jeff M. Phillips Suresh Venkatasubramanian

Abstract

For many problems in data analysis, the natural way to model objects is as a probability distribution over a finite and discrete domain. Probability distributions over such domains can be represented as points on a (high-dimensional) simplex, and thus many inference questions involving distributions can be viewed geometrically as manipulating points on a simplex. The dimensionality of these points makes analysis difficult, and thus a key goal is to reduce the dimensionality of the data *while still preserving the distributional structure*.

In this paper, we propose an algorithm for dimensionality reduction on the simplex,mapping a set of high-dimensional distributions to a space of lower-dimensional distributions, whilst approximately preserving the pairwise Hellinger distance between distributions. By introducing a restriction on the input data to distributions that are in some sense quite smooth, we can map n points on the d-simplex to the simplex of $O(\varepsilon^{-2}\log n)$ dimensions with ε-distortion with high probability. Our techniques rely on classical Johnson and Lindenstrauss dimensionality reduction methods for Euclidean point sets and require the same number of random bits as non-sparse methods proposed by Achlioptas for database-friendly dimensionality reduction.

1 Introduction

In many applications, data is represented natively not as a vector in a normed space, but as a distribution over a finite and discrete support. A document (or even a topic) is represented as a distribution over words [27, 21, 9], an image is represented as distribution over scale-invariant fingerprints [25, 12], and audio signals are represented as distributions over frequencies [16]. This *probabilistic* view of the data is important for learning and inference, as well as information-theoretic approaches to data mining.

These distributions are defined over very large supports – a document vector might have hundreds of dimensions corresponding to different words in a vocabulary. This high dimensionality poses the usual challenges for data analysis, and dimensionality reduction is a standard tool one might apply in order to process the data feasibly.

However, traditional dimensionality reduction cannot be applied in these settings for two reasons. First, the *distributional structure* of the data must be preserved. Techniques like the Johnson-Lindenstrauss transform, or more general metric embedding methods, can be used to embed high dimensional vectors in a Euclidean space (or a normed space in general). However, they do not ensure that the resulting objects are also probability distributions. This is important because the inference procedures we wish to apply (Bayesian methods, or information-theoretic analysis) heavily utilize the distributional nature of the data in addition to its metric structure, and losing the former renders these procedures invalid. Secondly, the natural distance measures used to compare distributions are not the traditional ℓ_p-induced distances. Rather, they are either the non-metric Bregman divergences, or (the object of study here), distances like the Hellinger distance, which can be interpreted statistically as capturing the information distance between probability distributions.

A probability distribution over finite support can be represented as a point on a (high-dimensional) simplex. Thus, the problem of dimensionality reduction for distributions can be phrased as the problem of doing dimensionality reduction on the simplex under the Hellinger distance. This is the problem we study in this paper.

1.1 Background

Metric embeddings and dimensionality reduction are crucial tools in the understanding of high dimensional data and have been objects of intense study for nearly two decades in theoretical computer science [17]. In particular, dimensionality reduction in Euclidean spaces has been studied and applied not just in the algorithms community, but also in machine learning, computer vision, and natural language processing. The Johnson-Lindenstrauss lemma [22] is the prototype for the result we prove in this paper; it was originally shown for finite subsets of a Hilbert space, and in addition to admitting multiple proofs, has also been extended to the sphere [2], as well as general manifolds [11, 7, 3]. *Structure-preserving* embeddings have also been studied. For example, there are results on embedding (shortest path metrics on) graphs of large genus onto graphs of small genus (or even planar graphs) [20, 10, 28].

Dimensionality reduction techniques for information distances have received little attention. Since the Hellinger distance maps isometrically into an ℓ_2 space (see Section 2), it is easy [8] to embed the simplex with Hellinger distance in a low-dimensional *Euclidean* space. However this embedding cannot be directly used to obtain a mapping to a low-dimensional simplex.

1.2 Overview and Paper Outline

Dimensionality reduction on the simplex can be reduced to a different question that itself is of interest: "Given a collection of points in the positive orthant of $\mathbb{R}^d$, embed them into the *positive orthant* of $\mathbb{R}^k, k \ll d$". Traditional dimensionality reduction proceeds by constructing a random projection from $\mathbb{R}^d$ to $\mathbb{R}^k$. This projection can be constructed in many different ways [23, 24, 26, 1, 18, 13, 6, 5, 4, 15, 13, 14], but a key feature is that it is expansive, spreading points all over the lower dimensional space. It turns out that this expansivity (captured by the fact that coordinates of the resulting points must be permitted to take both positive and negative values) is critical to guaranteeing an unbiased estimator of the distance (which in turn then yields a low-dimensional embedding via probability amplification).

When we move to the positive orthant, it is no longer clear how to achieve the twin goals of unbiasedness and "non-expansivity". Simple ideas like truncation or wrap-arounds introduce bias that is hard to analyze. Further, by work of Matousek[26] and Naor and Indyk[19], it is know that any family of projections that preserves distances must satisfy strong higher moment properties. Our solution is a compromise. We define a family that by construction guarantees positivity and that admits the "right" higher moment behavior using a delicate geometric argument. The price we pay is that the approach must be restricted to points that lie in an interior region of the simplex.

We lay out basic definitions in the next section. Theorem 3.1 is our main result on dimensionality reduction for the simplex and is presented in Section 3. Underpinning this result is a key lemma (Lemma 3.4) that characterizes the behavior of the family of projections from high-dimensional to low-dimensional positive orthants. We state this lemma and use it in Section 3, and prove it in Section 4.

