Random projection trees for vector quantization

Sanjoy Dasgupta and Yoav Freund

***Abstract*—A simple and computationally efficient scheme for tree-structured vector quantization is presented. Unlike previous methods, its quantization error depends only on the intrinsic dimension of the data distribution, rather than the apparent dimension of the space in which the data happen to lie.**

***Index Terms*—Vector quantization, source coding, random projection, manifolds, computational complexity.**

I. Introduction

We study algorithms for vector quantization codebook design, which we define as follows. The *input* to the algorithm is a set of n vectors $S = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R}^D$. The output of the algorithm is a set of k vectors $R = \{\mu_1, \ldots, \mu_k\}, \mu_i \in \mathbb{R}^D$, where k is much smaller than n. The set R is called the *codebook*. We say that R is a good codebook for S if for most $x \in S$ there is a *representative* $r \in R$ such that the Euclidean distance between x and r is small. We define the *average quantization error* of R with respect to S as:

$$\mathrm{Q}(R, S) = \mathbb{E}\left[\min_{1 \le j \le k} \|X - \mu_j\|^2\right] = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - \mu_j\|^2$$

where $\|\cdot\|$ denotes Euclidean norm and the expectation is over X drawn uniformly at random from S.[1] The goal of the algorithm is to construct a codebook R with a small average quantization error. The *k-optimal set of centers* is defined to be the codebook R of size k for which $\mathrm{Q}(R, S)$ is minimized; the task of finding such a codebook is sometimes called the *k-means* problem.

It is known that for general sets in $\mathbb{R}^D$ of diameter one, the average quantization error is roughly $k^{-2/D}$ for large k[8]. This is discouraging when D is high. For instance, if $D = 100$, and A is the average quantization error for k_1 codewords, then to guarantee a quantization error of $A/2$ one needs a codebook of size $k_2 \approx 2^{D/2}k_1$: that is, 2^{50} times as many codewords just to halve the error. In other words, vector quantization is susceptible to the same *curse of dimensionality* that has been the bane of other nonparametric statistical methods.

A recent positive development in statistics and machine learning has been the realization that quite often, datasets that are represented as collection of vectors in $\mathbb{R}^D$ for some large value of D, actually have low *intrinsic* dimension, in the sense of lying close to a manifold of dimension $d \ll D$. We will give several examples of this below. There has thus been increasing interest in algorithms that *learn* this manifold from data, with the intention that future data can then be transformed into this low-dimensional space, in which the usual nonparametric (and other) methods will work well [19], [17], [2].

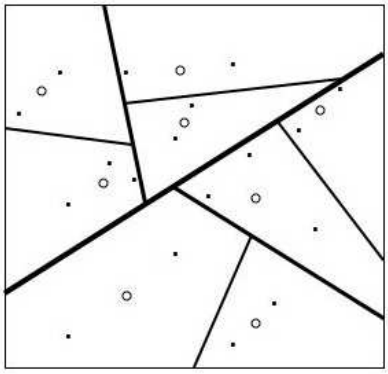


Fig. 1. Spatial partitioning of $\mathbb{R}^2$ induced by an RP tree with three levels. The dots are data points; each circle represents the mean of the vectors in one cell.

In this paper, we are interested in techniques that automatically adapt to intrinsic low dimensional structure without having to explicitly learn this structure. We describe an algorithm for designing a tree-structured vector quantizer whose quantization error is $k^{-1/O(d)}$ (times the quantization error induced by a single codeword); that is to say, its error rate depends only on the low intrinsic dimension rather than the high apparent dimension. The algorithm is based on a hierarchical decomposition of $\mathbb{R}^D$: first the entire space is split into two pieces, then each of these pieces is further split in two, and so on, until a desired average quantization error is reached. Each codeword is the average of the examples that belong to a single cell.

Tree-structured vector quantizers abound; the power of our approach comes from the particular splitting method. To divide a region S into two, we pick a random direction from the surface of the unit sphere in $\mathbb{R}^D$, and split S at the median of its projection onto this direction (Figure 1).[2] We call the resulting spatial partition a *random projection tree* or *RP tree*.

At first glance, it might seem that a better way to split a region is to find the 2-optimal set of centers for it. However, as we explain below, this is an NP-hard optimization problem, and is therefore unlikely to be computationally tractable. Although there are several algorithms that attempt to solve this problem, such as Lloyd's method [13], [12], they are not in general able to find the optimal solution. In fact, they are often far from optimal. A related option would be to use an approximation algorithm for 2-means: an algorithm

Both authors are with the Department of Computer Science and Engineering, University of California, San Diego. Email: {dasgupta,yfreund}@cs.ucsd.edu. This work was supported by the National Science Foundation under grants IIS-0347646, IIS-0713540, and IIS-0812598.

[1]The results presented in this paper generalize to the case where S is infinite and the expectation is taken with respect to a probability measure over S. However, as our focus is on algorithms whose input is a finite set, we assume, throughout the paper, that the set S is finite.

[2]There is also a second type of split that we occasionally use.

that is guaranteed to return a solution whose cost is at most $(1+\epsilon)$ times the optimal cost, for some $\epsilon > 0$. However, for our purposes, we would need $\epsilon \approx 1/d$, and the best known algorithm at this time [11] would require a prohibitive running time of $O(2^{d^{O(1)}} Dn)$.

For our random projection trees, we show that if the data have intrinsic dimension d (in a sense we make precise below), then each split pares off about a $1/d$ fraction of the quantization error. Thus, after $\log k$ levels of splitting, there are k cells and the multiplicative change in quantization error is of the form $(1-1/d)^{\log k} = k^{-1/O(d)}$. There is no dependence on the extrinsic dimensionality D.

II. Detailed overview

A. Low-dimensional manifolds

The increasing ubiquity of massive, high-dimensional data sets has focused the attention of the statistics and machine learning communities on the curse of dimensionality. A large part of this effort is based on exploiting the observation that many high-dimensional data sets have low *intrinsic dimension*. This is a loosely defined notion, which is typically used to mean that the data lie near a smooth low-dimensional manifold.

For instance, suppose that you wish to create realistic animations by collecting human motion data and then fitting models to it. A common method for collecting motion data is to have a person wear a skin-tight suit with high contrast reference points printed on it. Video cameras are used to track the 3D trajectories of the reference points as the person is walking or running. In order to ensure good coverage, a typical suit has about $N = 100$ reference points. The position and posture of the body at a particular point of time is represented by a $(3N)$-dimensional vector. However, despite this seeming high dimensionality, the number of degrees of freedom is small, corresponding to the dozen-or-so joint angles in the body. The positions of the reference points are more or less deterministic functions of these joint angles.

Interestingly, in this example the intrinsic dimension becomes even smaller if we *double* the dimension of the embedding space by including for each sensor its relative velocity vector. In this space of dimension $6N$ the measured points will lie very close to the *one* dimensional manifold describing the combinations of locations and speeds that the limbs go through during walking or running.

To take another example, a speech signal is commonly represented by a high-dimensional time series: the signal is broken into overlapping windows, and a variety of filters are applied within each window. Even richer representations can be obtained by using more filters, or by concatenating vectors corresponding to consecutive windows. Through all this, the intrinsic dimensionality remains small, because the system can be described by a few physical parameters describing the configuration of the speaker's vocal apparatus.

In machine learning and statistics, almost all the work on exploiting intrinsic low dimensionality consists of algorithms for learning the structure of these manifolds; or more precisely, for learning embeddings of these manifolds into low-dimensional Euclidean space. Our contribution is a simple and compact data structure that automatically exploits the low intrinsic dimensionality of data on a local level without having to explicitly learn the global manifold structure.

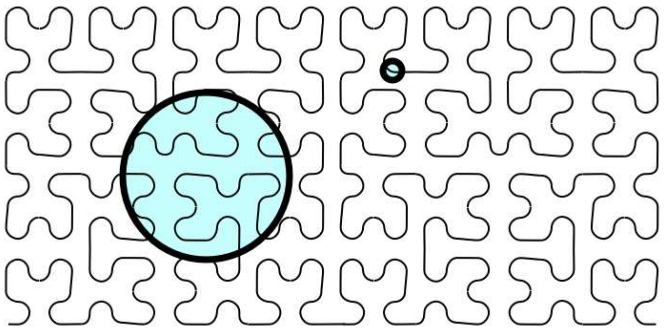


Fig. 2. Hilbert's space filling curve. Large neighborhoods look 2-dimensional, smaller neighborhoods look 1-dimensional, and even smaller neighborhoods would in practice consist mostly of measurement noise and would therefore again be 2-dimensional.

B. Defining intrinsic dimensionality

Low-dimensional manifolds are our inspiration and source of intuition, but when it comes to precisely defining intrinsic dimension for data analysis, the differential geometry concept of manifold is not entirely suitable. First of all, any data set lies on a one-dimensional manifold, as evidenced by the construction of space-filling curves. Therefore, some bound on curvature is implicitly needed. Second, and more important, it is unreasonable to expect data to lie *exactly* on a low-dimensional manifold. At a certain small resolution, measurement error and noise make any data set full-dimensional. The most we can hope is that the data distribution is concentrated *near* a low-dimensional manifold of bounded curvature. Figure 2 illustrates how dimension can vary across the different neighborhoods of a set, depending on the sizes of these neighborhoods and also on their locations.

We address these various concerns with a statistically-motivated notion of dimension: we say $T \subset S$ has *covariance dimension* (d, ϵ) if a $(1-\epsilon)$ fraction of its variance is concentrated in a d-dimensional subspace. To make this precise, let $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_D^2$ denote the eigenvalues of the covariance matrix of T (that is, the covariance matrix of the uniform distribution over the points in T); these are the variances in each of the eigenvector directions.

Definition 1: Set $T \subset \mathbb{R}^D$ has *covariance dimension* (d, ϵ) if the largest d eigenvalues of its covariance matrix satisfy

$$\sigma_1^2 + \cdots + \sigma_d^2 \;\geq\; (1-\epsilon) \cdot (\sigma_1^2 + \cdots + \sigma_D^2).$$

Put differently, this means that T is well-approximated by an affine subspace of dimension d, in the sense that its average squared distance from this subspace is at most ϵ times the overall variance of T.

It is often too much to hope that the entire data set S would have low covariance dimension. The case of interest is when this property holds *locally*, for neighborhoods of S.

