Database-friendly Random Projections

Dimitris Achlioptas*
Microsoft

ABSTRACT

A classic result of Johnson and Lindenstrauss asserts that any set of n points in d-dimensional Euclidean space can be embedded into k-dimensional Euclidean space — where k is logarithmic in n and independent of d — so that all pairwise distances are maintained within an arbitrarily small factor. All known constructions of such embeddings involve projecting the n points onto a random k-dimensional hyperplane. We give a novel construction of the embedding, suitable for database applications, which amounts to computing a simple aggregate over k random attribute partitions.

1. INTRODUCTION

Consider projecting the points of your favorite sculpture first onto a plane and then onto a single line. The result amply demonstrates the power of dimensionality.

Conversely, given a high-dimensional pointset it is natural to ask whether it exploits its full allotment of dimensionality or, rather, it could be embedded into a lower dimensional space without suffering great distortion.

In general, such questions involve a, perhaps infinite, collection of points endowed with some distance function (metric). In this paper, we will only deal with finite sets of points in Euclidean space (so the Euclidean distance is the metric). In particular, it will be convenient to think of n points in $\mathbb{R}^d$ as an $n \times d$ table (matrix) A with each point represented as a row (vector) with d attributes (coordinates).

Given such a matrix A, one of the most common embeddings is the one suggested by its Singular Value Decomposition. In particular, to embed the n points into $\mathbb{R}^k$ we project them onto the k-dimensional space spanned by the singular vectors corresponding to the k largest singular values of A. If one rewrites the result of this projection as a (rank k) $n \times d$ matrix A_k, we are guaranteed that for every rank k matrix D

$$|A - A_k| \leq |A - D| \ ,$$

for any unitarily invariant norm, such as the Frobenius or the L2 norm. Thus, distortion here amounts to a certain distance (norm) between the set of projected points, A_k, and the original set of points A. If we associate with each row (point) a vector corresponding to the difference between its original and its new position then, for example, under the Frobenius norm the distortion equals the sum of the squared lengths of these vectors. It is clear that such a notion of distortion captures a significant global property. At the same time, though, it does not offer any local guarantees. For example, the distance between a pair of points can be arbitrarily smaller than what it was in the original space, if that is advantageous to minimizing the total distortion.

The study of embeddings that respect local properties is a rich area of mathematics with deep and beautiful results. Such embeddings can guarantee, for example, that all distances between pairs of points are approximately maintained or, more generally, that for a given $q \geq 2$, a certain notion of "volume" is maintained for all collections of up to q points (thus capturing higher order local structure). The algorithmic uses of such embeddings were first considered in the seminal paper of Linial, London and Rabinovich [9] and have by now become an important part of modern algorithmic design. A real gem in this area has been the following result of Johnson and Lindenstrauss [7].

LEMMA 1 ([7]). *Given $\epsilon > 0$ and an integer n, let k be a positive integer such that $k \geq k_0 = O(\epsilon^{-2} \log n)$. For every set P of n points in $\mathbb{R}^d$ there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in P$*

$$(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+\epsilon)||u-v||^2 \ .$$

We will refer to embeddings providing a guarantee akin to that of Lemma 1 as JL-embeddings. In the last few years, JL-embeddings have been useful in solving a variety of problems. The rough idea is the following. By providing a low dimensional representation of the data, JL-embeddings speed up certain algorithms dramatically, in particular algorithms whose run-time depends exponentially in the dimension of the working space (there are a number of practical problems for which the best known algorithms have such behaviour). At the same time, the provided guarantee regarding pairwise distances is often enough to establish that the solution

*Address: Microsoft Corporation, One Microsoft Way, Redmond WA, 98052, U.S.A. Email: optas@microsoft.com

found by working in the low dimensional space is a good approximation to the optimal solution in the original space. We give a few examples below.

Papadimitriou, Raghavan, Tamaki and Vempala [10], proved that embedding the points of A in a low-dimensional space can significantly speed up the computation of a low rank approximation to A, without significantly affecting its quality. In [6], Indyk and Motwani showed that JL-embeddings are useful in solving the ε-approximate nearest neighbor problem, where (after some preprocessing of the pointset P) one is to answer queries of the following type: "Given an arbitrary point x, find a point $y \in P$, such that for every point $z \in P$, $||x - z|| \geq (1-\varepsilon)||x - y||$." In a different vein, Schulman [11] used JL-embeddings as part of an approximation algorithm for the version of clustering where we seek to minimize the sum of the squares of intracluster distances. Recently, Indyk [5] showed that JL-embeddings can also be used in the context of "data-stream" computation, where one has limited memory and is allowed only a single pass over the data (stream).

1.1 Our contribution

Over the years, the probabilistic method has allowed for the original proof of Johnson and Lindenstrauss to be greatly simplified and sharpened [4, 6, 3], while at the same time giving conceptually simple randomized algorithms for constructing the embedding. Roughly speaking, all such algorithms project the input points onto a spherically random hyperplane through the origin.

Performing such a projection, while conceptually simple, is non-trivial, especially in a database environment. Moreover, its computational cost can be prohibitive for certain applications. At the same time, JL-embeddings have become an important algorithmic design tool and in certain domains they are a desirable standard data processing step. With this in mind, it is natural to ask if we can compute such embeddings in a manner that is simpler and more efficient than the one suggested by the current methods.

Our main result, below, is a first step in this direction, asserting that one can replace projections onto random hyperplanes with much simpler and faster operations, requiring extremely simple probability distributions. In particular, these operations can be implemented readily using standard SQL primitives without any additional functionality. Moreover, somewhat surprisingly, this comes without *any* sacrifice in the quality of the embedding. In fact, we will see that for every fixed value of d we can get slightly better bounds than all current methods.

We describe the main result below in standard mathematical terminology. Following that, we give an example of how to compute the embedding using database operations. As in Lemma 1, the parameter ϵ controls the accuracy in distance preservation, while now β controls the probability of success.

