Learning Mixtures of Distributions

Kamalika Chaudhuri

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-124
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-124.html

October 13, 2007

Learning Mixtures of Distributions

by

Kamalika Chaudhuri

B.Tech., (Indian Institute of Technology, Kanpur) 2002

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Satish Rao, Chair
Professor Christos Papadimitriou
Professor Richard Karp
Professor Dorit Hochbaum

Fall 2007

The dissertation of Kamalika Chaudhuri is approved:

Chair

Date

Date

Date

Date

University of California, Berkeley

Fall 2007

Learning Mixtures of Distributions

by

Kamalika Chaudhuri

Abstract

Learning Mixtures of Distributions

by

Kamalika Chaudhuri

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Satish Rao, Chair

This thesis studies 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 *mixing weights*, $\{w_1, \ldots, w_T\}$ such that $\sum_i w_i = 1$. A sample from a mixture is generated by choosing i with probability w_i and choosing a sample from distribution D_i. Given samples from a mixture of distributions, the problem of learning the mixture is that of finding the parameters of the distributions comprising $\mathcal{D}$ and grouping the samples according to source distribution. A common theoretical framework for addressing the problem also assumes that we are given a separation condition, which is a promise that any two distributions in the mixture are sufficiently different.

In this thesis, we study three aspects of the problem. First, in Chapter 3,we focus on optimizing the separation condition while learning mixtures of distributions. The most common algorithms in practice are singular value decomposition based algorithms, which work when the separation is $\Theta(\frac{\sigma}{\sqrt{w_{\min}}})$, where σ is the maximum directional standard deviation of any distribution in the mixture, and $w_{\min}$ is the minimum mixing weight. We show an algorithm which successfully learns mixtures of distributions with a separation condition that depends only logarithmically on the skewed mixing weights. In particular, it succeeds for a separation between the centers that is $\Theta(\sigma\sqrt{T \log \Lambda})$, where T is the number of distributions, and Λ is polynomial in

T and the imbalance in the mixing weights. We require that the distance between the centers be *spread* across $\Theta(T \log \Lambda)$ coordinates. In addition, we show that if every vector in the subspace spanned by the centers has a small projection, of the order of $\frac{1}{T \log \Lambda}$ on each coordinate vector, then, our algorithm succeeds for a separation of only $\Omega(\sigma_* \sqrt{T \log \Lambda})$, where σ_* is the maximum directional standard deviation in the space containing the centers. Our algorithm works for *Binary Product Distributions* and *Axis-Aligned Gaussians*. The spreading condition above is implied by the separation condition for binary product distributions, and is necessary for algorithms that rely on linear correlations.

Motivated by the application in population genetics, in Chapter 4,we study the sample complexity of learning mixtures of binary product distributions. In this thesis, we take a step towards learning mixtures of binary product distributions with optimal sample complexity by providing an algorithm which learns a mixture of two binary product distributions with uniform mixing weights and low sample complexity. Our algorithm clusters all the samples correctly with high probability, so long as $d(\mu_1, \mu_2)$, the square of the Euclidean distance between the centers of distributions is at least polylogarithmic in s, the number of samples and the following trade-off holds between the separation and the number of samples:

$$s \cdot d^2(\mu_1, \mu_2) \geq a \cdot n \log s \log(ns)$$

for some constant a.

Finally, in Chapter 5, we study the problem of learning mixtures of heavy-tailed product distributions. To this end, we provide an embedding from $\mathbf{R}^n$ to $\{0,1\}^{n'}$, which maps random samples from distributions with medians that are far apart to random samples from distributions on $\{0,1\}^{n'}$, with centers that are far apart. The main application of our embedding is in designing an algorithm for learning mixtures of heavy-tailed distributions. We provide a polynomial-time algorithm, which learns mixtures of general product distributions, as long as the distribution of each coordinate satisfies two properties: symmetry about the median and $\frac{3}{4}$-radius upper-

bounded by R. The separation required by our algorithm to correctly classify a $1-\delta$ fraction of the samples is that the distance between the medians of any two distributions in the mixture should be $\Omega(R\sqrt{T\log\Lambda}+R\sqrt{T\log\frac{T}{\delta}})$, and this distance should be spread across $\Omega(T\log\Lambda+T\log\frac{T}{\delta})$ coordinates. A second application of our embedding is in designing algorithms for learning mixtures of distributions with finite variance, which work under a separation requirement of $\Omega(\sigma_*\sqrt{T\log\Lambda})$ and a spreading requirement of $\Omega(T\log\Lambda+T\log\frac{T}{\delta})$. This algorithm does not require the more stringent spreading condition needed by the algorithm which offers similar guarantees in Chapter 3.

Professor Satish Rao

Dissertation Committee Chair

Acknowledgements

Many students find it difficult to acquire one advisor. I was fortunate enough to be guided by two wonderful advisors – Satish Rao and Christos Papadimitriou. While Satish listened with great patience to all my unformed ideas, supported me in my every eccentric venture and then encouraged me, Christos instilled in me a love for mathematics, and inspired me to do theory. They are the ones who taught me the art of research, and not just the art of research. This thesis would not have been possible without Satish and Christos, and I would like to thank them for being there for me throughout my years of graduate school.

One of the advantages of being a student in Berkeley is that one gets a chance to work with many extraordinary people. I would like to thank everyone I worked with while in Berkeley – Kunal Talwar and Sam Riesenfeld, for the hours we spent poring over bounded-degree MSTs, Hoeteck Wee for the time we wrote code together, Kevin Chen for the times we worked hard to decode the mysteries of computational biology, Henry Lin for the times we spent on hard combinatorial problems, Elitza Maneva and Sam Riesenfeld for the hours we spent on our class projects, and Shuheng Zhou and Eran Halperin for the time we spent working on clusterings. Thanks to Richard Karp and Dorit Hochbaum for their guidance, and Umesh Vazirani, Alistair Sinclair and Luca Trevisan for their advice on life, the universe and everything.

My most precious moments in Berkeley were spent in the company of the wonderful friends I made. I would like to thank Kunal Talwar and Hoeteck Wee for their love and support, and Andrej Bogdanov for his love and friendship and for trying to teach me how to drive a stick-shift. Thanks to Kris Hildrum for her advice on graduate school, and to Kevin Chen, Elitza Maneva, Sam Riesenfeld and Kenji Obata for their friendship and their company in the grand quest for culinary delights in the bay area. Thanks to Kaushik Dutta, Henry Lin, Alex Fabrikant and Boriska Toth for making Soda Hall a fun place. Thanks to Kathryn Schild, my house-mate of three years, for

putting up with my eccentricities. Finally, thanks to all those I have missed in these acknowledgements, and I am sure I have missed many.

I would like to thank my parents for their love and support throughout my years of graduate school, indeed, throughout my life. This thesis is dedicated to them. I would like to thank my sister for her love and her text messages: it was thoughts of my family that kept me going when things got tough.

Finally, my husband Ranjit. Without his love and support, none of this would have been possible. I do not know how to thank him enough.

To my parents

Contents

List of Figures

Chapter 1

Introduction

Let us begin with a problem in population genetics. In recent years, a goal of population genetics is to perform large-scale association studies, in which a complex disease such as cancer, or Alzheimer's disease is associated with some genetic factors. Normally, in these studies, a set of cases (individuals carrying the disease) and controls (background populations) are collected and genotyped. Mathematically, the genotypes may simply be viewed as bit vectors which represent the genetic content of different positions in the genome (these positions are known as SNPs - single nucleotide polymorphisms). Then, the bit vectors of the cases and controls are compared to each other. A significant difference between the cases and the controls at one of these SNP positions implies that this SNP is correlated with the disease. This may be used as diagnostic tool for the disease, or simply as an insight into the process that leads for the disease.

Unfortunately, in practice things are not as simple as described above. One major obstacle is the fact that the result, although statistically significant, may be an artifact of the study design. In particular, consider the scenario in which the case population was collected in a hospital in city A, while the control population was collected in another city B, on the other side of the world. The significant difference between the two populations may very well be due to another characteristic (such as eye color, etc.) in which the two populations differ, and it is not necessarily due to the difference in carrying the disease. This problem is known as population substructure,

or population stratification. Given genetic data for a set of individuals belonging to two or more different hidden populations, the problem of population stratification is to characterize the parameters of each population as well as cluster the individuals according to source distribution.

A natural theoretical model for the data is a *random generative model*. In this model, the members of a specific population are assumed to be generated by some distribution over bit vectors. The members of different populations are generated by distributions with different parameters. The genetic data for the entire set of individuals is then set to be generated by *a mixture of distributions*.

More formally, 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 $\sum_i w_i = 1$. A sample from a mixture $\mathcal{D}$ is generated by choosing a distribution D_i with probability w_i and then selecting a random sample from D_i. For our application, each distribution D_i would correspond to a population present in the data, and w_i would represent the proportion of individuals in the sample set from population D_i.

For genetic data, a simple probability distribution for each population would be *a binary product distribution*, a distribution on $\{0,1\}^n$ where each coordinate is distributed independently of any others. The genetic data for the entire set of individuals can be thus modeled by a mixture of binary product distributions.

Under a random generative model for the data, the problem of population stratification reduces to the problem of *learning mixtures of distributions*, where the individual distributions in the mixture are binary product distributions. Given only samples generated from a mixture of distributions, the problem of learning the mixture is to learn the parameters of the individual distributions comprising the mixture through clustering the samples according to source distribution.

In this thesis, we study the problem of learning mixtures of distributions over high-dimensional spaces with special focus on product distributions. A *product distribution* over $\mathbf{R}^n$ is one in which each coordinate is distributed independently of any others.

Product distributions are used as simple models for a wide variety of data, including genetic data.

1.1 Optimization Criteria

If the distributions $D_1, \ldots, D_T$ in a mixture are very close to each other, then, even if we knew the parameters of the distributions, classifying the points would be hard in an information theoretic sense. For example, if $\mathcal{D} = \{D_1, D_2\}$, where $D_1 = \mathcal{N}(0, 1)$ and $D_2 = \mathcal{N}(1, 1)$, a constant fraction of the probability mass of D_1 and D_2 overlap. Thus, even if one knows the parameters of D_1 and D_2, but not the distribution generating each individual sample, it is impossible to classify the samples with an accuracy better than a certain threshold. Moreover, even if the overlap in probability mass is sufficient to allow correct classification for a large fraction of the samples, if the points from different distributions are close in space, the search problem of finding a classifier that will tease the distributions apart becomes harder.

To address this, Dasgupta [7] introduced the notion of using as an optimization criterion a *separation condition* on the mixture. A separation condition is a promise that the distributions comprising the mixture are sufficiently different under some measure. Given samples from a mixture of distributions and the separation condition, the goal of the algorithm is to compute the parameters of the individual distributions comprising the mixture and cluster a large fraction of the samples according to source distribution. The challenge of the algorithm designer is to design algorithms for learning mixtures of distributions which work with a less restrictive separation condition.

A commonly used separation measure for Gaussian mixtures is the distance between the centers of the distributions, parameterized by the maximum directional standard deviation, σ, of any distribution in the mixture. This optimization criterion is motivated by the fact that for mixtures of spherical Gaussians, if the separation is less than σ, the samples from the mixtures cannot be accurately clustered.

