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 $\varepsilon$-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 $\ell_p$-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 $\ell_2$ 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 known 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 $S^{d-1} \subset \mathbb{R}^d$ is the set of all points $\{(x_1, \ldots, x_d) \mid \sum x_i^2 = 1\}$. For notational convenience, we will use $\Delta$ 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 $S^{d-1}_+ = S^{d-1} \cap \mathbb{R}^+_d$. The Hellinger distance $d_H : \Delta \times \Delta \rightarrow \mathbb{R}^+$ is a commonly used distance measure between distributions, defined as

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

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

Let $\theta_o$ be the angle from $v x_c$ to a basis vector with the $i^{th}$ coordinate being 1, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, i.e. $e_i$ is a basis vector in the standard orthonormal basis of $\mathbb{R}^d$. Thus, $\cos \theta_o = x_c \cdot e_i = \frac{1}{\sqrt{d}}$.

Call the set of unit vectors at angle $\theta_o$ to $x_c$ the possible basis set, which we denote $\mathbb{B}^d \subset \mathbb{R}^d$.

Let $e_i'$ be the opposing vector to $e_i$, given by $e_i' = \frac{1}{\sqrt{d-1}} \sum_{j \neq i} e_j$ and let the angle $\theta_{IR}$ be the angle from $x_c$ to some $e_i'$, so that $\cos \theta_{IR} = x_c \cdot e_i' = (d-1) \cdot \frac{1}{\sqrt{d}} \cdot \frac{1}{\sqrt{d-1}} = \sqrt{\frac{d-1}{d}}$.

The inner region $\mathbb{I}^d \subset S^{d-1}$ is the set of all unit vectors of angle $\theta_{IR}$ or less to $x_c$. We also call all points $p \in \Delta^{d-1}$ that correspond to vectors $v p$ at angle at most $\theta_{IR}$ from $x_c$ the inner region of the simplex and denote the set as $\mathbb{I}^{(d-1)}$; 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 $u \in S^{d-1}$ be a vector. The set $X \subset S^{d-1}$ of vectors at angle $\theta$ to $u$ forms a sphere of dimension $d-2$ and radius $\sin \theta$ in the subspace orthogonal to $u$.

**Proof.** Let $u$ and $r$ be unit vectors with angle $\theta$ 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 $u = (1, 0, \ldots, 0)$ and $r = (r_1, \ldots, r_n)$. Thus in this basis we have $\cos \theta = u \cdot r = r_1$. Since $r$ is a unit vector, we know that $r_1^2 + \ldots + r_n^2 = 1$, so $r_1^2 + \ldots + r_n^2 = 1 - \cos^2 \theta = \sin^2 \theta$. 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 $\varepsilon$-distortion if there exists a constant $c$ such that

$$\forall p, p' \in P \quad (1 - \varepsilon)d_Y(f(p), f(p')) \leq c \cdot d_X(p, p') \leq (1 + \varepsilon)d_Y(f(p), f(p')).$$

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 \( P \subseteq \mathbb{I}(\Delta^{d-1}) \) be a set of \( n \) points in metric space \((\Delta^{d-1}, d_H)\). For constants \( 0 < \varepsilon < 2 \) and \( 0 < \delta < 1 \) and integer \( k = O\left( \frac{1}{\varepsilon^2} \ln \left( \frac{n^2}{\delta^2} \right) \right) \), 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 \( \varepsilon \)-distortion in \((\Delta^{k-1}, d_H)\) 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 \( \Delta^{d-1} \) to \( S^{d-1} \) 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 \( S^{k-1} \) using the normalization \( S(p) = \frac{p}{\|p\|} \). Finally, since the points now lie on \( S^{k-1}_+ \), we can apply the inverse mapping \( h^{-1} \) to the simplex \( \Delta^{k-1} \). 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 \( \varepsilon \) distortion, the mapping \( S \) introduces only a small overall distortion.

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

\[
(1 - \varepsilon/2)\|p - p'\| \leq \frac{1 - \varepsilon/4}{1 + \varepsilon/4} \|p - p'\| \leq \|\psi(p) - \psi(p')\| \leq \frac{1 + \varepsilon/4}{1 - \varepsilon/4} \|p - p'\| \leq (1 + \varepsilon)\|p - p'\|.
\]

The first and last inequality follow by \( 1 - 2\alpha \leq \frac{1 - \alpha}{1 + \alpha} \) and from \( \frac{1 + \alpha}{1 - \alpha} \leq 1 + 4\alpha \), 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 \( \alpha \) 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 \( \varepsilon \)-distortion under \( d_E \). Our strategy is to pick \( k \) random vectors \( r_1, \ldots, r_k \) from some distribution over \( \mathbb{B}^d \), and define the map \( f : \mathbb{R}^d \to \mathbb{R}^k \) by \( f(v) = \sqrt{\frac{1}{k}} \langle v \cdot r_1, \ldots, v \cdot r_k \rangle \). This strategy is justified by the following lemma.