2 Definitions

The (d-1)-dimensional *simplex* $\Delta^{d-1} \subset \mathbb{R}^d$ is the set of all points $\{(x_1, \ldots, x_d) \mid \sum x_i = 1, x_i \geq 0\}$. The (d-1)-dimensional *unit sphere* $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ is the set of all points $\{(x_1, \ldots x_d) \mid \sum x_i^2 = 1\}$. For notational convenience, we will use Δ and $\mathbb{S}$ to denote the simplex and unit sphere respectively when the dimension is either irrelevant or implied. We will also use $\mathbb{R}^d_+$ to denote the *positive orthant* of $\mathbb{R}^d$, i.e the set of all points $\{(x_1, \ldots x_d) \mid x_i \geq 0\}$. Let $\mathbb{S}^{d-1}_+ = \mathbb{S}^{d-1} \cap \mathbb{R}^d_+$. The *Hellinger distance* $d_H : \Delta \times \Delta \to \mathbb{R}^+$ is a commonly used distance measure between distributions, defined as

$$\mathrm{d_H}(p,q)^2 = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2.$$

There is a natural mapping between Δ^{d-1} and $\mathbb{S}_+^{d-1}$, which also establishes a connection between the Hellinger distance and Euclidean distance. Let the map $h : \Delta^{d-1} \to \mathbb{S}^{d-1}$ be defined by $h(p_1,\ldots,p_d) = (\sqrt{p_1},\ldots,\sqrt{p_d})$. By construction, $h(p)$ is a point on $\mathbb{S}_+^{d-1}$. Further, h is a bijection between Δ^{d-1} and $\mathbb{S}_+^{d}$ since it has an inverse $h^{-1} : \mathbb{S}_+^{d-1} \to \Delta^{d-1}$ given by $h^{-1}(q_1,\ldots,q_d) = (q_1^2,\ldots,q_d^2)$.

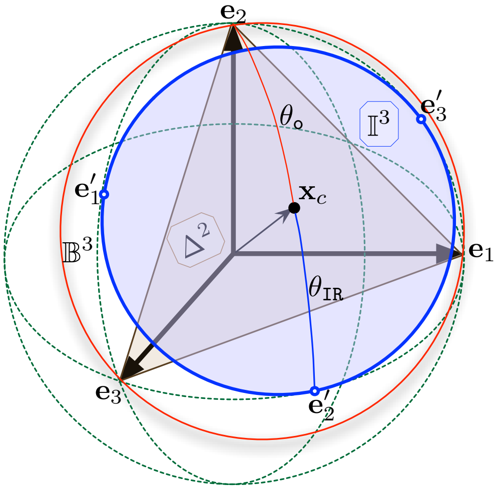

Figure 1: Illustration of Δ^2, $\mathbb{B}^3$, $\mathbb{I}^3$.

Inner region and possible basis set. We refer to the point $\mathbf{x}_c = \left(\frac{1}{\sqrt{d}},\ldots,\frac{1}{\sqrt{d}}\right)$ as the *center point* of the positive orthant. Note that $\mathbf{x}_c$ lies on $\mathbb{S}_+^{d-1}$.

Let θ_o be the angle from vx_c to a basis vector with the i^{th} coordinate being 1, $\mathbf{e}_i = (0,\ldots,0,1,0,\ldots,0)$, i.e. $\mathbf{e}_i$ is a basis vector in the standard orthonormal basis of $\mathbb{R}^d$. Thus, $\cos\theta_o = \mathbf{x}_c \cdot \mathbf{e}_i = \frac{1}{\sqrt{d}}$. Call the set of unit vectors at angle θ_o to $\mathbf{x}_c$ the *possible basis set*, which we denote $\mathbb{B}^d \subset \mathbb{R}^d$.

Let $\mathbf{e}_i'$ be the *opposing vector to* $\mathbf{e}_i$, given by $\mathbf{e}_i' = \frac{1}{\sqrt{d-1}}\sum_{j\neq i}\mathbf{e}_j$ and let the angle $\theta_{\mathtt{IR}}$ be the angle from $\mathbf{x}_c$ to some $\mathbf{e}_i'$, so that $\cos\theta_{\mathtt{IR}} = \mathbf{x}_c \cdot \mathbf{e}_i' = (d-1)\cdot\frac{1}{\sqrt{d}}\cdot\frac{1}{\sqrt{d-1}} = \sqrt{\frac{d-1}{d}}$. The *inner region* $\mathbb{I}^d \subset \mathbb{S}_+^{d-1}$ is the set of all unit vectors of angle $\theta_{\mathtt{IR}}$ or less to $\mathbf{x}_c$. We also call all points $\mathbf{p} \in \Delta^{d-1}$ that correspond to vectors $\mathbf{v}_p$ at angle at most $\theta_{\mathtt{IR}}$ from $\mathbf{x}_c$ the *inner region of the simplex* and denote the set as $\mathbb{I}(\Delta^{d-1})$; these are in general not unit vectors.

Both the boundary of the inner region $\mathbb{I}^d$ and the possible basis $\mathbb{B}^d$ have the shape of a lower-dimensional sphere, as given by the following lemma.

Lemma 2.1. *Let* $\mathbf{u} \in \mathbb{S}^{d-1}$ *be a vector. The set* $X \subset \mathbb{S}^{d-1}$ *of vectors at angle* θ *to* $\mathbf{u}$ *forms a sphere of dimension* $d-2$ *and radius* $\sin\theta$ *in the subspace orthogonal to* $\mathbf{u}$.

Proof. Let $\mathbf{u}$ and $\mathbf{r}$ be unit vectors with angle θ between them. The dot product does not depend on the orientation of the basis used to give the coordinates of the vectors, so w.l.o.g. we can chose any orthonormal basis such that $\mathbf{u} = (1,0,\ldots,0)$ and $\mathbf{r} = (r_1,\ldots,r_n)$. Thus in this basis we have $\cos\theta = \mathbf{u}\cdot\mathbf{r} = r_1$. Since $\mathbf{r}$ is a unit vector, we know that $r_1^2+\ldots+r_n^2 = 1$, so $r_2^2+\ldots+r_n^2 = 1-\cos^2\theta = \sin^2\theta$. So coordinates r_2 through r_n obey the equation of an (d-2)-sphere of radius $\sin\theta$. □

Distortion. One last definition is necessary before we start looking at results. We use this to express the degree of distortion on a set of points mapped from one space to another. Let $P \subset X$ be a set of points in a metric space (X,d_X). Let $f : X \to Y$ be a map to another metric space (Y,d_Y). We say that $Q = \{f(p)|p \in P\}$ has ε*-distortion* if there exists a constant c such that

$$\forall p,p' \in P \quad (1-\varepsilon)d_Y(f(p),f(p')) \le c\cdot d_X(p,p') \le (1+\varepsilon)d_Y(f(p),f(p')). \tag{2.1}$$

Our main theorem will be about metric spaces using Hellinger distance d_H, but all intermediate results will be with respect to the Euclidean distance d_E defined as $d_E(p,q) = \sqrt{\sum_{i=1}^{d}(p_i-q_i)^2}$.