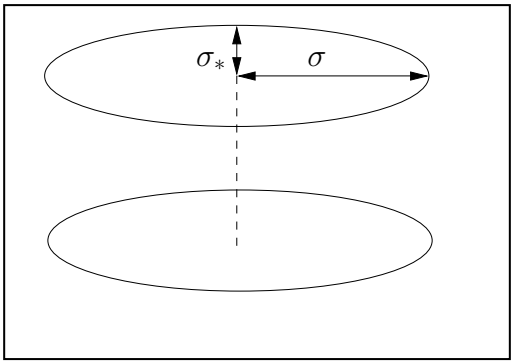

Figure 1.1: An Example where $\sigma >> \sigma_*$

A second optimization criterion for the problem of learning mixtures of distributions is the *sample complexity*, or the number of samples required by the algorithm to provide a correct clustering of the data. The sample complexity is very important in practical applications. For example, in the application to population genetics described above, a typical data set would consist of genetic data on a few thousands of genetic factors for a few hundreds of individuals. It is therefore essential that an algorithm which deals with such data is efficient in terms of the number of samples.

1.2 A Summary of Our Results

In this thesis, we study three aspects of the problem of learning mixtures of distributions.

First, we consider the problem of learning mixtures of distributions with *optimal separation*. The state-of-the-art algorithms for learning mixtures of distributions are singular-value decomposition-based algorithms, which require a separation of $\Omega(\frac{\sigma}{\sqrt{w_{\min}}})$ between the centers, where σ is the maximum directional standard deviation of any distribution in the mixture, and $w_{\min}$ is the lowest mixing weight. This bound is suboptimal because of two reasons.

First, given enough samples, mixing weights have no bearing on the separability of distributions. To illustrate this point, consider for example, two mixtures of any two

fixed distributions D_1 and D_2 with different mixing weights: in the first mixture, $w_1 = w_2 = 1/2$, and in the second mixture, $w_1 = 1/4$ and $w_2 = 3/4$. If we have unlimited computational resources, and we can learn the first mixture with 50 samples, we should be able to learn the second mixture with 100 samples, because in the latter case, we have at least as many samples from D_1 and only strictly more samples from D_2. This does not always hold for singular-value-decomposition based methods. Secondly, regardless of the value of σ, an algorithm, which has prior knowledge of the subspace containing the centers of the distribution, should be able to learn the mixture when the separation between the centers is proportional to σ_*, the maximum directional standard deviation of any distribution in the mixture in the subspace containing the centers.

In this thesis, we present an algorithm for learning mixtures of distributions, which is more stable in the presence of skewed mixing weights, and, under certain conditions, in the presence of high variance outside the subspace containing the centers. Our algorithm is motivated by the fact that in practice, mixtures with skewed mixing weights arise naturally – see [10; 21] for an example in population genetics. In particular, the dependence of the separation required by our algorithm on skewed mixing weights is only logarithmic. Additionally, with arbitrarily small separation, (*i.e.,* even when the separation is not enough for classification), with enough samples, we can approximate the subspace containing the centers. Previous techniques failed to do so for non-spherical distributions regardless of the number of samples, unless the separation was sufficiently large. Our algorithm works for *Binary Product Distributions* and *Axis-Aligned Gaussians*, both of which are product distributions. We require that the distance between the centers be *spread* across $\Theta(T \log \Lambda)$ coordinates, where Λ depends polynomially on the maximum distance between centers and w_{min}. For our algorithm to classify the samples correctly, we further need the separation between centers to be $\Theta(\sigma\sqrt{T \log \Lambda})$.

In addition, if a stronger version of the spreading condition is satisfied, then our

algorithm requires a separation of only $\Theta(\sigma_*\sqrt{T \log \Lambda})$ to ensure correct classification of the samples. The stronger spreading condition, which is discussed in more detail later, ensures that when we split the set of all coordinates randomly into two sets, the maximum directional variance of any distribution in the mixture along the projection of the subspace containing the centers into the subspaces spanned by the coordinate vectors in each set, is comparable to σ_*^2.

Motivated by the application in population genetics, we next study the sample complexity of learning mixtures of binary product distributions. The state of the art algorithms for learning mixtures of binary product distributions are SVD-based approaches, which require $\Omega(\frac{nT}{w_{\min}})$ samples when working with distributions in n-dimensional space. Typical population stratification applications involve data on a few thousand genetic factors for a few hundred individuals; the guarantees on the sample requirement of SVD-based algorithms is thus insufficient for such approaches.

In this thesis, we take a step towards learning mixtures of binary product distributions with optimal sample complexity by providing an algorithm which learns a mixture of two binary product distributions with uniform mixing weights and low sample complexity. Our algorithm clusters all the samples correctly with high probability, so long as $d(\mu_1, \mu_2)$, the square of the Euclidean distance between the centers of distributions is at least polylogarithmic in s, the number of samples and the following trade-off holds between the separation and the number of samples:

$$s \cdot d^2(\mu_1, \mu_2) \geq a \cdot n \log s \log(ns)$$

for some constant a. We note that in the worst case for our algorithm, when the separation $d(\mu_1, \mu_2)$ is logarithmic in s, the number of samples required is $\tilde{O}(n)$, which matches the sample complexity bounds of SVD-based approaches.

Finally, we study the problem of learning mixtures of heavy-tailed product distributions. This problem was introduced by Dasgupta *et. al* [6] who also introduced the notion of using as a separation measure the distance between the medians of the distributions as a function of the maximum half-radius of any distribution. The main

challenge in learning mixtures of heavy-tailed distributions is that spectral methods and distance concentration – two of the tools commonly used for learning mixtures of Gaussians and binary product distributions do not apply. As a result, previous methods which learnt mixtures of heavy tailed distributions with separation guarantees comparable to the guarantees in [1; 17] required a running time exponential in the dimension.

The main contribution in this thesis is to provide an embedding from $\mathbf{R}^n$ to $\{0,1\}^{n'}$ where $n' > n$. The embedding has the property that samples from distributions in $\mathbf{R}^n$ which satisfy certain conditions and have medians that are far apart, map to samples from distributions in $\{0,1\}^{n'}$ which have centers that are far apart. We then apply this embedding to design an algorithm for learning mixtures of heavy-tailed product distributions. Our algorithm learns mixtures of general product distributions, as long as the distribution of each coordinate satisfies the following two properties. First, the distribution of each coordinate is symmetric about its median, and second, 3/4 of the probability mass is present in an interval of length $2R$ around the median. The separation condition required to correctly classify a $1-\delta$ fraction of the samples is that the distance between the medians of any two distributions in the mixture is $\Omega(R\sqrt{T\log\Lambda} + R\sqrt{T\log(\frac{T}{\delta})}$, where Λ is polynomial in n and T. In addition, we require a spreading constraint, which states that the distance between the medians of any two distributions should be spread across $\Omega(T\log\Lambda + T\log\frac{T}{\delta})$ coordinates. The number of samples required by the algorithm is polynomial in n, T, and $\frac{1}{w_{\min}}$ and our algorithm is based on the algorithm in [5].

A second application of our embedding is in designing an algorithm for learning mixtures of product distributions with finite variance, but with separation proportional to σ_*, the maximum directional standard deviation of any distribution in the mixture along the subspace containing the centers. Given samples from a mixture of distributions with finite variance, our algorithm can learn the mixture provided the following conditions hold. The distance between every pair of centers is at least

$\Omega(\sigma_* \sqrt{T \log \Lambda})$ and this distance is spread along $\Omega(T \log \Lambda)$ coordinates, with no coordinate contributing more than a $\frac{1}{T \log \Lambda}$ fraction of the distance. This condition is weaker than, and is implied by the stronger spreading condition in Chapter 3, which states that every vector in the space spanned by the centers has high spread.

1.3 Bibliographic Notes

Chapters 3 and 5 are joint work with Satish Rao [5], [4]. Chapter 4 is joint work with Eran Halperin, Satish Rao, and Shuheng Zhou [3].

Chapter 2

Background

In this section, we briefly outline some previous work on learning mixtures of distributions with special focus on theoretical work. Since Gaussians are ubiquitous in statistical models, much of the previous work focuses on learning mixtures of Gaussians.

2.1 Early Work

The theoretical framework we use for learning mixtures of distributions was introduced by Dasgupta [7]. In this model, the input to the algorithm is a set of samples randomly generated from a mixture of distributions $\mathcal{D}$, along with the guarantee that every pair of distributions in the mixture is sufficiently different by some measure. This guarantee is called *a separation condition* and the measure of separation used by Dasgupta was the distance between the means of each pair of Gaussians, parameterized by the maximum directional standard deviation of any Gaussian in the mixture.

Dasgupta [7] also provided an algorithm for learning mixtures of spherical Gaussians with a separation of $\Theta(\sigma\sqrt{n})$, where n is the number of dimensions and σ is the directional standard deviation of each Gaussian. A spherical Gaussian is one which has the same standard deviation in every direction. The algorithm uses a random projection of the samples onto a space of dimension $\Theta(\log n)$ followed by a search for the centers in the low-dimensional space.

A variant of the Expectation-Maximization (EM) algorithm [9], which is extensively used in practice to learn mixture models was shown to work for a mixture of Gaussians by Dasgupta and Schulman [8]. Their algorithm works for mixtures of Gaussians with bounded eccentricities - which are Gaussians with a bounded ratio between the maximum and minimum directional variance. The algorithm uses two rounds of EM, and requires a separation of $\Theta(\sigma n^{1/4})$.

2.2 Distance Concentration

Distance concentration algorithms work on the principle that two samples from the same distribution are closer in space with high probability than two samples from different distributions. A series of algorithms with provable guarantees that use distance concentration was provided by Arora and Kannan [2].

The guarantees provided by the algorithms of Arora and Kannan are as follows. The radius of a Gaussian with variances $\sigma_1^2, \ldots, \sigma_n^2$ along its axes is defined as the quantity $(\sum_k \sigma_k^2)^{1/2}$. Arora and Kannan provided an algorithm which learns a mixture of general Gaussians, provided the following separation condition holds. For every i and j, and some t,

$$d(\mu_i, \mu_j) \geq (R_i^2 - R_j^2)^- + 500t(R_i + R_j)\sigma + 100t^2\sigma^2$$

Here, $(R_i^2 - R_j^2)^- = R_i^2 - R_j^2$ when $R_i < R_j$ and 0 otherwise, $d(\mu_i, \mu_j)$ is the square of the Euclidean distance between the means of D_i and D_j, R_i, R_j are the radii of the Gaussians D_i and D_j, and σ is the maximum standard deviation of D_i or D_j in any direction. The separation condition required by the algorithms in [2] was that $t = \Theta(\log |S|)$, where S is the set of samples to be clustered.

To compare this guarantee with the guarantees of [7] and [8], note that for spherical Gaussians, $R_i = R_j$, and the separation requirement is $\Theta(\sigma n^{1/4})$. For general Gaussians with equal radii, assuming that n is much larger than $\log |S|$, the separation requirement is $\Theta(\sqrt{R\sigma})$, where R is the radius of each Gaussian. We note

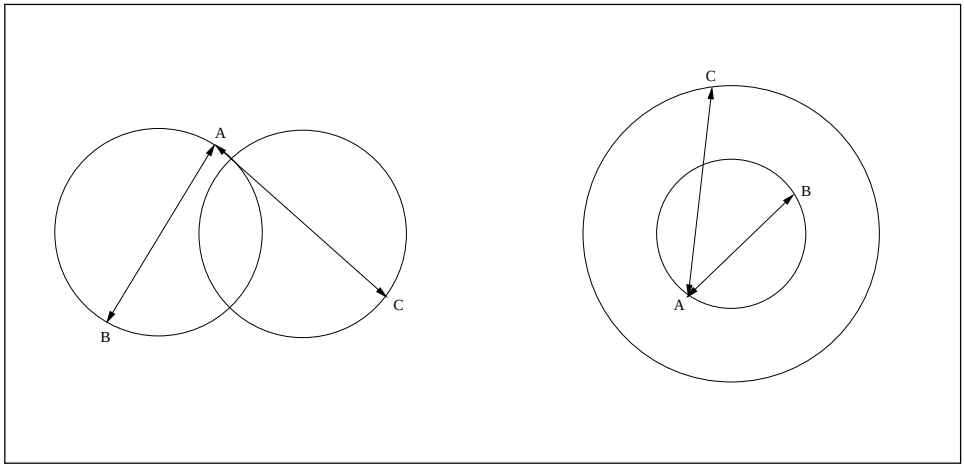

Figure 2.1: Distance Concentration for Gaussians

that this bound is better than $\Theta(\sigma n^{1/4})$ but worse than $\Theta(Rn^{-1/4})$. The algorithms also work when the Gaussians have widely differing radii, but have centers which are close together, for example, for concentric Gaussians with sufficiently different radii – a property which the other algorithms for learning mixtures of Gaussians do not possess.