THEOREM 2. *Let P be an arbitrary set of n points in $\mathbb{R}^d$, represented as an $n \times d$ matrix A. Given $\epsilon, \beta > 0$ let*

$$k_0 = \frac{4 + 2\beta}{\epsilon^2/2 - \epsilon^3/3} \log n \ .$$

For integer $k \geq k_0$, let R be a $d \times k$ random matrix with $R(i,j) = r_{ij}$, where $\{r_{ij}\}$ are independent random variables from either one of the following two probability distributions:

$$r_{ij} = \begin{cases} +1 & \textit{with probability} \quad 1/2 \\ -1 & \cdot\cdot \quad 1/2 \end{cases},$$

$$r_{ij} = \sqrt{3} \times \begin{cases} +1 & \textit{with probability} \quad 1/6 \\ 0 & \cdot\cdot \quad 2/3 \\ -1 & \cdot\cdot \quad 1/6 \end{cases}.$$

Let

$$E = \frac{1}{\sqrt{k}} A R \ .$$

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ map the i^{th} row of A to the i^{th} row of E.

With probability at least $1 - n^{-\beta}$, for all $u, v \in P$

$$(1-\epsilon)||u-v||^2 \leq ||f(u) - f(v)||^2 \leq (1+\epsilon)||u-v||^2 \ .$$

In a database system, all operations needed to compute $A\,R$ are very efficient and easy to implement. For example, with the second distribution above, the embedding amounts to generating k new attributes, each one formed by applying the same process: throw away 2/3 of all attributes at random; partition the remaining attributes randomly into two equal parts; for each partition, produce a new attribute equal to the sum of all attributes; take the difference of the two sum-attributes.

All in all, using Theorem 2, one needs very simple probability distributions, no floating point arithmetic, and all computation amounts to highly optimized database operations (aggregation). By using the second probability distribution, where $r_{ij} = 0$ with probability 2/3, we also get a threefold speedup as we only need to process a third of all attributes for each of the k coordinates. On the other hand, when $r_{ij} \in \{-1, +1\}$, conceptually the construction seems to be about as simple as one could hope for.

Looking a bit more closely into the matrix E we see that each row (vector) of A is projected onto k random vectors whose coordinates $\{r_{ij}\}$ are independent random variables with mean 0 and variance 1. If the $\{r_{ij}\}$ were independent Normal random variables with mean 0 and variance 1, it is well-known that the resulting vectors would point to uniformly random directions in space. Projections onto such random lines through the origin have been considered in a number of settings, including the work of Kleinberg on approximate nearest neighbors [8] and of Vempala on learning intersections of halfspaces [12]. More recently, such projections have also been used in learning mixture of Gaussians models, starting with the work of Dasgupta [2] and later with the work of Arora and Kannan [1].

Our proof will suggest that for any fixed vector α, the behavior of its projection onto a random vector c is mandated by the even moments of $||\alpha \cdot c||$. In fact, our result follows by showing that for every vector α, under our distributions for $\{r_{ij}\}$, these moments are dominated by the corresponding moments for the case where c is spherically symmetric. As a result, projecting onto vectors whose entries are distributed

like the columns of matrix R could replace projection onto random lines; it is computationally simpler and results in projections that are at least as nicely behaved.

Finally, we note that Theorem 2 allows one to use significantly fewer random bits than all previous methods for constructing JL-embeddings. While the amount of randomness needed is still quite large, such attempts for randomness reduction are of independent interest and our result can be viewed as a first step in that direction.

2. PREVIOUS WORK

Let us write $X \stackrel{D}{=} Y$ to denote that X is distributed as Y and recall that $N(0,1)$ denotes the standard Normal random variable having mean 0 and variance 1.

As we will see, in all methods for producing JL-embeddings, including ours, the heart of the matter is showing that for any vector, the squared length of its projection is sharply concentrated around its expected value. Armed with a sufficiently strong such concentration bound, one then proves the assertion of Lemma 1 for a collection of n points in $\mathbb{R}^d$ by applying the union bound for the $\binom{n}{2}$ events corresponding to each distance-vector being distorted by more than $(1 \pm \epsilon)$.

The original proof of Johnson and Lindenstrauss [7] uses quite heavy geometric approximation machinery to yield such a concentration bound when the projection is onto a uniformly random hyperplane through the origin. That proof was greatly simplified and sharpened by Frankl and Meahara [4] who considered a direct projection onto k random orthonormal vectors, yielding the following result.

THEOREM 3 ([4]). *For any $\epsilon \in (0, 1/2)$, any sufficiently large set $P \in \mathbb{R}^d$, and $k \geq k_0 = \lceil 9(\epsilon^2 - 2\epsilon^3/3)^{-1} \log |P| \rceil + 1$, there exists a map $f : P \to \mathbb{R}^k$ such that for all $u, v \in P$,*

$$(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+\epsilon)||u-v||^2 \ .$$

The next great simplification of the proof of Lemma 1 was given, independently, by Indyk and Motwani [6] and Dasgupta and Gupta [3], the latter also giving a slight sharpening of the bound for k_0. Below we state our rendition of how this simplification was achieved.

Assume that we try to implement the scheme of Frankl and Maehara [4] but we are lazy about enforcing either normality (unit length) or orthogonality among our k vectors. Instead, we just pick our k vectors independently, in a spherically symmetric manner. As we saw earlier, we can achieve this by taking as the coordinates of each vector independent $N(0,1)$ random variables. We then merely scale each vector by $1/\sqrt{d}$ so that its expected length is 1.