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

Proof. The center point does not change with a rotation of the orthonormal basis about \( x_i \); \( \mathbb{B}^d \) and \( \mathbb{I}^d \) are also invariant. Thus, w.l.o.g. consider \( e_1 = v \), rather than a general vector from \( \mathbb{B}^d \). The angle between \( e_1 \) and \( e'_i = \pi/2 \) since \( e_1 \cdot e'_i = 0 \). The set of all vectors \( t \) that have angle with \( e_1 \) 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 \( \mathbb{I}^d \subseteq \mathbb{S}^{d-1}_+ \), this concludes the proof.

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

Theorem 3.2. Let \( P \subseteq \mathbb{R}^d \) be an arbitrary set of \( n \) points. For any \( 0 < \varepsilon < 1/2 \) and \( 0 < \delta < 1 \) and for integer \( k \geq O\left( \frac{1}{\varepsilon^2} \ln \left( \frac{n^2}{\delta^2} \right) \right) \), 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 \( \varepsilon \)-distortion in \((\mathbb{R}^d, d_H)\) with probability at least \( 1 - \delta \).

Furthermore, for any point \( p \in \mathbb{I}^d \subseteq \mathbb{R}^d_+ \), then \( f(p) \in \mathbb{R}^k_+ \).
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 \( \mathbb{E} [(\mathbf{u} \cdot \mathbf{r})^2] = 1/d \) and for integer \( m \geq 2 \)

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

where \((2m-1)!! = (2m\,)!/m!m!\) 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 \( f : \mathbb{R}^d \rightarrow \mathbb{R}^k \) 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 \( \mathbf{x}_e = (1,0,0,\ldots,0) \) (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 \( \mathbf{r} = \left( \cos \theta_0, \sin \theta_0 \cdot \frac{X_0}{\sqrt{d-1}}, \ldots, \sin \theta_0 \cdot \frac{X_d}{\sqrt{d-1}} \right) \), where the \( X_i \) are i.i.d. from \( \{-1,+1\} \). The vector \( \mathbf{r} \in \mathbb{R}^d \) and for any unit vector \( \mathbf{u} \in \mathbb{R}^d \) satisfies \( \mathbb{E} [(\mathbf{u} \cdot \mathbf{r})^2 \cdot d] = 1 \) and for integer \( m \geq 2 \)

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

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 \( f : \mathbb{R}^d \rightarrow \mathbb{R}^k \) be defined as in lemma 3.3 using this \( \mathbf{r} \). Since \( \mathbf{r} \in \mathbb{R}^d \), if a vector \( \mathbf{v} \in \mathbb{R}^d \), then for each \( \mathbf{r}_i \) of \( f, \mathbf{v} \cdot \mathbf{r}_i \geq 0 \). Consequently, any point \( \mathbf{p} \in \mathbb{R}^d \subset \mathbb{R}^d \) 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 \( \delta \). 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. \( \square \)
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

\[
\frac{1}{(\varepsilon')^2} \cdot \ln \left( \frac{n+1}{\delta} \right) = \Theta \left( \frac{1}{\varepsilon^2} \ln \left( \frac{n}{\delta} \right) \right),
\]

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 \( \varepsilon \) with the same probability by Lemma 3.1. Since \( h \) maps points from \( \mathbb{I}(\Delta^{d-1}) \) to \( \mathbb{R}^d \), and \( f \) maps points in \( \mathbb{R}^d \) 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 \( \mathbb{R}_+^{k-1} \). 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 \( \Delta^{k-1} \), and has \( \varepsilon \)-distortion with probability at least \( 1 - \delta \).

\[ \square \]

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 \( x_c \) 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 \( \theta_o \), this construction has a very nice property, as expressed in the following lemma.