Although distance-based algorithms, when applied directly to Gaussians in high dimensions, require a large separation between the means, because of their simplicity these algorithms are often used in conjunction with dimension-reduction techniques such as singular value decompositions. In this case, the dimension of the space in which distance concentration is applied is very low, which results in a lower separation requirement.

2.3 Spectral Methods

A popular class of methods for learning mixtures of distributions used extensively in practice are the spectral methods.

The goal of these methods is to isolate a low dimensional subspace, such that the samples from different distributions are sharply separated when projected to such a subspace. If the number of clusters is small, one such subspace is the subspace

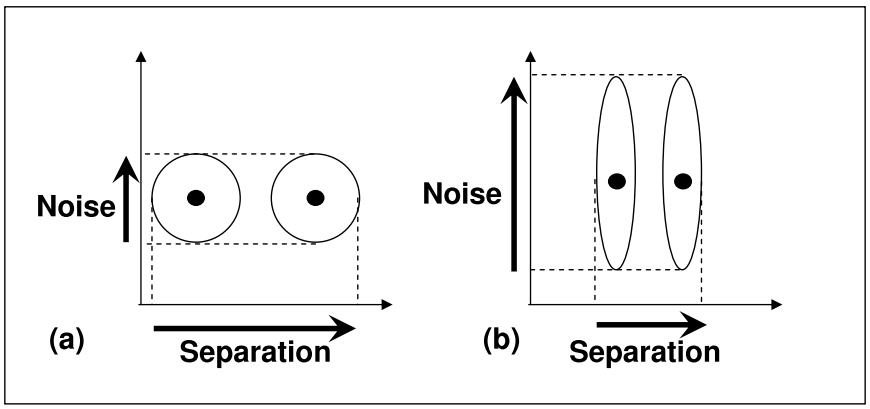

Figure 2.2: Maximum Variance Directions for Spherical and General Gaussians

spanned by the centers of the distributions, and therefore the goal is to find the subspace, or a subspace very close to the subspace which is spanned by the centers of the distributions.

Spectral approaches typically use the vectors corresponding to the top few directions of *highest variance* of the mixture as an approximation to the subspace containing the centers. This is computed by a *Singular Value Decomposition*(SVD) of the matrix of samples, and hence these methods are also called SVD-based methods.

SVD-based approaches have been theoretically analyzed by Vempala and Wang [22] for spherical distributions, and for more general distributions by [17; 1]. For spherical distributions, such approaches work extremely well. As shown in Figure 2.2(a), for spherical distributions, the mixture has higher variance along a direction that contributes to the separation between the centers than along a direction which has no such contribution. As a result,provided there are sufficiently many samples, the subspace containing the centers of the distributions can always be computed by SVD-based approaches. In addition, if the minimum separation between any two distributions is $\Theta(\sigma T^{1/4})$, the samples can be correctly clustered by the algorithm of Vempala and Wang [22].

For general distributions however, SVD-based approaches do not work so well. The directions of maximum variance may well not be the directions in which the

centers are separated, but instead may be the directions of very high noise, as illustrated in Figure 2.2(b). Even if there is no such direction of very high noise, and the distributions are approximately spherical, if the mixing weights are skewed, the directions of maximum variance may well not be the directions along which the centers are separated. This is because a distribution with low mixing weight diminishes the contribution to the variance along a direction that separates the centers: if a distribution has a low mixing weight, there are fewer samples from it in the mixture, which in turn contribute only a small amount to the variance along a direction that separates its center from the center of a distribution with a high mixing weight. On the other hand, given enough samples, the mixing weights have no bearing on the separability of distributions in an information theoretic sense. To illustrate this point, consider for example, two mixtures of any two fixed distributions D_1 and D_2 with different mixing weights: in the first mixture, $w_1 = w_2 = 1/2$, and in the second mixture, $w_1 = 1/4$ and $w_2 = 3/4$. If we have unlimited computational resources, and we can learn the first mixture with 50 samples, we should be able to learn the second mixture with 100 samples, because in the latter case, we have at least as many samples from D_1 and only strictly more samples from D_2.

Moreover, in practice, mixtures with skewed mixing weights arise naturally. See [10; 21], for an example in population genetics. In fact, the results of [17; 1] show that for general Gaussians, the maximum variance directions are the interesting directions when the separation is $\Theta(\frac{\sigma}{\sqrt{w_{\min}}})$, where $w_{\min}$ is the smallest mixing weight of any distribution. This is the best possible result for this approach: [1] presents an example which shows that this condition is required for SVD-based algorithms to succeed.

2.4 Other Models

A second model has been used for learning mixtures of product distributions by Freund and Mansour [14] and Feldman, O'Donnell, and Servedio [11; 12].

In this model, the algorithm is supplied as input samples from a mixture of dis-

tributions and a parameter ϵ, and the goal is to learn within an ϵ-approximation, the parameters of the mixture that produce the samples. More specifically, given samples from a mixture $\mathcal{D}$, an algorithm is required to produce the parameters of a mixture of distributions $\mathcal{D}'$, such that the KL-divergence between $\mathcal{D}$ and $\mathcal{D}'$ is at most ϵ. Recall that the KL-divergence between two distributions with probability density functions P and Q is defined as:

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

and is a measure of the dissimilarity between two distributions.

This model is different from the model of [7] in that there is no separation condition. On the other hand, in this model we are satisfied with an estimate of the parameters of the distributions comprising the mixture, and we do not require reliable classification. As we show in Sections 4.4 and 3.6, even with unlimited computational resources, reliable classification requires that every pair of distributions in the mixture have some minimum separation.

Freund and Mansour [14] provide an algorithm which learns a mixture of two binary product distributions in this model. Their algorithm works by estimating the line joining the two centers of the distributions in the mixture. The running time of the algorithm is $\Theta(\frac{n^2}{\epsilon^2})$, and the number of samples required is also $\Theta(\frac{n^2}{\epsilon^2})$.

Feldman, O'Donnell, and Servedio [11] explore the problem of learning mixtures of T product distributions. Their algorithm works as follows. For every set $\mathcal{F}$ of T coordinates, their algorithm first guesses the values of the centers of each distribution restricted to the coordinates in $\mathcal{F}$. Then, it uses samples to estimate the rest of the parameters of the distributions, and verify that the estimated mixture is indeed close to the input mixture. Their algorithm requires $n^{O(T^3)}$ samples and has a running time polynomial in $\frac{1}{\epsilon}$ and $n^{O(T^3)}$. Their algorithm is extended to axis-aligned Gaussians in [12].

Chapter 3

Learning Mixtures with Small Separation

3.1 Overview

In this chapter, we address the problem of optimizing separation bounds while learning mixtures of distributions. Recall from chapter 2 that the state-of-the-art algorithms for learning mixtures of distributions are singular-value decomposition-based algorithms, which require a separation of $\Omega(\frac{\sigma}{\sqrt{w_{\min}}})$ between the centers, where σ is the maximum directional standard deviation of any distribution in the mixture, and $w_{\min}$ is the lowest mixing weight. This bound is suboptimal because of two reasons. First, given enough samples, mixing weights have no bearing on the separability of distributions. To illustrate this point, consider for example, two mixtures of any two fixed distributions D_1 and D_2 with different mixing weights: in the first mixture, $w_1 = w_2 = 1/2$, and in the second mixture, $w_1 = 1/4$ and $w_2 = 3/4$. If we have unlimited computational resources, and we can learn the first mixture with 50 samples, we should be able to learn the second mixture with 100 samples, because in the latter case, we have at least as many samples from D_1 and only strictly more samples from D_2. This does not always hold for singular-value-decomposition based methods. Secondly, regardless of the value of σ, an algorithm, which has prior knowledge of the subspace containing the centers of the distribution, should be able to learn the mixture when the separation between the centers is proportional to σ_*, the maximum

directional standard deviation of any distribution in the mixture in the subspace containing the centers.

In this chapter, we present an algorithm for learning mixtures of distributions, which is stable in the presence of skewed mixing weights, and, under certain conditions, in the presence of high variance outside the subspace containing the centers. Our algorithm is motivated by the fact that in practice, mixtures with skewed mixing weights arise naturally – see [10; 21] for an example in population genetics. In particular, the dependence of the separation required by our algorithm on skewed mixing weights is only logarithmic. Additionally, with arbitrarily small separation, (*i.e.,* even when the separation is not enough for classification), with enough samples, we can approximate the subspace containing the centers. Previous techniques failed to do so for non-spherical distributions regardless of the number of samples, unless the separation was sufficiently large. Our algorithm works for *Binary Product Distributions* and *Axis-Aligned Gaussians*, both of which possess the coordinate independence property. We require that the distance between the centers be *spread* across $\Theta(T \log \Lambda)$ coordinates, where Λ depends polynomially on the maximum distance between centers and w_{min}. For our algorithm to classify the samples correctly, we further need the separation between centers to be $\Theta(\sigma\sqrt{T \log \Lambda})$.

In addition, if a stronger version of the spreading condition is satisfied, then our algorithm requires a separation of only $\Theta(\sigma_*\sqrt{T \log \Lambda})$ to ensure correct classification of the samples. The stronger spreading condition, which is discussed in more detail later, ensures that when we split the set of all coordinates randomly into two sets, the maximum directional variance of any distribution in the mixture along the projection of the subspace containing the centers into the subspaces spanned by the coordinate vectors in each set, is comparable to σ_*^2.

The spreading condition, which we define precisely in the technical section, is necessary for any method which only has access to the covariance matrix of samples from the mixture. As we discuss in section 3.6.2, the subspace containing the centers

of the distributions is invisible to any such method when the spread is across fewer than T dimensions. The spreading condition follows from the separation condition for binary product distributions. For Gaussian mixtures, this condition fails to hold only when the centers are highly separated in a few coordinates.

Our algorithm is based upon two key insights. The first insight is that if the centers are separated along several coordinates, then many of these coordinates are *correlated* with each other. To exploit this observation, we choose half the coordinates randomly, and search the space of this half for directions of high variance. We use the remaining half of coordinates to *filter* the found directions. If a found direction separates the centers, it is likely to have some correlation with coordinates in the remaining half, and therefore is preserved by the filter. If, on the other hand, the direction found is due to noise, coordinate independence ensures that there will be no correlation with the second half of coordinates, and therefore such directions get filtered away.

The second insight is that the tasks of searching for and filtering the directions can be simultaneously accomplished via a singular value decomposition of the matrix of covariances between the two halves of coordinates.In particular, we show that the top few directions of maximum variance of the covariance matrix approximately capture the subspace containing the centers. Moreover, we show that the covariance matrix has low singular value along any noise direction. By combining these ideas, we obtain an algorithm that is almost insensitive to mixing weights, a property essential for applications like population stratification [3], and which can be implemented using the heavily optimized and thus, efficient, SVD procedure, and which works with a separation condition closer to the information theoretic bound.