Figure 3 depicts a set $S \subset \mathbb{R}^2$ that lies close to a one dimensional manifold. We can imagine that S was generated

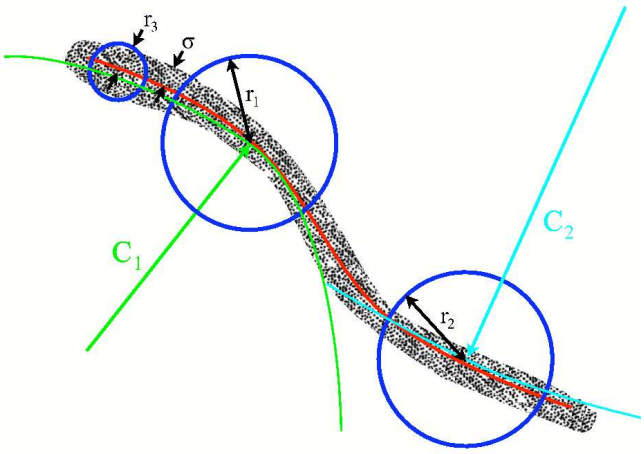


Fig. 3. A data set that lies close to a one-dimensional manifold. Three neighborhoods are shown, indicated by disks. r_i indicate the radii of the disks, C_i indicate the curvature of the set in the neighborhood. σ indicates the standard deviation of the noise added to the manifold.

by selecting points on the manifold according to some distribution and then adding spherical Gaussian noise with standard deviation σ. Consider the restriction of S to a neighborhood defined by a ball B_i of radius r_i (three such disks are shown). The radii of the first two neighborhoods (r_1, r_2) are significantly larger than the noise level σ and significantly smaller than the local curvature radii of the manifold (C_1, C_2). As a result $S \cap B_1$ and $S \cap B_2$ have covariance dimension $(1, \epsilon)$ for $\epsilon \approx (\sigma/r_i)^2 + (r_i/C_i)^2$. On the other hand, $r_3 \approx \sigma$ and therefore the covariance dimension of $S \cap B_3$ is two. In Appendix III, we formally prove a statement of this form for arbitrary d-dimensional manifolds of bounded curvature.

The local covariance dimension captures the essence of intrinsic dimension without being overly sensitive to noise in the dataset. On the other hand, the notion lacks some of the intuitions that we associate with manifolds. In particular, the fact that a set S has covariance dimension (d, ϵ) does *not* imply that subsets of S have low dimension. Covariance dimension is a natural way to characterize finite point sets, but not a good way to characterize differentiable manifolds.

C. Random projection trees

Our new data structure, the random projection tree, is built by recursive binary splits. The core tree-building algorithm is called MAKETREE, which takes as input a data set $S \subset \mathbb{R}^D$, and repeatedly calls a splitting subroutine CHOOSERULE.

procedure MAKETREE(S)
 if $|S| < MinSize$ **then return** $(Leaf)$
 $Rule \leftarrow$ CHOOSERULE(S)
 $LeftTree \leftarrow$ MAKETREE($\{x \in S : Rule(x) = \text{true}\}$)
 $RightTree \leftarrow$ MAKETREE($\{x \in S : Rule(x) = \text{false}\}$)
 return $([Rule, LeftTree, RightTree])$

The RP tree has two types of split. Typically, a direction is chosen uniformly at random from surface of the unit sphere and the cell is split at its median, by a hyperplane orthogonal to this direction. Although this generally works well in terms of decreasing vector quantization error, there are certain situations in which it is inadequate. The prototypical such situation is as follows: the data in the cell lie almost entirely in a dense spherical cluster around the mean, but there is also a concentric shell of points much farther away. This outer shell has the effect of making the quantization error fairly large, and any median split along a hyperplane creates two hemispherical cells with the same balance of inner and outer points, and thus roughly the same quantization error; so the split is not very helpful. To see how such a data configuration might arise in practice, consider a data set consisting of image patches. The vast majority of patches are empty, forming the dense cluster near the mean; the rest are much farther away.

The failure case for the hyperplane split is easy to characterize: it happens only if the average interpoint distance within the cell is much smaller than the diameter of the cell. In this event, we use a different type of split, in which the cell is partitioned into two pieces based on distance from the mean.

procedure CHOOSERULE(S)
 if $\Delta^2(S) \leq c \cdot \Delta_A^2(S)$
 then choose a random unit direction v
 $Rule(x) := x \cdot v \leq \text{median}_{z \in S}(z \cdot v)$
 else $Rule(x) :=$
 $\|x - \text{mean}(S)\| \leq \text{median}_{z \in S}(\|z - \text{mean}(S)\|)$
 return $(Rule)$

In the code, c is a constant, $\Delta(S)$ is the diameter of S (the distance between the two furthest points in the set), and $\Delta_A(S)$ is the *average* diameter, that is, the average distance between points of S:

$$\Delta_A^2(S) = \frac{1}{|S|^2} \sum_{x,y \in S} \|x - y\|^2.$$

D. Main result

Recall that an RP tree has two different types of split; let's call them splits *by distance* and splits *by projection*.

Theorem 2: There are constants $0 < c_1, c_2, c_3 < 1$ with the following property. Suppose an RP tree is built using data set $S \subset \mathbb{R}^D$. Consider any cell C such that $S \cap C$ has covariance dimension (d, ϵ), where $\epsilon < c_1$. Pick $x \in S \cap C$ at random, and let C' be the cell containing it at the next level down.

- If C is split by distance, $\mathbb{E}\left[\Delta^2(S \cap C')\right] \leq c_2 \Delta^2(S \cap C)$.
- If C is split by projection, then $\mathbb{E}\left[\Delta_A^2(S \cap C')\right] \leq (1 - (c_3/d))\, \Delta_A^2(S \cap C)$.

In both cases, the expectation is over the randomization in splitting C and the choice of $x \in S \cap C$.

To translate Theorem 2 into a statement about vector quantization error, we combine the two notions of diameter into a single quantity: $\Phi(S) = \Delta_A^2(S) + (1/cd)\Delta^2(S)$. Then Theorem 2 immediately implies that (under the given conditions) there is a constant $c_4 = \min\{(1 - c_2)/2, c_3/2\}$ such that for either split,

$$\mathbb{E}\left[\Phi(S \cap C')\right] \leq (1 - c_4/d)\Phi(S).$$

Suppose we now built a tree T to $l = \log k$ levels, and that the (d, ϵ) upper bound on covariance dimension holds throughout. For a point X chosen at random from S, let $C(X)$ denote the leaf cell (of the $2^l = k$ possibilities) into which it falls. As we will see later (Corollary 6), the quantization error within this cell is precisely $\frac{1}{2}\Delta_A^2(C(X))$. Thus,

$$\begin{aligned}
&\mathbb{E}_T\,[k\text{-quantization error}] \\
&= \frac{1}{2}\mathbb{E}_T\mathbb{E}_X\left[\Delta_A^2(S \cap C(X))\right] \\
&\leq \frac{1}{2}\mathbb{E}_T\mathbb{E}_X\left[\Phi(S \cap C(X))\right] \\
&\leq \frac{1}{2}\left(1 - \frac{c_4}{d}\right)^l \cdot \Phi(S) \\
&\leq \frac{1}{2} \cdot k^{-c_4/d} \cdot \left(\Delta_A^2(S) + (1/cd)\Delta^2(S)\right)
\end{aligned}$$

where $\mathbb{E}_T$ denotes expectation over the randomness in the tree construction.

E. The hardness of finding optimal centers

Given a data set, the optimization problem of finding a k-optimal set of centers is called the k-means problem. Here is the formal definition.

> k-MEANS CLUSTERING
> *Input:* Set of points $x_1, \ldots, x_n \in \mathbb{R}^D$; integer k.
> *Output:* A partition of the points into clusters $C_1, \ldots, C_k$, along with a center μ_j for each cluster, so as to minimize
>
> $$\sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \mu_j\|^2.$$

The typical method of approaching this task is to apply Lloyd's algorithm [13], [12], and usually this algorithm is itself called k-means. The distinction between the two is particularly important to make because Lloyd's algorithm is a heuristic that often returns a suboptimal solution to the k-means problem. Indeed, its solution is often very far from optimal.

What's worse, this suboptimality is not just a problem with Lloyd's algorithm, but an inherent difficulty in the optimization task. k-MEANS CLUSTERING is an NP-hard optimization problem, which means that it is very unlikely that there exists an efficient algorithm for it. To explain this a bit more clearly, we delve briefly into the theory of computational complexity.

The running time of an algorithm is typically measured as a function of its input/output size. In the case of k-means, for instance, it would be given as a function of n, k, and D. An efficient algorithm is one whose running time scales *polynomially* with the problem size. For instance, there are algorithms for sorting n numbers which take time proportional to $n \log n$; these qualify as efficient because $n \log n$ is bounded above by a polynomial in n.

For some optimization problems, the best algorithms we know take time *exponential* in problem size. The famous traveling salesman problem (given distances between n cities, plan a circular route through them so that each city is visited once and the overall tour length is minimized) is one of these. There are various algorithms for it that take time proportional to 2^n (or worse): this means each additional city causes the running time to be doubled. Even small graphs are therefore hard to solve.

This lack of an efficient algorithm is not limited to just a few pathological optimization problems, but recurs across the entire spectrum of computational tasks. Moreover, it has been shown that the fates of these diverse problems (called *NP-complete* problems) are linked: either *all* of them admit efficient algorithms, or none of them do. The mathematical community strongly believes the latter to be the case, although it is has not been proved. Resolving this question is one of the seven "grand challenges" of the new millenium identified by the Clay Institute.

In Appendix II, we show the following.

Theorem 3: k-MEANS CLUSTERING is an NP-hard optimization problem, even if k is restricted to 2.

Thus we cannot expect to be able to find a k-optimal set of centers; the best we can hope is to find some set of centers that achieves roughly the optimal quantization error.

F. Related work

Quantization: The literature on vector quantization is substantial; see the wonderful survey of Gray and Neuhoff [9] for a comprehensive overview. In the most basic setup, there is some distribution P over $\mathbb{R}^D$ from which random vectors are generated and observed, and the goal is to pick a finite codebook $C \subset \mathbb{R}^D$ and an encoding function $\alpha : \mathbb{R}^D \to C$ such that $x \approx \alpha(x)$ for typical vectors x. The quantization error is usually measured by squared loss, $\mathbb{E}\|X - \alpha(X)\|^2$. An obvious choice is to let $\alpha(x)$ be the nearest neighbor of x in C. However, the number of codewords is often so enormous that this nearest neighbor computation cannot be performed in real time. A more efficient scheme is to have the codewords arranged in a tree [4].

The asymptotic behavior of quantization error, assuming optimal quantizers and under various conditions on P, has been studied in great detail. A nice overview is presented in the recent monograph of Graf and Luschgy [8]. The rates obtained for k-optimal quantizers are generally of the form $k^{-2/D}$. There is also work on the special case of data that lie *exactly* on a manifold, and whose distribution is within some constant factor of uniform; in such cases, rates of the form $k^{-2/d}$ are obtained, where d is the dimension of the manifold. Our setting is considerably more general than this: we do not assume optimal quantization (which is NP-hard), we have a broad notion of intrinsic dimension that allows points to merely be close to a manifold rather than on it, and we make no other assumptions about the distribution P.

