Beyond Gaussians: Spectral Methods for Learning Mixtures of Heavy-Tailed Product Distributions

Kamalika Chaudhuri
Information Theory and Applications, UC San Diego
kamalika@soe.ucsd.edu

Satish Rao
Computer Science Division, UC Berkeley
satishr@cs.berkeley.edu

Abstract

We study the problem of learning mixtures of distributions, a natural formalization of clustering. A mixture of distributions is a collection of distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$ and weights $w_1, \ldots, w_T$. A sample from a mixture is drawn by selecting D_i with probability w_i and then selecting a sample from D_i. The goal, in learning a mixture, is to learn the parameters of the distributions comprising the mixture, given only samples from the mixture.

In this paper, we focus on learning mixtures of heavy-tailed product distributions, which was studied by [DHKS05]. The challenge in learning such mixtures is that the techniques developed for learning mixture-models, such as spectral methods and distance concentration, do not apply. The previous algorithm for this problem was due to [DHKS05], which achieved performance comparable to the algorithms of [AM05, KSV05, CR08] given a mixture of Gaussians, but took time exponential in the dimension. We provide an algorithm which has the same performance, but runs in polynomial time.

Our main contribution is an embedding which transforms a mixture of heavy-tailed product distributions into a mixture of distributions over the hypercube in a higher dimension, while still maintaining separability. Combining this embedding with standard spectral techniques results in algorithms that can learn mixtures of heavy-tailed distributions with separation comparable to the guarantees of [DHKS05]. Our algorithm runs in time polynomial in the dimension, number of clusters, and imbalance in the weights.

1 Introduction

We study the problem of learning mixtures of distributions, a natural formalization of clustering. A *mixture of distributions* is a collection of T distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$ over $\mathbf{R}^n$ and mixing weights $w_1, \ldots, w_T$ such that $\sum_{i=1}^{T} w_i = 1$. A sample from a mixture is drawn by first selecting i with probability w_i, and then choosing a random sample from D_i. The goal, in learning a mixture, is to learn the parameters of the distributions comprising the mixture, and to classify the samples according to source distribution, given only the ability to sample from the mixture.

Learning mixtures of distributions frequently arise in many applications in machine learning, and a fair amount of empirical work has been devoted to the problem. On the theoretical side, all work (except for the work of [DHKS05]) has focussed on learning mixtures of distributions with one of the following characteristics: either the distributions in question have exponentially-decaying tails, for example, mixtures of Gaussians [Das99, DS00, AM05, KSV05, AK01, VW02], or they have severely bounded range, for example, mixtures of binary product distributions [FOS05, CR08]. In the latter case, the bounds deteriorate with the maximum range of values taken by any coordinate of a sample drawn from the mixture.

In this paper, we focus our attention to learning mixtures of more general distributions. In particular, we study learning mixtures of heavy-tailed product distributions, which was introduced by Dasgupta *et. al* [DHKS05].

If the distributions comprising a mixture are very close together, in the sense that they have a high overlap in probability mass, then, even if we knew the parameters of the distributions comprising the mixture, the samples would be hard to classify. To address this, Dasgupta [Das99] introduced the notion of a *separation condition*.A separation condition is a promise that the distributions comprising a mixture are sufficiently different according to some measure, and the goal of the algorithm is to learn correctly a mixture which obeys a certain separation condition. Naturally, the less stringent a separation condition is, the harder it is to learn a mixture, and therefore, a line of theoretical research has focussed on learning mixtures of distributions under less and less restrictive separation conditions. For mixtures of Gaussians, the common measure of separation used is the minimum distance between the means of any two distributions in the mixture, parameterized by the maximum directional standard deviation of any distribution in the mixture. However, this is not a good measure for the type of distributions considered here, as the directional standard deviation may be infinite; following [DHKS05], we therefore use as a measure of separation the minimum distance between the *medians* of any two distributions in the mixture, as parameterized by the maximum $\frac{3}{4}$-*radius*. Recall that given $0 < \beta \leq 1$, the β-radius of a one-dimensional distribution D with median $m(D)$ is the minimum number R_β such that the probability mass of D in the interval $[m(D) - R_\beta, m(D) + R_\beta]$ is at least β.

The major challenge in learning mixtures of heavy-tailed distributions is that none of the tools developed in the literature for learning mixtures of Gaussians or binary product distributions work when the mixture consists of more general distributions. The key ingredients of such algorithms for learning mixtures are: (1) a singular value decomposition of part [CR08] or whole [VW02, KSV05, AM05] of the covariance matrix of the samples and (2) distance-thresholding based clustering algorithms. Singular value decompositions of the covariance matrix do not converge if the distributions have infinite variance. Even for mixtures of distributions with finite variance, distance concentration, which works on the principle that two samples from the same distribution are closer in space than two samples from different distributions, does not work unless the distributions have light tails or a very small range. The previous algorithm for the problem is due to [DHKS05], which learns mixtures of heavy-tailed distributions with performance comparable to the performance of algorithms in [AM05, KSV05, CR08] given a mixture of Gaussians; however, it involves an exhaustive search over all partitions of $\Omega(n)$ samples, where n is the number of dimensions, and hence takes time exponential in the dimension.

In this paper, we show a general procedure for transforming mixtures of heavy-tailed product distributions into mixtures which are more well-behaved, while preserving the separability of the distributions in the mixture. In particular, we provide an efficiently computable embedding from $\mathbf{R}^n$ to $\{0,1\}^{O(n^{3/2})}$. Our embedding, when applied to a mixture of heavy-tailed product distributions which have certain conditions comparable to those in [DHKS05], produces a mixture of distributions in $\{0,1\}^{O(n^{3/2})}$ with centers that are far apart. In addition, we show that the resulting mixture has good properties such that standard algorithms for learning mixtures of binary product distributions – such as the SVD-based algorithms of [AM05, KSV05] and the correlations-based algorithm of [CR08] can be applied to learn it, leading to efficient algorithms for learning mixtures of heavy-tailed product distributions.

More specifically, our results are as follows. Given a mixture of general product distributions, such that each distribution is symmetric about its median, and has $\frac{3}{4}$-radius upper-bounded by R, our embedding transforms it into a mixture of distributions over $\{0,1\}^{O(n^{3/2})}$, while preserving the distance between the centers in a certain sense which is explained in Theorem 1. We can now apply either SVD-based clustering algorithms [KSV05, AM05], and in this case, for sucess with probability $1 - \delta$, we require that (a) the separation between the medians of distributions D_i and D_j be $\Omega(R(w_i^{-1/2} + w_j^{-1/2}) + R\sqrt{T \log \frac{nT}{\delta}})$ and (b) this separation be spread across $\Omega((w_i^{-1/2} + w_j^{-1/2})^2 + T \log \frac{nT}{\delta})$ coordinates. Alternatively, we can apply the correlations-based algorithm of [CR08] on the transformed mixture, to get a logarithmic dependence on the mixing weights. In this case, to learn the mixture with probability $1 - \delta$, we require that (a) the minimum distance between the medians of any two distributions in the mixture to be $\Omega(R\sqrt{T \log \Lambda} + R\sqrt{T \log(nT/\delta)})$ and (b) that this separation to be spread across $\Omega(T \log \Lambda + T \log(nT/\delta))$ coordinates, where Λ is polynomial in n,T and $\frac{1}{w_{\min}}$.

We note that conditions comparable to all these four conditions are required by [DHKS05] for learning mixtures of heavy-tailed distributions; our work improves on their results by providing a polynomial-time algorithm for the problem, as opposed to an exponential-time algorithm. In addition, we also do not need the restriction, needed by [DHKS05], that the probability density function should be decreasing with distance from the median. We also note that the guarantees of our algorithms are comparable to the guarantees of [AM05, KSV05, CR08] when the input is a mixture of axis-aligned Gaussians.