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

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

\[
\mathbb{E}[(u \cdot r)^2] = \mathbb{E}[(\cos \theta_o \cos \gamma + \sin \theta_o \sin \gamma (u' \cdot r'))^2]
\]

\[
= \mathbb{E} \left[ \left( \frac{1}{\sqrt{d}} \cos \gamma + \sqrt{\frac{d-1}{d}} \sin \gamma (u' \cdot r') \right)^2 \right]
\]

\[
= \frac{1}{d} \mathbb{E} \left[ \cos \gamma + \sqrt{d-1} \sin \gamma (u' \cdot r')^2 \right]
\]

\[
= \frac{1}{d} \left( \cos \gamma + (d-1) \sin \gamma \mathbb{E}[(u' \cdot r')^2] + 2 \cos \gamma \sin \gamma \sqrt{d-1} \mathbb{E}[(u' \cdot r')] \right)
\]

\[
= \frac{1}{d} \left( \cos \gamma + (d-1) \sin \gamma \frac{1}{d-1} \right) = \frac{1}{d}
\]

The condition \( \mathbb{E}[(u' \cdot r')^2] = \frac{1}{d-1} \) is saying that \( \sqrt{d-1} u' \cdot r' \) is an unbiased estimator of \( ||u'||^2 \) - and lemma 4.1 then tells us that \( \sqrt{d} u \cdot r' \) is an unbiased estimator of \( ||u||^2 \). “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 $r'$ and $r$ as in Lemma 4.1. Let $T^D \overset{D}{\sim} N(0, \frac{1}{d-1})$, and $T^D \overset{D}{\sim} N(0, \frac{1}{2})$. For any positive integer $m$ and for any unit vector $u'$ in the subspace $V$ such that $u' \cdot r'$ is symmetrically distributed about zero and such that

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

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

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

Proof. For any $\gamma$ we can expand

$$E \left[ (u \cdot r)^{2m} \right] = E \left[ (\cos \theta_o \cos \gamma + \sin \theta_o \sin \gamma (u' \cdot r'))^{2m} \right]$$

$$= \frac{1}{d^m} E \left[ \left( \cos \gamma + \sqrt{d-1} \sin \gamma (u' \cdot r') \right)^{2m} \right]$$

$$= \frac{1}{d^m} E \left[ \sum_{i=0}^{2m} \binom{2m}{i} \cos^{2m-i} \gamma (d-1)^{i/2} \sin^i \gamma (u' \cdot r')^i \right]$$

$$= \frac{1}{d^m} \sum_{i=0}^{2m} \binom{2m}{i} \cos^{2m-i} \gamma (d-1)^{i/2} \sin^i \gamma E \left[ (u' \cdot r')^i \right]$$

$$= \frac{1}{d^m} \sum_{i=0}^{2m} \binom{2m}{i} \cos^{2m-i} \gamma (d-1)^{i/2} \sin^i \gamma \gamma \sin^2 j \gamma (d-1)^j E \left[ (u' \cdot r')^{2j} \right]$$

$$\leq \frac{1}{d^m} \sum_{j=0}^{m} \binom{2m}{2j} \gamma \sin^2 j \gamma (2j-1)!!.$$

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

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

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. $\gamma$ 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 $g_j(m, \gamma) = \binom{2m}{2j} \gamma \sin^2 j \gamma (2j-1)!!$. **
Then
\[
\frac{\partial}{\partial \gamma} g_j(m, \gamma) = \binom{2m}{2j} (2j - 1)!! \frac{\partial}{\partial \gamma} \left( \cos^{2(m-j)} \gamma \sin^{2j} \gamma \right)
\]
\[
= \binom{2m}{2j} (2j - 1)!! \cdot \left( -2(m-j) \cos^{2(m-j)-1} \gamma \sin^{2j+1} \gamma + 2j \cos^{2(m-j)+1} \gamma \sin^{2j-1} \gamma \right) \tag{4.2}
\]

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 \(A_j \leq B_{j+1}\) for \(j < m\).

\[
A_j \leq B_{j+1}
\]
\[
\binom{2m}{2j} (2j - 1)!! \cdot 2(m-j) \cos^{2(m-j)-1} \gamma \sin^{2j+1} \gamma \leq \binom{2m}{2(j+1)} (2j+1)!! \cdot 2(j+1) \cos^{2(m-j)-1} \gamma \sin^{2j+1} \gamma
\]
\[
\binom{2m}{2j} (m-j) \leq \binom{2m}{2(j+1)} (j+1)
\]
\[
\frac{(2m)!}{(2j)!(2(m-j))!} (m-j) \leq \frac{(2m)!}{(2j+2)!(2(m-j)-2)!} (j+1) \leq \frac{1}{(2j+2) \cdot (2j+1)} \leq 1
\]
\[
\frac{1}{(2(m-j)-1)} \leq m-j \leq m-1
\]

So \(A_j \leq B_{j+1}\) 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. \(\gamma\) is \((2m-1)!!\), as described above. Thus \(E \left[ (u \cdot r)^{2m} \right] \leq (2m-1)!!/d^m\).