Compressed sensing: The field of compressed sensing has grown out of the surprising realization that high-dimensional sparse data can be accurately reconstructed from just a few random projections [3], [5]. The central premise of this research area is that the original data thus never even needs to be collected: all one ever sees are the random projections.

RP trees are similar in spirit and entirely compatible with this viewpoint. Theorem 2 holds even if the random projections are forced to be the same across each entire level of the tree.

For a tree of depth k, this means only k random projections are ever needed, and these can be computed beforehand (the split-by-distance can be reworked to operate in the projected space rather than the high-dimensional space). The data are not accessed in any other way.

III. An RP tree adapts to intrinsic dimension

An RP tree has two varieties of split. If a cell C has much larger diameter than average-diameter (average interpoint distance), then it is split according to the distances of points from the mean. Otherwise, a random projection is used.

The first type of split is particularly easy to analyze.

A. Splitting by distance from the mean

This option is invoked when the points in the current cell, call them S, satisfy $\Delta^2(S) > c\Delta_A^2(S)$; recall that $\Delta(S)$ is the diameter of S while $\Delta_A^2(S)$ is the average interpoint distance.

Lemma 4: Suppose that $\Delta^2(S) > c\Delta_A^2(S)$. Let S_1 denote the points in S whose distance to mean(S) is less than or equal to the median distance, and let S_2 be the remaining points. Then the expected squared diameter after the split is

$$\frac{|S_1|}{|S|}\Delta^2(S_1) + \frac{|S_2|}{|S|}\Delta^2(S_2) \;\leq\; \left(\frac{1}{2} + \frac{2}{c}\right)\Delta^2(S).$$

The proof of this lemma is deferred to the Appendix, as are all other proofs in this paper.

B. Splitting by projection: proof outline

Suppose the current cell contains a set of points $S \subset \mathbb{R}^D$ for which $\Delta^2(S) \leq c\Delta_A^2(S)$. We will show that a split by projection has a constant probability of reducing the average squared diameter $\Delta_A^2(S)$ by $\Omega(\Delta_A^2(S)/d)$. Our proof has three parts:

I. Suppose S is split into S_1 and S_2, with means μ_1 and μ_2. Then the reduction in average diameter can be expressed in a remarkably simple form, as a multiple of $\|\mu_1 - \mu_2\|^2$.

II. Next, we give a lower bound on the distance between the *projected* means, $(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2$. We show that the distribution of the projected points is subgaussian with variance $O(\Delta_A^2(S)/D)$. This well-behavedness implies that $(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2 = \Omega(\Delta_A^2(S)/D)$.

III. We finish by showing that, approximately, $\|\mu_1 - \mu_2\|^2 \geq (D/d)(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2$. This is because $\mu_1 - \mu_2$ lies close to the subspace spanned by the top d eigenvectors of the covariance matrix of S; and with high probability, *every* vector in this subspace shrinks by $O(\sqrt{d/D})$ when projected on a random line.

We now tackle these three parts of the proof in order.

C. Quantifying the reduction in average diameter

The average squared diameter $\Delta_A^2(S)$ has certain reformulations that make it convenient to work with. These properties are consequences of the following two observations, the first of which the reader may recognize as a standard "bias-variance" decomposition of statistics.

Lemma 5: Let X, Y be independent and identically distributed random variables in $\mathbb{R}^n$, and let $z \in \mathbb{R}^n$ be any fixed vector.

(a) $\mathbb{E}\left[\|X - z\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] + \|z - \mathbb{E}X\|^2$.
(b) $\mathbb{E}\left[\|X - Y\|^2\right] = 2\,\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right]$.

This can be used to show that the averaged squared diameter, $\Delta_A^2(S)$, is twice the average squared distance of points in S from their mean.

Corollary 6: The average squared diameter of a set S can also be written as:

$$\Delta_A^2(S) = \frac{2}{|S|}\sum_{x \in S}\|x - \text{mean}(S)\|^2.$$

At each successive level of the tree, the current cell is split into two, either by a random projection or according to distance from the mean. Suppose the points in the current cell are S, and that they are split into sets S_1 and S_2. It is obvious that the expected diameter is nonincreasing:

$$\Delta(S) \;\geq\; \frac{|S_1|}{|S|}\Delta(S_1) + \frac{|S_2|}{|S|}\Delta(S_2).$$

This is also true of the expected average diameter. In fact, we can precisely characterize how much it decreases on account of the split.

Lemma 7: Suppose set S is partitioned (in any manner) into S_1 and S_2. Then

$$\Delta_A^2(S) - \left\{\frac{|S_1|}{|S|}\Delta_A^2(S_1) + \frac{|S_2|}{|S|}\Delta_A^2(S_2)\right\} = \frac{2|S_1|\cdot|S_2|}{|S|^2}\,\|\text{mean}(S_1) - \text{mean}(S_2)\|^2.$$

This completes part I of the proof outline.

D. Properties of random projections

Our quantization scheme depends heavily upon certain regularity properties of random projections. We now review these properties, which are critical for parts II and III of our proof.

The most obvious way to pick a random projection from $\mathbb{R}^D$ to $\mathbb{R}$ is to choose a projection direction u uniformly at random from the surface of the unit sphere S^{D-1}, and to send $x \mapsto u \cdot x$.

Another common option is to select the projection vector from a multivariate Gaussian distribution, $u \sim N(0, (1/D)I_D)$. This gives almost the same distribution as before, and is slightly easier to work with in terms of the algorithm and analysis. We will therefore use this type of projection, bearing in mind that all proofs carry over to the other variety as well, with slight changes in constants.

The key property of a random projection from $\mathbb{R}^D$ to $\mathbb{R}$ is that it approximately preserves the lengths of vectors, modulo a scaling factor of $\sqrt{D}$. This is summarized in the lemma below.

Lemma 8: Fix any $x \in \mathbb{R}^D$. Pick a random vector $U \sim N(0, (1/D)I_D)$. Then for any $\alpha, \beta > 0$:

(a) $\mathbb{P}\left[|U \cdot x| \leq \alpha \cdot \frac{\|x\|}{\sqrt{D}}\right] \;\leq\; \sqrt{\frac{2}{\pi}}\,\alpha$

(b) $\mathbb{P}\left[|U \cdot x| \geq \beta \cdot \frac{\|x\|}{\sqrt{D}}\right] \leq \frac{2}{\beta} e^{-\beta^2/2}$

Lemma 8 applies to any individual vector. Thus it also applies, in expectation, to a vector chosen at random from a set $S \subset \mathbb{R}^D$. Applying Markov's inequality, we can then conclude that when S is projected onto a random direction, *most* of the projected points will be close together, in a *central interval* of size $O(\Delta(S)/\sqrt{D})$.

Lemma 9: Suppose $S \subset \mathbb{R}^D$ lies within some ball $B(x_0, \Delta)$. Pick any $0 < \delta, \epsilon \leq 1$ such that $\delta\epsilon \leq 1/e^2$. Let ν be any measure on S. Then with probability $> 1 - \delta$ over the choice of random projection U onto $\mathbb{R}$, all but an ϵ fraction of $U \cdot S$ (measured according to ν) lies within distance $\sqrt{2 \ln \frac{1}{\delta\epsilon}} \cdot \frac{\Delta}{\sqrt{D}}$ of $U \cdot x_0$.

As a corollary (taking ν to be the uniform distribution over S and $\epsilon = 1/2$), the median of the projected points must also lie within this central interval.

Corollary 10: Under the hypotheses of Lemma 9, for any $0 < \delta < 2/e^2$, the following holds with probability at least $1 - \delta$ over the choice of projection:

$$|\text{median}(U \cdot S) - U \cdot x_0| \leq \frac{\Delta}{\sqrt{D}} \cdot \sqrt{2 \ln \frac{2}{\delta}}.$$

Finally, we examine what happens when the set S is a d-dimensional subspace of $\mathbb{R}^D$. Lemma 8 tells us that the projection of any *specific* vector $x \in S$ is unlikely to have length too much greater than $\|x\|/\sqrt{D}$, with high probability. A slightly weaker bound can be shown to hold for all of S simultaneously; the proof technique has appeared before in several contexts, including [15], [1].

Lemma 11: There exists a constant κ_1 with the following property. Fix any $\delta > 0$ and any d-dimensional subspace $H \subset \mathbb{R}^D$. Pick a random projection $U \sim N(0, (1/D)I_D)$. Then with probability at least $1 - \delta$ over the choice of U,

$$\sup_{x \in H} \frac{|x \cdot U|^2}{\|x\|^2} \leq \kappa_1 \cdot \frac{d + \ln 1/\delta}{D}.$$

E. Properties of the projected data

Projection from $\mathbb{R}^D$ into $\mathbb{R}^1$ shrinks the average squared diameter of a data set by roughly D. To see this, we start with the fact that when a data set with covariance A is projected onto a vector U, the projected data have variance $U^T A U$. We now show that for random U, such quadratic forms are concentrated about their expected values.

Lemma 12: Suppose A is an $n \times n$ positive semidefinite matrix, and $U \sim N(0, (1/n)I_n)$. Then for any $\alpha, \beta > 0$:

(a) $\mathbb{P}[U^T A U < \alpha \cdot \mathbb{E}[U^T A U]] \leq e^{-((1/2)-\alpha)/2}$, and
(b) $\mathbb{P}[U^T A U > \beta \cdot \mathbb{E}[U^T A U]] \leq e^{-(\beta-2)/4}$.

Lemma 13: Pick $U \sim N(0, (1/D)I_D)$. Then for any $S \subset \mathbb{R}^D$, with probability at least $1/10$, the projection of S onto U has average squared diameter

$$\Delta_A^2(S \cdot U) \geq \frac{\Delta_A^2(S)}{4D}.$$

Next, we examine the overall distribution of the projected points. When $S \subset \mathbb{R}^D$ has diameter Δ, its projection into the line can have diameter upto Δ, but as we saw in Lemma 9, most of it will lie within a central interval of size $O(\Delta/\sqrt{D})$. What can be said about points that fall outside this interval?

We can apply Lemma 9 to larger intervals of the form $[-k\Delta/\sqrt{D}, k\Delta/\sqrt{D}]$, with failure probability $\delta/2^k$. Taking a union bound over all such intervals with integral k, we get the following.

Lemma 14: Suppose $S \subset B(0, \Delta) \subset \mathbb{R}^D$. Pick any $\delta > 0$ and choose $U \sim N(0, (1/D)I_D)$. Then with probability at least $1 - \delta$ over the choice of U, the projection $S \cdot U = \{x \cdot U : x \in S\}$ satisfies the following property for all positive integers k.