Our Techniques

An initial approach for converting a mixture of general product distributions to a mixture of distributions with better properties is to remove the *outlier points*, which lie very far from the other samples. However, for the types of distributions we consider, a sample may be an outlier along each coordinate with constant $(1/4)$ probability, and since there are n coordinates, with high probability, every point is an outlier. Another approach could be to try to round the outlier points along each dimension; however, since the different mixture components may have different mixing weights, given samples from the mixture, it is hard to determine which of the samples are outliers along a specific coordinate.

To address these issues, we use techniques from metric embeddings [Ind01]. The main idea behind our embedding is to use many random *cutting points* to divide the real line into intervals of length $\Omega(R)$; points which fall into the even intervals are then mapped to 0 and those which fall into the odd intervals are mapped to 1. Although this process does not preserve distances between all pairs of points, we show that this succeeds in separating the centers of two distributions which have medians that are far apart compared to their $3/4$-radius R. Our techniques are related to techniques in metric-embedding [Ind01]; however, so far as we know, this is the first time they have been applied to learning mixtures of distributions. Combining our embedding with existing standard algorithms for learning mixtures of distributions, we get efficient algorithms for learning mixtures of heavy-tailed distributions.

2 Related Work

Heavy-Tailed Mixtures

The work most related to ours is the work of Dasgupta, Hopcroft, Kleinberg and Sandler [DHKS05]. Dasgupta *et. al* [DHKS05] introduced the problem of learning mixtures of heavy-tailed distributions and the notion of using the distance between the medians, parameterized by the half-radius, as a measure of separation between such distributions. Their work deals with the class of all product distributions in which the distribution of each coordinate has the following properties: (a) symmetry around the median (b) decreasing probability density with distance from the median and (c) $\frac{1}{2}$-radius upper bounded by R'. In contrast, we require the distribution of each coordinate to be symmetric about its median and have $\frac{3}{4}$-radius upper bounded by R, and do not require the second assumption of [DHKS05].

[DHKS05] provide two algorithms for learning such mixtures. First, they provide an algorithm which requires a separation of $\Omega(R'\sqrt{\frac{T}{\delta}})$ and a spreading condition that the distance between the medians of any two distributions in the mixture should be spread over $\Theta(T/\delta)$ coordinates, to classify a $1-\delta$ fraction of the samples correctly. This algorithm works by performing an exhaustive search over all partitions of $\Theta(\frac{n\log(nT)}{w_{\min}})$ samples, and therefore has a running time exponential in $\Theta(\frac{n\log(nT)}{w_{\min}})$. In contrast, our algorithms work with similar separation and spreading conditions, and only take time polynomial in n.

Second, they provide an algorithm which works with a stronger separation requirement of $\Omega(R'\sqrt{n})$ and a spreading condition that the distance between the medians of any two distributions in the mixture be spread over $\Theta(T/\delta)$ coordinates. Typically, for such problems, the dimension n is much larger than the number of clusters T, and hence the separation needed here is much larger than the separation needed by the previous algorithm and our algorithms. This algorithm works by performing an exhaustive search over all partitions of $\Theta(\frac{\log(nT)}{w_{\min}})$ samples, and therefore has a running time exponential in $\Theta(\frac{\log(nT)}{w_{\min}})$. Since $w_{\min}$ is at most $\frac{1}{T}$, this may be polynomial in n but remains exponential in T. In contrast, the running times of our algorithms are polynomial in n, T, and $\frac{1}{w_{\min}}$, and for distributions in which the $\frac{3}{4}$-radius is comparable with the half-radius, our algorithms work with separation and spreading constraints comparable to algorithm (1) of [DHKS05].

[DHKS05] also works with a second class of distributions, which have mildly decaying tails. In this case, they provide an algorithm which clusters correctly $1-\delta$ fraction of the samples in time exponential in n, so long as the separation between any two distributions is $\Omega(R'T^{5/2}/\delta^2)$.

Other Mixture Models

There has been a long line of theoretical work on learning mixtures of Gaussians. For this problem, the separation condition is usually expressed in terms of n, the number of dimensions, σ, the maximum directional standard deviation of any distribution in the mixture, and T, the number of clusters. In [Das99], Dasgupta provided an algorithm which learns mixtures of spherical Gaussians when the centers of each pair of distributions is separated by $\Omega(\sigma\sqrt{n})$. In [DS00], Dasgupta and Schulman provided an algorithm which applied to more situations and required a separation of $\Omega(\sigma n^{1/4})$. [AK01] showed how to learn mixtures of arbitrary Gaussians with a separation of $\Omega(\sigma n^{1/4})$ using distance concentration. In addition to the usual separation between the centers, their results apply to other situations, for example, to concentric Gaussians with sufficiently different variance.

The first algorithm that removed the dependence on n was due to Vempala and Wang [VW02], who gave a singular value decomposition based algorithm for learning mixtures of spherical Gaussians with a separation of $\Omega(T^{1/4}\sigma)$. Their algorithm applies a singular value decomposition of the matrix of samples to compute a T-dimensional subspace which approximates the subspace containing the centers, and then uses distance concentration to cluster the samples projected on this low-dimensional space. In further work, [KSV05] and [AM05] showed how to use singular value decomposition based algorithms to learn mixtures of general Gaussians when the separation between the centers of distributions D_i and D_j is $\Omega(\sigma(w_i^{-1/2}+w_j^{-1/2})+\sigma\sqrt{T\log(\frac{T}{\delta})})$. The algorithm of [AM05] was shown to apply to f-convergent and g-concentrated distributions, with bounds that vary with the nature of the distributions. Their algorithm also applies to product distributions on binary vectors. However, their algorithm does not apply to distributions with infinite variance. Even for distributions with finite variance, unless the distribution has rapidly decaying tails, their algorithm yields poor guarantees, proportional to the maximum range of the distribution of each coordinate.

More recently, [CR08] show an algorithm which, under certain conditions, learns mixtures of binary product distributions and axis-aligned Gaussians when the centers are separated by $\Omega(\sigma_*(\sqrt{T\log\Lambda}+\sqrt{T\log(\frac{T}{\delta})}))$ where σ_* is the maximum directional variance in the space containing the centers, and Λ is polynomial in n, T and $\frac{1}{w_{\min}}$. Their algorithm also does not work for distributions with infinite variance and yields poor guarantees for mixtures of heavy-tailed product distributions.

3 A Summary of our Results

We begin with some definitions about distributions over high-dimensional spaces.

Mixture of Distributions. A mixture of distributions is a collection of distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$ and mixing weights $w_1, \ldots, w_T$ such that $\sum_{i=1}^{T} w_i = 1$. A sample from a mixture is drawn by selecting D_i with probability w_i and then choosing a sample from D_i.

Median. We say that a distribution D on $\mathbf{R}$ has median $m(D)$ if the probability that a sample drawn from D is less than or equal to $m(D)$ is $1/2$. We say that a distribution D on $\mathbf{R}^n$ has median $m(D) = (m_1, \ldots, m_n)$ if the projection of D on the f-th coordinate axis has median m_f, for $1 \leq f \leq n$. For a distribution D, we write $m(D)$ to denote the median of D.

Center. We say that a distribution D on $\mathbf{R}^n$ has center $(c_1, \ldots, c_n)$ if the projection of D on the f-th coordinate axis has expectation c_f, for $1 \leq f \leq n$.

β-Radius. For $0 < \beta \leq 1$, the β-Radius of a distribution D on $\mathbf{R}$ with median $m(D)$ is the smallest R_β such that

$$\Pr_{x \sim D}[m(D) - R_\beta \leq x \leq m(D) + R_\beta] \geq \beta$$

Effective Distance. To better describe our results, we need to define the concept of *effective distance*. The effective distance between two points x and y in $\mathbf{R}^n$ at scale R, denoted by $d_R(x, y)$ is defined as:

$$d_R(x, y) = \sqrt{\sum_{f=1}^{n} \min(R^2, (x^f - y^f)^2)}$$

The effective distance between two points x and y at scale R is thus high if many coordinates contribute to the distance between the points.