3.2 The Model and Results

3.2.1 Notation

We begin with some preliminary definitions about distributions drawn over n dimensional spaces. In the sequel, we use the terms coordinate and feature interchangeably. We use $f, g, \ldots$ to range over coordinates, and $i, j, \ldots$ to range over distributions. For any $x \in \mathbf{R}^n$, we write x^f for the f-th coordinate of x. For any space $\mathcal{H}$ (resp. vector v), we use $\bar{\mathcal{H}}$ (resp. $\bar{v}$) to denote the orthogonal complement of $\mathcal{H}$ (resp. v). For a subspace $\mathcal{H}$ and a vector v, we write $\mathbf{P}_{\mathcal{H}}(v)$ for the projection of v onto the subspace $\mathcal{H}$. For any vector x, we use $||x||$ to denote the norm of x. For any two vectors, x and y, we use $\langle x, y \rangle$ to denote the dot-product of x and y. We use the following Lemma which is similar to Markov's inequality.

Lemma 1 *Let X be a random variable such that $0 \leq X \leq \Gamma$. Then,* $\Pr[X \leq \mathbf{E}[X]/2] \leq \frac{2\Gamma - 2\mathbf{E}[X]}{2\Gamma - \mathbf{E}[X]}$.

Proof: Let $Z = \Gamma - X$. Then, $Z > 0$, $\mathbf{E}[Z] = \Gamma - \mathbf{E}[X]$, and when $X < \mathbf{E}[X]/2$, $Z > \Gamma - \mathbf{E}[X]/2$. We can apply Markov's Inequality on Z to conclude that for any α,

$$\Pr[Z \geq \alpha \mathbf{E}[Z]] \leq \frac{1}{\alpha}$$

The lemma follows by plugging in $\alpha = \frac{\Gamma - \mathbf{E}[X]/2}{\Gamma - \mathbf{E}[X]}$. □

Mixtures of Distributions. A *mixture of distributions* $\mathcal{D}$, is a collection of distributions, $\{D_1, \ldots, D_T\}$, over points in n dimensions, and a set of mixing weights $w_1, \ldots, w_T$ such that $\sum_i w_i = 1$. In the sequel, n is assumed to be much larger than T. When working with a mixture of binary product distributions, we assume that the f-th coordinate of a point drawn from distribution D_i is 1 with probability μ_i^f, and 0 with probability $1 - \mu_i^f$. When working with a mixture of axis-aligned Gaussian distributions, we assume that the f-th coordinate of a point drawn from distribution D_i is distributed as a Gaussian with mean μ_i^f and standard deviation σ_i^f.

Centers. We define the *center* of a distribution i as the vector μ_i, and the *center of mass of the mixture* as the vector $\bar{\mu}$ where $\bar{\mu}^f$ is the mean of the mixture for the coordinate f.

Directional Variance. Let $\mathcal{D}$ be a mixture of distributions. We define σ^2 as the maximum variance of any distribution of the mixture, along any direction. We define σ_*^2 as the maximum variance of any distribution of the mixture, along any direction in the subspace containing the centers of the distributions. We write $\sigma_{\max}^2$ as the maximum variance of the entire mixture in any direction. This may be more than σ^2 due to contribution from the separation between the centers.

Spread. We say that a unit vector v in $\mathbf{R}^n$ has spread $\mathcal{S}$ if

$$\sum_f (v^f)^2 \geq \mathcal{S} \cdot \max_f (v^f)^2$$

Effective Distance and Effective Aspect Ratio. Given a subspace $\mathcal{K}$ of $\mathbf{R}^n$ and two points x, y in $\mathbf{R}^n$, we write $d_{\mathcal{K}}(x, y)$ for the square of the Euclidean distance between x and y projected along the subspace $\mathcal{K}$.

Given a pair of centers μ_i, μ_j, we choose c_{ij} and a parameter Λ as follows. $\Lambda > \frac{\sigma_{\max} T \log^2 n}{w_{\min} \cdot (\min_{i,j} c_{ij}^2)}$ and c_{ij} is the maximum value such that there are $49T \log \Lambda$ coordinates f with $|\mu_i^f - \mu_j^f| > c_{ij}$. As $\sigma_{\max}$ is the maximum directional standard deviation of the mixture, and $w_{\min} \cdot (\min_{i,j} c_{ij}^2)$ is a lower bound on its minimum directional standard deviation along a space that contains a lot of the distance between the centers, Λ is, in some sense, the *effective aspect ratio* of the mixture, when we ignore the contributions to the distance from the top few coordinates. We note that Λ is bounded by a polynomial in $T, \sigma_*, 1/w_{min}$ and logarithmic in n. We also note that $\Lambda \leq \min_{i,j} \frac{\sigma_{\max} T \log^2 n}{w_{\min} \min_{i,j} \max_f |\mu_i - \mu_j|^2}$; however, this upper bound may be very loose if there are a few coordinates f for which $|\mu_i^f - \mu_j^f|$ is very high.

We define $c_{\min}$ to be the minimum over all pairs i, j of c_{ij}. Given a pair of centers i and j, let Δ_{ij} be the set of coordinates f such that $|\mu_i^f - \mu_j^f| > c_{ij}$, and let ν_{ij} be defined as: $\nu_{ij}^f = \mu_i^f - \mu_j^f$, if $f \notin \Delta_{ij}$, and $\nu_{ij}^f = c_{ij}$ otherwise. We define $\bar{d}(\mu_i, \mu_j)$,

the effective distance between μ_i and μ_j to be the square of the L_2 norm of ν_{ij}. In contrast, the square of the norm of the vector $\mu_i - \mu_j$ is the actual distance between centers μ_i and μ_j, and is always greater than or equal to the effective distance between μ_i and μ_j. Moreover, given i and j and the subspace $\mathcal{K}$, we define $\bar{d}_{\mathcal{K}}(\mu_i, \mu_j)$ as the square of the norm of the vector ν_{ij} projected onto the subspace $\mathcal{K}$.

3.2.2 A Summary of Our Results

In this chapter, we present Algorithm Correlation-Cluster, a correlation-based algorithm for learning mixtures of binary product distributions and axis-aligned Gaussians. The inputs to the algorithm are samples from a mixture of distributions, and the output is a clustering of the samples. The main component of Algorithm Correlation-Cluster is Algorithm Correlation-Subspace, which, given samples from a mixture of distributions, computes an approximation to the subspace containing the centers of the distributions.

The properties of these algorithms are summarized in Theorem 2 and Theorem 1.

Theorem 1 (Spanning centers) *Suppose we are given a mixture of distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$, with mixing weights $w_1, \ldots, w_T$. Then with at least constant probability, Algorithm* Correlation-Subspace *has the following properties.*

1. *If, for all i and j, $\bar{d}(\mu_i, \mu_j) \geq 49 c_{ij}^2 T \log \Lambda$, then in polynomial time we can find a subspace $\mathcal{K}$ of dimension at most $2T$ such that, for all pairs i, j,*

$$d_{\mathcal{K}}(\mu_i, \mu_j) \geq \frac{99}{100}(\bar{d}(\mu_i, \mu_j) - 49 T c_{ij}^2 \log \Lambda)$$

2. *If, in addition, every vector in $\mathcal{C}$ has spread $32T \log \frac{\sigma}{\sigma_*}$, then, with at least constant probability, the maximum directional variance in $\mathcal{K}$ of any distribution D_i in the mixture is at most $11\sigma_*^2$.*

The number of samples required by Algorithm Correlation-Subspace *is polynomial in $\frac{\sigma}{\sigma_*}$, T, n,σ and $\frac{1}{w_{\min}}$, and the algorithm runs in time polynomial in n, T, and the number of samples.*

The subspace $\mathcal{K}$ described above approximates the subspace in which the centers μ_i of the distributions lie in the sense that the centers when projected onto $\mathcal{K}$ are far apart.

Theorem 2 (Clustering) *Suppose we are given a mixture of distributions $\mathcal{D} = \{D_1, \ldots, D_T\}$, with mixing weights $w_1, \ldots, w_T$. Then, Algorithm* Correlation-Cluster *has the following properties.*

1. *If for all i and j, $\bar{d}(\mu_i, \mu_j) \geq 49Tc_{ij}^2 \log \Lambda$, and for all i, j we have:*

$$\begin{aligned} \bar{d}(\mu_i, \mu_j) &> 59\sigma^2 T(\log \Lambda + \log n) \quad \textit{(for axis-aligned Gaussians)} \\ \bar{d}(\mu_i, \mu_j) &> 59T(\log \Lambda + \log n) \quad \textit{(for binary product distributions)} \end{aligned}$$

   then with probability $1 - \frac{1}{n}$ over the samples and with constant probability over the random choices made by the algorithm, Algorithm Correlation-Cluster *computes a correct clustering of the sample points.*

2. *If every vector in $\mathcal{C}$ has spread at least $32T \log \frac{\sigma}{\sigma_*}$, and for all i, j, we have, for axis-aligned Gaussians:*

$$\bar{d}(\mu_i, \mu_j) \geq 150\sigma_*^2 T(\log \Lambda + \log n)$$

   then, with constant probability over the randomness in the algorithm, and with probability $1 - \frac{1}{n}$ over the samples, Algorithm Correlation-Cluster *computes a correct clustering of the sample points.*

Algorithm Correlation-Cluster *runs in time polynomial in n and the number of samples required by Algorithm* Correlation-Cluster *is polynomial in $\frac{\sigma}{\sigma_*}$, T, n, σ and $\frac{1}{w_{\min}}$.*

The second theorem follows from the first and distance concentration Lemmas of [1] as described in Section 3.4.4. The Lemmas show that once the points are projected onto the subspace computed in Theorem 1, a distance-based clustering method suffices to correctly cluster the points.

CORRELATION-CLUSTER(S)

1. Partition S into S_A and S_B uniformly at random.
2. Compute: $\mathcal{K}_A$ = FIND-SUBSPACE(S_A), $\mathcal{K}_B$ = FIND-SUBSPACE(S_B)
3. Project each point in S_B (resp. S_A) on the subspace $\mathcal{K}_A$ (resp. $\mathcal{K}_B$).
4. Use a distance-based clustering algorithm [2] to partition the points in S_A and S_B after projection.

Figure 3.1: Algorithm CORRELATION-CLUSTER

3.3 Algorithm CORRELATION-CLUSTER

Our clustering algorithm follows the same basic framework as the SVD-based algorithms of [22; 17; 1]. This framework is described in Figure 3.1. The input to the algorithm is a set S of samples, and the output is a pair of clusterings of the samples according to source distribution.

The first step in the algorithm is to find a subspace $\mathcal{K}$ such that the distances between the centers of each pair of distributions in the mixture are large when the centers are projected onto $\mathcal{K}$. [22; 17; 1] find such a subspace via a SVD of the samples; our algorithm uses a different procedures which we describe in the following section. After computing $\mathcal{K}$, the samples are projected onto $\mathcal{K}$ and finally, a distance-based clustering algorithm is used to find the clusters.

We note that in order to preserve independence the samples we project onto $\mathcal{K}$ should be distinct from the ones we use to compute $\mathcal{K}$. A clustering of the complete set of points can then be computed by partitioning the samples into two sets A and B. We use A to compute $\mathcal{K}_A$, which is used to cluster B and vice-versa.