An immediate gain of this approach is that now, for any fixed vector α, the length of its projection onto each of our vectors is also a Normal random variable. This is due to a powerful and deep fact, namely the 2-stability of the Gaussian distribution: for any real numbers $\alpha_1, \alpha_2, \ldots, \alpha_d$, if $\{Z_i\}_{i=1}^d$ is a family of independent Normal random variables and $X = \sum_{i=1}^d \alpha_i Z_i$, then $X \stackrel{D}{=} c\,N(0,1)$, where $c = (\alpha_1^2 + \cdots + \alpha_d^2)^{1/2}$. As a result, if we interpret each of the k projection lengths as a coordinate in $\mathbb{R}^k$, then the squared length of the resulting vector follows the Chi-square distribution for which strong concentration bounds are readily available.

And what have we lost? Surprisingly little. While we did not insist upon either orthogonality, or normality, with high probability, the resulting k vectors come very close to having both these properties. In particular, the length of each of the k vectors is sharply concentrated (around 1) as the sum of d independent random variables. Moreover, since the k vectors point in uniformly random directions in $\mathbb{R}^d$, they get rapidly closer to being orthogonal as d grows.

Unlike Indyk and Motwani [6], Dasgupta and Gupta [3] exploited spherical symmetry without appealing directly to the 2-stability of the Gaussian distribution. Instead they observe that, by symmetry, the projection of any unit vector α on a random hyperplane through the origin is distributed exactly like the projection of a random point from the surface of the d-dimensional sphere onto a fixed subspace of dimension k. Such a projection can be studied readily, though, as now each coordinate is a scaled Normal random variable. Their analysis gave the strongest known bound, namely $k \geq k_0 = 4(\epsilon^2/2 - \epsilon^3/3)^{-1}$. Note that this is exactly the same as our bound in Theorem 2 as β tends to 0.

3. SOME INTUITION

By combining the analysis of [3] with the viewpoint of [6] it is in fact not hard to show that Theorem 2 holds if for all i, j, $r_{ij} \stackrel{D}{=} N(0,1)$. Thus, our contribution essentially begins with the realization that spherical symmetry, while making life extremely comfortable, is not essential. What is essential is concentration. So, at least in principle, one is free to consider other candidate distributions for the $\{r_{ij}\}$, if perhaps at the expense of comfort.

As we saw earlier, each column of our matrix R will give us a coordinate of the projection in $\mathbb{R}^k$. Moreover, the squared length of the projection is merely the sum of the squares of these coordinates. So, effectively, each column acts as an estimator of the original vector's length (by taking its inner product with it), while in the end we take the consensus estimate (sum) over our k estimators. From this point of view, requiring our k vectors to be orthonormal has the pleasant statistical interpretation of "greatest efficiency". In any case, though, as long as each column is an unbiased, bounded variance estimator the Central Limit Theorem asserts that by taking enough columns we can get an arbitrarily good estimate of the original length. Naturally, how many estimators are "enough" depends solely on the variance of the estimators.

So, already we see that the key issue is the concentration of the projection of an arbitrary fixed vector α onto a single random vector. The main technical difficulty that results from giving up spherical symmetry is that this concentration can now depend on α. Our main technical contribution lies in determining probability distributions for $\{r_{ij}\}$ for which this concentration, for all vectors, is as good as when $r_{ij} \stackrel{D}{=} N(0,1)$. In fact, it will turn out that for every *fixed* value of d, we can get a (minuscule) improvement

over the concentration for that case. Thus, for every fixed d, we can actually get a *strictly better* bound for k, albeit marginally, than by taking spherically random vectors.

The reader might be wondering "how can it be that perfect spherical symmetry does not buy us anything?" (and is in fact slightly worse for each fixed d). At a high level, an answer to this question might go as follows. Given that we do not have spherical symmetry anymore, an adversary could try to pick a vector α so that the length of its projection is as variable as possible. It is clear that not all vectors α are equal with respect to this variability. What then does a worst-case vector w look like? How much are we exposing to the adversary by committing to pick our column vectors among lattice points rather than arbitrary points in $\mathbb{R}^d$?

As we will see, the worst-case vector is $w = (1/\sqrt{d})\ (1, \ldots, 1)$ (and all 2^d vectors resulting by sign-flipping w's coordinates). So, the worst-case vector turns out to be a more or less "typical" vector, at least in terms of the fluctuations in its coordinates, unlike say $(1, 0, \ldots, 0)$. As a result it is not hard to believe that the adversary would not fare much worse by picking a random vector. But in that case the adversary does not benefit at all from our commitment.

To get a more satisfactory answer, it seems like one has to delve into the proof. In particular, both for the spherically random case and for our distributions, the bound on k is mandated by the probability of overestimating the projected length. Thus, the "bad events" amount to the spanning vectors being too "well-aligned" with α. As a result, for any fixed d one has to consider the tradeoff between the probability and the extent of alignment.

For example, let us consider the projection onto a single random vector when $d = 2$ and $r_{ij} \in \{-1, +1\}$. As we said above, the worst case vector is $w = (1/\sqrt{2})(1, 1)$. So, it's easy to see that with probability $1/2$ we have perfect alignment (when our random vector is $\pm w$) and with probability $1/2$ we have orthogonality. On the other hand, for the spherically symmetric case, we have to consider the integral over all points on the plane, weighted by their probability under the two-dimensional Gaussian distribution. By a convexity argument it turns out that for every fixed d, the even moments of the projected length are (marginally) greater in the spherically symmetric case. This leads to a (marginally) weaker probability bound for that case. As one might guess, the two bounds coincide as d tends to infinity.

4. PRELIMINARIES

Let $x \cdot y$ denote the inner product of vectors x, y. To simplify notation in the calculations, we will work with matrix R scaled by $1/\sqrt{d}$. Thus, R is a random $d \times k$ matrix with $R(i,j) = r_{ij}/\sqrt{d}$, where the $\{r_{ij}\}$ are distributed as in Theorem 2. As a result, to get E we need to scale $A \times R$ by $\sqrt{d/k}$ rather than $1/\sqrt{k}$. Therefore, if c_j denotes the j^{th} column of R, then $\{c_j\}_{j=1}^k$ is a family of k i.i.d. random unit vectors in $\mathbb{R}^d$ and for all $\alpha \in \mathbb{R}^d$, $f(\alpha) = \sqrt{d/k}\ (\alpha \cdot c_1, \ldots, \alpha \cdot c_d)$.