Notation. We use subscripts i, j to index over distributions in the mixture and subscripts f, g to index over coordinates in $\mathbf{R}^n$. Moreover, we use subscripts $(f, k), \ldots$ to index over coordinates in the transformed space. We use R to denote the maximum $\frac{3}{4}$-radius of any coordinate of any distribution in the mixture. For each distribution D_i in the mixture, and each coordinate f, we use D_i^f to denote the projection of D_i on the f-th coordinate axis. For any i, we use $\tilde{D}_i$ to denote the distribution induced by applying our embedding on D_i. Similarly, for any i and any f, we use $\tilde{D}_i^f$ to denote the distribution induced by applying our embedding on D_i^f. Moreover, we use $\tilde{\mu}_i$ to denote the center of $\tilde{D}_i$ and $\tilde{\mu}_i^f$ to denote the center of $\tilde{D}_i^f$.

We use $||x||$ to denote the L_2 norm of a vector x. We use n to denote the number of dimensions and s to denote the number of samples. For a point x, and subspace $\mathcal{H}$, we use $\mathbf{P}_{\mathcal{H}}(x)$ to denote the projection of x on $\mathcal{H}$.

3.1 Our Results

The main contribution of this paper is an embedding from $\mathbf{R}^n$ to $\{0,1\}^{n'}$, where $n' > n$. The embedding has the property that samples from two product distributions on $\mathbf{R}^n$ which have medians that are far apart map to samples from distributions on $\{0,1\}^{n'}$ with centers which are also far apart. In particular, let $\mathcal{D} = \{D_1, \ldots, D_T\}$ be a mixture of product distributions such that each coordinate f of each distribution D_i in the mixture satisfies the following properties:

1. *Symmetry* about the median.
2. $\frac{3}{4}$-radius upper bounded by R.

In particular, this allows the distribution of each coordinate to have infinite variance. Then the properties of our embedding can be summarized by the following theorems.

Theorem 1 *Suppose we are given access to samples from a mixture of product distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$ over $\mathbf{R}^n$ such that for every i and f, D_i^f satisfies properties (1) and (2). Moreover, let for any i, $\tilde{\mu}_i$ denote the center of the distribution $\tilde{D}_i$ obtained by applying our embedding Φ on D_i. If, for some constant c_1,*

$$d_R(m(D_i), m(D_j)) \geq c_1 R$$

, then, there exists a constant c_2, such that

$$||\tilde{\mu}_i - \tilde{\mu}_j|| \geq c_2 n^{1/4} T^{1/2} (\log n \log T)^{1/2} \times \frac{d_R(m(D_i), m(D_j))}{R}$$

with probability $1 - \frac{1}{n}$ over the randomness in computing Φ. Moreover, for any i, any k, k' and any $f \neq f'$, coordinates (f, k) and (f', k') of $\tilde{D}_i$ are independently distributed.

Our embedding can be combined with the SVD-based clustering algorithms of [KSV05, AM05] to provide an efficient algorithm for learning mixtures of heavy-tailed distributions. The resulting clustering algorithm has the following guarantees.

Theorem 2 *Suppose we are given access to samples from a mixture of product distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$ over $\mathbf{R}^n$ such that for every i and f, D_i^f satisfies properties (1) and (2). If, for some constant c_3,*

$$d_R(m(D_i), m(D_j)) \geq c_3 R(w_i^{-1/2} + w_j^{-1/2} + \sqrt{T \log \frac{nT}{\delta}})$$

Then, Algorithm HT-SVD *clusters the samples correctly with probability* $1-\delta$ *over the samples, and with probability* $1-\frac{1}{n}$ *over the randomness in the algorithm. The algorithm runs in time polynomial in* n *and* T*, and the number of samples required by the algorithm is* $\tilde{O}(\frac{n^{3/2}T}{w_{\min}})$*.*

Alternatively, we can also combine our algorithm with the more recent correlation-based clustering algorithm of [CR08]. The result is an efficient algorithm with the following guarantees.

Theorem 3 *Suppose we are given* s *samples from a mixture of product distributions* $\mathcal{D} = \{D_1, \ldots, D_T\}$ *over* $\mathbf{R}^n$ *such that for every* i *and* f*,* D_i^f *satisfies properties (1) and (2). If, for some constant* c_3*,*

$$d_R(m(D_i), m(D_j)) \geq c_3 R(\sqrt{T \log \Lambda} + \sqrt{T \log \frac{nT}{\delta}})$$

where $\Lambda = \Theta(\frac{T\sqrt{n}\log^2 n}{w_{\min}})$*. Then, Algorithm* HT-CORRELATIONS *clusters the samples correctly with probability* $1-\delta$ *over the samples, and with at least constant probability over the randomness in the algorithm. The algorithm runs in time polynomial in* n *and* T*, and the number of samples required by the algorithm is polynomial in* n*,* T*, and* $\frac{1}{w_{\min}}$*.*

The condition imposed on the centers of the distributions states that every pair of centers is sufficiently far apart in space, and the distance between every pair of centers is spread across $\Omega\left(T \log \Lambda + T \log \frac{nT}{\delta}\right)$ coordinates.

3.2 Discussions

Symmetry. Our embedding still seems to work when the distributions do not have perfect symmetry, but satisfy an approximate symmetry condition. However, we illustrate by an example that we need at least a weak version of the symmetry condition for our embedding to work. Let D_1 and D_2 be the following distributions over $\mathbf{R}$, where M is a very large number. For D_1 the probability density function is:

$$\begin{aligned} f_1(x) &= \frac{3}{8R}, \quad -R \leq x \leq R \\ &= \frac{1}{8MR}, \quad MR \leq x \leq 2MR \\ &= \frac{1}{8MR}, \quad -2MR \leq x \leq -MR \end{aligned}$$

The density function for D_2 is:

$$\begin{aligned} f_2(x) &= \frac{3}{8R}, \quad -R \leq x \leq R \\ &= \frac{1}{4MR}, \quad -2MR \leq x \leq -MR \end{aligned}$$

We note that although the medians of D_1 and D_2 are $R/3$ distance apart, the overlap in their probability mass in any interval of size $2R$ is very high. Therefore, since our embedding relies on the fact that two distributions which have medians that are far apart, and $\frac{3}{4}$-radius bounded by R, have low overlap in probability mass in a region of size $\Omega(R)$ around the median, it does not work for distributions like D_1 and D_2.

Spreading Condition. We note that our spreading condition, while similar to the *slope* requirement of [DHKS05], is weaker; while they require the total contribution to the distance between any two medians from all the coordinates to be large with respect to the contribution from the maximum coordinate, we only require that the contribution come from a few coordinates, regardless of what the maximum contribution from a coordinate is.

4 Embedding Distributions onto the Hamming Cube

In this section, we describe an embedding which maps points in $\mathbf{R}^n$ to points on a Hamming Cube of higher dimension. The embedding has the following property. If for any i and j, D_i and D_j are product distributions on $\mathbf{R}^n$ with properties (1) and (2) such that their medians are far apart, then, the distributions induced on the Hamming cube by applying the embedding on points from D_i and D_j respectively also have centers which are far apart.

The building blocks of our embedding are embeddings $\{\Phi_f\}$, one for each coordinate, f in $\{1, \ldots, n\}$. The final embedding Φ is a concatenation of the maps Φ_f for $1 \leq f \leq n$. We describe more precisely how to put together the maps Φ_f in Section 4.3; for now, we focus on the individual embeddings Φ_f.