3 Main Results

In this section, we establish the main result of this paper.

Theorem 3.1. *Let $P \subset \mathbb{I}(\Delta^{d-1})$ be a set of n points in metric space (Δ^{d-1}, d_H). For constants $0 < \varepsilon < 2$ and $0 < \delta < 1$ and integer $k = O\left(\frac{1}{\varepsilon^2}\ln\left(\frac{n}{\delta}\right)\right)$, there exists a set of maps $\mathcal{G}$ such that when $g : \Delta^{d-1} \to \Delta^{k-1} \in \mathcal{G}$ is picked uniformly from $\mathcal{G}$, the image of P under g has ε-distortion in (Δ^{k-1}, d_H) with probability at least $1 - \delta$.*

3.1 Overview Of Proof

The construction of the family $\mathcal{G}$ proceeds in four steps. First, points are mapped from Δ^{d-1} to $\mathbb{S}^{d-1}$ via the mapping h. Note that the resulting points lie in $\mathbb{R}^d_+$. In the second step, we (randomly) construct a mapping f that projects the points from $\mathbb{R}^{d+1}_+$ to $\mathbb{R}^{k+1}_+$. Third, these points are then projected onto $\mathbb{S}^{k-1}$ using the normalization $S(p) = \frac{p}{\|p\|}$. Finally, since the points now lie on $\mathbb{S}^{k-1}_+$, we can apply the inverse mapping h^{-1} to the simplex Δ^{k-1}. The overall mapping is thus the composition $g = h^{-1} \circ S \circ f \circ h$.

The mappings h and h^{-1} are isometric and introduce no distortion on the distances. The following lemma shows that under the assumption that f only introduces ε distortion, the mapping S introduces only a small overall distortion.

Lemma 3.1. *For $P \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$ and $0 \le \varepsilon \le 1/2$, let $\pi : \mathbb{R}^d \to \mathbb{R}^k$ be an $(\varepsilon/4)$-distortion embedding of $(P \cup \{0\}, d_E)$. Then $\psi = S \circ \pi : \mathbb{S}^{d-1} \to \mathbb{S}^{k-1}$ has ε-distortion on (P, d_E).*

Proof. For all $p, p' \in P$, the maximum value of $||\psi(p) - \psi(p')||$ occurs when $||\pi(p)||$ and $||\pi(p')||$ are as small as possible and $||\pi(p) - \pi(p')||$ is as large as possible [2]. Similarly, the smallest value of $||\psi(p) - \psi(p')||$ occurs when $||\pi(p)||$ and $||\pi(p')||$ are as large as possible and $||\pi(p) - \pi(p')||$ is as small as possible. Thus

$$(1-\varepsilon/2)\|p-p'\| \le \frac{1-\varepsilon/4}{1+\varepsilon/4}\|p-p'\| \le \|\psi(p)-\psi(p')\| \le \frac{1+\varepsilon/4}{1-\varepsilon/4}\|p-p'\| \le (1+\varepsilon)\|p-p'\|.$$

The first and last inequality follow by $1 - 2\alpha \le \frac{1-\alpha}{1+\alpha}$ and from $\frac{1+\alpha}{1-\alpha} \le 1 + 4\alpha$, respectively, for $0 < \alpha < 1/2$ and setting $\alpha = \varepsilon/4$. These are easily proven by showing that $\frac{1+\alpha}{1-\alpha}$ is a convex function in α which intersects the line $1 + 2\alpha$ only at $\alpha = 0$ and the line $1 + 4\alpha$ at $\alpha = 0$ and $\alpha = 1/2$. □

3.2 Projections that preserve positivity

To complete the proof of Theorem 3.1, we need to construct a mapping $f : \mathbb{R}^d_+ \to \mathbb{R}^k_+$ that has ε-distortion under d_E. Our strategy is to pick k random vectors $\mathbf{r}_1, \ldots, \mathbf{r}_k$ from some distribution over $\mathbb{B}^d$, and define the map $f : \mathbb{R}^d \to \mathbb{R}^k$ by $f(\mathbf{v}) = \sqrt{\frac{d}{k}}(\mathbf{v} \cdot \mathbf{r}_1, \ldots, \mathbf{v} \cdot \mathbf{r}_k)$. This strategy is justified by the following lemma.

Lemma 3.2. *The angle between any $\mathbf{v} \in \mathbb{B}^d$ and any vector $\mathbf{u} \in \mathbb{I}^d$ is at most $\pi/2$.*

Proof. The center point does not change with a rotation of the orthonormal basis about $\mathbf{x}_c$; $\mathbb{B}^d$ and $\mathbb{I}^d$ are also invariant. Thus, w.l.o.g. consider $\mathbf{e}_1 = \mathbf{v}$, rather than a general vector from $\mathbb{B}^d$. The angle between $\mathbf{e}_1$ and $\mathbf{e}_1'$ is $\pi/2$ since $\mathbf{e}_1 \cdot \mathbf{e}_1' = 0$. The set of all vectors $\mathbf{t}$ that have angle with $\mathbf{e}_1$ less than or equal to $\pi/2$ form a hemisphere (hemihypersphere), and the positive orthant $\mathbb{S}^{d-1}_+$ is contained in this hemisphere; see Figure 1. Since $\mathbb{I}^d \subset \mathbb{S}^{d-1}_+$, this concludes the proof. □

Lemma 3.2 proves that if $\mathbf{u} \in \mathbb{I}^d$, then $f(\mathbf{u}) \in \mathbb{R}^k_+$, since every dot product $\mathbf{v} \cdot \mathbf{r}_i$ will be non-negative. This will lead us to a proof of the following theorem:

Theorem 3.2. *Let $P \subset \mathbb{R}^d$ be an arbitrary set of n points. For any $0 < \varepsilon < 1/2$ and $0 < \delta < 1$ and for integer $k \ge O\left(\frac{1}{\varepsilon^2}\ln\left(\frac{n}{\delta}\right)\right)$, there exists a set of maps $\mathcal{F}$ such that when $f : \mathbb{R}^d \to \mathbb{R}^k \in \mathcal{F}$ is picked uniformly from $\mathcal{F}$, the image of P under f has ε-distortion with probability at least $1 - \delta$.*

Furthermore, for any point $\mathbf{p} \in \mathbb{I}^d \subset \mathbb{R}^d_+$, then $f(\mathbf{p}) \in \mathbb{R}^k_+$.