We now present our algorithm which computes a basis for the subspace $\mathcal{K}$. With slight abuse of notation we use $\mathcal{K}$ to denote the set of vectors that form the basis for the subspace $\mathcal{K}$.The input to CORRELATION-SUBSPACE is a set S of samples, and

the output is a subspace $\mathcal{K}$ of dimension at most $2T$.

Algorithm CORRELATION-SUBSPACE**:**

Step 1: Initialize and Split Initialize the basis $\mathcal{K}$ with the empty set of vectors. Randomly partition the coordinates into two sets, $\mathcal{F}$ and $\mathcal{G}$, each of size $n/2$. Order the coordinates as those in $\mathcal{F}$ first, followed by those in $\mathcal{G}$.

Step 2: Sample Translate each sample point so that the center of mass of the set of sample points is at the origin. Let F (respectively G) be the matrix which contains a row for each sample point, and a column for each coordinate in $\mathcal{F}$ (respectively $\mathcal{G}$). For each matrix, the entry at row x, column f is the value of the f-th coordinate of the sample point x divided by $\sqrt{|S|}$.

Step 3: Compute Singular Space For the matrix $F^{\mathbf{T}}G$, compute $\{v_1, \ldots, v_T\}$, the top T left singular vectors, $\{y_1, \ldots, y_T\}$, the top T right singular vectors, and $\{\lambda_1, \ldots, \lambda_T\}$, the top T singular values.

Step 4: Expand Basis For each i, we abuse notation and use v_i (y_i respectively) to denote the vector obtained by concatenating v_i with the 0 vector in $n/2$ dimensions (0 vector in $n/2$ dimensions concatenated with y_i respectively). For each i, if the singular value λ_i is more than a threshold $\tau = O\left(\frac{w_{\min} c_{ij}^2}{T \log^2 n} \cdot \sqrt{\log \Lambda}\right)$, we add v_i and y_i to $\mathcal{K}$.

Step 5: Output Output the set of vectors $\mathcal{K}$.

The main idea behind our algorithm is to use half the coordinates (the set $\mathcal{F}$) to compute a subspace which approximates the subspace containing the centers, and the remaining half (the set $\mathcal{G}$) to validate that the subspace computed is indeed a good approximation. We critically use the coordinate independence property to make this validation possible.

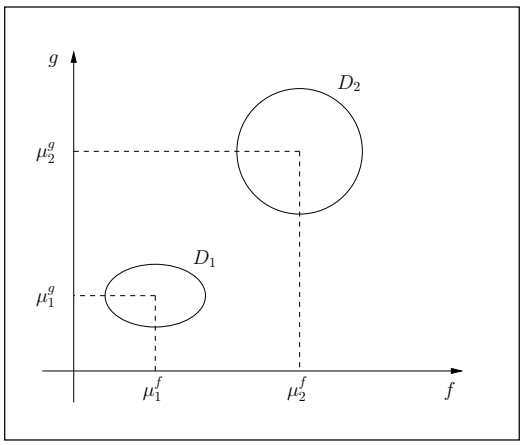

Figure 3.2: Projections on Two Coordinates Separating Centers

3.4 Analysis of Algorithm Correlation-Cluster

This section is devoted to proving Theorems 1, and 2. The following notation is used consistently for the rest of this section.

Notation. We write $\mathcal{F}$-space (resp. $\mathcal{G}$-space) for the $n/2$ dimensional subspace of $\mathbf{R}^n$ spanned by the coordinate vectors $\{e_f \mid f \in \mathcal{F}\}$ (resp. $\{e_g \mid g \in \mathcal{G}\}$). We write $\mathcal{C}$ for the subspace spanned by the set of vectors μ_i. We write $\mathcal{C}_\mathcal{F}$ for the space spanned by the set of vectors $\mathbf{P}_\mathcal{F}(\mu_i)$. We write $\mathbf{P}_\mathcal{F}(\bar{\mathcal{C}}_\mathcal{F})$ for the orthogonal complement of $\mathcal{C}_\mathcal{F}$ in the $\mathcal{F}$-space. Moreover, we write $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$ for the subspace of dimension $2T$ spanned by the union of a basis of $\mathcal{C}_\mathcal{F}$ and a basis of $\mathcal{C}_\mathcal{G}$.

Next, we define a key ingredient of the analysis.

Covariance Matrix. Let N be a large number. We define $\hat{F}$ (resp. $\hat{G}$), the *perfect sample matrix* with respect to $\mathcal{F}$ (resp. $\mathcal{G}$) as the $N \times n/2$ matrix whose rows from $(w_1 + \ldots + w_{i-1})N + 1$ through $(w_1 + \ldots + w_i)N$ are equal to the vector $\mathbf{P}_\mathcal{F}(\mu_i)/\sqrt{N}$ (resp. $\mathbf{P}_\mathcal{G}(\mu_i)/\sqrt{N}$). For a coordinate f, let X_f be a random variable which is distributed as the f-th coordinate of the mixture $\mathcal{D}$. As the entry in row f and column g in the matrix $\hat{F}^\mathbf{T}\hat{G}$ is equal to $\mathbf{Cov}(X_f, X_g)$, the covariance of X_f and X_g, we call the matrix $\hat{F}^\mathbf{T}\hat{G}$ the *covariance matrix* of $\mathcal{F}$ and $\mathcal{G}$.

Proof Structure. The overall structure of our proof is as follows. First, we show

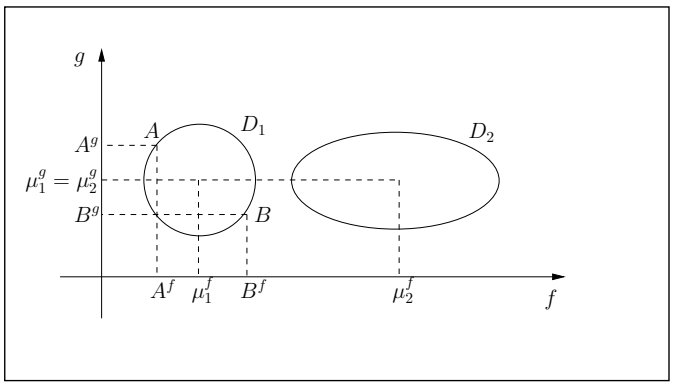

Figure 3.3: Projections on a Coordinate Separating Centers and a Noise Coordinate

that the centers of the distributions in the mixture have a high projection on the subspace of highest correlation between the coordinates. To do this, we first assume, in Section 3.4.1 that the input to the algorithm in Step 2 are the perfect sample matrices $\hat{F}$ and $\hat{G}$. Of course, we cannot directly feed in the matrices $\hat{F}, \hat{G}$, as the values of the centers are not known in advance. Next, we show in Section 3.4.2 that this holds even when the matrices F and G in Step 2 of Algorithm Correlation-Subspace are obtained by sampling. In Section 3.4.3, we combine these two results and prove Theorem 1. Finally, in Section 3.4.4, we show that distance concentration algorithms work in the low-dimensional subspace produced by Algorithm Correlation-Cluster, and complete the analysis by proving Theorem 2.

3.4.1 The Perfect Sample Matrix

The goal of this section is to prove Lemmas 2 and 7, which establish a relationship between directions of high correlation and directions which contain a lot of distance between centers. Lemma 2 shows that a direction which contains a lot of effective distance between some pair of centers, is also a direction of high correlation. The intuition for this lemma is shown in Figure 3.2.

Lemma 7 shows that a direction $v \in \mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$, which is perpendicular to the space containing the centers, is a direction with 0 correlation. The intuition for this lemma is shown in Figure 3.3. In addition, we show in Lemma 8, another property

of the perfect sample matrix – the covariance matrix constructed from the perfect sample matrix has rank at most T. We conclude this section by showing in Lemma 9 that when every vector in $\mathcal{C}$ has high spread, then the directional variance of any distribution in the mixture along $\mathcal{F}$-space or $\mathcal{G}$-space is of the order of σ_*^2.

We begin by showing that the directions which contain a lot of the distance between the centers, also have high correlations between coordinates between most splits. This means that with high probability over the splits, such directions are directions of high correlation.

Lemma 2 *Let v be any vector in $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$ such that for some i and j, $\bar{d}_v(\mu_i,\mu_j) \geq 49Tc_{ij}^2\log\Lambda$. If $v_\mathcal{F}$ and $v_\mathcal{G}$ are the normalized projections of v to $\mathcal{F}$-space and $\mathcal{G}$-space respectively, then, with probability at least $1-\frac{1}{T}$ over the splitting step, for all such v, $v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}v_\mathcal{G} \geq \tau$ where $\tau = O\left(\frac{w_{\min}c_{ij}^2}{T\log^2 n}\cdot\sqrt{\log\Lambda}\right)$.*

PROOF: From Lemma 3, for a fixed vector $v\in\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$, $v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}v_\mathcal{G}\geq 2\tau$, with probability $1-\Lambda^{-2T}$. Let u be a vector, and $u_\mathcal{F}$ and $u_\mathcal{G}$ be its normalized projection on $\mathcal{F}$-space and $\mathcal{G}$-space respectively, such that $||u_\mathcal{F}-v_\mathcal{F}|| < \frac{\delta}{2}$ and $||u_\mathcal{G}-v_\mathcal{G}|| < \frac{\delta}{2}$. Then, $u_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}u_\mathcal{G} = v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}v_\mathcal{G} + (u_\mathcal{F}^\mathbf{T}-v_\mathcal{F}^\mathbf{T})\hat{F}^\mathbf{T}\hat{G}u_\mathcal{G} + v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}(u_\mathcal{G}-v_\mathcal{G}) \geq v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}v_\mathcal{G} - \delta\sigma_{\max}$, as $\sigma_{\max}$ is the maximum value of $x^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}y$, over all unit vectors x and y. Since $v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}v_\mathcal{G} > 2\tau$, if $\delta < \frac{\tau}{\sigma_{\max}}$, $u_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}u_\mathcal{G} > \tau$. To show that the lemma holds for all $v\in\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$, it is therefore sufficient to show that the statement holds for all unit vectors in a $(\frac{\tau}{\sigma_{\max}})$-cover of $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$. Since $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$ has dimension at most $2T$, there are at most $(\frac{\tau}{\sigma_{\max}})^{2T}$ vectors in such a cover. As $\Lambda > \frac{\sigma_{\max}}{\tau}$, the lemma follows from Lemma 3 and a union bound over the vectors in a $(\frac{\tau}{\sigma_{\max}})$-cover of $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$. □

Lemma 3 *Let v be a fixed vector in $\mathcal{C}$ such that for some i and j, $\bar{d}_v(\mu_i,\mu_j) \geq 49Tc_{ij}^2\log\Lambda$. If $v_\mathcal{F}$ and $v_\mathcal{G}$ are the projections of v to $\mathcal{F}$-space and $\mathcal{G}$-space respectively, then, with probability at least $1-\Lambda^{-2T}$ over the splitting step, $v_\mathcal{F}^\mathbf{T}\hat{F}^\mathbf{T}\hat{G}v_\mathcal{G}\geq 2\tau$ where $\tau = O\left(\frac{w_{\min}c_{ij}^2}{T\log^2 n}\cdot\sqrt{\log\Lambda}\right)$.*

Let $\hat{F}_v$ ($\hat{G}_v$ respectively) be the $s \times n/2$ matrix obtained by projecting each row of $\hat{F}$ (respectively $\hat{G}$) on $v_{\mathcal{F}}$ (respectively $v_{\mathcal{G}}$). Then,

$$v_{\mathcal{F}}^{\mathbf{T}} \hat{F}_v^{\mathbf{T}} \hat{G}_v v_{\mathcal{G}} = \sum_i w_i \langle v_{\mathcal{F}}, \mathbf{P}_{v_{\mathcal{F}}}(\mu_i - \bar{\mu}) \rangle \langle v_{\mathcal{G}}, \mathbf{P}_{v_{\mathcal{G}}}(\mu_i - \bar{\mu}) \rangle = v_{\mathcal{F}}^{\mathbf{T}} \hat{F}^{\mathbf{T}} \hat{G} v_{\mathcal{G}}$$

Moreover, for any pair of vectors x in $\mathcal{F}$-space and y in $\mathcal{G}$-space such that $\langle x, v_{\mathcal{F}} \rangle = 0$ and $\langle y, v_{\mathcal{G}} \rangle = 0$,

$$x^{\mathbf{T}} \hat{F}_v^{\mathbf{T}} \hat{G}_v y = \sum_i w_i \langle x, \mathbf{P}_{v_{\mathcal{F}}}(\mu_i - \bar{\mu}) \rangle \langle y, \mathbf{P}_{v_{\mathcal{G}}}(\mu_i - \bar{\mu}) \rangle = 0$$

Therefore, $\hat{F}_v^{\mathbf{T}} \hat{G}_v$ has rank at most 1.