Each embedding Φ_f, in its turn, is a concatenation of two embeddings. The first one ensures that, for any i and j, if D_i^f and D_j^f are two distributions with properties (1) and (2) such that $|m(D_i^f) - m(D_j^f)|$ is smaller than (or in the same range as) R, then, the expected distance between the centers of the distributions induced by applying the embedding on points from D_i^f and D_j^f is $\Omega(\frac{|m(D_i^f)-m(D_j^f)|}{R})$. Unfortunately, this embedding does not provide good guarantees when $|m(D_i^f)-m(D_j^f)|$ is large with respect to R. To address this, we use our second embedding, which guarantees that when $|m(D_i^f)-m(D_j^f)|$ is large with respect to R, the centers of the two distributions induced by applying the embedding on points from D_i^f and D_j^f are at least constant distance apart. By concatenating these two embeddings, we ensure that in either case, the centers of the induced distributions obtained by applying Φ_f on D_i^f and D_j^f are far apart.

4.1 Embedding Distributions with Small Separation

In this section, we describe an embedding with the following property. If, for any i, j, and f, D_i^f and D_j^f have properties (1) and (2) and $|m(D_i^f) - m(D_j^f)| < 8R$, then the distance between the centers of the distributions induced by applying ψ to points generated from D_i^f and D_j^f, is proportional to $\frac{|m(D_i^f)-m(D_j^f)|}{8R}$.

The embedding is as follows. Given a parameter R_1, and $r \in [0, R_1)$, we define, for a point $x \in \mathbf{R}$,

$$\begin{aligned} \psi_r(x) &= 0, \text{if } \lfloor \frac{x-r}{R_1} \rfloor \text{is even} \\ &= 1, \text{otherwise} \end{aligned}$$

In other words, we divide the real line into intervals of length R_1 and assign label 0 to the even intervals and label 1 to the odd intervals. The value of $\psi_r(x)$ is then the label of the interval containing $x - r$.

The properties of this embedding can be summarized as follows.

Theorem 4 *For any i, j, and f, if D_i^f and D_j^f have properties (1) and (2), and if r is drawn uniformly at random from $[0, R_1)$ and $R_1 > 2R+3|m(D_i^f)-m(D_j^f)|$, then,*

$$\mathbf{E}[| \Pr_{x\sim D_i^f}[\psi_r(x) = 0] - \Pr_{x\sim D_j^f}[\psi_r(x) = 0]|] \geq \frac{|m(D_i^f) - m(D_j^f)|}{2R_1}$$

Here the expectation is taken over the distribution of r.

Notation For $i = 1, \ldots, T$, we write φ_i^f as the probability density function of distribution D_i^f centered at 0, and F_i^f as the cumulative density function of distribution D_i^f centered at 0. For a real number $r \in [0, R_1)$, and for $i = 1, \ldots, T$, we define

$$\alpha_i^f(r) = \sum_{\lambda=-\infty}^{\infty} (F_i^f(r+(2\lambda+1)R_1) - F_i^f(r+2\lambda R_1))$$

More specifically, $\alpha_i^f(r)$ is the sum of the probability mass of the distribution D_i in the even intervals when the shift is r, which is again the probability that a point drawn from D_i is mapped to 0 by the embedding ψ_r. In the sequel, we use Δ to denote $|m(D_i^f) - m(D_j^f)|$. We also assume without loss of generality that $m(D_j^f) \leq m(D_i^f)$, and $m(D_i^f) = 0$. Then, the left-hand side of the equation in Theorem 4 can be written as follows.

$$\begin{aligned} &\mathbf{E}[| \Pr_{x\sim D_i^f}[\psi_r(x) = 0] - \Pr_{x\sim D_j^f}[\psi_r(x) = 0]|] \\ &= \frac{1}{R_1} \int_{r=-R_1/2}^{R_1/2} |\alpha_j^f(r+\Delta) - \alpha_i^f(r)|\mathbf{dr} \qquad (1) \end{aligned}$$

The proof of Theorem 4 follows in two steps. First, we show that if D_i^f were a shifted version of D_j^f, a slightly stronger version of Theorem 4 would hold. This is shown in Lemma 5. Next,Lemma 8 shows that even if D_i^f is not a shifted version of D_j^f, the statements in Theorem 4 still hold.

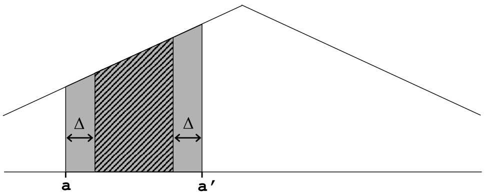

Figure 1: Proof of Lemma 6

Lemma 5 *For any Δ, if $R_1 > 3\Delta + 2R$, then, for any i,*

$$\int_{r=-R_1/2}^{R_1/2} (\alpha_i^f(r) - \alpha_i^f(\Delta + r))\mathbf{dr} \geq \frac{\Delta}{2}$$

Note that the difference between the statement of Theorem 4 and Lemma 5 is that the left-hand side of the equation in Theorem 4 has an absolute value, and hence Lemma 5 makes a stronger statement (under stronger assumptions).

Before we prove Lemma 5, we need the following lemma.

Lemma 6 *Let $[a, a']$ be any interval of length more than 2Δ. Then, for any i,*

$$\begin{aligned} \Delta \cdot \int_a^{a'} \varphi_i^f(r)\mathbf{dr} &\geq \int_{r=a}^{a'} (F_i^f(r+\Delta) - F_i^f(r))\mathbf{dr} \\ &\geq \Delta \cdot \int_{r=a+\Delta}^{a'-\Delta} \varphi_i^f(r)\mathbf{dr} \end{aligned}$$

Proof: For any r,

$$F_i^f(r+\Delta) - F_i^f(r) = \int_{t=r}^{r+\Delta} \varphi_i^f(t)\mathbf{dt}$$

We divide the interval $[a, a']$ into infinitesimal intervals of length $\bar{\delta}$. The probability mass of distribution D_i in an interval $[t, t+\bar{\delta}]$ is $\bar{\delta} \cdot \varphi_i^f(t)$.

Note that in the expression

$$\int_{r=a}^{a'} (F_i^f(r+\Delta) - F_i^f(r))\mathbf{dr}$$

the probability mass of each interval $[t, t+\bar{\delta}]$ where t lies in $[a+\Delta, a'-\Delta]$ is counted exactly $\frac{\Delta}{\bar{\delta}}$ times, and the probability mass of D_i in an interval $[t, t+\bar{\delta}]$, where t lies in the interval $[a, a+\Delta) \cup (a'-\Delta, a']$ is counted at most $\frac{\Delta}{\bar{\delta}}$ times – see Figure 1. Since $\varphi_i^f(t) \geq 0$ for all t, the lemma follows in the limit when $\bar{\delta} \to 0$. □

Proof:(Of Lemma 5) The shaded area in Figure 2 shows the value of $\alpha_i^f(r) - \alpha_i^f(r+\Delta)$ for a distribution D_i.