To prove this, we use the following two lemmas. Lemma 3.3 is based heavily on a lemma by Achlioptas [1, Lemma 5.1]; we state it in a slightly more general form, and for the lower bound we simplify the proof a little using techniques developed by Matousek [26]. The upper bound proof remains unchanged. The proof of Lemma 3.3 can be found in Appendix A.

Lemma 3.3. *Let* $T \overset{D}{=} N(0,1/d)$ *(normally distributed with* $1/d$ *standard deviation) and let* $\mathbf{r}$ *be a random vector which for any unit vector* $\mathbf{u} \in \mathbb{R}^d$ *satisfies* $\mathrm{E}\left[(\mathbf{u}\cdot\mathbf{r})^2\right] = \frac{1}{d}$ *and for integer* $m \geq 2$

$$\mathrm{E}\left[(\mathbf{u}\cdot\mathbf{r})^{2m}\right] \leq \mathrm{E}\left[T^{2m}\right] = (2m-1)!!/d^m,$$

where $(2m-1)!! = \frac{(2m)!}{m!2^m}$ *is the double factorial. Pick k random, independently distributed vectors* $\mathbf{r}_1,\ldots\mathbf{r}_k$ *from some distribution satisfying this condition, and define the map* $f:\mathbb{R}^d \to \mathbb{R}^k$ *by* $f(\mathbf{v}) = \sqrt{\frac{d}{k}}(\mathbf{v}\cdot\mathbf{r}_1,\ldots,\mathbf{v}\cdot\mathbf{r}_k)$*. Then for* $0 < \varepsilon < 1/2$

$$\Pr\left[||f(\mathbf{v})||^2 < (1-\varepsilon)||\mathbf{v}||^2\right] \leq \exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)\right)$$

and

$$\Pr\left[||f(\mathbf{v})||^2 > (1+\varepsilon)||\mathbf{v}||^2\right] \leq \exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)\right).$$

The key lemma in this paper shows how to construct a distribution over the possible basis set that satisfies the requirements of Lemma 3.3. An orthonormal basis where $\mathbf{x}_c = (1,0,0,\ldots 0)$ (not the standard basis of $\mathbb{R}^d$) is called a *central basis*.

Lemma 3.4. *In a* central basis *we define a random vector* $\mathbf{r} = \left(\cos\theta_o, \sin\theta_o\cdot\frac{X_2}{\sqrt{d-1}}, \ldots,\sin\theta_o\cdot\frac{X_d}{\sqrt{d-1}}\right)$*, where the* X_i *are i.i.d. from* $\{-1,+1\}$*. The vector* $\mathbf{r}\in\mathbb{B}^d$ *and for any unit vector* $\mathbf{u}\in\mathbb{R}^d$ *satisfies* $\mathrm{E}\left[(\mathbf{u}\cdot\mathbf{r})^2\cdot d\right] = 1$ *and for integer* $m \geq 2$

$$\mathrm{E}\left[(\mathbf{u}\cdot\mathbf{r})^{2m}\right] \leq \mathrm{E}\left[T^{2m}\right] = (2m-1)!!/d^m.$$

The proof of this lemma is delicate, requiring a careful balancing of moments. It is proved in Section 4. We now show how the above lemmas can be used to prove Theorem 3.2 by a straightforward application of the tail bounds given in Lemma 3.3.

Proof of Theorem 3.2. Take $\mathbf{r}$ to be a random vector as defined in Lemma 3.4 and let the map $f:\mathbb{R}^d \to \mathbb{R}^k$ be defined as in lemma 3.3 using this $\mathbf{r}$. Since $\mathbf{r}\in\mathbb{B}^d$, if a vector $\mathbf{v}\in\mathbb{I}^d$, then for each $\mathbf{r}_i$ of f, $\mathbf{v}\cdot\mathbf{r}_i \geq 0$. Consequently, any point $p\in\mathbb{I}^d\subset\mathbb{R}^d$ is mapped to $\mathbb{R}^k_+$ by f.

Also, by lemma 3.3, we can bound the probability that equation 2.1 does not hold for all pairwise distances by using a trivial union bound over the pairwise distance vectors. There are $\binom{n}{2}$ such vectors, and the total probability that 2.1 fails should be at most δ. Hence

$$\binom{n}{2}\cdot 2\exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)\right) \leq 2n^2\exp\left(-\frac{k\varepsilon^2}{12}\right) \leq \delta$$

Setting $k = (12/\varepsilon^2)\ln(2n^2/\delta) = O((1/\varepsilon^2)\ln(n/\delta))$ completes the proof. □

3.3 Putting It All Together

Assembling all the pieces, we can now prove Theorem 3.1.

Proof of Theorem 3.1. Let $\varepsilon' = \varepsilon/4$. Thus, when $0 < \varepsilon < 2$, we get $0 < \varepsilon' < 1/2$. Thus by Theorem 3.2 there exists constants such that for integer k where

$$k \geq O\left(\frac{1}{(\varepsilon')^2} \ln\left(\frac{n+1}{\delta}\right)\right) = O\left(\frac{1}{\varepsilon^2} \ln\left(\frac{n}{\delta}\right)\right),$$

the image of $h(P) \cup \{0\}$ under $f : \Delta^{d-1} \to \mathbb{R}^k$ has $\varepsilon/4$-distortion with probability $1-\delta$, since h does not give any distortion. This in turn implies that the distortion of P under $S \circ f \circ h$ is at most ε with the same probability by Lemma 3.1. Since h maps points from $\mathbb{I}(\Delta^{d-1})$ to $\mathbb{I}^d$, and f maps points in $\mathbb{I}^d$ to $\mathbb{R}^k_+$, and S does not change the sign of any coordinates of a vector, we see that $S \circ f \circ h$ maps points from $\mathbb{I}(\Delta^{d-1})$ to $\mathbb{S}^{k-1}_+$. Hence, image of P under $S \circ f \circ h$ lies in the domain of h^{-1} and so $h^{-1} \circ S \circ f \circ h : \mathbb{I}(\Delta^{d-1}) \to \Delta^{k-1}$ maps points from the inner region of the $(d-1)$-simplex $\mathbb{I}(\Delta^{d-1})$ to the $(k-1)$-simplex Δ^{k-1}, and has ε-distortion with probability at least $1-\delta$. □