Finally, to prove Lemma 3.4, we need two lemmas by Achlioptas [1] which give us a distribution over the subspace orthogonal to \(x_c\) that satisfies the requirements of Lemmas 4.1 and 4.2.

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

**Proof.** The vector \(r'\) is always a unit vector, since \(||r'||^2 = \frac{1}{d-1} \sum_{i=1}^{d-1} X_i^2 = 1\). 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 \(w = \frac{1}{\sqrt{d-1}} (1, \ldots, 1)\) be a unit vector in the subspace \(V\) orthogonal to \(x_c\), i.e. with \(d-1\) coordinates. Let \(T' \equiv N(0, \frac{1}{d-1})\). Define \(r'\) as in Lemma 4.3. Then for every unit vector \(u' \in V\), and for any non-negative integer \(m\)
\[
E \left[ (u' \cdot r')^{2m} \right] \leq E \left[ (w \cdot r')^{2m} \right] \leq E \left[ r'^{2m} \right].
\]

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

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

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

and this completes the proof.
References


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 the same proof as Achlioptas [1].

Let $\mathbf{v}$ be some fixed vector and $\tilde{\mathbf{v}} = \mathbf{v}/||\mathbf{v}||$, then

\[
\Pr \left[ ||f(\mathbf{v})||^2 < (1 - \varepsilon)||\mathbf{v}||^2 \right] = \Pr \left[ \sum_{i=1}^{d} (\mathbf{v} \cdot \mathbf{r}_i)^2 < (1 - \varepsilon) \sum_{j=1}^{d} v_j^2 \right]
\]

\[
= \Pr \left[ \sum_{i=1}^{d} (\tilde{\mathbf{v}} \cdot \mathbf{r}_i)^2 < k(1 - \varepsilon) \right]
\]

\[
= \Pr \left[ \exp \left\{ \lambda d \sum_{i=1}^{k} (\tilde{\mathbf{v}} \cdot \mathbf{r}_i)^2 \right\} < \exp \{ \lambda k (1 - \varepsilon) \} \right], \lambda > 0
\]

\[
\leq \mathbb{E} \left[ \exp \left\{ -\lambda d \sum_{i=1}^{k} (\tilde{\mathbf{v}} \cdot \mathbf{r}_i)^2 \right\} \right] e^{\lambda k (1 - \varepsilon)}
\]

\[
= \mathbb{E} \left[ \exp \left\{ -\lambda d (\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^2 \right\} \right]^k e^{\lambda k (1 - \varepsilon)}
\]

\[
\leq \mathbb{E} \left[ 1 - \lambda d (\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^2 + \frac{\lambda^2 d^2 (\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^4}{2} \right]^k e^{\lambda k (1 - \varepsilon)}
\]

\[
\leq \left( 1 - \lambda + \frac{3}{2} \lambda^2 \right)^k e^{\lambda k (1 - \varepsilon)}.
\]

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

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

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

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

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

\[
\mathbb{E} \left[ \exp \left\{ d \cdot \lambda (\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^2 \right\} \right] \leq \sum_{i=0}^{\infty} \frac{d^i \lambda^i}{i!} \mathbb{E} \left[ (\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^{2i} \right]
\]

\[
= \sum_{i=0}^{\infty} \frac{\lambda^i d^i}{i!} \mathbb{E} \left[ (\tilde{\mathbf{v}} \cdot \mathbf{r}_1)^{2i} \right]
\]

\[
\leq \sum_{i=0}^{\infty} \frac{\lambda^i d^i}{i!} \mathbb{E} \left[ T^{2i} \right] = \mathbb{E} \left[ \exp \left( d \cdot \lambda T^2 \right) \right] = \frac{1}{\sqrt{1 - 2\lambda}}, \lambda < 1/2.
\]
Where \( E[\exp\{d \cdot \lambda T^2\}] = \frac{1}{\sqrt{1 - 2\lambda}} \) is a standard and easily proven result. This implies that

\[
\Pr\left[ ||f(v)||^2 > (1 - \varepsilon)||v||^2 \right] \leq E\left[ \exp\left( \lambda d (\tilde{v} \cdot r_1)^2 \right) \right]^k e^{-\lambda k(1 + \varepsilon)}, \quad \lambda > 0
\]

\[
\leq \left( \frac{1}{\sqrt{1 - 2\lambda}} \right)^k e^{-\lambda k(1 + \varepsilon)}, \quad \lambda < 1/2
\]

\[
\leq ((1 + \varepsilon \exp(-\varepsilon))^{k/2}
\]

\[
\leq \exp\left( - \frac{k}{2} \left( \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right) \right), \quad (A.4)
\]

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