In practice, of course, such scaling can be postponed until after the matrix multiplication (projection) has been performed, so that we maintain the advantage of only having $\{-1, 0, +1\}$ in the projection matrix.

Let us start by computing $\mathbf{E}\left(||f(\alpha)||^2\right)$ for an arbitrary vector $\alpha \in \mathbb{R}^d$. Let $\{Q_j\}_{j=1}^k$ be defined as

$$Q_j = \alpha \cdot c_j \ .$$

Then

$$\mathbf{E}\left(Q_j\right) = \mathbf{E}\left(\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \alpha_i r_{ij}\right) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \alpha_i \mathbf{E}\left(r_{ij}\right) = 0 \ , \quad (1)$$

and

$$\begin{aligned}
\mathbf{E}\left(Q_j^2\right) &= \mathbf{E}\left(\left(\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \alpha_i r_{ij}\right)^2\right) \\
&= \frac{1}{d}\mathbf{E}\left(\sum_{i=1}^{d} \left(\alpha_i r_{ij}\right)^2 + \sum_{l=1}^{d}\sum_{m=1}^{d} 2\alpha_l \alpha_m r_{lj} r_{mj}\right) \\
&= \frac{1}{d} \sum_{i=1}^{d} \alpha_i^2\, \mathbf{E}\left(r_{ij}^2\right) + \frac{1}{d} \sum_{l=1}^{d}\sum_{m=1}^{d} 2\alpha_l \alpha_m \mathbf{E}\left(r_{lj}\right) \mathbf{E}\left(r_{mj}\right) \\
&= \frac{1}{d} \times ||\alpha||^2 \ . \quad (2)
\end{aligned}$$

Note that to get (1) and (2) we only used that $\{r_{ij}\}$ are independent, $\mathbf{E}(r_{ij}) = 0$ and $\mathrm{Var}(r_{ij}) = 1$. Using (2) we get

$$\mathbf{E}\left(||f(\alpha)||^2\right) = \frac{d}{k} \times \sum_{j=1}^{k} \mathbf{E}\left(Q_j^2\right) = ||\alpha||^2 \ .$$

That is, $\mathbf{E}\left(||f(\alpha)||^2\right) = ||\alpha||^2$ for *any* independent family of $\{r_{ij}\}$ with $\mathbf{E}(r_{ij}) = 0$ and $\mathrm{Var}(r_{ij}) = 1$.

From the above we see that any distribution where $\mathbf{E}(r_{ij}) = 0$ and $\mathrm{Var}(r_{ij}) = 1$ is, in principle, a candidate for the entries of R. In fact, in [13], Arriaga and Vempala independently suggested the possibility of getting JL-embeddings by projecting onto a matrix where $r_{ij} \in \{-1, +1\}$ but did not give any bounds on the necessary value of k.

As we mentioned earlier, having a JL-embedding amounts to the following: for each of the $\binom{n}{2}$ pairs $u, v \in P$, the squared norm of the vector $u - v$, is maintained within a factor of $1 \pm \epsilon$. Therefore, if we can prove that for some $\beta > 0$ and every vector $\alpha \in \mathbb{R}^d$,

$$\Pr[(1-\epsilon)||\alpha||^2 \le ||f(\alpha)||^2 \le (1+\epsilon)||\alpha||^2] \ge 1 - \frac{2}{n^{2+\beta}} \ , \quad (3)$$

then the probability that our projection does not yield a JL-embedding is bounded by $\binom{n}{2} \times 2/n^{2+\beta} < 1/n^{\beta}$.

Let us note that since for a fixed projection matrix, $||f(\alpha)||^2$ is proportional to $||\alpha||$, it suffices to consider probability bounds for arbitrary *unit* vectors. Moreover, note that when $\mathbf{E}(||f(\alpha)||^2) = ||\alpha||^2$, inequality (3) merely asserts that the random variable $||f(\alpha)||^2$ is concentrated around its expectation. Before considering this point for our distributions for $\{r_{ij}\}$, let us first wrap up the spherically random case.

Getting a concentration inequality for $||f(\alpha)||^2$ when $r_{ij} \stackrel{D}{=} N(0,1)$ is straightforward. Due to the 2-stability of the Normal distribution, for *every* unit vector α, we have $||f(\alpha)||^2 \stackrel{D}{=}$

$\chi^2(k)/k$, where $\chi^2(k)$ denotes the Chi-square distribution with k degrees of freedom. The fact that we get the same distribution for every vector α corresponds to the intuition that "all vectors are the same" with respect to projection onto a spherically random vector. Standard tail-bounds for the Chi-square distribution readily yield the following.

LEMMA 4. *Let $r_{ij} \stackrel{D}{=} N(0,1)$ for all i,j. Then, for any $\epsilon > 0$ and any unit-vector $\alpha \in \mathbb{R}^d$,*

$$\begin{aligned}\Pr\left[||f(\alpha)||^2 \geq (1+\epsilon)k/d\right] &< \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \ , \\ \Pr\left[||f(\alpha)||^2 \leq (1-\epsilon)k/d\right] &< \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \ .\end{aligned}$$

Thus, to get a JL-embedding we need only require

$$2 \times \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \leq \frac{2}{n^{2+\beta}} \ ,$$

which holds for

$$k \geq \frac{4+2\beta}{\epsilon^2/2-\epsilon^3/3} \ \log n \ .$$

Let us note that the bound on the upper tail of $||f(\alpha)||^2$ above is *tight* (up to lower order terms). As a result, as long as the union bound is used, one cannot hope for a better bound on k while using spherically random vectors.