4 Moment Bounds

In this section, we prove lemma 3.4. Before doing so, we consider the motivation behind the construction of the random vector used in the lemma. The *central basis* lets us decompose the space $\mathbb{R}^d$ as one dimension parallel to $\mathbf{x}_c$ and an orthogonal $(d-1)$-dimensional subspace. By generating random unit vectors in this subspace, we can construct random unit vectors in the *possible basis set* $\mathbb{B}^d$ very easily, and given our particular choice of θ_o, this construction has a very nice property, as expressed in the following lemma.

Lemma 4.1. *If* $\mathbf{r}' = (r_1, \ldots r_{d-1})$ *is a random unit vector in the subspace V orthogonal to* $\mathbf{x}_c$*, then* $\mathbf{r} = (\cos\theta_o, \sin\theta_o r_1, \ldots, \sin\theta_o r_{d-1})$ *is a random unit vector in* $\mathbb{B}^d$*. For any unit vector* $\mathbf{u}' \in V$*, if* $\mathrm{E}[\mathbf{u}' \cdot \mathbf{r}'] = 0$ *and* $\mathrm{E}\left[(\mathbf{u}' \cdot \mathbf{r}')^2\right] = \frac{1}{d-1}$*, then for an* $\mathbf{u} \in \mathbb{S}^{d-1}$ $\mathrm{E}\left[(\mathbf{u} \cdot \mathbf{r})^2\right] = \frac{1}{d}$*.*

Proof. The first statement is true, since $\mathbf{r} \cdot \mathbf{x}_c = \cos\theta_o$ and $||\mathbf{r}||^2 = 1$. For the second statement, note that for some γ any unit vector $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ can be rewritten as $\mathbf{u} = (\cos\gamma, \sin\gamma u'_1, \ldots, \sin\gamma u'_{d-1})$ where $\mathbf{u}' = (u'_1, \ldots, u'_{d-1})$ is a unit vector in the subspace V orthogonal to $\mathbf{x}_c$. Hence

$$\begin{aligned}
\mathrm{E}\left[(\mathbf{u} \cdot \mathbf{r})^2\right] =& \mathrm{E}\left[\left(\cos\theta_o \cos\gamma + \sin\theta_o \sin\gamma(\mathbf{u}' \cdot \mathbf{r}')\right)^2\right] \\
=& \mathrm{E}\left[\left(\frac{1}{\sqrt{d}}\cos\gamma + \sqrt{\frac{d-1}{d}}\sin\gamma(\mathbf{u}' \cdot \mathbf{r}')\right)^2\right] \\
=& \frac{1}{d}\mathrm{E}\left[\left(\cos\gamma + \sqrt{d-1}\sin\gamma(\mathbf{u}' \cdot \mathbf{r}')\right)^2\right] \\
=& \frac{1}{d}\left(\cos^2\gamma + (d-1)\sin^2\gamma \mathrm{E}\left[(\mathbf{u}' \cdot \mathbf{r}')^2\right] + 2\cos\gamma \sin\gamma\sqrt{d-1}\,\mathrm{E}\left[(\mathbf{u}' \cdot \mathbf{r}')\right]\right) \\
=& \frac{1}{d}\left(\cos^2\gamma + (d-1)\sin^2\gamma\frac{1}{d-1}\right) = \frac{1}{d}
\end{aligned}$$

□

The condition $\mathrm{E}\left[(\mathbf{u}' \cdot \mathbf{r}')^2\right] = \frac{1}{d-1}$ is saying that $(\sqrt{d-1}\mathbf{u}' \cdot \mathbf{r}')^2$ is an unbiased estimator of $||\mathbf{u}'||^2$ - and lemma 4.1 then tells us that $(\sqrt{d}\mathbf{u} \cdot \mathbf{r})^2$ is an unbiased estimator of $||\mathbf{u}||^2$. "Good behavior" in the subspace thus gives "good behavior" in the entire space. The following lemma extends this to all even moments of the dot product considered, and is one of the main technical results in this paper.

Lemma 4.2. *Define* $\mathbf{r}'$ *and* $\mathbf{r}$ *as in Lemma 4.1. Let* $T' \stackrel{D}{=} N(0, \frac{1}{d-1})$*, and* $T \stackrel{D}{=} N(0, \frac{1}{d})$*. For any positive integer m and for any unit vector* $\mathbf{u}'$ *in the subspace V such that* $\mathbf{u}' \cdot \mathbf{r}'$ *is symmetrically distributed about zero and such that*

$$\mathrm{E}\left[\left(\mathbf{u}' \cdot \mathbf{r}'\right)^{2m}\right] \leq \mathrm{E}\left[T'^{2m}\right] = (2m-1)!!\left(\frac{1}{d-1}\right)^{m},$$

then for any unit vector $\mathbf{u} \in \mathbb{R}^d$

$$\mathrm{E}\left[(\mathbf{u} \cdot \mathbf{r})^{2m}\right] \leq \mathrm{E}\left[T^{2m}\right] = (2m-1)!!\frac{1}{d^m}.$$