The proof strategy is to show that if $d_v(\mu_i, \mu_j)$ is large then the matrix $\hat{F}_v^{\mathbf{T}} \hat{G}_v$ has high norm. We require the following notation. For each coordinate f we define a T-dimensional vector z_f as $z_f = [\sqrt{w_1}\mathbf{P}_v(\mu_1^f - \bar{\mu}^f), \ldots, \sqrt{w_T}\mathbf{P}_v(\mu_T^f - \bar{\mu}^f)]$. Notice that for any two coordinates f,g: $\langle z_f, z_g \rangle = \mathbf{Cov}(\mathbf{P}_v(X_f), \mathbf{P}_v(X_g))$, computed over the entire mixture. We also observe that $\sum_f ||z_f||^2 = \sum_i w_i \cdot d_v(\mu_i, \bar{\mu})$. The RHS of this equality is the weighted sum of the squares of the Euclidean distances between the centers of the distributions and the center of mass. By the triangle inequality, this quantity is at least $49 w_{\min} c_{ij}^2 T \log \Lambda$. To prove Lemma 3, we require the following three Lemmas.

Lemma 4 *If A is a set of coordinates whose cardinality $|A|$ is greater than T, such that for each $f \in A$, the norms $||z_f||$ are equal and $\sum_{f \in A} ||z_f||^2 = D$, then*

$$\sum_{f,g \in A,\ f \neq g} \langle z_f, z_g \rangle^2 \geq \frac{D^2}{2T}(1 - \frac{T}{|A|})$$

PROOF: The proof uses the following strategy. First, we group the coordinates into sets $B_1, B_2, \ldots$ such that for each B_k, the set $\{z_f \mid f \in B_k\}$ approximately form a basis of the T dimensional space spanned by the set of vectors $\{z_f \mid f \in A\}$. Next, for each basis B_k we estimate the value of $\sum_{f \in B_k} \sum_{g \notin B_k} \langle z_f, z_g \rangle^2$ and finally, we prove the Lemma, by using the estimates, and the fact that: $\sum_{f,g \in A,\ f \neq g} \langle z_f, z_g \rangle^2$ can be written as $\sum_{B_k} \sum_{f \in B_k} \sum_{g \notin B_k} \langle z_f, z_g \rangle^2$. Consider the following procedure which returns a series of sets of vectors, $B_1, B_2, \ldots$.

$S \leftarrow A;\ k \leftarrow 0$
while $S \neq \emptyset$
  $k \leftarrow k+1;\ B_k \leftarrow \emptyset;$
  for each $z \in S$
    if $||z||/2 \geq$ projection of z onto B_k **then**
      remove z from S, add z to B_k

We observe that by construction, for each k,

$$\sum_{f \in B_k} \sum_{g \in B_{k'},\ k'>k} \langle z_f, z_g \rangle^2 \geq \frac{D}{|A|} \sum_{g \in B_{k'},\ k'>k} \frac{||z_g||^2}{2}$$

This inequality follows since the basis is formed of vectors of length $|D|/|A|$ and the fact that the projection of each z_g on B_k preserves at least half of its norm.

Since the vectors are in T dimensions there are at least $|A|/T$ sets B_k. Notice that for all $f \in A$, we have $||z_f||^2 = D/|A|$, as all the $||z_f||$ are equal. Hence, the average contribution to the sum from each B_k is at least $\frac{|A|-T}{2} \cdot \frac{D^2}{|A|^2}$, from which the Lemma follows. □

Lemma 5 *Let A be a set of coordinates with cardinality more than $144T^2 \log \Lambda$ such that for each $f \in A$, $||z_f||$ is equal and $\sum_{f \in A} ||z_f||^2 = D$. Then,*

1. $\sum_{f,g \in A, f \neq g} \langle z_f, z_g \rangle^2 \geq \frac{D^2}{288T^2 \log \Lambda}$
2. *with probability $1 - \Lambda^{-2T}$ over the splitting of coordinates in Step 1,*

$$\sum_{f \in \mathcal{F} \cap A, g \in \mathcal{G} \cap A} \langle z_f, z_g \rangle^2 \geq \frac{D^2}{1152T^2 \log \Lambda}$$

PROOF: We can partition A into $72T \log \Lambda$ groups $A_1, A_2, \ldots$ each with at least $2T$ coordinates, such that for each group A_k, $\sum_{f \in A_k} ||z_f||^2 \geq \frac{D}{72T \log \Lambda}$. From Lemma 4, for each such group A_k,

$$\sum_{f,f' \in A_k, f \neq f'} \langle z_f, z_g \rangle^2 \geq \left(\frac{D}{72T \log \Lambda} \right)^2 \cdot \frac{1}{4T} \geq \frac{D^2}{20736T^3 \log^2 \Lambda}$$

Summing over all the groups,

$$\sum_{f,f' \in A, f \neq f'} \langle z_f, z_{f'} \rangle^2 \geq \sum_k \sum_{f,f' \in A_k, f \neq f'} \langle z_f, z_g \rangle^2 \geq \frac{D^2}{288T^2 \log \Lambda}$$

from which the first part of the lemma follows.

For each group A_k, $\sum_{f\in\mathcal{F}\cap A_k}\sum_{g\in\mathcal{G}\cap A_k}\langle z_f, z_g\rangle^2$ is a random variable whose value depends on the outcome of the splitting in Step 1 of the algorithm. The maximum value of this random variable is $\sum_{f,g\in A_k, f\neq g}\langle z_f, z_g\rangle^2$, and the expected value is $\frac{1}{2}\sum_{f,g\in A_k, f\neq g}\langle z_f, z_g\rangle^2$. Therefore, by Lemma 1, with probability $\frac{1}{3}$,

$$\sum_{f\in\mathcal{F}\cap A_k}\sum_{g\in\mathcal{G}\cap A_k}\langle z_f, z_g\rangle^2 \geq \frac{1}{4}\sum_{f,g\in A_k, f\neq g}\langle z_f, z_g\rangle^2 \geq \frac{D^2}{82944T^3\log\Lambda}$$

Moreover, as each group A_k is disjoint, the splitting process in each group is independent. Since there are $72T\log\Lambda$ groups, using the Chernoff bounds, we can conclude that with probability $1-\Lambda^{-2T}$, for at least $\frac{1}{6}$ fraction of the groups A_k, $\sum_{f\in\mathcal{F}\cap A_k}\sum_{g\in\mathcal{G}\cap A_k}\langle z_f, z_g\rangle^2 \geq \frac{D^2}{82944T^3\log\Lambda}$, from which the second part of the lemma follows. □

Lemma 6 *Let A be a set of coordinates such that for each $f\in A$, $||z_f||$ is equal and $\sum_{f\in A}||z_f||^2 = D$. If $48T\log\Lambda + T < |A| \leq 144T^2\log\Lambda$,*

1. $\sum_{f,g\in A, f\neq g}\langle z_f, z_g\rangle^2 \geq \frac{D^2}{1152T^4\log\Lambda}$
2. *With probability $1-\Lambda^{-2T}$ over the splitting in Step 1,*

$$\sum_{f\in\mathcal{F}\cap A, g\in\mathcal{G}\cap A}\langle z_f, z_g\rangle^2 \geq \frac{D^2}{4608T^4\log\Lambda}$$

PROOF: The proof strategy here is to group the coordinates in A into tuples $A_k = \{f_k, f_k'\}$ such that for all but T such tuples, $\langle z_{f_k}, z_{f_k'}\rangle^2$ is high. For each A_k we estimate the quantity $\langle z_{f_k}, z_{f_k'}\rangle^2$. The proof of the first part of the lemma follows by using the estimates and the fact that $\sum_{f,f'\in A, f\neq f'}\langle z_f, z_{f'}\rangle^2 \geq \sum_k\langle z_{f_k}, z_{f_k'}\rangle^2$. Consider the following procedure, which returns a series of sets of vectors $A_1, A_2, \ldots$.

$S \leftarrow A;\ k \leftarrow 0$
while $S \neq \emptyset$
  $k \leftarrow k+1;\ A_k \leftarrow \emptyset;$
  if there exists $z_{f_k}, z_{f_k'} \in S$ with $\langle z_{f_k}, z_{f_k'}\rangle \geq \frac{D}{2T|A|}$, **then**
    $A_k = \{f_k, f_k'\}$
    remove $z_{f_k}, z_{f_k'}$ from S

We observe that in some iteration k, if Step 4 of the procedure succeeds, then, $\langle z_{f_k}, z_{f'_k}\rangle^2 \geq \frac{D^2}{4T^2|A|^2}$. We now show that the step will succeed if there are more than T vectors in S.

Suppose for contradiction that Step 4 of the procedure fails when $|S| > T$. Then, we can build a set of vectors B, which approximately forms a basis of S as follows.

$B \leftarrow \emptyset$
for each $z \in S$
  if projection of z onto B is less than $\frac{D}{2|A|}$ **then**
    remove z from S, add z to B

Since the vectors in S have dimension at most T, if $|S| > T$, there is at least one vector $z \notin B$. By construction, the projection of any such z on B preserves at least half its norm. Therefore, there exists some vector $z_f \in B$ such that $\langle z, z_f\rangle \geq \frac{D}{2T|A|}$, which is a contradiction to the failure of Step 4 of our procedure. Thus,

$$\sum_k \langle z_{f_k}, z_{f'_k}\rangle^2 \geq \frac{D^2}{4T^2|A|^2} \cdot \left(\frac{|A|-T}{2}\right) \geq \frac{D^2}{8T^2|A|} \geq \frac{D^2}{1152T^4 \log \Lambda}$$

from which the first part of the lemma follows.

With probability $\frac{1}{2}$ over the splitting in Step 1 of the algorithm, for any tuple A_k, f_k and f'_k will belong to different sets, and thus contribute $\frac{D^2}{4T^2|A|}$ to $\sum_{f\in\mathcal{F}\cap A, g\in\mathcal{G}\cap A}\langle z_f, z_g\rangle^2$. Since the groups are disjoint, the splitting process in each group is independent. As there are $\frac{|A|-T}{2}$ groups, and $|A| \geq 48Tc_{ij}^2 \log\Lambda$, we can use the Chernoff Bounds to conclude that with probability at least $1 - \Lambda^{-2T}$, at least $1/4$ fraction of the tuples contribute to the sum, from which the second part of the lemma follows. □

PROOF:(Of Lemma 3) From the definition of effective distance, if the condition: $\bar{d}_v(\mu_i, \mu_j) > 49c_{ij}^2 T \log\Lambda$ holds then there are at least $49T\log\Lambda$ vectors z_f with total squared norm at least $98w_{\min}{c_{ij}}^2 T\log\Lambda$. In the sequel we will scale down each vector z_f with norm greater than $c_{ij}\sqrt{w_{\min}}$ so that its norm is exactly $c_{ij}\sqrt{w_{\min}}$. We divide the vectors into $\log n$ groups as follows: group B_k contains vectors which have norm between $\frac{c_{ij}\sqrt{w_{\min}}}{2^k}$ and $\frac{c_{ij}\sqrt{w_{\min}}}{2^{k-1}}$.