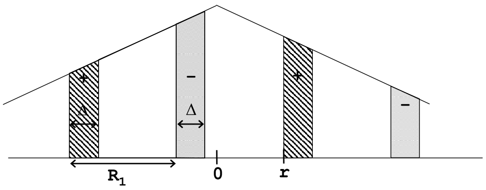

Figure 2: Proof of Lemma 5

We can write:

$$
\begin{aligned}
& \int_{r=-R_1/2}^{R_1/2} (\alpha_i^f(r) - \alpha_i^f(r+\Delta))\mathbf{dr} \\
= & \int_{r=-R_1/2}^{R_1/2} \sum_{\lambda=-\infty}^{\infty} [(F_i^f(r+(2\lambda+1)R_1) \\
& -F_i^f(r+2\lambda R_1) - (F_i^f(r+\Delta+(2\lambda+1)R_1) \\
& -F_i^f(r+\Delta+2\lambda R_1))]\mathbf{dr} \\
= & \int_{r=-R_1/2}^{R_1/2} \sum_{\lambda=-\infty}^{\infty} [(F_i^f(r+(2\lambda+1)R_1) \\
& -F_i^f(r+\Delta+(2\lambda+1)R_1) - (F_i^f(r+2\lambda R_1) \\
& -F_i^f(r+\Delta+2\lambda R_1))]\mathbf{dr} \\
= & \int_{r=-R_1/2}^{R_1/2} \sum_{\lambda=-\infty}^{\infty} [(F_i^f(r+2\lambda R_1+\Delta) \\
& -F_i^f(r+2\lambda R_1)) - (F_i^f(r+(2\lambda+1)R_1+\Delta) \\
& -F_i^f(r+(2\lambda+1)R_1))]\mathbf{dr}
\end{aligned}
$$

From Lemma 6, the first term is at least

$$\Delta \cdot \sum_{\lambda=-\infty}^{\infty} \int_{r=-R_1/2+\Delta}^{R_1/2-\Delta} \varphi_i^f(r+2\lambda R_1)\mathbf{dr}$$

This is Δ times the total probability mass of D_i in the intervals $[2\lambda R_1 - R_1/2 + \Delta, 2\lambda R_1 + R_1/2 - \Delta]$, for all λ. Since $R_1 > 2\Delta + 2R$, this includes the interval $[-R, R]$, and as the median of D_i is at 0 and D_i has $\frac{3}{4}$-radius less than or equal to R, the value of the first term is at least $\frac{3\Delta}{4}$.

From Lemma 6, the second term is at most

$$\Delta \cdot \sum_{\lambda=-\infty}^{\infty} \int_{r=-R_1/2}^{R_1/2} \varphi_i^f(r+(2\lambda+1)R_1)\mathbf{dr}$$

This is the total probability mass of D_i in the intervals $[(2\lambda+1)R_1 - R_1/2, (2\lambda+1)R_1 + R_1/2]$, for all λ. Since $R_1 > 3\Delta + 2R$, none of these intervals have any intersection with $[-R, R]$. The total probability mass in these intervals is therefore at most $\frac{1}{4}$, and therefore the value of the second term is at most $\frac{\Delta}{4}$. The lemma follows. □

Next we show that Theorem 4 holds even if distribution D_i^f is not a shifted version of distribution D_j^f. This is shown by a combination of Lemmas 7 and 8, which are both consequences of the symmetry of the distributions D_i^f and D_j^f.

Lemma 7 *Suppose that for any i, j, and f, D_i^f,D_j^f have property (1) and median* 0*. Then, for any r,*

$$\alpha_i^f(r) - \alpha_j^f(r) = \alpha_j^f(-r) - \alpha_i^f(-r)$$

Proof: We define

$$\bar{\alpha}_i^f(r) = \sum_{\lambda=-\infty}^{\infty} F_i^f(r+2\lambda R_1) - F_i^f(r+(2\lambda-1)R_1)$$

Thus, $\bar{\alpha}_i^f(r)$ is the probability mass of D_i in the odd intervals, which is again the probability that ψ_r maps a random point from D_i to 1 when the shift chosen is r. Therefore, $\bar{\alpha}_i^f(r) = 1 - \alpha_i^f(r)$. Since D_i is symmetric with median 0, for any interval $[a, a']$, $a' > a > 0$, $F_i^f(a') - F_i^f(a) = F_i^f(-a) - F_i^f(-a')$. Therefore,

$$
\begin{aligned}
& \alpha_i^f(-r) \\
= & \sum_{\lambda=-\infty}^{\infty} F_i^f(-r+(2\lambda+1)R_1) - F_i^f(-r+2\lambda R_1) \\
= & \sum_{\lambda=-\infty}^{\infty} F_i^f(r-2\lambda R_1) - F_i^f(r-(2\lambda+1)R_1) \\
= & \bar{\alpha}_i^f(r)
\end{aligned}
$$

The lemma follows because

$$\bar{\alpha}_i^f(r) - \bar{\alpha}_j^f(r) = \alpha_j^f(r) - \alpha_i^f(r)$$

□

Lemma 8 *For any i and j, if D_i^f and D_j^f have properties (1) and (2), then,*

$$
\begin{aligned}
& \int_{r=-R_1/2}^{R_1/2} |\alpha_j^f(r+\Delta) - \alpha_i^f(r)|\mathbf{dr} \\
\geq & \int_{r=-R_1/2}^{R_1/2} (\alpha_j^f(r) - \alpha_j^f(r+\Delta))\mathbf{dr}
\end{aligned}
$$

Proof: By Lemma 7, for every $r \in [-R_1/2, R_1/2]$, there is a unique $r' = -r$ such that $\alpha_i^f(r) - \alpha_j^f(r) = \alpha_j^f(r') - \alpha_i^f(r')$. We claim that for every such pair r, r',

$$
\begin{aligned}
& |\alpha_j^f(r+\Delta) - \alpha_i^f(r)| + |\alpha_j^f(r'+\Delta) - \alpha_i^f(r')| \\
\geq & (\alpha_j^f(r) - \alpha_j^f(r+\Delta)) + (\alpha_j^f(r') - \alpha_j^f(r'+\Delta))
\end{aligned}
$$

We note that for a fixed pair (r, r'),

$$\begin{aligned} & |\alpha_j^f(r+\Delta) - \alpha_i^f(r)| + |\alpha_j^f(r'+\Delta) - \alpha_i^f(r')| \\ = \; & |\alpha_j^f(r+\Delta) - \alpha_i^f(r)| + |\alpha_j^f(r'+\Delta) + \alpha_j^f(r) \\ & -\alpha_j^f(r) - \alpha_i^f(r')| \\ \geq \; & |\alpha_j^f(r+\Delta) - \alpha_i^f(r) + \alpha_j^f(r'+\Delta) + \alpha_j^f(r) \\ & -\alpha_j^f(r) - \alpha_i^f(r')| \\ \geq \; & |(\alpha_j^f(r+\Delta) - \alpha_j^f(r)) + (\alpha_j^f(r'+\Delta) - \alpha_j^f(r')) \\ & +(\alpha_j^f(r) + \alpha_j^f(r') - \alpha_i^f(r) - \alpha_i^f(r'))| \end{aligned}$$

The lemma follows by summing over all such pairs (r, r'). □

Proof: (Of Theorem 4) From Equation 1 and Lemma 5,

$$\mathbf{E}[|\Pr_{x \sim D_i^f}[\psi_r(x) = 0] - \Pr_{x \sim D_j^f}[\psi_r(x) = 0]|]$$

$$\frac{1}{R_1}\int_{-R_1/2}^{R_1/2} |\alpha_j^f(r+\Delta) - \alpha_i^f(r)|\mathbf{dr} \geq \frac{\Delta}{2R_1}$$

The second step follows from Lemma 5. □

4.2 Embedding Distributions with Large Separation

In this section, we describe an embedding with the following property. For any i, j, and f, if D_i^f and D_j^f have properties (1) and (2), and $|m(D_i^f) - m(D_j^f)| \geq 8R$, then, the expected gap between the centers of the distributions induced by applying the embeddings on points from D_i^f and D_j^f is at least a constant.

The embedding is as follows. Given a random $\zeta = \{\rho, \{\varepsilon_k\}_{k \in \mathbf{Z}}\}$ where ρ is a number in $[0, R_2)$ and $\{\varepsilon_k\}$ is an infinite sequence of bits, we define $\phi_\zeta : \mathbf{R} \to \{0, 1\}$ as follows.

$$\phi_\zeta(x) = \varepsilon_{k(x)}, \text{where} \quad k(x) = \lfloor \frac{x - \rho}{R_2} \rfloor \qquad (2)$$

In other words, if $x - \rho$ lies in the interval $[8kR, 8(k+1)R)$, then $\phi_\zeta(x) = \varepsilon_k$.