> The fraction of points outside the interval $\left(-\frac{k\Delta}{\sqrt{D}}, +\frac{k\Delta}{\sqrt{D}}\right)$ is at most $\frac{2^k}{\delta} \cdot e^{-k^2/2}$.

F. Distance between the projected means

We are dealing with the case when $\Delta^2(S) \leq c \cdot \Delta_A^2(S)$, that is, the diameter of set S is at most a constant factor times the average interpoint distance. If S is projected onto a random direction, the projected points will have variance about $\Delta_A^2(S)/D$, by Lemma 13; and by Lemma 14, it isn't too far from the truth to think of these points as having roughly a Gaussian distribution. Thus, if the projected points are split into two groups at the mean, we would expect the means of these two groups to be separated by a distance of about $\Delta_A(S)/\sqrt{D}$. Indeed, this is the case. The same holds if we split at the median, which isn't all that different from the mean for close-to-Gaussian distributions.

Lemma 15: There is a constant κ_2 for which the following holds. Pick any $0 < \delta < 1/16c$. Pick $U \sim N(0, (1/D)I_D)$ and split S into two pieces:

$$S_1 = \{x \in S : x \cdot U < s\} \quad \text{and} \quad S_2 = \{x \in S : x \cdot U \geq s\},$$

where s is either $\text{mean}(S \cdot U)$ or $\text{median}(S \cdot U)$. Write $p = |S_1|/|S|$, and let $\widetilde{\mu}_1$ and $\widetilde{\mu}_2$ denote the means of $S_1 \cdot U$ and $S_2 \cdot U$, respectively. Then with probability at least $1/10 - \delta$,

$$(\widetilde{\mu}_2 - \widetilde{\mu}_1)^2 \geq \kappa_2 \cdot \frac{1}{(p(1-p))^2} \cdot \frac{\Delta_A^2(S)}{D} \cdot \frac{1}{c \log(1/\delta)}.$$

G. Distance between the high-dimensional means

Split S into two pieces as in the setting of Lemma 15, and let μ_1 and μ_2 denote the means of S_1 and S_2, respectively. We already have a lower bound on the distance between the projected means, $\widetilde{\mu}_2 - \widetilde{\mu}_1$; we will now show that $\|\mu_2 - \mu_1\|$ is larger than this by a factor of about $\sqrt{D/d}$. The main technical difficulty here is the dependence between the μ_i and the projection U. Incidentally, this is the only part of the entire argument that exploits intrinsic dimensionality.

Lemma 16: There exists a constant κ_3 with the following property. Suppose set $S \subset \mathbb{R}^D$ is such that the top d eigenvalues of $\text{cov}(S)$ account for more than $1 - \epsilon$ of its trace. Pick a random vector $U \sim N(0, (1/D)I_D)$, and split S into

two pieces, S_1 and S_2, in any fashion (which may depend upon U). Let $p = |S_1|/|S|$. Let μ_1 and μ_2 be the means of S_1 and S_2, and let $\widetilde{\mu}_1$ and $\widetilde{\mu}_2$ be the means of $S_1 \cdot U$ and $S_2 \cdot U$.

Then, for any $\delta > 0$, with probability at least $1-\delta$ over the choice of U,

$$\|\mu_2 - \mu_1\|^2 \geq \frac{\kappa_3 D}{d + \ln 1/\delta}\left((\widetilde{\mu}_2 - \widetilde{\mu}_1)^2 - \frac{4}{p(1-p)}\frac{\epsilon \Delta_A^2(S)}{\delta D}\right).$$

We can now finish off the proof of Theorem 2.

Theorem 17: Fix any $\epsilon \leq O(1/c)$. Suppose set $S \subset \mathbb{R}^D$ has the property that the top d eigenvalues of $\mathrm{cov}(S)$ account for more than $1-\epsilon$ of its trace. Pick a random vector $U \sim N(0, (1/D)I_D)$ and split S into two parts,

$$S_1 = \{x \in S : x \cdot U < s\} \quad \text{and} \quad S_2 = \{x \in S : x \cdot U \geq s\},$$

where s is either $\mathrm{mean}(S \cdot U)$ or $\mathrm{median}(S \cdot U)$. Then with probability $\Omega(1)$, the expected average diameter shrinks by $\Omega(\Delta_A^2(S)/cd)$.

Proof: By Lemma 7, the reduction in expected average diameter is

$$\Delta_A^2(S) - \left\{\frac{|S_1|}{|S|}\Delta_A^2(S_1) + \frac{|S_2|}{|S|}\Delta_A^2(S_2)\right\} = \frac{2|S_1| \cdot |S_2|}{|S|^2}\|\mathrm{mean}(S_1) - \mathrm{mean}(S_2)\|^2,$$

or $2p(1-p)\|\mu_1 - \mu_2\|^2$ in the language of Lemmas 15 and 16. The rest follows from those two lemmas. ■

IV. Using RP trees

RP trees are easily adapted to the setting of *streaming* data: situations where data points arrive one at a time, and must be processed quickly and then discarded from memory. In such a model, an RP tree can be built gradually, starting with a single cell and splitting cells when there is enough data to accurately assess the diameter, the mean, and the median. These quantities can be computed approximately from streaming data, without having to store all the data points.

In terms of choosing projection directions, the theorems say that a random projection has a constant probability of being useful. To significantly boost this probability of success, we recommend maintaining a small dictionary of directions, chosen randomly from the unit sphere at the very outset; when it is time to split a cell, all these directions can be tried, and the best one – the direction that most reduces the average quantization error – can be chosen.

If a lot more time were available, a natural alternative to random projection would be to split the cell along the principal component direction, that is, along the primary eigenvector of its covariance matrix. The intuition behind this choice is that it is optimal for Gaussian data.

Theorem 18: Suppose the data distribution is a multivariate Gaussian $N(0, \Sigma)$. Any direction $v \in \mathbb{R}^D$ defines a split into two half-spaces, $\{x : x \cdot v < 0\}$ and $\{x : x \cdot v \geq 0\}$. Of all such splits, the smallest average quantization error (for two cells) is attained when v is the principal eigenvector of Σ.

The proof, as always, is in Appendix I.

A practical alternative to an expensive eigenvalue computation is to compute an approximate principal eigenvector by stochastic gradient descent or some other such incremental optimization method. [21] suggests initializing a vector $v_0 \in \mathbb{R}^D$ randomly, and then performing an update

$$v_{t+1} = (1-\gamma_t)v_t + \gamma_t X_t X_t^T \frac{v_t}{\|v_t\|}$$

when a new data point X_t is seen. We have found experimentally that for step size $\gamma_t = 1/t$, vector v_t converges rapidly to a good split direction.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2008.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

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

[4] P.A. Chou, T. Lookabaugh, and R.M. Gray. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory*, 35(2):299–315, 1989.

[5] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[6] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56:9–33, 2004.

[7] R. Durrett. *Probability: Theory and Examples*. Duxbury, second edition, 1995.

[8] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer, 2000.

[9] R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

[10] J.B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage University Paper series on Quantitative Application in the Social Sciences, 07-011. 1978.

[11] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *IEEE Symposium on Foundations of Computer Science*, 2004.

[12] S.P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[13] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[14] J. Matousek. *Lectures on Discrete Geometry*. Springer, 2002.

[15] V.D. Milman. A new proof of the theorem of a. dvoretsky on sections of convex bodies. *Functional Analysis and its Applications*, 5(4):28–37, 1971.

[16] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 2006.

[17] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, (290):2323–2326, 2000.

[18] I.J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44:522–553, 1938.

[19] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[20] S. Vempala. *Private communication*, 2007.

[21] J. Weng, Y. Zhang, and W.-S. Hwang. Candid covariance-free incremental principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 2003.

V. Appendix I: Proofs of main theorem

A. Proof of Lemma 4

Let random variable X be distributed uniformly over S. Then

$$\mathbb{P}\left[\|X - \mathbb{E}X\|^2 \;\geq\; \text{median}(\|X - \mathbb{E}X\|^2)\right] \;\geq\; \frac{1}{2}$$

by definition of median, so $\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] \geq \text{median}(\|X - \mathbb{E}X\|^2)/2$. It follows from Corollary 6 that

$$\text{median}(\|X - \mathbb{E}X\|^2) \leq 2\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] = \Delta_A^2(S).$$

Set S_1 has squared diameter $\Delta^2(S_1) \leq (2\,\text{median}(\|X - \mathbb{E}X\|))^2 \leq 4\Delta_A^2(S)$. Meanwhile, S_2 has squared diameter at most $\Delta^2(S)$. Therefore,

$$\frac{|S_1|}{|S|}\Delta^2(S_1) + \frac{|S_2|}{|S|}\Delta^2(S_2) \;\leq\; \frac{1}{2}\cdot 4\Delta_A^2(S) + \frac{1}{2}\Delta^2(S)$$

and the lemma follows by using $\Delta^2(S) > c\Delta_A^2(S)$.

B. Proofs of Lemma 5 and Corollary 6

Part (a) of Lemma 5 is immediate when both sides are expanded. For (b), we use part (a) to assert that for any fixed y, we have $\mathbb{E}\left[\|X - y\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] + \|y - \mathbb{E}X\|^2$. We then take expectation over $Y = y$.

Corollary 6 follows by observing that $\Delta_A^2(S)$ is simply $\mathbb{E}\left[\|X - Y\|^2\right]$, when X, Y are i.i.d. draws from the uniform distribution over S.

C. Proof of Lemma 7

Let μ, μ_1, μ_2 denote the means of S, S_1, and S_2. Using Corollary 6 and Lemma 5(a), we have

$$\begin{aligned}
&\Delta_A^2(S) - \frac{|S_1|}{|S|}\Delta_A^2(S_1) - \frac{|S_2|}{|S|}\Delta_A^2(S_2) \\
&= \frac{2}{|S|}\sum_S \|x-\mu\|^2 - \frac{|S_1|}{|S|}\cdot\frac{2}{|S_1|}\sum_{S_1}\|x-\mu_1\|^2 \\
&\quad - \frac{|S_2|}{|S|}\cdot\frac{2}{|S_2|}\sum_{S_2}\|x-\mu_2\|^2 \\
&= \frac{2}{|S|}\left\{\sum_{S_1}\left(\|x-\mu\|^2 - \|x-\mu_1\|^2\right)\right. \\
&\quad \left. + \sum_{S_2}\left(\|x-\mu\|^2 - \|x-\mu_2\|^2\right)\right\} \\
&= \frac{2|S_1|}{|S|}\|\mu_1-\mu\|^2 + \frac{2|S_2|}{|S|}\|\mu_2-\mu\|^2.
\end{aligned}$$

Writing μ as a weighted average of μ_1 and μ_2 then completes the proof.