Proof. For any γ we can expand

$$\begin{aligned}
E\left[(\mathbf{u} \cdot \mathbf{r})^{2m}\right] &= E\left[\left(\cos\theta_o \cos\gamma + \sin\theta_o \sin\gamma(\mathbf{u}' \cdot \mathbf{r}')\right)^{2m}\right] \\
&= \frac{1}{d^m}\mathrm{E}\left[\left(\cos\gamma + \sqrt{d-1}\sin\gamma(\mathbf{u}' \cdot \mathbf{r}')\right)^{2m}\right] \\
&= \frac{1}{d^m}E\left[\sum_{i=0}^{2m}\binom{2m}{i}\cos^{2m-i}\gamma(d-1)^{i/2}\sin^i\gamma(\mathbf{u}' \cdot \mathbf{r}')^i\right] \\
&= \frac{1}{d^m}\sum_{i=0}^{2m}\binom{2m}{i}\cos^{2m-i}\gamma(d-1)^{i/2}\sin^i\gamma E\left[(\mathbf{u}' \cdot \mathbf{r}')^i\right] \\
&= \frac{1}{d^m}\sum_{j=0}^{m}\binom{2m}{2j}\cos^{2(m-j)}\gamma\sin^{2j}\gamma(d-1)^j E\left[(\mathbf{u}' \cdot \mathbf{r}')^{2j}\right] \qquad (4.1) \\
&\leq \frac{1}{d^m}\sum_{j=0}^{m}\binom{2m}{2j}\cos^{2(m-j)}\gamma\sin^{2j}\gamma \cdot (2j-1)!!.
\end{aligned}$$

In getting to line 4.1, we used that all the odd moments of $\mathbf{u}' \cdot \mathbf{r}'$ are zero to eliminate the terms with odd i, since $\mathbf{u}' \cdot \mathbf{r}'$ is symmetrically distributed about zero. Our goal is now to show that this upper bound is in turn upper bounded by $\mathrm{E}\left[T^{2m}\right] = (2m-1)!!\frac{1}{d^m}$ (this equality is a standard result), and that this value occurs when $\gamma = \pi/2$. Let

$$g(m,\gamma) = \sum_{j=0}^{m}\binom{2m}{2j}\cos^{2(m-j)}\gamma\sin^{2j}\gamma(2j-1)!! = \mathrm{E}\left[(\mathbf{u} \cdot \mathbf{r})^{2m}\right] \cdot d^m.$$

First observe that all the powers of sines and cosines in this expression are even. This implies that the sign of any sine or cosine factor has no effect on the value of the function, which in turn implies that the function reaches both its maximum and minimum w.r.t. γ in the interval $\gamma \in [0,\pi/2]$, since it ranges over the same values in the intervals $[\pi/2,\pi]$, $[\pi,3\pi/2]$ and $[3\pi/2,2\pi]$. In the following it will be shown that $\frac{\partial}{\partial\gamma}g(m,\gamma)$ can be written as a sum of terms where every term has factor with both positive integer powers $\sin\gamma$ and $\cos\gamma$ and a positive (integer) coefficient. In the range $\gamma \in [0,\pi/2]$, every term in such a sum takes non-negative values, and hence it can only be zero if every single term is zero. This implies that $\cos\gamma = 0$ or $\sin\gamma = 0$. But in that case, these solutions must be the locations of the maxima and minima of $g(m,\gamma)$. Hence the minimum is $g(m,0) = 1$ and the maximum is $g(m,\pi/2) = (2m-1)!!$.

We now prove that $\frac{\partial}{\partial\gamma}g(m,\gamma)$ has the claimed properties. Let $g_j(m,\gamma) = \binom{2m}{2j}\cos^{2(m-j)}\gamma\sin^{2j}\gamma(2j-1)!!$.

Then

$$\begin{aligned}\frac{\partial}{\partial\gamma}g_j(m,\gamma) &= \binom{2m}{2j}(2j-1)!!\frac{\partial}{\partial\gamma}\left(\cos^{2(m-j)}\gamma\sin^{2j}\gamma\right)\\ &= \binom{2m}{2j}(2j-1)!! \\ &\quad\cdot\left(-2(m-j)\cos^{2(m-j)-1}\gamma\sin^{2j+1}\gamma+2j\cos^{2(m-j)+1}\gamma\sin^{2j-1}\gamma\right)\end{aligned} \tag{4.2}$$

For $0<j<m$, we see $g_j(m,\gamma)$ contributes a positive and a negative term to $\frac{\partial}{\partial\gamma}g(m,\gamma)$. For $j=0$, $g_j(m,\gamma)$ contributes only a term with a negative coefficient, and for $j=m$, $g_j(m,\gamma)$ contributes only a positive term. Let $-A_j$ be the negative term of $\frac{\partial}{\partial\gamma}g_j(m,\gamma)$ and B_j the positive term. We now show that $A_j\le B_{j+1}$ for $j<m$.

$$\begin{array}{rcl} A_j &\le& B_{j+1}\\ \binom{2m}{2j}(2j-1)!!\cdot 2(m-j)\cos^{2(m-j)-1}\gamma\sin^{2j+1}\gamma &\le& \binom{2m}{2(j+1)}(2j+1)!!\cdot 2(j+1)\cos^{2(m-j)-1}\gamma\sin^{2j+1}\gamma\\ \binom{2m}{2j}(m-j) &\le& \binom{2m}{2(j+1)}(2j+1)\cdot(j+1)\\ \frac{(2m)!}{(2j)!(2(m-j))!}(m-j) &\le& \frac{(2m)!}{(2j+2)!(2(m-j)-2)!}(2j+1)\cdot(j+1)\\ \frac{m-j}{2(m-j)\cdot(2(m-j)-1)} &\le& \frac{(2j+1)\cdot(j+1)}{(2j+2)\cdot(2j+1)}\\ \frac{1}{(2(m-j)-1)} &\le& 1\\ 1 &\le& m-j\\ j &\le& m-1 \end{array}$$

So $A_j\le B_{j+1}$ holds when $j<m$. Since $A_m=0$ and $B_0=0$, this means that $\frac{\partial}{\partial\gamma}g(m,\gamma)=\sum_{j=0}^{m}\frac{\partial}{\partial\gamma}g_j(m,\gamma)=\sum_{j=0}^{m-1}B_{j+1}-A_j$. This proves the claim that $\frac{\partial}{\partial\gamma}g(m,\gamma)$ can be written as a sum of terms where every term has factor with both positive integer powers of $\sin\gamma$ and $\cos\gamma$ and a positive (integer) coefficient. Which in turn proves that the maximum of $g(m,\gamma)$ w.r.t. γ is $(2m-1)!!$, as described above. Thus $E\left[(\mathbf{u}\cdot\mathbf{r})^{2m}\right]\le(2m-1)!!/d^m$. □

Finally, to prove Lemma 3.4, we need two lemmas by Achlioptas [1] which give us a distribution over the subspace orthogonal to $\mathbf{x}_c$ that satisfies the requirements of Lemmas 4.1 and 4.2.