The properties of the embedding ϕ_ζ can be summarized as follows.

Theorem 9 *For any i, j, and f, let D_i^f and D_j^f have properties (1) and (2), and let $|m(D_i^f) - m(D_j^f)| \geq 8R$. If $R_2 \geq 8R$, and if ρ is generated uniformly at random from the interval $[0, R_2)$, and each ε_k is generated by an independent toss of a fair coin, then,*

$$\mathbf{E}[|\Pr_{x \sim D_i^f}[\phi_\zeta(x) = 0] - \Pr_{x \sim D_j^f}[\phi_\zeta(x) = 0]|] \geq \frac{1}{8}$$

where the expectation is taken over the distribution of ζ.

Proof: We say that an interval $[a, a']$ of length $8R$ or less is *cut* by the embedding if there exists some $y \in [a, a']$ such that $\frac{y-r}{8R}$ is an integer. If $[a, a']$ is cut at y, then, with probability $\frac{1}{2}$ over the choice of $\{\varepsilon_k\}$, any point x in the interval $[a, y]$ has a different value of $\phi_\zeta(x)$ than any point in $(y, a']$. If an interval is not cut, then all points in the interval have the same value of ϕ_ζ with probability 1 over the choice of $\{\varepsilon_k\}$.

Since the intervals $[m(D_i^f) - R, m(D_i^f) + R]$ and $[m(D_j^f) - R, m(D_j^f) + R]$ have length at least $2R$,

$$\Pr[[m(D_i^f) - R, m(D_i^f) + R], [m(D_j^f) - R, m(D_j^f) + R] \text{ are not cut}] \geq 1 - \frac{2R + 2R}{8R} \geq \frac{1}{2}$$

If none of the intervals $[m(D_i^f) - R, m(D_i^f) + R]$ and $[m(D_j^f) - R, m(D_j^f) + R]$ are cut,

$$\Pr[\phi_\zeta(m(D_i^f) - R) \neq \phi_\zeta(m(D_j^f) - R)] = \frac{1}{2}$$

Let us assume that the intervals $[m(D_i^f) - R, m(D_i^f) + R]$ and $[m(D_j^f) - R, m(D_j^f) + R]$ are not cut and

$$\phi_\zeta(m(D_i^f) - R) \neq \phi_\zeta(m(D_j^f) - R)$$

. From the two equations above, the probability of this event is at least $\frac{1}{4}$. Also suppose without loss of generality that $\phi_\zeta(m(D_i^f) - R) = 0$. Then, since R is an upper bound on the $\frac{3}{4}$-radius of the distributions D_i^f and D_j^f, the probability mass of D_i^f that maps to 0 is at least $\frac{3}{4}$, and the probability mass of D_j^f that maps to 0 is at most $\frac{1}{4}$. Therefore, with probability at least $\frac{1}{4}$,

$$|\Pr_{x \sim D_i^f}[\phi_\zeta(x) = 0] - \Pr_{x \sim D_j^f}[\phi_\zeta(x) = 0]| \geq \frac{1}{2}$$

The theorem follows. □

4.3 Combining the Embeddings

In this section, we show how to combine the embeddings of Sections 4.1 and 4.2 to provide a map Φ which obeys the guarantees of Theorem 1. Given parameters R_1, R_2, and q, we define Φ_f for a coordinate f as follows.

$$\Phi_f(x) = (\phi_{\zeta_1}(x^f), \ldots, \phi_{\zeta_q}(x^f), \psi_{r_1}(x^f), \ldots, \psi_{r_q}(x^f)) \qquad (3)$$

Here, $\zeta_1, \ldots, \zeta_q$ are q independent random values of $\zeta = (\rho, \{\varepsilon_k\}_{k \in \mathbf{Z}})$, where ρ is drawn uniformly at random from the interval $[0, R_2)$, and ε_k, for all k, are generated by independent tosses of an unbiased coin. $r_1, \ldots, r_q$ are q independent random values of r, where r is drawn uniformly at random from the interval $[0, R_1)$. Finally, the embedding Φ is defined as:

$$\Phi(x) = \Phi_1(x) \oplus \ldots \oplus \Phi_n(x) \qquad (4)$$

The properties of the embedding Φ are summarized in Theorem 1. Next, we prove Theorem 1. We begin with the following lemma, which demonstrates the properties of each Φ_f.

Lemma 10 *Let $R_1 \geq 26R$, $R_2 \geq 8R$, and $q = 4\sqrt{n}T \log n \log T$, and suppose we are given samples from a mixture of product distributions which satisfy conditions (1) and (2). Then, for all i and j, the embedding $\Phi = \oplus_f \Phi_f$ defined in Equation 3 satisfies the following conditions. With probability at least $1 - \frac{1}{n}$ over the randomness in the embedding, for each coordinate f,*

1. *If $|m(D_i^f) - m(D_j^f)| > 8R$, then, for some constant c_5,*

$$||\mathbf{E}_{x\sim D_i^f}[\Phi_f(x)] - \mathbf{E}_{x\sim D_j^f}[\Phi_f(x)]|| \geq c_5 n^{1/4}T^{1/2} \times (\log n \log T)^{1/2}$$

2. *If $\frac{R}{\sqrt{n}} \leq |m(D_i^f) - m(D_j^f)| \leq 8R$, then, for some constant c_6,*

$$||\mathbf{E}_{x\sim D_i^f}[\Phi_f(x)] - \mathbf{E}_{x\sim D_j^f}[\Phi_f(x)]|| \geq c_6 n^{1/4}T^{1/2} \times (\log n \log T)^{1/2} \frac{|m(D_i^f) - m(D_j^f)|}{R}$$

Proof:(Of Lemma 10) The first part of the lemma follows by Theorem 9, along with an application of the Chernoff Bounds, followed by a Union Bound over all i, j, f. The second part follows similarly by an application of Theorem 4. □

Proof:(Of Theorem 1) We call a coordinate f *very low* for distributions i and j if $|m(D_i^f) - m(D_j^f)| \leq \frac{R}{\sqrt{n}}$, *low* if $\frac{R}{\sqrt{n}} \leq |m(D_i^f) - m(D_j^f)| < 8R$, and *high* otherwise. Let $V_{i,j}$, $L_{i,j}$ and $H_{i,j}$ respectively denote the set of very low, low and high coordinates for distributions D_i and D_j. Then,

$$||\tilde{\mu}_i - \tilde{\mu}_j||^2 = \sum_{f \in V_{i,j}} ||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2 + \sum_{f \in L_{i,j}} ||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2 + \sum_{f \in H_{i,j}} ||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2$$

From Lemma 10,this sum is at least

$$\sum_{f \in L_{i,j}} c_6 n^{1/2} T \log n \log T \frac{|m(D_i^f) - m(D_j^f)|^2}{R^2} + \sum_{f \in H_{i,j}} c_5 n^{1/2} T \log n \log T$$

which, by the definition of effective distance is at least

$$c_7 n^{1/2} T \log n \log T \Big(\frac{d_R^2(m(D_i), m(D_j))}{R^2} - \frac{\sum_{f \in V_{i,j}} (m(D_i^f) - m(D_j^f))^2}{R^2} \Big)$$

where c_7 is some constant. Now the contribution from the very low coordinates to the distance between $m(D_i)$ and $m(D_j)$ is at most $\sqrt{\sum_f R^2/n} = R$. Since

$$d_R(m(D_i), m(D_j)) \geq 2R$$

, this contribution is at most $\frac{1}{2}$ the total distance. The first part of the theorem therefore follows.