D. Proof of Lemma 8

Since U has a Gaussian distribution, and any linear combination of independent Gaussians is a Gaussian, it follows that the projection $U \cdot x$ is also Gaussian. Its mean and variance are easily seen to be zero and $\|x\|^2/D$, respectively. Therefore, writing

$$Z \;=\; \frac{\sqrt{D}}{\|x\|}(U \cdot x)$$

we have that $Z \sim N(0,1)$. The bounds stated in the lemma now follow from properties of the standard normal. In particular, $N(0,1)$ is roughly flat in the range $[-1,1]$ and then drops off rapidly; the two cases in the lemma statement correspond to these two regimes.

The highest density achieved by the standard normal is $1/\sqrt{2\pi}$. Thus the probability mass it assigns to the interval $[-\alpha, \alpha]$ is at most $2\alpha/\sqrt{2\pi}$; this takes care of (a). For (b), we use a standard tail bound for the normal, $\mathbb{P}(|Z| \geq \beta) \leq (2/\beta)e^{-\beta^2/2}$; see, for instance, page 7 of [7].

E. Proof of Lemma 9

Set $c = \sqrt{2\ln 1/(\delta\epsilon)} \geq 2$.

Fix any point x, and randomly choose a projection U. Let $\widetilde{x} = U \cdot x$ (and likewise, let $\widetilde{S} = U \cdot S$). What is the chance that $\widetilde{x}$ lands far from $\widetilde{x}_0$? Define the bad event to be $F_x = \mathbf{1}(|\widetilde{x} - \widetilde{x}_0| \geq c\Delta/\sqrt{D})$. By Lemma 8(b), we have

$$\mathbb{E}_U[F_x] \;\leq\; \mathbb{P}_U\left[|\widetilde{x} - \widetilde{x}_0| \geq c\cdot\frac{\|x - x_0\|}{\sqrt{D}}\right] \;\leq\; \frac{2}{c}\,e^{-c^2/2} \;\leq\; \delta\epsilon.$$

Since this holds for any $x \in S$, it also holds in expectation over x drawn from ν. We are interested in bounding the probability (over the choice of U) that more than an ϵ fraction of ν falls far from $\widetilde{x}_0$. Using Markov's inequality and then Fubini's theorem, we have

$$\mathbb{P}_U\left[\mathbb{E}_\mu[F_x] \geq \epsilon\right] \;\leq\; \frac{\mathbb{E}_U[\mathbb{E}_\mu[F_x]]}{\epsilon} \;=\; \frac{\mathbb{E}_\mu[\mathbb{E}_U[F_x]]}{\epsilon} \;\leq\; \delta,$$

as claimed.

F. Proof of Lemma 11

It is enough to show that the inequality holds for $S = H \cap$ (surface of the unit sphere in $\mathbb{R}^D$). Let N be any $(1/2)$-cover of this set (that is, $\sup_{z\in S}\inf_{x\in N}\|x - z\| \leq 1/2$); it is possible to achieve $|N| \leq 10^d$ [14]. Apply Lemma 8, along with a union bound, to conclude that with probability at least $1-\delta$ over the choice of projection U,

$$\sup_{x\in N}|x\cdot U|^2 \;\leq\; 2\cdot\frac{\ln|N| + \ln 1/\delta}{D}.$$

Now, define C by

$$C \;=\; \sup_{x\in S}\left(|x\cdot U|^2\cdot\frac{D}{\ln|N| + \ln 1/\delta}\right).$$

We'll complete the proof by showing $C \leq 8$. To this end, pick the $x^* \in S$ for which the supremum is realized (note S

is compact), and choose $y \in N$ whose distance to x^* is at most $1/2$. Then,

$$\begin{aligned} |x^* \cdot U| &\leq |y \cdot U| + |(x^* - y) \cdot U| \\ &\leq \sqrt{\frac{\ln|N| + \ln 1/\delta}{D}} \left(\frac{1}{2}\sqrt{C} + \sqrt{2}\right) \end{aligned}$$

From the definition of x^*, it follows that $\sqrt{C} \leq \sqrt{2} + \sqrt{C}/2$ and thus $C \leq 8$.

G. Proof of Lemma 12

This follows by examining the moment-generating function of $U^T AU$. Since the distribution of U is spherically symmetric, we can work in the eigenbasis of A and assume without loss of generality that $A = \text{diag}(a_1, \ldots, a_n)$, where $a_1, \ldots, a_n$ are the eigenvalues. Moreover, for convenience we take $\sum a_i = 1$.

Let $U_1, \ldots, U_n$ denote the individual coordinates of U. We can rewrite them as $U_i = Z_i/\sqrt{n}$, where $Z_1, \ldots, Z_n$ are i.i.d. standard normal random variables. Thus

$$U^T AU \;=\; \sum_i a_i U_i^2 \;=\; \frac{1}{n}\sum_i a_i Z_i^2.$$

This tells us immediately that $\mathbb{E}[U^T AU] = 1/n$.

We use Chernoff's bounding method for both parts. For (a), for any $t > 0$,

$$\begin{aligned} &\mathbb{P}\left[U^T AU < \alpha \cdot \mathbb{E}[U^T AU]\right] \\ &= \mathbb{P}\left[\sum_i a_i Z_i^2 < \alpha\right] = \mathbb{P}\left[e^{-t\sum_i a_i Z_i^2} > e^{-t\alpha}\right] \\ &\leq \frac{\mathbb{E}\left[e^{-t\sum_i a_i Z_i^2}\right]}{e^{-t\alpha}} = e^{t\alpha}\prod_i \mathbb{E}\left[e^{-t a_i Z_i^2}\right] \\ &= e^{t\alpha}\prod_i \left(\frac{1}{1+2ta_i}\right)^{1/2} \end{aligned}$$

and the rest follows by using $t = 1/2$ along with the inequality $1/(1+x) \leq e^{-x/2}$ for $0 < x \leq 1$. Similarly for (b), for $0 < t < 1/2$,

$$\begin{aligned} &\mathbb{P}\left[U^T AU > \beta \cdot \mathbb{E}[U^T AU]\right] \\ &= \mathbb{P}\left[\sum_i a_i Z_i^2 > \beta\right] = \mathbb{P}\left[e^{t\sum_i a_i Z_i^2} > e^{t\beta}\right] \\ &\leq \frac{\mathbb{E}\left[e^{t\sum_i a_i Z_i^2}\right]}{e^{t\beta}} = e^{-t\beta}\prod_i \mathbb{E}\left[e^{t a_i Z_i^2}\right] \\ &= e^{-t\beta}\prod_i \left(\frac{1}{1-2ta_i}\right)^{1/2} \end{aligned}$$

and it is adequate to choose $t = 1/4$ and invoke the inequality $1/(1-x) \leq e^{2x}$ for $0 < x \leq 1/2$.

H. Proof of Lemma 13

By Corollary 6,

$$\Delta_A^2(S\cdot U) \;=\; \frac{2}{|S|}\sum_{x\in S}((x - \text{mean}(S))\cdot U)^2 \;=\; 2U^T\text{cov}(S)U.$$

where $\text{cov}(S)$ is the covariance of data set S. This quadratic term has expectation (over choice of U)

$$\begin{aligned} \mathbb{E}[2U^T\text{cov}(S)U] &= 2\sum_{i,j}\mathbb{E}[U_iU_j]\text{cov}(S)_{ij} \\ &= \frac{2}{D}\sum_i \text{cov}(S)_{ii} \;=\; \frac{\Delta_A^2(S)}{D}. \end{aligned}$$

Lemma 12(a) then bounds the probability that it is much smaller than its expected value.

I. Proof of Lemma 15

Let the random variable $\widetilde{X}$ denote a uniform-random draw from the projected points $S \cdot U$. Without loss of generality $\text{mean}(S) = 0$, so that $\mathbb{E}\widetilde{X} = 0$ and thus $p\widetilde{\mu}_1 + (1-p)\widetilde{\mu}_2 = 0$. Rearranging, we get $\widetilde{\mu}_1 = -(1-p)(\widetilde{\mu}_2 - \widetilde{\mu}_1)$ and $\widetilde{\mu}_2 = p(\widetilde{\mu}_2 - \widetilde{\mu}_1)$.

We already know from Lemma 13 (and Corollary 6) that with probability at least $1/10$, the variance of the projected points is significant: $\text{var}(\widetilde{X}) \geq \Delta_A^2(S)/8D$. We'll show this implies a similar lower bound on $(\widetilde{\mu}_2 - \widetilde{\mu}_1)^2$.

Using $\mathbf{1}(\cdot)$ to denote $0-1$ indicator variables,

$$\begin{aligned} \text{var}(\widetilde{X}) &\leq \mathbb{E}[(\widetilde{X} - s)^2] \\ &\leq \mathbb{E}[2t|\widetilde{X} - s| + (|\widetilde{X} - s| - t)^2 \cdot \mathbf{1}(|\widetilde{X} - s| \geq t)] \end{aligned}$$

for any $t > 0$. This is a convenient formulation since the linear term gives us $\widetilde{\mu}_2 - \widetilde{\mu}_1$:

$$\begin{aligned} \mathbb{E}[2t|\widetilde{X} - s|] &= 2t(p(s - \widetilde{\mu}_1) + (1-p)(\widetilde{\mu}_2 - s)) \\ &= 4t \cdot p(1-p)\cdot(\widetilde{\mu}_2 - \widetilde{\mu}_1) + 2ts(2p-1). \end{aligned}$$

The last term vanishes since the split is either at the mean of the projected points, in which case $s = 0$, or at the median, in which case $p = 1/2$.

Next, we'll choose

$$t \;=\; t_o\frac{\Delta(S)}{\sqrt{D}}\cdot\sqrt{\log\frac{1}{\delta}}$$

for some suitable constant t_o, so that the quadratic term in $\text{var}(\widetilde{X})$ can be bounded using Lemma 14 and (if the split point is the median) Corollary 10: with probability at least $1 - \delta$,

$$E[(|\widetilde{X} - s| - t)^2 \cdot \mathbf{1}(|\widetilde{X} - s| \geq t)] \;\leq\; \delta \cdot \frac{\Delta^2(S)}{D}$$

(this is a simple integration). Putting the pieces together, we have

$$\frac{\Delta_A^2(S)}{8D} \;\leq\; \text{var}(\widetilde{X}) \;\leq\; 4t\cdot p(1-p)\cdot(\widetilde{\mu}_2 - \widetilde{\mu}_1) + \delta\cdot\frac{\Delta^2(S)}{D}.$$

The result now follows immediately by algebraic manipulation, using the relation $\Delta^2(S) \leq c\Delta_A^2(S)$.