Lemma 4.3. *In the subspace V orthogonal to $\mathbf{x}_c$, define a random vector $\mathbf{r}'=\frac{1}{\sqrt{d-1}}(X_1,\ldots,X_{d-1})$, where the X_i are i.i.d. from $\{-1,+1\}$. Then $\mathbf{r}'$ is always a unit vector,* $\mathrm{E}[\mathbf{u}'\cdot\mathbf{r}']=0$, *and* $\mathrm{E}\left[(\mathbf{u}'\cdot\mathbf{r}')^2\right]=\frac{1}{d-1}$.

Proof. The vector $\mathbf{r}'$ is always a unit vector, since $||\mathbf{r}'||^2=\frac{1}{d-1}\sum_{i=1}^{d-1}X_i^2=1$. The other two claims are proven in Achlioptas' preliminary statements [1], equation (3) on p. 677 and equation (4) on p. 679. Notice that he is considering a space of dimension d, but here the space is of dimension $d-1$. □

Lemma 4.4. *Let $\mathbf{w}=\frac{1}{\sqrt{d-1}}(1,\ldots,1)$ be a unit vector in the subspace V orthogonal to $\mathbf{x}_c$, i.e. with d-1 coordinates. Let $T'\overset{D}{=}N(0,\frac{1}{d-1})$. Define $\mathbf{r}'$ as in Lemma 4.3. Then for every unit vector $\mathbf{u}'\in V$, and for any non-negative integer m*

$$\mathrm{E}\left[(\mathbf{u}'\cdot\mathbf{r}')^{2m}\right]\le\mathrm{E}\left[(\mathbf{w}\cdot\mathbf{r}')^{2m}\right]\le\mathrm{E}\left[T'^{2m}\right].$$

Proof. The first inequality follows by Achlioptas' [1] Lemma 6.1, and the second by his Lemma 6.2. □

Finally, we may use the random vector defined in Lemma 4.3 to prove Lemma 3.4.

Proof of Lemma 3.4. Define $\mathbf{r}'$ as in 4.3, and using this $\mathbf{r}'$, define $\mathbf{r}$ as in Lemma 4.1. Then by Lemma 4.3, $\mathbf{r}'$ satisfies the requirements of Lemma 4.1, so $\mathbf{r}$ is a random unit vector from $\mathbb{B}^d$, and for any fixed unit vector $\mathbf{u} \in \mathbb{R}^d$, we have $\mathrm{E}\left[(\mathbf{u} \cdot \mathbf{r})^2\right] = \frac{1}{d}$. Also, by Lemma 4.4, $\mathbf{r}'$ satisfies the requirements of Lemma 4.2, so for all non-negative integer m

$$\mathrm{E}\left[(\mathbf{u} \cdot \mathbf{r})^{2m}\right] \leq \mathrm{E}\left[T^{2m}\right] = (2m-1)!!\frac{1}{d^m},$$

and this completes the proof. □

References

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

[2] A. Agarwal, J. Phillips, and S. Venkatasubramanian. Universal multidimensional scaling. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

[3] P. K. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in euclidean space. In *Proceedings of the twenty-third annual symposium on Computational geometry*, SCG '07, pages 381–389, New York, NY, USA, 2007. ACM.

[4] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, New York, NY, USA, 2006. ACM.

[5] N. Ailon and E. Liberty. Fast dimension reduction using rademacher series on dual bch codes. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1–9, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[6] N. Ailon and E. Liberty. Almost optimal unrestricted fast johnson-lindenstrauss transform. *CoRR*, abs/1005.5513, 2010.

[7] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

[8] A. Bhattacharya, P. Kar, and M. Pal. On low distortion embeddings of statistical distance measures into low dimensional spaces. In S. S. Bhowmick, J. Küng, and R. Wagner, editors, *DEXA*, volume 5690 of *Lecture Notes in Computer Science*, pages 164–172. Springer, 2009.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[10] G. Borradaile, J. R. Lee, and A. Sidiropoulos. Randomly removing g handles at once. In *Proceedings of the 25th annual symposium on Computational geometry*, SCG '09, pages 371–376, New York, NY, USA, 2009. ACM.

[11] K. L. Clarkson. Tighter bounds for random projections of manifolds. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, SCG '08, pages 39–48, New York, NY, USA, 2008. ACM.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[13] A. Dasgupta, R. Kumar, and T. Sarlos. A sparse johnson: Lindenstrauss transform. In *STOC '10: Proceedings of the 42nd ACM symposium on Theory of computing*, pages 341–350, New York, NY, USA, 2010. ACM.

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

[15] P. Frankl and H. Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory Ser. A*, 44(3):355–362, 1987.

[16] R. Gray, A. Buzo, A. Gray Jr, and Y. Matsuyama. Distortion measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):367–376, Aug 1980.

[17] P. Indyk and J. Matousek. Low-distortion embeddings of finite metric spaces. In *Handbook of Discrete and Computational Geometry*, pages 177–196. CRC Press, 2004.

[18] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, New York, NY, USA, 1998. ACM.

[19] P. Indyk and A. Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3, August 2007.

[20] P. Indyk and A. Sidiropoulos. Probabilistic embeddings of bounded genus graphs into planar graphs. In *Proceedings of the twenty-third annual symposium on Computational geometry*, SCG '07, pages 204–209, New York, NY, USA, 2007. ACM.

[21] T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms.* Kluwer/Springer, 2002.

[22] 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.

[23] D. M. Kane and J. Nelson. A derandomized sparse johnson-lindenstrauss transform. *CoRR*, abs/1006.3585, 2010.