To prove our result we use the exact same approach, arguing that for every unit vector $\alpha \in \mathbb{R}^d$, the random variable $||\alpha c||^2$ is sharply concentrated around its expectation, where c is a column of our projection matrix R. In the next section we state a lemma analogous to Lemma 4 above and show how it follows from bounds on certain moments of $||\alpha c||^2$. We prove those bounds in Section 6.

5. PROBABILITY BOUNDS

To simplify notation let us define for an arbitrary vector α,

$$S = S(\alpha) = \sum_{j=1}^{k} (\alpha \cdot c_j)^2 = \sum_{j=1}^{k} Q_j^2(\alpha) \ ,$$

where c_j is the j^{th} column of R, so that $||f(\alpha)||^2 = S \times d/k$.

LEMMA 5. *Let r_{ij} have any of the two distributions in Theorem 2. Then, for any $\epsilon > 0$ and any unit vector $\alpha \in \mathbb{R}^d$,*

$$\begin{aligned}\Pr[S > (1+\epsilon)k/d] &< \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \ , \\ \Pr[S < (1-\epsilon)k/d] &< \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \ .\end{aligned}$$

In proving Lemma 5 we will generally omit the dependence of probabilities on α, making it explicit only when it affects our calculations.

We will use the standard technique of applying Markov's inequality to the moment generating function of S. In particular, for arbitrary $h > 0$ we write

$$\begin{aligned}\Pr\left[S > (1+\epsilon)\frac{k}{d}\right] &= \Pr\left[\exp(hS) > \exp\left(h(1+\epsilon)\frac{k}{d}\right)\right] \\ &< \mathbf{E}\left(\exp\left(hS\right)\right) \exp\left(-h(1+\epsilon)\frac{k}{d}\right) \ .\end{aligned}$$

Since $\{Q_j\}_{j=1}^k$ are i.i.d. we get

$$\begin{aligned}\mathbf{E}\left(\exp\left(hS\right)\right) &= \mathbf{E}\left(\prod_{j=1}^{k} \exp\left(hQ_j^2\right)\right) &(4)\\ &= \prod_{j=1}^{k} \mathbf{E}\left(\exp\left(hQ_j^2\right)\right) &(5)\\ &= \left(\mathbf{E}\left(\exp\left(hQ_1^2\right)\right)\right)^k \ , &(6)\end{aligned}$$

where passing from (4) to (5) uses that the $\{Q_j\}_{j=1}^k$ are independent, while passing from (5) to (6) uses that they are identically distributed. Thus, for any $\epsilon > 0$

$$\begin{aligned}&\Pr\left[S > (1+\epsilon)\frac{k}{d}\right] \\ &< \left(\mathbf{E}\left(\exp\left(hQ_1^2\right)\right)\right)^k \exp\left(-h(1+\epsilon)\frac{k}{d}\right) \ . &(7)\end{aligned}$$

We will get a tight bound on $\mathbf{E}\left(\exp\left(hQ_1^2\right)\right)$ from Lemma 6 below.

Similarly, but this time considering $\exp(-hS)$ for arbitrary $h > 0$, we get that for any $\epsilon > 0$

$$\begin{aligned}&\Pr\left[S < (1-\epsilon)\frac{k}{d}\right] \\ &< \left(\mathbf{E}\left(\exp\left(-hQ_1^2\right)\right)\right)^k \exp\left(h(1-\epsilon)\frac{k}{d}\right) \ . &(8)\end{aligned}$$

Rather than bounding $\mathbf{E}\left(\exp\left(-hQ_1^2\right)\right)$ directly, this time we will expand $\exp(hQ_1^2)$ to get

$$\begin{aligned}&\Pr\left[S < (1-\epsilon)\frac{k}{d}\right] &(9)\\ &< \left(\mathbf{E}\left(1 - hQ_1^2 + \frac{\left(-hQ_1^2\right)^2}{2!}\right)\right)^k \exp\left(h(1-\epsilon)\frac{k}{d}\right) \\ &= \left(1 - \frac{h}{d} + \frac{h^2}{2}\mathbf{E}\left(Q_1^4\right)\right)^k \exp\left(h(1-\epsilon)\frac{k}{d}\right) \ , &(10)\end{aligned}$$

where $\mathbf{E}(Q_1^2)$ was given by (2).

We will get a tight bound on $\mathbf{E}\left(Q_1^4\right)$ from Lemma 6 below.

LEMMA 6. *For all $h \in [0, d/2)$ and all $d \geq 1$,*

$$\begin{aligned}\mathbf{E}\left(\exp\left(hQ_1^2\right)\right) &\leq \frac{1}{\sqrt{1-2h/d}} \ , &(11)\\ \mathbf{E}\left(Q_1^4\right) &\leq \frac{3}{d^2} \ . &(12)\end{aligned}$$

The proof of Lemma 6 will comprise Section 6. Below we show how it implies Lemma 5 and thus Theorem 2.

Proof of Lemma 5. Substituting (11) in (7) we get (13). To optimize the bound we set the derivative in (13) with respect to h to 0. This gives $h = \frac{d}{2}\frac{\epsilon}{1+\epsilon} < \frac{d}{2}$. Substituting this value of h we get (14) and series expansion yields (15).

$$\begin{aligned}
&\Pr\left[S > (1+\epsilon)\frac{k}{d}\right] \\
&\quad\leq \left(\frac{1}{\sqrt{1-2h/d}}\right)^k \exp\left(-h(1+\epsilon)\frac{k}{d}\right) && (13)\\
&\quad= \left((1+\epsilon)\exp(-\epsilon)\right)^{k/2} && (14)\\
&\quad< \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \ . && (15)
\end{aligned}$$

Similarly, substituting (12) in (8) we get (16). This time taking $h = \frac{d}{2}\frac{\epsilon}{1+\epsilon}$ is not optimal but it is "good enough", giving (17). Again, series expansion yields (18).

$$\begin{aligned}
&\Pr\left[S < (1-\epsilon)\frac{k}{d}\right] \\
&\leq \left(1-\frac{h}{d}+\frac{3}{2}\left(\frac{h}{d}\right)^2\right)^k \exp\left(h(1-\epsilon)\frac{k}{d}\right) && (16)\\
&= \left(1-\frac{\epsilon}{2(1+\epsilon)}+\frac{3\epsilon^2}{8(1+\epsilon)^2}\right)^k \exp\left(\frac{\epsilon(1-\epsilon)k}{2(1+\epsilon)}\right) && (17)\\
&< \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \ . && (18)
\end{aligned}$$

□

6. MOMENT BOUNDS

Here we prove bounds on certain moments of Q_1. To simplify notation, we drop the subscript, writing it as Q.