J. Proof of Lemma 16

Assume without loss of generality that S has zero mean. Let H denote the subspace spanned by the top d eigenvectors of the covariance matrix of S, and let $H^\perp$ be its orthogonal subspace. Write any point $x \in \mathbb{R}^D$ as $x_H + x_\perp$, where each component is seen as a vector in $\mathbb{R}^D$ that lies in the respective subspace.

Pick the random vector U; with probability $\geq 1-\delta$ it satisfies the following two properties.

Property 1: For some constant $\kappa' > 0$, for every $x \in \mathbb{R}^D$

$$|x_H \cdot U|^2 \;\leq\; \|x_H\|^2 \cdot \kappa' \cdot \frac{d + \ln 1/\delta}{D} \;\leq\; \|x\|^2 \cdot \kappa' \cdot \frac{d + \ln 1/\delta}{D}.$$

This holds (with probability $1 - \delta/2$) by Lemma 11.

Property 2: Letting X denote a uniform-random draw from S, we have

$$\begin{aligned}
\mathbb{E}_X[(X_\perp \cdot U)^2] &\leq \frac{2}{\delta} \cdot \mathbb{E}_U \mathbb{E}_X[(X_\perp \cdot U)^2] \\
&= \frac{2}{\delta} \cdot \mathbb{E}_X \mathbb{E}_U[(X_\perp \cdot U)^2] \\
&= \frac{2}{\delta D} \cdot \mathbb{E}_X[\|X_\perp\|^2] \;\leq\; \frac{\epsilon \Delta_A^2(S)}{\delta D}.
\end{aligned}$$

The first step is Markov's inequality, and holds with probability $1-\delta/2$. The last inequality comes from the local covariance condition.

So assume the two properties hold. Writing $\mu_2 - \mu_1$ as $(\mu_{2H} - \mu_{1H}) + (\mu_{2\perp} - \mu_{1\perp})$,

$$\begin{aligned}
(\widetilde{\mu}_2 - \widetilde{\mu}_1)^2 &= ((\mu_{2H} - \mu_{1H}) \cdot U + (\mu_{2\perp} - \mu_{1\perp}) \cdot U)^2 \\
&\leq 2((\mu_{2H} - \mu_{1H}) \cdot U)^2 + 2((\mu_{2\perp} - \mu_{1\perp}) \cdot U)^2.
\end{aligned}$$

The first term can be bounded by Property 1:

$$((\mu_{2H} - \mu_{1H}) \cdot U)^2 \;\leq\; \|\mu_2 - \mu_1\|^2 \cdot \kappa' \cdot \frac{d + \ln 1/\delta}{D}.$$

For the second term, let $\mathbb{E}_X$ denote expectation over X chosen uniformly at random from S. Then

$$\begin{aligned}
&((\mu_{2\perp} - \mu_{1\perp}) \cdot U)^2 \\
&\leq 2(\mu_{2\perp} \cdot U)^2 + 2(\mu_{1\perp} \cdot U)^2 \\
&= 2(\mathbb{E}_X[X_\perp \cdot U \mid X \in S_2])^2 + 2(\mathbb{E}_X[X_\perp \cdot U \mid X \in S_1])^2 \\
&\leq 2\mathbb{E}_X[(X_\perp \cdot U)^2 \mid X \in S_2] + 2\mathbb{E}_X[(X_\perp \cdot U)^2 \mid X \in S_1] \\
&\leq \frac{2}{1-p} \cdot \mathbb{E}_X[(X_\perp \cdot U)^2] + \frac{2}{p} \cdot \mathbb{E}_X[(X_\perp \cdot U)^2] \\
&= \frac{2}{p(1-p)} \mathbb{E}_X[(X_\perp \cdot U)^2] \;\leq\; \frac{2}{p(1-p)} \cdot \frac{\epsilon \Delta_A^2(S)}{\delta D}.
\end{aligned}$$

by Property 2. The lemma follows by putting the various pieces together.

K. Proof of Theorem 18

Let $X \in \mathbb{R}^D$ be a random data vector; by assumption, it is drawn from a multivariate Gaussian $N(0, \Sigma)$. Let $u_1, \ldots, u_D$ denote the eigenvectors of Σ, with corresponding eigenvalues $\lambda_1 \geq \cdots \geq \lambda_D$. The principal eigenvector is u_1 (without loss of generality), and by Lemma 7 the reduction in quantization error from splitting along this direction is $\|\mu_{\text{pca}}\|^2$, where

$$\mu_{\text{pca}} = \mathbb{E}[X \mid X \cdot u_1 \geq 0].$$

Now consider any other split direction (unit vector) v and the corresponding half-space $\{x : x \cdot v \geq 0\}$. The reduction in quantization error it induces is $\|\mu\|^2$, for $\mu = \mathbb{E}[X \mid X \cdot v \geq 0]$. We'll show $\|\mu\| \leq \|\mu_{\text{pca}}\|$.

Our proof technique is to show that $\|\mu\|$ can be written in the form $\mathbb{E}[X \cdot u \mid X \cdot v \geq 0]$ for some unit direction u, and to argue that this is maximized when $v = u$. Thereupon, $\|\mu\| = \mathbb{E}[X \cdot u \mid X \cdot u \geq 0]$, which is maximized for $u = u_1$.

In what follows, define $\rho = \mathbb{E}[Z \mid Z \geq 0]$ where Z is a standard normal. It is easy to compute ρ, but we will not need its numerical value; what matters is that the reduction in quantization error for different split directions turns out to be a multiple of ρ.

Lemma 19: Pick any unit direction u. For $X \sim N(0, \Sigma)$,

$$\mathbb{E}[X \cdot u \mid X \cdot u \geq 0] \;=\; \rho\sqrt{u^T \Sigma u}.$$

Proof: $X \cdot u$ has distribution $N(0, u^T \Sigma u)$, which in turn is equal in distribution to a standard normal times $\sqrt{u^T \Sigma u}$. ■

Lemma 20: For any direction v, let $\mu = \mathbb{E}[X \mid X \cdot v \geq 0]$. Then $\|\mu\| \leq \|\mu_{\text{pca}}\|$.

Proof: By symmetry, μ_{pca} lies in the direction of the principal eigenvector u_1. Thus,

$$\begin{aligned}
\|\mu_{\text{pca}}\| &= \mu_{\text{pca}} \cdot u_1 \\
&= \mathbb{E}[X \cdot u_1 \mid X \cdot u_1 \geq 0] \\
&= \rho\sqrt{u_1^T \Sigma u_1} \;=\; \rho\sqrt{\lambda_1}
\end{aligned}$$

Now, suppose μ lies in some (unit) direction u. Then

$$\begin{aligned}
\|\mu\| &= \mu \cdot u \\
&= \mathbb{E}[X \cdot u \mid X \cdot v \geq 0] \\
&\leq \mathbb{E}[X \cdot u \mid X \cdot u \geq 0] \\
&= \rho\sqrt{u^T \Sigma u} \;\leq\; \rho\sqrt{\lambda_1}.
\end{aligned}$$

The second-last inequality follows from Lemma 21 and the last is a consequence of λ_1 being the largest eigenvalue. ■

Lemma 21: Suppose random vector $Y \in \mathbb{R}^D$ is drawn from a symmetric density (that is, Y has the same distribution as $-Y$). Consider any two unit directions $u, v \in \mathbb{R}^D$. Then

$$\mathbb{E}[Y \cdot u \mid Y \cdot v \geq 0] \;\leq\; \mathbb{E}[Y \cdot u \mid Y \cdot u \geq 0].$$

Proof: Let P denote the distribution of Y. Consider the regions

$$\begin{aligned}
A &= \{y : y \cdot u \geq 0, y \cdot v \geq 0\} \\
A^+ &= \{y : y \cdot u \geq 0, y \cdot v < 0\} \\
A^- &= \{y : y \cdot u < 0, y \cdot v \geq 0\}
\end{aligned}$$

Since $A^+ = -A^-$ (upto sets of measure zero), it follows by symmetry that $P(A^+) = P(A^-)$. Likewise by symmetry, we

have that $P(A \cup A^+) = P(A \cup A^-) = 1/2$. Thus

$$\begin{aligned}
&\mathbb{E}[Y \cdot u \mid Y \cdot u \geq 0] - \mathbb{E}[Y \cdot u \mid Y \cdot v \geq 0] \\
&= \frac{\mathbb{E}[Y \cdot u \mid Y \in A]P(A) + \mathbb{E}[Y \cdot u \mid Y \in A^+]P(A^+)}{1/2} \\
&\quad - \frac{\mathbb{E}[Y \cdot u \mid Y \in A]P(A) + \mathbb{E}[Y \cdot u \mid Y \in A^-]P(A^-)}{1/2} \\
&= 2\left(\mathbb{E}[Y \cdot u \mid Y \in A^+] - \mathbb{E}[Y \cdot u \mid Y \in A^-]\right) P(A^+) \\
&\geq 0
\end{aligned}$$

where the last inequality follows by observing that the first term in the parenthesis is nonnegative while the second is negative. ■

VI. Appendix II: The hardness of k-means clustering

k-MEANS CLUSTERING
Input: Set of points $x_1, \ldots, x_n \in \mathbb{R}^d$; integer k.
Output: A partition of the points into clusters $C_1, \ldots, C_k$, along with a center μ_j for each cluster, so as to minimize

$$\sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \mu_j\|^2.$$

(Here $\|\cdot\|$ is Euclidean distance.) It can be checked that in any optimal solution, μ_j is the mean of the points in C_j. Thus the $\{\mu_j\}$ can be removed entirely from the formulation of the problem. From Lemma 5(b),

$$\sum_{i \in C_j} \|x_i - \mu_j\|^2 = \frac{1}{2|C_j|} \sum_{i,i' \in C_j} \|x_i - x_{i'}\|^2.$$

Therefore, the k-means cost function can equivalently be rewritten as

$$\sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{i,i' \in C_j} \|x_i - x_{i'}\|^2.$$

We consider the specific case when k is fixed to 2.

Theorem 22: 2-means clustering is an NP-hard optimization problem.

This was recently asserted in [6], but the proof was flawed. One of the authors privately communicated an alternative argument to us [20], but since this is as yet unpublished, we give our own proof here.

We establish hardness by a sequence of reductions. Our starting point is a standard restriction of 3SAT that is well known to be NP-complete.

3SAT
Input: A Boolean formula in 3CNF, where each clause has exactly three literals and each variable appears at least twice.
Output: `true` if formula is satisfiable, `false` if not.

By a standard reduction from 3SAT, we show that a special case of NOT-ALL-EQUAL 3SAT is also hard. For completeness, the details are laid out in the next section.