For any sample x from any D_i in the mixture, and any k, k', coordinates (f, k) and (f', k') of $\Phi(x)$ are function of x^f and $x^{f'}$ respectively. As for $f \neq f'$, x^f and $x^{f'}$ are independently distributed, the second part of the theorem follows. □

5 Applications: Learning Mixtures

In this section, we show how our embedding in Theorem 1 can be combined with standard algorithm for learning mixture models to yield algorithms than can learn mixtures of heavy-tailed distributions. First, in Section 5.1, we show how to combine our embedding with SVD-based algorithms of [KSV05, AM05]; in Section 5.2, we show how to combine our embedding with the more recent algorithm of [CR08].

5.1 Clustering using SVD

In this section, we present Algorithm HT-SVD– a combination of SVD-based algorithms of [AM05, KSV05] with our embedding in Theorem 1. The input to the algorithm is a set S of samples, and the output is a partitioning of the samples. The algorithm is described in Figure 3.

The properties of Algorithm HT-SVD are summarized by Theorem 2, which we prove for the rest of this section. The two main steps in the proof are as follows: first, we show that after applying our embedding, the tranformed distributions have good properties, such as low directional variance and distance-concentration. Next, we show that these properties imply that SVD-based algorithms, such as those of [KSV05, AM05] can learn these mixtures effectively. The following lemma shows that the maximum directional variance of the transformed distributions in the mixture is high; this fact is later used crucially in demonstrating that SVD-based algorithms can effectively cluster the mixture.

Lemma 11 *For any i, the maximum directional variance of the transformed distribution $\tilde{D}_i$ is at most $O(n^{1/2}T \log n \log T)$.*

Proof: Let v be any unit vector in the transformed space. The variance of the transformed distribution $\tilde{D}_i$ along v

HT-SVD**(S)**

1. Let $R_1 = 26R$, $R_2 = 8R$, and $q = 4\sqrt{n}T \log n \log T$. Compute $\tilde{S} = \{\Phi(x) | x \in S\}$. Partition $\tilde{S}$ into $\tilde{S}_A$ and $\tilde{S}_B$ uniformly at random.
2. Construct the $\frac{s}{2} \times nq$ matrix $\bar{S}_A$ (respectively $\bar{S}_B$) in which the entry at row l and column l' is the l'-th coordinate of the l-th sample point in $\tilde{S}_A$ ($\tilde{S}_B$ respectively).
3. Let $\{v_{1,A}, \dots, v_{T,A}\}$ (resp. $\{v_{1,B}, \dots, v_{T,B}\}$) be the top T singular values of $\bar{S}_A$ (resp. $\bar{S}_B$). Project each point in $\tilde{S}_B$ (resp. $\tilde{S}_A$) on the subspace $\mathcal{K}_A$ (resp. $\mathcal{K}_B$) spanned by $v_{1,A}, \dots, v_{T,A}$ (resp. $v_{1,B}, \dots, v_{T,B}$).
4. Use a distance-based clustering algorithm as in [AK01] to partition the points in $\tilde{S}_A$ and $\tilde{S}_B$ after projection.

Figure 3: Algorithm Using SVDs

can be written as:

$$\begin{aligned}
& \mathbf{E}_{\tilde{x} \sim \tilde{D}_i}[\langle v, \tilde{x} - \mathbf{E}[\tilde{x}]\rangle^2] \\
= \; & \mathbf{E}_{\tilde{x} \sim \tilde{D}_i}[\sum_{(f,k)} (v^{f,k})^2 \cdot (\tilde{x}^{f,k} - \mathbf{E}[\tilde{x}^{f,k}])^2 \\
& +2 \sum_{(f,k),(f',k')} v^{f,k} \cdot v^{f',k'} \cdot (\tilde{x}^{f,k} - \mathbf{E}[\tilde{x}^{f,k}]) \\
& \times (\tilde{x}^{f',k'} - \mathbf{E}[\tilde{x}^{f',k'}])] \\
\leq \; & \mathbf{E}_{\tilde{x} \sim \tilde{D}_i}[\sum_{(f,k)} (v^{f,k})^2 + 2 \sum_{(f,k),(f',k')} v^{f,k} \cdot v^{f',k'} \\
& \times (\tilde{x}^{f,k} - \mathbf{E}[\tilde{x}^{f,k}]) \cdot (\tilde{x}^{f',k'} - \mathbf{E}[\tilde{x}^{f',k'}])] \\
\leq \; & \mathbf{E}_{\tilde{x} \sim \tilde{D}_i}[\sum_{(f,k)} (v^{f,k})^2 + 2 \sum_{f} \sum_{k,k'} v^{f,k} v^{f,k'} \\
& \times (\tilde{x}^{f,k} - \mathbf{E}[\tilde{x}^{f,k}]) \cdot (\tilde{x}^{f,k'} - \mathbf{E}[\tilde{x}^{f,k'}])] \\
\leq \; & \mathbf{E}_{\tilde{x} \sim \tilde{D}_i}[\sum_{f} (\sum_{k} v^{f,k})^2]
\end{aligned}$$

As $\tilde{x}^{f,k}$ is distributed independently of $\tilde{x}^{f',k'}$ when $f \neq f'$, in this case,

$$\mathbf{E}_{\tilde{x} \sim \tilde{D}_i}[(\tilde{x}^{f,k} - \mathbf{E}[\tilde{x}^{f,k}]) \cdot (\tilde{x}^{f',k'} - \mathbf{E}[\tilde{x}^{f',k'}])] = 0$$

The lemma follows as $|(\tilde{x}^{f,k} - \mathbf{E}[\tilde{x}^{f,k}])| \leq 1$ for any f and k, and there are at most $O(n^{1/2} T \log n \log T)$ coordinates corresponding to a single f. □

Next we show that the transformed distributions also possess some distance-concentration properties.

Lemma 12 *Let $\mathcal{H}$ be a d-dimensional subspace of $\{0,1\}^{4n^{3/2} T \log T \log n}$. Then for any i,*

$$\Pr_{\tilde{x} \sim \tilde{D}_i}[||\mathbf{P}_{\mathcal{H}}(\tilde{x} - \mathbf{E}[\tilde{x}])|| < 4n^{1/4} T^{1/2} (\log n \log T)^{1/2} \times \sqrt{d \log(d/\delta)}] \geq 1 - \delta$$

Proof: Let $q = 4n^{1/2} T \log n \log T$. Let $v_1, \dots, v_d$ be an orthonormal basis of $\mathcal{H}$. As

$$||\mathbf{P}_{\mathcal{H}}(\tilde{x})||^2 = \sum_{l=1}^{d} (\langle v_l, \tilde{x} \rangle)^2$$

we apply the Method of Bounded Differences to bound the value of each $\langle v_l, \tilde{x} \rangle$.

$$\langle v_l, \tilde{x} \rangle = \sum_f \sum_k v_l^{f,k} \cdot \tilde{x}^{f,k}$$

As changing each coordinate of the original sample point x will change at most q coordinates of $\tilde{x}$, γ_f, the change in $\langle v_l, \tilde{x} \rangle$ when we change a coordinate f of the original sample point is at most $(\sum_k v_l^{f,k})^2$. Therefore, $\gamma = \sum_f \gamma_f^2 = \sum_f (\sum_k v_l^{f,k})^2$. Since v_l is a unit vector, $\gamma \leq q$. Thus, for any l,

$$\Pr[|\langle v_l, \tilde{x} \rangle - \langle v_l, \mathbf{E}[\tilde{x}] \rangle| > \sqrt{q \log(d/\delta)}] \leq \frac{\delta}{d}$$

As $||\mathbf{P}_{\mathcal{H}}(\tilde{x} - \mathbf{E}[\tilde{x}])||^2 = \sum_l \langle v_l, \tilde{x} - \mathbf{E}[\tilde{x}] \rangle^2$, the lemma follows by applying a Union Bound over each vector v_l. □

We are now ready to prove Theorem 2. The main tool in our proof is the following lemma, due to [AM05], which shows that if the separation between the transformed centers is large, then, Step 3 of the algorithm will find a subspace in which the transformed centers are far apart.