[24] E. Liberty, N. Ailon, and A. Singer. Dense fast random projections and lean walsh transforms. In *APPROX '08 / RANDOM '08: Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop, RANDOM 2008 on Approximation, Randomization and Combinatorial Optimization*, pages 512–522, Berlin, Heidelberg, 2008. Springer-Verlag.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2), 2004.

[26] J. Matousek. On variants of the johnson–lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[27] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proc. 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

[28] A. Sidiropoulos. Optimal stochastic planarization. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 163–170, Washington, DC, USA, 2010. IEEE Computer Society.

A Proving tail bounds from moments

For the reader's convenience we replicate the following proofs by Achlioptas [1]. The only difference is the slightly more general statement of the proofs. The technical details are the same.

Proof of Lemma 3.3. First we prove the lower bound. We use essentially use the same proof as Achlioptas [1]. Let $\mathbf{v}$ be some fixed vector and $\tilde{\mathbf{v}} = \mathbf{v}/||\mathbf{v}|$, then

$$\begin{align}
\Pr\left[||f(\mathbf{v})||^2 < (1-\varepsilon)||\mathbf{v}||^2\right] &= \Pr\left[\frac{d}{k}\sum_{i=1}^{k}(\mathbf{v}\cdot\mathbf{r}_i)^2 < (1-\varepsilon)\sum_{j=1}^{d} v_j^2\right] \notag\\
&= \Pr\left[d\sum_{i=1}^{k}(\tilde{\mathbf{v}}\cdot\mathbf{r}_i)^2 < k(1-\varepsilon)\right] \notag\\
&= \Pr\left[\exp\left\{\lambda d\sum_{i=1}^{k}(\tilde{\mathbf{v}}\cdot\mathbf{r}_i)^2\right\} < \exp\left\{\lambda k(1-\varepsilon)\right\}\right], \lambda > 0 \notag\\
&\le \mathrm{E}\left[\exp\left\{-\lambda d\sum_{i=1}^{k}(\tilde{\mathbf{v}}\cdot\mathbf{r}_i)^2\right\}\right] e^{\lambda k(1-\varepsilon)} \tag{A.1}\\
&= \mathrm{E}\left[\exp\left\{-\lambda d\,(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^2\right\}\right]^k e^{\lambda k(1-\varepsilon)} \notag\\
&\le \mathrm{E}\left[1-\lambda d\,(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^2 + \frac{\lambda^2 d^2\,(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^4}{2}\right]^k e^{\lambda k(1-\varepsilon)} \tag{A.2}\\
&\le \left(1-\lambda+\frac{3}{2}\lambda^2\right)^k e^{\lambda k(1-\varepsilon)}. \tag{A.3}
\end{align}$$

In line A.1, we used a standard Markov bound. In line A.2, we used that $e^x \le 1+x+x^2/2$ for all $x \le 0$ and that $-\lambda d\,(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^2 < 0$. In line A.3, we used the assumption in the statement of the lemma that $\mathrm{E}\left[(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^2\right] = 1/d$ and $\mathrm{E}\left[(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^2\right] \le 3/d^2$. Next, we set $\lambda = \frac{\varepsilon}{2(1+\varepsilon)}$, which we can do since $\varepsilon > 0$. This gives

$$\Pr\left[||f(\mathbf{v})||^2 < (1-\varepsilon)||\mathbf{v}||^2\right] \le \left(1-\frac{\varepsilon}{2(1+\varepsilon)}+\frac{3}{8}\frac{\varepsilon^2}{(1+\varepsilon)^2}\right)^k \exp\left(\frac{k\varepsilon(1-\varepsilon)}{2(1+\varepsilon)}\right).$$

Finally, a comparison of power series shows that for $\varepsilon \le 1/2$, we get

$$\Pr\left[||f(\mathbf{v})||^2 < (1-\varepsilon)||\mathbf{v}||^2\right] \le \left(1-\frac{\varepsilon}{2(1+\varepsilon)}+\frac{3}{8}\frac{\varepsilon^2}{(1+\varepsilon)^2}\right)^k \exp\left(\frac{k\varepsilon(1-\varepsilon)}{2(1+\varepsilon)}\right) \le \exp\left(-\frac{k}{2}(\varepsilon^2/2-\varepsilon^3/3)\right).$$

This proves the lower bound. For the upper bound we replicate parts of two proofs by Achlioptas [1], his Lemma 5.1 and 5.2. We start by showing that

$$\begin{aligned}
\mathrm{E}\left[\exp\left\{d\cdot\lambda\,(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^2\right\}\right] &\le \mathrm{E}\left[\sum_{i=0}^{\infty}\frac{d^i\lambda^i\,(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^{2i}}{i!}\right]\\
&= \sum_{i=0}^{\infty}\frac{\lambda^i d^i}{i!}\mathrm{E}\left[(\tilde{\mathbf{v}}\cdot\mathbf{r}_1)^{2i}\right]\\
&\le \sum_{i=0}^{\infty}\frac{\lambda^i d^i}{i!}\mathrm{E}\left[T^{2i}\right] = \mathrm{E}\left[\exp\left(d\cdot\lambda T^2\right)\right] = \frac{1}{\sqrt{1-2\lambda}},\ \lambda < 1/2.
\end{aligned}$$

Where $\mathrm{E}\left[\exp\left\{d \cdot \lambda T^2\right\}\right] = \frac{1}{\sqrt{1-2\lambda}}$ is a standard and easily proven result. This implies that

$$\begin{aligned}
\Pr\left[||f(\mathbf{v})||^2 > (1-\varepsilon)||\mathbf{v}||^2\right] \leq & \mathrm{E}\left[\exp\left(\lambda d\,(\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^2\right)\right]^k e^{-\lambda k(1+\varepsilon)},\ \lambda > 0 \\
\leq & \left(\frac{1}{\sqrt{1-2\lambda}}\right)^k e^{-\lambda k(1+\varepsilon)},\ \lambda < 1/2 \\
\leq & ((1+\varepsilon)\exp(-\varepsilon))^{k/2} \\
\leq & \exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)\right), \qquad \text{(A.4)}
\end{aligned}$$

where line A.4 is found to hold for $0 < \varepsilon < 1/2$ by comparison of power series. □