NAESAT*
Input: A Boolean formula $\phi(x_1, \ldots, x_n)$ in 3CNF, such that (i) every clause contains exactly three literals, and (ii) each pair of variables x_i, x_j appears together in at most two clauses, once as either $\{x_i, x_j\}$ or $\{\overline{x}_i, \overline{x}_j\}$, and once as either $\{\overline{x}_i, x_j\}$ or $\{x_i, \overline{x}_j\}$.
Output: `true` if there exists an assignment in which each clause contains exactly one or two satisfied literals; `false` otherwise.

Finally, we get to a generalization of 2-MEANS.

GENERALIZED 2-MEANS
Input: An $n \times n$ matrix of interpoint distances D_{ij}.
Output: A partition of the points into two clusters C_1 and C_2, so as to minimize

$$\sum_{j=1}^{2} \frac{1}{2|C_j|} \sum_{i,i' \in C_j} D_{ii'}.$$

We reduce NAESAT* to GENERALIZED 2-MEANS. For any input ϕ to NAESAT*, we show how to efficiently produce a distance matrix $D(\phi)$ and a threshold $c(\phi)$ such that ϕ satisfies NAESAT* if and only if $D(\phi)$ admits a generalized 2-means clustering of cost $\leq c(\phi)$.

Thus GENERALIZED 2-MEANS CLUSTERING is hard. To get back to 2-MEANS (and thus establish Theorem 22), we prove that the distance matrix $D(\phi)$ can in fact be realized by squared Euclidean distances. This existential fact is also constructive, because in such cases, the embedding can be obtained in cubic time by classical multidimensional scaling [10].

A. Hardness of NAESAT*

Suppose we are given an input $\phi(x_1, \ldots, x_n)$ to 3SAT. If some variable appears just once in the formula, it (and its containing clause) can be trivially removed, so we assume there are no such variables. We construct an intermediate formula ϕ' that is satisfiable if and only if ϕ is, and additionally has exactly three occurrences of each variable: one in a clause of size three, and two in clauses of size two. This ϕ' is then used to produce an input ϕ'' to NAESAT*.

1) Constructing ϕ'.
   Suppose variable x_i appears $k \geq 2$ times in ϕ. Create k variables $x_{i1}, \ldots, x_{ik}$ for use in ϕ': use the same clauses, but replace each occurrence of x_i by one of the x_{ij}. To enforce agreement between the different copies x_{ij}, add k additional clauses $(\overline{x}_{i1} \vee x_{i2}), (\overline{x}_{i2} \vee x_{i3}), \ldots, (\overline{x}_{ik}, x_{i1})$. These correspond to the implications $x_1 \Rightarrow x_2, x_2 \Rightarrow x_3, \ldots, x_k \Rightarrow x_1$.
   By design, ϕ is satisfiable if and only if ϕ' is satisfiable.
2) Constructing ϕ''.
   Now we construct an input ϕ'' for NAESAT*. Suppose ϕ' has m clauses with three literals and m' clauses with two literals. Create $2m + m' + 1$ new variables: $s_1, \ldots, s_m$ and $f_1, \ldots, f_{m+m'}$ and f.
   If the jth three-literal clause in ϕ' is $(\alpha \vee \beta \vee \gamma)$, replace it with two clauses in ϕ'': $(\alpha \vee \beta \vee s_j)$ and $(\overline{s}_j \vee \gamma \vee f_j)$.
   If the jth two-literal clause in ϕ' is $(\alpha \vee \beta)$, replace it

with $(\alpha \vee \beta \vee f_{m+j})$ in ϕ''. Finally, add $m+m'$ clauses that enforce agreement among the f_i: $(\overline{f}_1 \vee f_2 \vee f), (\overline{f}_2 \vee f_3 \vee f), \ldots, (\overline{f}_{m+m'} \vee f_1 \vee f)$.

All clauses in ϕ'' have exactly three literals. Moreover, the only pairs of variables that occur together (in clauses) more than once are $\{f_i, f\}$ pairs. Each such pair occurs twice, as $\{f_i, f\}$ and $\{\overline{f}_i, f\}$.

Lemma 23: ϕ' is satisfiable if and only if ϕ'' is not-all-equal satisfiable.

Proof: First suppose that ϕ' is satisfiable. Use the same settings of the variables for ϕ''. Set $f = f_1 = \cdots = f_{m+m'} =$ `false`. For the jth three-literal clause $(\alpha \vee \beta \vee \gamma)$ of ϕ', if $\alpha = \beta =$ `false` then set s_j to `true`, otherwise set s_j to `false`. The resulting assignment satisfies exactly one or two literals of each clause in ϕ''.

Conversely, suppose ϕ'' is not-all-equal satisfiable. Without loss of generality, the satisfying assignment has f set to `false` (otherwise flip all assignments). The clauses of the form $(\overline{f}_i \vee f_{i+1} \vee f)$ then enforce agreement among all the f_i variables. We can assume they are all `false` (otherwise, once again, flip all assignments; this would make f `true`, but it wouldn't matter, since we will henceforth consider only the clauses that don't contain f). This means the two-literal clauses of ϕ' must be satisfied. Finally, consider any three-literal clause $(\alpha \vee \beta \vee \gamma)$ of ϕ'. This was replaced by $(\alpha \vee \beta \vee s_j)$ and $(\overline{s}_j \vee \gamma \vee f_j)$ in ϕ''. Since f_j is `false`, it follows that one of the literals α, β, γ must be satisfied. Thus ϕ' is satisfied. ■

B. Hardness of GENERALIZED 2-MEANS

Given an instance $\phi(x_1, \ldots, x_n)$ of NAESAT*, we construct a $2n \times 2n$ distance matrix $D = D(\phi)$ where the (implicit) $2n$ points correspond to literals. Entries of this matrix will be indexed as $D_{\alpha,\beta}$, for $\alpha, \beta \in \{x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_n\}$. Another bit of notation: we write $\alpha \sim \beta$ to mean that either α and β occur together in a clause or $\overline{\alpha}$ and $\overline{\beta}$ occur together in a clause. For instance, the clause $(x \vee \overline{y} \vee z)$ allows one to assert $\overline{x} \sim y$ but not $x \sim y$. The input restrictions on NAESAT* ensure that every relationship $\alpha \sim \beta$ is generated by a unique clause; it is not possible to have two different clauses that both contain either $\{\alpha, \beta\}$ or $\{\overline{\alpha}, \overline{\beta}\}$.

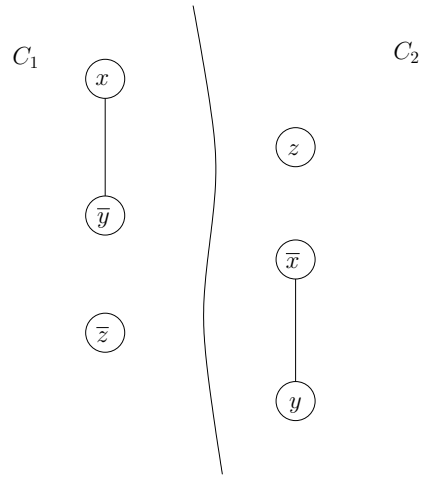
Define

$$D_{\alpha,\beta} = \begin{cases} 0 & \text{if } \alpha = \beta \\ 1+\Delta & \text{if } \alpha = \overline{\beta} \\ 1+\delta & \text{if } \alpha \sim \beta \\ 1 & \text{otherwise} \end{cases}$$

Here $0 < \delta < \Delta < 1$ are constants such that $4\delta m < \Delta \leq 1 - 2\delta n$, where m is the number of clauses of ϕ. One valid setting is $\delta = 1/(5m+2n)$ and $\Delta = 5\delta m$.

Lemma 24: If ϕ is a satisfiable instance of NAESAT*, then $D(\phi)$ admits a generalized 2-means clustering of cost $c(\phi) = n - 1 + 2\delta m/n$, where m is the number of clauses of ϕ.

Proof: The obvious clustering is to make one cluster (say C_1) consist of the positive literals in the satisfying not-all-equal assignment and the other cluster (C_2) the negative literals. Each cluster has n points, and the distance between any two distinct points α, β within a cluster is either 1 or, if $\alpha \sim \beta$, $1+\delta$. Each clause of ϕ has at least one literal in C_1 and at least one literal in C_2, since it is a not-all-equal assignment. Hence it contributes exactly one $\sim$ pair to C_1 and one $\sim$ pair to C_2. The figure below shows an example with a clause $(x \vee \overline{y} \vee z)$ and assignment $x =$ `true`, $y = z =$ `false`.



Thus the clustering cost is

$$\begin{aligned} \frac{1}{2n}\sum_{i,i' \in C_1} D_{ii'} + \frac{1}{2n}\sum_{i,i' \in C_2} D_{ii'} &= 2 \cdot \frac{1}{n}\left(\binom{n}{2} + m\delta\right) \\ &= n - 1 + \frac{2\delta m}{n}. \end{aligned}$$

■

Lemma 25: Let C_1, C_2 be any 2-clustering of $D(\phi)$. If C_1 contains both a variable and its negation, then the cost of this clustering is at least $n - 1 + \Delta/(2n) > c(\phi)$.

Proof: Suppose C_1 has n' points while C_2 has $2n - n'$ points. Since all distances are at least 1, and since C_1 contains a pair of points at distance $1 + \Delta$, the total clustering cost is at least

$$\begin{aligned} &\frac{1}{n'}\left(\binom{n'}{2} + \Delta\right) + \frac{1}{2n-n'}\binom{2n-n'}{2} \\ &= n - 1 + \frac{\Delta}{n'} \geq n - 1 + \frac{\Delta}{2n}. \end{aligned}$$

Since $\Delta > 4\delta m$, this is always more than $c(\phi)$. ■

Lemma 26: If $D(\phi)$ admits a 2-clustering of cost $\leq c(\phi)$, then ϕ is a satisfiable instance of NAESAT*.

Proof: Let C_1, C_2 be a 2-clustering of cost $\leq c(\phi)$. By the previous lemma, neither C_1 nor C_2 contain both a variable and its negation. Thus $|C_1| = |C_2| = n$. The cost of the clustering can be written as

$$\frac{2}{n}\left(\binom{n}{2} + \delta \sum_{\text{clauses}} \left\{ \begin{array}{ll} 1 & \text{if clause split between } C_1,\, C_2 \\ 3 & \text{otherwise} \end{array} \right\}\right)$$

Since the cost is $\leq c(\phi)$, it follows that *all* clauses are split between C_1 and C_2, that is, every clause has at least one literal in C_1 and one literal in C_2. Therefore, the assignment that sets all of C_1 to `true` and all of C_2 to `false` is a valid NAESAT* assignment for ϕ. ■

C. Embeddability of $D(\phi)$

We now show that $D(\phi)$ can be embedded into l_2^2, in the sense that there exist points $x_\alpha \in \mathbb{R}^{2n}$ such that $D_{\alpha,\beta} =$

$\|x_\alpha - x_\beta\|^2$ for all α, β. We rely upon the following classical result [18].