Lemma 13 *Let, for each i, $c_{i,A}$ be the empirical centers of $\tilde{D}_i$ computed from the points in $\tilde{S}_A$, and let σ be the maximum directional standard deviation of any $\tilde{D}_i$. Then,*

$$\begin{aligned}
||\mathbf{P}_{\mathcal{K}_B}(c_{i,A} - c_{j,A})|| \; \geq \; & ||c_{i,A} - c_{j,A}|| \\
& - \sigma(w_i^{-1/2} + w_j^{-1/2})
\end{aligned}$$

Proof: (Of Theorem 2) Let $q = 4n^{1/2}T\log n\log T$. When the distributions in the input mixture obey the separation conditions of Theorem 2, from Theorem 1, for each i and j, the distance between the transformed centers $\tilde{\mu}_i$ and $\tilde{\mu}_j$ is at least :

$$\Omega(\sqrt{q})\cdot(w_i^{-1/2}+w_j^{-1/2}+\sqrt{T\log(Tn/\delta)})$$

Since the number of samples is at least $\Omega(\frac{n^{3/2}}{w_{min}})$, the distance between the sample means and actual means of the transformed distributions are at most $O(1)$. Therefore, from Theorem 13,

$$||\mathbf{P}_{\mathcal{K}_B}(c_{i,A}-c_{j,A})|| \geq c_8\sqrt{qT\log(Tn/\delta)}$$

where $c_{i,A}$ and $c_{j,A}$ are the empirical centers of the transformed distributions, and c_8 is some constant. As $\mathcal{K}_B$ has dimension at most T, from Lemma 12 and a union bound over all pairs of samples, with probability $1-\delta$, all pairs of samples drawn from a distribution D_i have distance at most

$$2n^{1/4}T^{1/2}(\log n\log T)^{1/2}\sqrt{2T\log(nT/\delta)}$$

in the subspace $\mathcal{K}_B$. On the other hand, for some constant a', a sample drawn from D_i and a sample drawn from D_j are at least

$$a'n^{1/4}T^{1/2}(\log n\log T)^{1/2}\sqrt{T\log(nT/\delta)}$$

apart in $\mathcal{K}_B$. Algorithm HT-SVD therefore works for $a' > 2\sqrt{2}$. □

5.2 Clustering Using Correlations

In this section, we present Algorithm HT-CORRELATIONS which is a combination of our embedding with the correlations-based clustering algorithm of [CR08]. Algorithm HT-CORRELATIONS is described in Figure 4. The input to the algorithm is a set S of s samples, and the output is a partitioning of the samples.

The properties of Algorithm HT-CORRELATIONS are described in Theorem 3. This section is devoted to proving Theorem 3. The proof proceeds in three steps. First, we deduce from Theorem 1 that if the distributions satisfy the conditions in Theorem 3, then the transformed distributions satisfy the separation and spreading requirements of Theorem 1 in [CR08]. We can then apply Theorem 1 to show that the centers of the transformed distributions are far apart in $\mathcal{K}_A$ and $\mathcal{K}_B$, the subspaces computed in Step 4 of
Algorithm HT-CORRELATIONS. Finally, we use this fact along with Lemmas 11 and 12 to show that distance concentration algorithms work in these output subspaces.

Proof:(Of Theorem 3) Let $q = 4n^{1/2}T\log n\log T$. From Theorem 1 and Conditions (1) and (2), for each i and j, the distance between the transformed centers $\tilde{\mu}_i$ and $\tilde{\mu}_j$ is at least

$$\Omega(\sqrt{q})(\sqrt{T\log\Lambda}+\sqrt{T\log(nT/\delta)})$$

We note that the proof of Theorem 1 in [CR08] requires only that for each distribution, the coordinates in $\mathcal{F}$ are independently distributed from the coordinates in $\mathcal{G}$. Since the distribution of any coordinate in $\mathcal{F}$ is independent of the distribution in $\mathcal{G}$ (although the coordinates within $\mathcal{F}$ or $\mathcal{G}$ are not necessarily independently distributed), we can apply Theorem 1 in [CR08] to conclude that for each i and j, there exists some constant a such that:

$$\begin{aligned} d_{\mathcal{K}_B}(\tilde{\mu}_i,\tilde{\mu}_j) &\geq \Omega(d(\tilde{\mu}_i,\tilde{\mu}_j)) \\ &\geq a(\sqrt{qT\log\Lambda}+\sqrt{qT\log(nT/\delta)}) \end{aligned}$$

As $\mathcal{K}_B$ has dimension at most $2T$, from Lemma 12 and a union bound, with probability $1-\delta$, all pairs of samples drawn from a distribution D_i have distance at most $2\sqrt{qT\log(Tn/\delta)}$ in the subspace $\mathcal{K}_B$. On the other hand, a sample drawn from D_i and a sample drawn from D_j are at least $(a_1-2)\sqrt{2qT\log(Tn/\delta)}$ apart in $\mathcal{K}_B$. Algorithm HT-CORRELATIONS therefore works. □

References

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 247–257, 2001.

[AM05] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 458–469, 2005.

[CR08] K. Chaudhuri and S. Rao. Learning mixtures of distributions using correlations and independence. In *21st Annual Conference on Learning Theory*, 2008.

[Das99] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer S cience*, pages 634–644, 1999.

[DHKS05] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 491–500, 2005.

[DS00] S. Dasgupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proceedings of FOCS*, 2005.

[Ind01] Piotr Indyk. Algorithmic applications of low-distortion geometric embeddings. In *FOCS*, pages 10–33, 2001.

HT-CORRELATIONS(S)

1. Partition the set of coordinates into $\mathcal{F}$ and $\mathcal{G}$ uniformly at random.

2. Partition S uniformly at random into S_A and S_B. Let $R_1 = 26R$, $R_2 = 8R$, and $q = 4\sqrt{n}T \log n \log T$. Compute $\tilde{S}_A = \{\Phi(x) | x \in S_A\}$ and $\tilde{S}_B = \{\Phi(x) | x \in S_B\}$.

3. Construct the $\frac{nq}{2} \times \frac{nq}{2}$ covariance matrix M_A (respectively M_B), which has a row for each tuple (f, k), $f \in \mathcal{F}$, $k \in [q]$, and a column for each tuple (g, k), $g \in \mathcal{G}$, $k \in [q]$. The entry at row (f, k) and column (g, k') is the covariance between coordinate (f, k) and (g, k') of the transformed points over all samples in S_A (S_B respectively).

4. Let $\{v_{1,A}, \ldots, v_{T,A}\}$ and $\{y_{1,A}, \ldots, y_{T,A}\}$ ($\{v_{1,B}, \ldots, v_{T,B}\}$ and $\{y_{1,B}, \ldots, y_{T,B}\}$ respectively) be the top T left and right singular vectors of M_A (resp. M_B). Project each point in $\tilde{S}_B$ (resp. $\tilde{S}_A$) on the subspace $\mathcal{K}_A$ (resp. $\mathcal{K}_B$) spanned by $\{v_{1,A}, \ldots, v_{T,A}\} \cup \{y_{1,A}, \ldots, y_{T,A}\}$ (resp. $\{v_{1,B}, \ldots, v_{T,B}\} \cup \{y_{1,B}, \ldots, y_{T,B}\}$).

5. Use a distance-based clustering algorithm [AK01] to partition the points in $\tilde{S}_A$ and $\tilde{S}_B$ after projection.

Figure 4: Algorithm Using Correlations

[KSV05] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.

[VW02] V. Vempala and G. Wang. A spectral algorithm of learning mixtures of distributions. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science*, pages 113–123, 2002.