It should be clear that the distribution of Q depends on α, i.e., $Q = Q(\alpha)$. This is precisely what we give up by not projecting onto spherically symmetric vectors. Our strategy for giving bounds on the moments of Q will be to determine a "worst case" unit vector w and consider $Q(w)$. Our precise claim is the following.

LEMMA 7. *Let*

$$w = \frac{1}{\sqrt{d}}(1,\ldots,1) \ .$$

For every unit vector $\alpha \in \mathbb{R}^d$, and for all $k = 0, 1, \ldots$

$$\mathbf{E}\left(Q(\alpha)^{2k}\right) \leq \mathbf{E}\left(Q(w)^{2k}\right) \ . \qquad (19)$$

Moreover, we will prove that the even moments of $Q(w)$ are dominated by the even moments of an appropriately scaled Normal random variable, i.e., the corresponding moments from the spherically symmetric case.

LEMMA 8. *Let*

$$T \stackrel{D}{=} N(0, 1/d) \ .$$

For all $d \geq 1$ and all $k = 0, 1, \ldots$

$$\mathbf{E}\left(Q(w)^{2k}\right) \leq \mathbf{E}\left(T^{2k}\right) \ . \qquad (20)$$

Postponing the proof of Lemmata 7 and 8 for a moment, let us use them to prove Lemma 6.

Proof of Lemma 6. We start by observing that

$$\mathbf{E}\left(T^4\right) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2)\left(\frac{\lambda^4}{d^2}\right) d\lambda = \frac{3}{d^2} \ .$$

Along with (19) and (20) this readily implies (12).

For any real-valued random variable U, the Monotone Convergence Theorem (MCT) implies

$$\mathbf{E}\left(\exp\left(hU^2\right)\right) = \mathbf{E}\left(\sum_{k=0}^{\infty}\frac{(hU^2)^k}{k!}\right) = \sum_{k=0}^{\infty}\frac{h^k}{k!}\mathbf{E}\left(U^{2k}\right)$$

for all h such that $\mathbf{E}\left(\exp\left(hU^2\right)\right)$ is bounded.

For $\mathbf{E}\left(\exp\left(hT^2\right)\right)$, below, taking $h \in [0, d/2)$ makes the integral converge, giving (21). Thus, for such h we can apply the MCT to get (22). Now, applying (19) and (20) to (22) gives (23). Applying the MCT once more gives (24).

$$\begin{aligned}
\mathbf{E}\left(\exp\left(hT^2\right)\right) &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp(-\lambda^2/2)\exp\left(h\frac{\lambda^2}{d}\right)d\lambda \\
&= \frac{1}{\sqrt{1-2h/d}} && (21)\\
&= \sum_{k=0}^{\infty}\frac{h^k}{k!}\mathbf{E}\left(T^{2k}\right) && (22)\\
&\geq \sum_{k=0}^{\infty}\frac{h^k}{k!}\mathbf{E}\left(Q^{2k}\right) && (23)\\
&= \mathbf{E}\left(\exp\left(hQ^2\right)\right) \ . && (24)
\end{aligned}$$

Thus, $\mathbf{E}\left(\exp\left(hQ^2\right)\right) \leq 1/\sqrt{1-2h/d}$ for $h \in [0, d/2)$, which is precisely inequality (11). □

To prove Lemma 7 we need the following lemma. It's proof appears in the Appendix.

LEMMA 9. *Let r_1, r_2 be i.i.d. r.v. having one of the following two probability distributions: $r_i \in \{-1, +1\}$, each value having probability 1/2, or, $r_i \in \{-\sqrt{3}, 0, +\sqrt{3}\}$ with 0 having probability 2/3 and $\pm\sqrt{3}$ being equiprobable.*

For real numbers a, b let $c = \sqrt{(a^2+b^2)/2}$. Then, for all T and all $k = 0, 1, \ldots$

$$\mathbf{E}\left((T + ar_1 + br_2)^{2k}\right) \leq \mathbf{E}\left((T + cr_1 + cr_2)^{2k}\right) \ .$$

Proof of Lemma 7. Recall that for any vector α, $Q(\alpha) = Q_1(\alpha) = \alpha \cdot c_1$ where

$$c_1 = \frac{1}{\sqrt{d}}(r_{11}, \ldots, r_{d1}) \ .$$

If $\alpha = (\alpha_1, \ldots, \alpha_d)$ is such that $\alpha_i^2 = \alpha_j^2$ for all i, j, then by symmetry, $Q(\alpha)$ and $Q(w)$ are identically distributed and the lemma holds trivially. Otherwise, we can assume without loss of generality, that $\alpha_1^2 \neq \alpha_2^2$ and consider the

"more balanced" unit vector $\theta = (c, c, \alpha_3, \ldots, \alpha_d)$, where $c = \sqrt{(\alpha_1^2 + \alpha_2^2)/2}$. We will prove that

$$\mathbf{E}\left(Q(\alpha)^{2k}\right) \leq \mathbf{E}\left(Q(\theta)^{2k}\right) \quad . \tag{25}$$

Applying this argument repeatedly yields the lemma, as θ eventually becomes w.