We will call a vector *small* if its norm is less than $\frac{\sqrt{w_{\min}}c_{ij}}{2\sqrt{\log n}}$, and otherwise, we call the vector *big*. We observe that there exists a set of vector B with the following properties: (1) the cardinality of B is more than $49T\log\Lambda$, (2) the total sum of squares of the norm of the vectors in B is greater than $\frac{49T\log\Lambda w_{\min}c_{ij}^2}{\log n}$, and, (3) the ratio of the norms of any two vectors in B is at most $2\sqrt{\log n}$.

Case 1: Suppose there exists a group B_k of small vectors the squares of whose norms sum to a value greater than $\frac{49Tw_{\min}c_{ij}^2\log\Lambda}{\log n}$. By definition, such a group has more than $49T\log\Lambda$ vectors, and the ratio is at most 2.

Case 2: Otherwise, there are at least $49T\log\Lambda$ big vectors. By definition, the sum of the squares of their norms exceeds $\frac{49Tw_{\min}c_{ij}^2\log\Lambda}{\log n}$. Due to the scaling, the ratio is at most $2\sqrt{\log n}$.

We scale down the vectors in B so that each vector has squared norm $\frac{w_{\min}c_{ij}^2}{2^k}$ in case 1, and, squared norm $\frac{w_{\min}c_{ij}^2}{4\log n}$ in case 2. Due to (2) and (3), the total squared norm of the scaled vectors is at least $\frac{49Tw_{\min}c_{ij}^2\log\Lambda}{4\log^2 n}$.

Due to (1), we can now apply Lemmas 5 and 6 on the vectors to conclude that for some constant a_1, with probability $1-\Lambda^{-2T}$, $\sum_{f\in\mathcal{F},g\in\mathcal{G}}\langle z_f,z_g\rangle^2 \geq a_1\cdot\left(\frac{w_{\min}^2c_{ij}^4\log\Lambda}{T^2\log^4 n}\right)$. The above sum is the square of the Frobenius norm $|\hat{F}_v^{\mathbf{T}}\hat{G}_v|_{\mathbf{F}}$ of the matrix $\hat{F}_v^{\mathbf{T}}\hat{G}_v$. Since $\hat{F}_v^{\mathbf{T}}\hat{G}_v$ has rank at most 1, an application of Lemma 14 completes the proof, for $\tau = O\left(\frac{w_{\min}c_{ij}^2}{T\log^2 n}\cdot\sqrt{\log\Lambda}\right)$. □

Next we show that a vector $x\in\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$ is a direction of 0 correlation. A similar statement holds for a vector $y\in\mathbf{P}_{\mathcal{G}}(\bar{\mathcal{C}}_{\mathcal{G}})$.

Lemma 7 *If at Step 2 of Algorithm* CORRELATION-SUBSPACE*, the values of F and G are respectively $\hat{F}$ and $\hat{G}$, and for some k,the top k-th left singular vector is v_k and the corresponding singular value λ_k is more than τ, then for any vector x in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$, $\langle v_k,x\rangle = 0$.*

PROOF: We first show that for any x in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$, and any y, $x^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y = 0$.

$$x^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y = \sum_{i=1}^{T} w_i\langle\mathbf{P}_{\mathcal{F}}(\mu_i),x\rangle\cdot\langle\mathbf{P}_{\mathcal{G}}(\mu_i),y\rangle$$

Since x is in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$, $\langle \mathbf{P}_{\mathcal{F}}(\mu_i), x\rangle = 0$, for all i, and hence $x^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y = 0$ for all x in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$. We now prove the Lemma by induction on k.

Base case ($k = 1$). Let $v_1 = u_1 + x_1$, where $u_1 \in \mathcal{C}_{\mathcal{F}}$ and $x_1 \in \mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$. Let y_1 be the top right singular vector of $\hat{F}^{\mathbf{T}}\hat{G}$, and let $|x_1| > 0$. Then, $v_1^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_1 = u_1^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_1$, and $u_1/|u_1|$ is a vector of norm 1 such that $\frac{1}{|u_1|}u_1^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_1 > v_1^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_1$, which contradicts the fact that v_1 is the top left singular vector of $\hat{F}^{\mathbf{T}}\hat{G}$.

Inductive case. Let $v_k = u_k + x_k$, where $u_k \in \mathcal{C}_{\mathcal{F}}$ and $x_k \in \mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$. Let y_k be the top k-th right singular vector of $\hat{F}^{\mathbf{T}}\hat{G}$, and let $|x_k| > 0$. We first show that u_k is orthogonal to each of the vectors $v_1, \ldots, v_{k-1}$. Otherwise, suppose there is some j, $1 \leq j \leq k-1$, such that $\langle u_k, v_j\rangle \neq 0$. Then, $\langle v_k, v_j\rangle = \langle x_k, v_j\rangle + \langle u_k, v_j\rangle = \langle u_k, v_j\rangle \neq 0$. This contradicts the fact that v_k is a left singular vector of $\hat{F}^{\mathbf{T}}\hat{G}$. Therefore, $v_k^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_k = u_k^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_k$, and $u_k/|u_k|$ is a vector of norm 1, orthogonal to $v_1, \ldots, v_{k-1}$ such that $\frac{1}{|u_k|}u_k^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_k > v_k^{\mathbf{T}}\hat{F}^{\mathbf{T}}\hat{G}y_k$. This contradicts the fact that v_k is the top k-th left singular vector of $\hat{F}^{\mathbf{T}}\hat{G}$. The Lemma follows.

□

Now we show that the covariance matrix constructed out of the perfect sample matrices has rank at most T.

Lemma 8 *The covariance matrix $\hat{F}^{\mathbf{T}}\hat{G}$ has rank at most T.*

PROOF: For each distribution i, define $y_i^{\mathcal{F}}$ as an $n/2$ dimensional vector, whose f-th element is $\sqrt{w_i}\mathbf{P}_{\bar{\mathcal{K}}}(\mu_i^f - \bar{\mu}^f)$, *i.e.,* the i-th element of z_f. Similarly, for each distribution i, we define $y_i^{\mathcal{G}}$ as an $n/2$ dimensional vector, whose g-th element is $\sqrt{w_i}\mathbf{P}_{\bar{\mathcal{K}}}(\mu_i^g - \bar{\mu}^g)$, *i.e.,* the i-th element of z_g. We observe that the value of $\hat{F}^{\mathbf{T}}\hat{G}$ equals $\sum_{i=1}^{T} y_i^{\mathcal{F}} \cdot (y_i^{\mathcal{G}})^{\mathbf{T}}$. As each outer product of the sum is a rank 1 matrix, the sum, *i.e.,* the covariance matrix, has rank at most T. □

Finally, we show that if the spread of every vector in $\mathcal{C}$ is high, then with high probability over the splitting of coordinates in Step 1 of Algorithm CORRELATION-SUBSPACE, the maximum directional variances of any distribution D_i in $\mathcal{C}_{\mathcal{F}}$ and $\mathcal{C}_{\mathcal{G}}$

are high. This means that there is enough information in both $\mathcal{F}$-space and $\mathcal{G}$-space for correctly clustering the distributions through distance concentration.

Lemma 9 *If every vector $v \in \mathcal{C}$ has spread at least $32T \log \frac{\sigma}{\sigma_*}$, then, with constant probability over the splitting of coordinates in Step 1 of Algorithm* Correlation-Subspace*, the maximum variance along any direction in $\mathcal{C}_\mathcal{F}$ or $\mathcal{C}_\mathcal{G}$ is at most $5\sigma_*^2$.*

Proof: Let v and v' be two unit vectors in $\mathcal{C}$, and let $v_\mathcal{F}$ (resp. $v'_\mathcal{F}$) and $v_\mathcal{G}$ (resp. $v'_\mathcal{G}$ denote the normalized projections of v (resp. v') on $\mathcal{F}$-space and $\mathcal{G}$-space respectively. If $||v_\mathcal{F} - v'_\mathcal{F}|| < \frac{\sigma_*}{\sigma}$, then, the directional variance of any D_i in the mixture along $v'_\mathcal{F}$ can be written as:

$$\begin{aligned}
\mathbf{E}[\langle v'_\mathcal{F}, x - \mathbf{E}[x]\rangle^2] &= \mathbf{E}[\langle v_\mathcal{F}, x - \mathbf{E}[x]\rangle^2] + \mathbf{E}[\langle v'_\mathcal{F} - v_\mathcal{F}, x - \mathbf{E}[x]\rangle^2] \\
&\quad +2\mathbf{E}[\langle v_\mathcal{F}, x - \mathbf{E}[x]\rangle]\mathbf{E}[\langle v'_\mathcal{F} - v_\mathcal{F}, x - \mathbf{E}[x]\rangle] \\
&\leq \mathbf{E}[\langle v_\mathcal{F}, x - \mathbf{E}[x]\rangle^2] + ||v_\mathcal{F} - v'_\mathcal{F}||^2\sigma^2
\end{aligned}$$

Thus, the directional variance of any distribution in the mixture along v' is at most the directional variance along v, plus an additional σ_*^2. Therefore, to show this lemma, we need to show that if v is any vector on a $\frac{\sigma_*}{\sigma}$-cover of $\mathcal{C}$, then with high probability over the splitting of coordinates in Step 1 of Algorithm Correlation-Subspace, the directional variances of any D_i in the mixture along $v_\mathcal{F}$ and $v_\mathcal{G}$ are at most $4\sigma_*^2$.

We show this in two steps. First we show that for any v in a $\frac{\sigma_*}{\sigma}$-cover of $\mathcal{C}$, $\frac{1}{4} \leq \sum_{f\in\mathcal{F}}(v^f)^2 \leq \frac{3}{4}$. Then, we show that this condition means that for this vector v, the maximum directional variances along $v_\mathcal{F}$ and $v_\mathcal{G}$ are at most $4\sigma_*^2$.

Let v be any fixed unit vector in $\mathcal{C}$. We first show that with probability $1 - \left(\frac{\sigma_*}{\sigma}\right)^{2T}$ over the splitting of coordinates in Step 1 of Algorithm Correlation-Subspace, $\frac{1}{4} \leq \sum_{f\in\mathcal{F}}(v^f)^2 \leq \frac{3}{4}$. To show this bound, we apply the Method of Bounded Difference (Theorem 20 in the Appendix). Since we split the coordinates into $\mathcal{F}$ and $\mathcal{G}$ uniformly at random, $\mathbf{E}[\sum_{f\in\mathcal{F}}(v^f)^2] = \frac{1}{2}$. Let γ_f be the change in $\sum_{f\in\mathcal{F}}(v^f)^2$ when the inclusion or exclusion of coordinate f in the set $\mathcal{F}$ changes. Then, $\gamma_f = (v^f)^2$ and $\gamma = \sum_f \gamma_f^2$.