Theorem 27 (Schoenberg): Let H denote the matrix $I - (1/N)\mathbf{1}\mathbf{1}^T$. An $N \times N$ symmetric matrix D can be embedded into l_2^2 if and only if $-HDH$ is positive semidefinite.

The following corollary is immediate.

Corollary 28: An $N \times N$ symmetric matrix D can be embedded into l_2^2 if and only if $u^T D u \leq 0$ for all $u \in \mathbb{R}^N$ with $u \cdot \mathbf{1} = 0$.

Proof: Since the range of the map $v \mapsto Hv$ is precisely $\{u \in \mathbb{R}^N : u \cdot \mathbf{1} = 0\}$, we have

$$\begin{aligned} &-HDH \text{ is positive semidefinite} \\ &\quad \Leftrightarrow \quad v^T HDHv \leq 0 \text{ for all } v \in \mathbb{R}^N \\ &\quad \Leftrightarrow \quad u^T D u \leq 0 \text{ for all } u \in \mathbb{R}^N \text{ with } u \cdot \mathbf{1} = 0. \end{aligned}$$

■

Lemma 29: $D(\phi)$ can be embedded into l_2^2.

Proof: If ϕ is a formula with variables $x_1, \ldots, x_n$, then $D = D(\phi)$ is a $2n \times 2n$ matrix whose first n rows/columns correspond to $x_1, \ldots, x_n$ and remaining rows/columns correspond to $\overline{x}_1, \ldots, \overline{x}_n$. The entry for literals (α, β) is

$$D_{\alpha\beta} = 1 - \mathbf{1}(\alpha = \beta) + \Delta \cdot \mathbf{1}(\alpha = \overline{\beta}) + \delta \cdot \mathbf{1}(\alpha \sim \beta),$$

where $\mathbf{1}(\cdot)$ denotes the indicator function.

Now, pick any $u \in \mathbb{R}^{2n}$ with $u \cdot \mathbf{1} = 0$. Let u^+ denote the first n coordinates of u and u^- the last n coordinates.

$$\begin{aligned} & u^T D u \\ &= \sum_{\alpha,\beta} D_{\alpha\beta} u_\alpha u_\beta \\ &= \sum_{\alpha,\beta} u_\alpha u_\beta - \sum_\alpha u_\alpha^2 + \Delta \sum_\alpha u_\alpha u_{\overline{\alpha}} + \\ &\quad \delta \sum_{\alpha,\beta} u_\alpha u_\beta \mathbf{1}(\alpha \sim \beta) \\ &\leq \left(\sum_\alpha u_\alpha \right)^2 - \|u\|^2 + 2\Delta(u^+ \cdot u^-) + \delta \sum_{\alpha,\beta} |u_\alpha||u_\beta| \\ &\leq -\|u\|^2 + \Delta(\|u^+\|^2 + \|u^-\|^2) + \delta \left(\sum_\alpha |u_\alpha| \right)^2 \\ &\leq -(1-\Delta)\|u\|^2 + 2\delta \|u\|^2 n \end{aligned}$$

where the last step uses the Cauchy-Schwarz inequality. Since $2\delta n \leq 1 - \Delta$, this quantity is always ≤ 0. ■

VII. Appendix III: Covariance dimension of a smooth manifold

Here we show that that a distribution concentrated near a smooth d-dimensional manifold of bounded curvature has neighborhoods of covariance dimension d. The first step towards establishing this result is to generalize our notion of covariance dimension slightly, from finite point sets to arbitrary distributions:

Definition 30: Probability measure ν over $\mathbb{R}^D$ has *covariance dimension* (d, ϵ) if the eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_D^2$ of its covariance matrix satisfy

$$\sigma_1^2 + \cdots + \sigma_d^2 \;\geq\; (1-\epsilon) \cdot (\sigma_1^2 + \cdots + \sigma_D^2).$$

In the same way, the notion of average diameter extends naturally to distributions:

$$\Delta_A^2(\nu) = \mathbb{E}_{X,Y \sim \nu} \|X - Y\|^2.$$

A. Curvature and covariance dimension

Suppose M is a d-dimensional Riemannian submanifold of $\mathbb{R}^D$. We would like to show that sufficiently small neighborhoods $M \cap B(x, r)$ (where $B(x, r)$ is the Euclidean ball of radius r centered at x) have covariance dimension (d, ϵ) for very small ϵ.

The relation of ϵ to r depends crucially on how curved the manifold M is locally. A convenient notion of curvature was recently introduced in [16]:

Definition 31: The *condition number* of M is defined to be $1/\tau$, where τ is the largest number such that: the open normal bundle about M of radius r is imbedded in $\mathbb{R}^D$ for all $r < \tau$.

With this notion in hand, we can state our main result.

Theorem 32: Let M be a d-dimensional Riemannian submanifold of $\mathbb{R}^D$ with finite condition number $1/\tau$, and let ν be any probability measure on M.

Pick any $r > 0$ and any point $p \in M$. Then the restriction of ν to the neighborhood $N = M \cap B(p, r)$ has covariance dimension $(d, 2(r/\tau)^2)$.

Proof: Let $\nu|_N$ be a shorthand for the restriction of ν to N. Clearly, there exists some $x_o \in N$ such that

$$\mathbb{E}_{X,X' \sim \nu|_N}[\|X - X'\|^2] \;\geq\; \mathbb{E}_{X \sim \nu|_N}[\|X - x_o\|^2].$$

Let T denote the tangent plane to M at x_o; this is a d-dimensional affine subspace. Consider the projection from N onto this plane, $f : N \to T$. Lemmas 5.3 and 5.4 of [16] (implicitly) assert that if $r \leq \tau/2$, then (i) f is $1-1$ and (ii) for any $x \in N$,

$$\|x - f(x)\| \leq \frac{r}{\tau} \cdot \|x - x_o\|.$$

If X is drawn from ν restricted to N, its expected squared distance from T is quite small:

$$\begin{aligned} \mathbb{E}_{X \sim \nu|_N}[\text{dist}(X,T)^2] &= \mathbb{E}_{X \sim \nu|_N}[\|X - f(X)\|^2] \\ &\leq \frac{r^2}{\tau^2} \cdot \mathbb{E}_{X \sim \nu|_N}[\|X - x_o\|^2] \\ &\leq \frac{r^2}{\tau^2} \cdot \mathbb{E}_{X,X' \sim \nu|_N}[\|X - X'\|^2] \\ &= \frac{r^2}{\tau^2} \cdot \Delta_A^2(\nu|_N). \end{aligned}$$

Thus $\nu|_N$ is well-approximated by a d-dimensional affine subspace. The bound on its covariance dimension then follows from Lemma 33. ■

What if the distribution of interest does not lie exactly on a low-dimensional manifold M, but close to it? One way to formalize this situation is to imagine that there is an underlying distribution ν on M, but that we only get to observe noisy vectors $X + Z$, where $X \sim \nu$ and $\mathbb{E}[Z \mid X] = 0, \mathbb{E}[\|Z\|^2 \mid X] \leq \sigma^2$. In such situations, Theorem 32 continues to hold, although the covariance dimension becomes $(d, 2((r/\tau)^2 + (\sigma^2/\Delta_A^2(\nu|_N))))$; the proof is exactly as before, except that the expected squared distance of each point $X + Z$ from the tangent plane T increases by (at most) σ^2.

B. Covariance dimension and approximating subspaces

Earlier, we asserted that a set has low covariance dimension if it is well approximated by a low-dimensional affine subspace. We now formalize this intuition.

Lemma 33: A probability measure ν over $\mathbb{R}^D$ has covariance dimension (d, ϵ) if and only if there exists an affine d-dimensional subspace $T \subset \mathbb{R}^D$ such that

$$\mathbb{E}_{X\sim\nu}[(\text{distance from } X \text{ to } T)^2] \leq \frac{\epsilon}{2}\Delta_A^2(\nu).$$

Proof: Assume without loss of generality that ν has mean zero and covariance $\text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$, where $\sigma_1^2 \geq \cdots \geq \sigma_D^2$. The d-dimensional subspace chosen by principal component analysis (PCA) would then be that spanned by the first d coordinate axes; call this T^*. Then

$$\begin{aligned}
& \nu \text{ has covariance dimension } (d, \epsilon) \\
\Leftrightarrow\ & (\sigma_1^2 + \cdots + \sigma_d^2) \geq (1-\epsilon)(\sigma_1^2 + \cdots + \sigma_D^2) \\
\Leftrightarrow\ & (\sigma_{d+1}^2 + \cdots + \sigma_D^2) \leq \epsilon(\sigma_1^2 + \cdots + \sigma_D^2) \\
\Leftrightarrow\ & \mathbb{E}_{X\sim\nu}[X_{d+1}^2 + \cdots + X_D^2] \leq \epsilon\mathbb{E}_{X\sim\nu}[\|X\|^2] \\
\Leftrightarrow\ & \mathbb{E}_{X\sim\nu}[\text{dist}(X, T^*)^2] \leq \frac{\epsilon}{2}\Delta_A^2(\nu) \\
\Leftrightarrow\ & \mathbb{E}_{X\sim\nu}[\text{dist}(X, T)^2] \leq \frac{\epsilon}{2}\Delta_A^2(\nu) \text{ for some affine} \\
& \text{subspace } T \text{ of dimension } d.
\end{aligned}$$

The last implication follows from the well-known fact that $\mathbb{E}_{X\sim\nu}[\text{dist}(X, T)^2]$ is minimized by the PCA subspace. ■

Sanjoy Dasgupta obtained his B.A. in 1993 from Harvard University and his Ph.D. in 2000 from the University of California, Berkeley, both in computer science.

From 2000 until 2002, he was a member of the technical staff at AT&T Labs. Since 2002 he has been in the Department of Computer Science and Engineering at the University of California, San Diego, where he is now Associate Professor. His work is in the area of algorithmic statistics, with a particular focus on minimally supervised learning.

Yoav Freund obtained his B.Sc. in mathematics and physics in 1982 and his M.Sc. in computer science in 1989, both from the Hebrew University in Jerusalem. He obtained a Ph.D. in computer science in 1993 from the University of California, Santa Cruz.

From 1993 until 2001, he was a member of the technical staff at Bell Labs and AT&T Labs, and from 2003 until 2005 was a research scientist at Columbia University. Since 2005 he has been a Professor of Computer Science and Engineering at the University of California, San Diego. His work is in the area of machine learning, computational statistics, and their applications. He is best known for his joint work with Robert Schapire on the Adaboost algorithm. For this work they were awarded the Gdel prize in Theoretical Computer Science in 2003, the Kanellakis Prize in 2004, and the AAAI Fellowship in 2008.