To prove (25), below we first express $\mathbf{E}\left(Q(\alpha)^{2k}\right)$ as a sum of averages over r_{11}, r_{21}. We then apply Lemma 9 to get that each term (average) in the sum, is bounded by the corresponding average for vector θ. More precisely,

$$\begin{aligned}
&\mathbf{E}\left(Q(\alpha)^{2k}\right) \\
&= \frac{1}{d^k} \sum_R \mathbf{E}\left((R + \alpha_1 r_{11} + \alpha_2 r_{21})^{2k}\right) \Pr\left[\sum_{i=3}^{d} \alpha_i r_{i1} = \frac{R}{\sqrt{d}}\right] \\
&\leq \frac{1}{d^k} \sum_R \mathbf{E}\left((R + c r_{11} + c r_{21})^{2k}\right) \Pr\left[\sum_{i=3}^{d} \alpha_i r_{i1} = \frac{R}{\sqrt{d}}\right] \\
&= \mathbf{E}\left(Q(\theta)^{2k}\right) \quad .
\end{aligned}$$

□

Proof of Lemma 8. Recall that $T \stackrel{D}{=} N(0, 1/d)$. We will first express T as the scaled sum of d independent standard Normal random variables. This will allow for a direct comparison of the terms in each of the two expectations.

Specifically, let $\{T_i\}_{i=1}^d$ be a family of i.i.d. standard Normal random variables. Then $\sum_{i=1}^d T_i$ is a Normal random variable with variance d. Therefore,

$$T \stackrel{D}{=} \frac{1}{d} \sum_{i=1}^{d} T_i \quad .$$

Recall also that $Q(w) = Q_1(w) = w \cdot c_1$ where

$$c_1 = \frac{1}{\sqrt{d}} (r_{11}, \ldots, r_{d1}) \quad .$$

To simplify notation let us write $r_{i1} = Y_i$ and let us also drop the dependence of Q on w. Thus,

$$Q = \frac{1}{d} \sum_{i=1}^{d} Y_i \quad ,$$

where $\{Y_i\}_{i=1}^d$ are i.i.d. r.v. having one of the following two distributions: $Y_i \in \{-1, +1\}$, each value having probability $1/2$, or $Y_i \in \{-\sqrt{3}, 0, +\sqrt{3}\}$ with 0 having probability $2/3$ and $\pm\sqrt{3}$ being equiprobable.

We are now ready to compare $\mathbf{E}\left(Q^{2k}\right)$ with $\mathbf{E}\left(T^{2k}\right)$. We first observe that for every $k = 0, 1, \ldots$

$$\begin{aligned}
\mathbf{E}\left(T^{2k}\right) &= \frac{1}{d^{2k}} \sum_{i_1=1}^{d} \cdots \sum_{i_{2k}=1}^{d} \mathbf{E}\left(T_{i_1} \cdots T_{i_{2k}}\right) \quad , \text{ and} \\
\mathbf{E}\left(Q^{2k}\right) &= \frac{1}{d^{2k}} \sum_{i_1=1}^{d} \cdots \sum_{i_{2k}=1}^{d} \mathbf{E}\left(Y_{i_1} \cdots Y_{i_{2k}}\right) \quad .
\end{aligned}$$

To prove the lemma we will show that for every value assignment to the indices $i_1, \ldots, i_{2k}$,

$$\mathbf{E}\left(Y_{i_1} \cdots Y_{i_{2k}}\right) \leq \mathbf{E}\left(T_{i_1} \cdots T_{i_{2k}}\right) \quad . \tag{26}$$

Let $V = \langle v_1, v_2, \ldots, v_{2k} \rangle$ be the value assignment considered. For $i \in \{1, \ldots, d\}$, let $C_V(i)$ be the number of times that i appears in V. Observe that if for some i, $c_V(i)$ is odd then both expectations appearing in (26) are 0, since both $\{Y_i\}_{i=1}^d$ and $\{T_i\}_{i=1}^d$ are independent families and $\mathbf{E}(Y_i) = \mathbf{E}(T_i) = 0$ for all i. Thus, we can assume that there exists a set $\{j_1, j_2, \ldots, j_p\}$ of indices and corresponding values $\ell_1, \ell_2, \ldots, \ell_p$ such that

$$\begin{aligned}
\mathbf{E}\left(Y_{i_1} \cdots Y_{i_{2k}}\right) &= \mathbf{E}\left(Y_{j_1}^{2\ell_1} Y_{j_2}^{2\ell_2} \cdots Y_{j_p}^{2\ell_p}\right), \quad \text{and} \\
\mathbf{E}\left(T_{i_1} \cdots T_{i_{2k}}\right) &= \mathbf{E}\left(T_{j_1}^{2\ell_1} T_{j_2}^{2\ell_2} \cdots T_{j_p}^{2\ell_p}\right) \quad .
\end{aligned}$$

Note now that since the indices $j_1, j_2, \ldots, j_p$ are distinct, $\{Y_{j_t}\}_{t=1}^p$ and $\{T_{j_t}\}_{t=1}^p$ are families of i.i.d. r.v. Therefore,

$$\begin{aligned}
\mathbf{E}\left(Y_{i_1} \cdots Y_{i_{2k}}\right) &= \mathbf{E}\left(Y_{j_1}^{2\ell_1}\right) \times \cdots \times \mathbf{E}\left(Y_{j_p}^{2\ell_p}\right), \quad \text{and} \\
\mathbf{E}\left(T_{i_1} \cdots T_{i_{2k}}\right) &= \mathbf{E}\left(T_{j_1}^{2\ell_1}\right) \times \cdots \times \mathbf{E}\left(T_{j_p}^{2\ell_p}\right) \quad .
\end{aligned}$$

So, without loss of generality, in order to prove (26) it suffices to prove that for every $\ell = 0, 1, \ldots$

$$\mathbf{E}\left(Y_1^{2\ell}\right) \leq \mathbf{E}\left(T_1^{2\ell}\right) \quad . \tag{27}$$

This, though, is completely trivial. Moreover, along with Lemma 9, it is the only point were we need to use properties of the distribution for the r_{ij} (here called Y_i).

Let us first recall the well-known fact that the (2ℓ)th moment of $N(0,1)$ is $(2\ell-1)!! = (2\ell)!/(\ell! 2^\ell) \geq 1$. Furthermore:
– If $Y_1 \in \{-1, +1\}$ then $\mathbf{E}\left(Y_1^{2\ell}\right) = 1$.
– If $Y_1 \in \{-\sqrt{3}, 0, +\sqrt{3}\}$ then $\mathbf{E}(Y_1^{2\ell}) = 3^{\ell-1} \leq (2\ell)!/(\ell! 2^\ell)$, where the last inequality follows by an easy induction. □

Let us note that since $\mathbf{E}\left(Y_1^{2\ell}\right) < \mathbf{E}\left(T_1^{2\ell}\right)$ for certain l, one can get that for each fixed d, both inequalities in Lemma 6 are actually strict, yielding slightly better tails bounds for S and a correspondingly better bound for k_0.

As a last remark we note that by using Jensen's inequality one can get a direct bound for $\mathbf{E}(Q^{2k})$ when $Y_i \in \{-1, +1\}$, i.e., without comparing it to $\mathbf{E}(T^{2k})$. That simplifies the proof for that case and shows that taking $Y_i \in \{-1, +1\}$ is the minimizer of $\mathbf{E}\left(\exp\left(hQ^2\right)\right)$ for all h.

Acknowledgments

I am grateful to Marek Biskup for his help with the proof of Lemma 8 and to Jeong Han Kim for suggesting the approach of equation (10). Many thanks also to Paul Bradley, Anna Karlin, Elias Koutsoupias and Piotr Indyk for comments on earlier versions of the paper and useful discussions.

7. REFERENCES

[1] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. Submitted, 2000.

[2] S. Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (New York, NY, 1999)*, pages 634–644. IEEE Comput. Soc. Press, Los Alamitos, CA, 1999.

[3] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Technical report 99-006, UC Berkeley, March 1999.

[4] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Combin. Theory Ser. B*, 44(3):355–362, 1988.

[5] P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000)*, pages 189–197. IEEE Comput. Soc. Press, Los Alamitos, CA, 2000.

[6] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *30th Annual ACM Symposium on Theory of Computing (Dallas, TX)*, pages 604–613. ACM, New York, 1998.

[7] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, pages 189–206. Amer. Math. Soc., Providence, R.I., 1984.

[8] J. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *29th Annual ACM Symposium on Theory of Computing (El Paso, TX, 1997)*, pages 599–608. ACM, New York, 1997.

[9] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[10] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *17th Annual Symposium on Principles of Database Systems (Seattle, WA, 1998)*, pages 159–168, 1998.

[11] L. J. Schulman. Clustering for edge-cost minimization. In *32nd Annual ACM Symposium on Theory of Computing (Portland, OR, 2000)*, pages 547–555. ACM, New York, 2000.

[12] S. Vempala. A random sampling based algorithm for learning the intersection of half-spaces. In *38th Annual Symposium on Foundations of Computer Science (Miami, FL, 1997)*, pages 508–513. IEEE Comput. Soc. Press, Los Alamitos, CA, 1997.

[13] S. Vempala and R. I. Arriaga. An algorithmic theory of learning: robust concepts and random projection. In *40th Annual Symposium on Foundations of Computer Science (New York, NY, 1999)*, pages 616–623. IEEE Comput. Soc. Press, Los Alamitos, CA, 1999.

APPENDIX

Proof of Lemma 9. Let us first consider the case where $r_i \in \{-1, +1\}$, each value having probability 1/2.

If $a^2 = b^2$ then $a = c$ and the lemma holds with equality. Otherwise, let us write

$$\mathbf{E}\left((T + cr_1 + cr_2)^{2k}\right) - \mathbf{E}\left((T + ar_1 + br_2)^{2k}\right) = \frac{S_k}{4}$$

where

$$\begin{aligned} S_k &= (T+2c)^{2k} + 2T^{2k} + (T-2c)^{2k} - (T+a+b)^{2k} \\ &\quad -(T+a-b)^{2k} - (T-a+b)^{2k} - (T-a-b)^{2k} \ . \end{aligned}$$

We will show that $S_k \geq 0$ for all $k \geq 0$.

Since $a^2 \neq b^2$ we can use the binomial theorem to expand every term other than $2T^{2k}$ in S_k and get

$$S_k = 2T^{2k} + \sum_{i=0}^{2k} \binom{2k}{i} T^{2k-i} D_i \ ,$$

where

$$D_i = (2c)^i + (-2c)^i - (a+b)^i - (a-b)^i - (-a+b)^i - (-a-b)^i \ .$$

Observe now that for odd i, $D_i = 0$. Moreover, we claim that $D_{2j} \geq 0$ for all $j \geq 1$. To see this claim observe that $(2a^2 + 2b^2) = (a+b)^2 + (a-b)^2$ and that for all $j \geq 1$ and $x, y \geq 0$, $(x+y)^j \geq x^j + y^j$. Thus,

$$\begin{aligned} S_k &= 2T^{2k} + \sum_{j=0}^{k} \binom{2k}{2j} T^{2(k-j)} D_{2j} \\ &= \sum_{j=1}^{k} \binom{2k}{2j} T^{2(k-j)} D_{2j} \\ &\geq 0 \ . \end{aligned}$$

The proof for the case where $r_i \in \{-\sqrt{3}, 0, +\sqrt{3}\}$ is merely a more cumbersome version of the proof above, so we omit it. That proof, though, brings forward an interesting point. If one tries to take $r_i = 0$ with probability greater than 2/3, while maintaining a range of size 3 and variance 1, the lemma fails. In other words, 2/3 is tight in terms of how much probability mass we can put to $r_i = 0$ and still have the all-ones vector be the worst-case one. □