Since the spread of vector v is at least $32T \log \frac{\sigma}{\sigma_*}$, $\gamma = \sum_f (v^f)^4 \leq \frac{1}{32T \log \frac{\sigma}{\sigma_*}}$, and from the Method of Bounded Differences,

$$\Pr[|\sum_{f \in \mathcal{F}} (v^f)^2 - \mathbf{E}[\sum_{f \in \mathcal{F}} (v^f)^2]| > \frac{1}{4}] \leq e^{-1/32\gamma} \leq \left(\frac{\sigma_*}{\sigma}\right)^{2T}$$

By taking an union bound over all v on a $\frac{\sigma *}{\sigma}$-cover of $\mathcal{C}$, we deduce that for any such v, $\frac{1}{4} \leq \sum_{f \in \mathcal{F}} (v^f)^2 \leq \frac{3}{4}$.

Since the maximum directional variance of any distribution D_i in the mixture in $\mathcal{C}$ is at most σ_*^2,

$$\sum_f (v^f)^2 (\sigma_i^f)^2 \leq \sigma_*^2$$

Therefore the maximum variance along $v_{\mathcal{F}}$ as well as $v_{\mathcal{G}}$ can be computed as:

$$\frac{1}{||v_{\mathcal{F}}||^2} \sum_{f \in \mathcal{F}} (v^f)^2 (\sigma_i^f)^2 \leq \frac{1}{||v_{\mathcal{F}}||^2} \sum_f (v^f)^2 (\sigma_i^f)^2 \leq 4\sigma_*^2$$

The lemma follows. □

3.4.2 Working with Real Samples

In this section, we show that given sufficient samples, the properties of the matrix $F^{\mathbf{T}}G$, where F and G are generated by sampling in Step 2 of Algorithm Correlation-Cluster are very close to the properties of the matrix $\hat{F}^{\mathbf{T}}\hat{G}$. The central lemma of this section is Lemma 10, which shows that, if there are sufficiently many samples, for any set of $2m$ vectors, $\{v_1, \ldots, v_m\}$ and $\{y_1, \ldots, y_m\}$, $\sum_k v_k^{\mathbf{T}} F^{\mathbf{T}} G y_k$ and $\sum_k v_k^{\mathbf{T}} \hat{F}^{\mathbf{T}} \hat{G} y_k$ are very close. This lemma is then used to prove Lemmas 11 and 12. Lemma 11 shows that the top few singular vectors of $F^{\mathbf{T}}G$ output by Algorithm Correlation-Subspace have very low projection on $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$ or $\mathbf{P}_{\mathcal{G}}(\bar{\mathcal{C}}_{\mathcal{G}})$. Lemma 12 shows that the rank of the matrix $F^{\mathbf{T}}G$ is almost T, in the sense that the $T+1$-th singular value of this matrix is very low.

Lemma 10 *Let $U = \{u_1, \ldots, u_m\}$ and $Y = \{y_1, \ldots, y_m\}$ be any two sets of orthonormal vectors, and let F and G be the matrices generated by sampling in Step 2 of the algorithm. If the number of samples $|S|$ is greater than $\Omega(\frac{m^3 n^2 \log n \log(\sigma_{\max}/\delta)}{\delta^2})$ (for Binary*

Product Distributions), and $\Omega\left(\max\left(\frac{\sigma^4 m^4 n^2 \log^2 n \log^2(\sigma_{\max}/\delta)}{\delta^2}, \frac{\sigma^2\sigma_{\max}^2 m^3 n \log n \log(\sigma_{\max}/\delta)}{\delta^2}\right)\right)$ *(for axis-aligned Gaussians), then, with probability at least* $1 - 1/n$, $|\sum_k u_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| \leq \delta$.

PROOF: Let u, y and u', y' be two pairs of unit vectors such that $||u - u'|| < \frac{\delta}{8\sigma_{\max}}$, and $||y - y'|| < \frac{\delta}{8\sigma_{\max}}$. Then, if $\delta < \frac{\sigma_{\max}}{8}$,

$$\begin{aligned} |u^{\mathbf{T}}F^{\mathbf{T}}Gy - u'^{\mathbf{T}}F^{\mathbf{T}}Gy'| &\leq (u-u')^{\mathbf{T}}F^{\mathbf{T}}Gy + u'^{\mathbf{T}}F^{\mathbf{T}}G(y-y') \\ &\leq ||u-u'||\sigma_{\max} + ||y-y'||\sigma_{\max} < \delta/2 \end{aligned}$$

The second line follows as $\sigma_{\max} = \max_{u,y} u^{\mathbf{T}}F^{\mathbf{T}}Gy$. It is therefore sufficient to show that the event $|\sum_k u_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| \leq \delta$ holds for all sets of m unit vectors U and Y on a $\frac{\delta}{8\sigma_{\max}}$-cover of $\mathbf{R}^n$. We show this by showing that if $|S|$ is large enough, the event $|\sum_k u_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| \leq \delta/2$ occurs with high probability for all U and Y where U and Y contain vectors from a $(\frac{\delta}{8\sigma_{\max}})$-cover of $\mathbf{R}^n$.

Let us consider a fixed set of vectors U and Y. We can write

$$\sum_k u_k^{\mathbf{T}}F^{\mathbf{T}}Gy_k = \frac{1}{|S|}\sum_k \sum_{x\in S} \langle u_k, \mathbf{P}_{\mathcal{F}}(x)\rangle \cdot \langle y_k, \mathbf{P}_{\mathcal{G}}(x)\rangle$$

We now apply concentration inequalities to bound the deviation of $\sum_k u_k^{\mathbf{T}}F^{\mathbf{T}}Gy_k$ from its mean. For Binary Product Distributions, we apply the Method of Bounded Differences (Theorem 20 in the Appendix) to evaluate this expression. Let $\gamma_{x,f}$ be the maximum change in $\sum_k u_k^{\mathbf{T}}F^{\mathbf{T}}Gy_k$ when we change coordinate f of sample point x. Then,

$$\gamma_{x,f}^2 = \frac{1}{|S|^2}(\sum_k u_k^f \langle y_k, \mathbf{P}_{\mathcal{G}}(x)\rangle)^2 \leq \frac{n}{|S|^2}(\sum_k u_k^f)^2$$

The second step follows because $\langle y_k, \mathbf{P}_{\mathcal{G}}(x)\rangle \leq \sqrt{n}$, for all x, as y_k is a unit vector. And, $\gamma = \sum_{x,f}\gamma_{x,f}^2 \leq \frac{n}{|S|^2}\cdot\sum_f(\sum_k u_k^f)^2 \cdot |S| \leq \frac{nm^2}{|S|}$. This follows because each u_k is a unit vector, and hence $\sum_{k=1}^m u_k$ has norm at most m. Now we can apply the Method of Bounded Differences to conclude that

$$\Pr[|\sum_k u_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| > t\sqrt{\frac{m^2 n}{|S|}}] \leq e^{-t^2/2}$$

Plugging in $t = \sqrt{8mn\log(\frac{\sigma_{\max}}{\delta})\log n}$, we get

$$\Pr[|\sum_k u_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| > \sqrt{\frac{8m^3n^2\log(\frac{\sigma_{\max}}{\delta})\log n}{|S|}}] \leq \frac{1}{n}\cdot\left(\frac{\delta}{8\sigma_{\max}}\right)^{2mn}$$

The expression on the left hand side is at most δ when $|S| > \Omega\left(\frac{m^3n^2\log(\sigma_{\max}/\delta)\log n}{\delta^2}\right)$.

For axis-aligned Gaussians, we use a similar proof involving the Gaussian Concentration of Measure Theorem (Theorem 24 in the Appendix).

If a sample x is generated by distribution D_i,

$$\langle y_k, \mathbf{P}_{\mathcal{G}}(x)\rangle = \langle y_k, \mathbf{P}_{\mathcal{G}}(\mu_i)\rangle + \langle y_k, \mathbf{P}_{\mathcal{G}}(x-\mu_i)\rangle$$

Since y_k is a unit vector, and the maximum directional variance of any D_i is at most σ^2, $\langle y_k, \mathbf{P}_{\mathcal{G}}(x-\mu_i)\rangle$ is distributed as a Gaussian with mean 0 and standard deviation at most σ. Therefore for each $x \in S$,

$$\Pr[|\langle y_k, \mathbf{P}_{\mathcal{G}}(x-\mu_i)\rangle| > \sigma\sqrt{4mn\log(\sigma_{\max}/\delta)\log n}] \leq \left(\frac{\delta}{8\sigma_{\max}}\right)^{2mn\log n}$$

and the probability that this happens for all $x \in S$ is at most $\frac{1}{n}\cdot\left(\frac{\delta}{8\sigma_{\max}}\right)^{2mn}$, provided $\delta < \sigma_{\max}$ and $|S|$ is polynomial in n. Thus, for any $x \in S$ generated from distribution D_i, and any k, $\langle y_k, \mathbf{P}_{\mathcal{G}}(x)\rangle^2 \leq 2\langle y_k, \mu_i\rangle^2 + 8\sigma^2mn\log(\sigma_{\max}/\delta)\log n$.

Therefore, for a specific $x \in S$, as $\mathbf{P}_{\mathcal{F}}(x)$ and $\mathbf{P}_{\mathcal{G}}(x)$ are independently distributed, except with very low probability, the distribution of $\langle u_k, \mathbf{P}_{\mathcal{F}}(x)\rangle\langle y_k, \mathbf{P}_{\mathcal{G}}(x)\rangle$ is dominated by the distribution of $\sqrt{2\langle y_k, \mu_i\rangle^2 + 8\sigma^2mn\log(\sigma_{\max}/\delta)\log n}\cdot\langle u_k, \mathbf{P}_{\mathcal{F}}(x)\rangle$. Note that $\langle u_k, \mathbf{P}_{\mathcal{F}}(x)\rangle$ is distributed as a Gaussian with mean $\langle u_k, \mu_i\rangle$ and variance at most σ^2.

Let γ_x be the derivative of $\frac{1}{|S|}\sum_{x\in S}\sum_k \sqrt{2\langle y_k, \mu_i\rangle^2 + 8\sigma^2mn\log(\sigma_{\max}/\delta)\log n}\cdot\langle u_k, \mathbf{P}_{\mathcal{F}}(x)\rangle$ with respect to the value of sample x in S. Then, the gradient γ of this expression is at most:

$$\begin{aligned}
\gamma^2 = \sum_x \gamma_x^2 &\leq \frac{1}{|S|^2}\sum_{x\in S} 2\sigma^2m^2(\langle y_k, \mu_i\rangle^2 + 4\sigma^2mn\log(\sigma_{\max}/\delta)\log n) \\
&\leq \frac{1}{|S|^2}\cdot(8\sigma^4m^3n\log(8\sigma_{\max}/\delta)\log n + m^2\sigma^2\sigma_{\max}^2)|S| \\
&\leq \frac{m^2\sigma^2}{|S|}\cdot(8\sigma^2mn\log(8\sigma_{\max}/\delta)\log n + 2\sigma_{\max}^2)
\end{aligned}$$

Applying the Gaussian Concentration of Measure Theorem,

$$\begin{aligned}\Pr[|\sum_k u_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| &> t\sqrt{\frac{m^2\sigma^2}{|S|}\cdot(2\sigma^2 mn\log(8\sigma_{\max}/\delta)\log n + \sigma_{\max}^2)}] \\ &\leq e^{-t^2/2}\end{aligned}$$

Plugging in $t = \sqrt{2mn\log(\sigma_{\max}/\delta)\log n}$, we see that the quantity on the left-hand side is at most δ when

$$|S| > \Omega\left(\max\left(\frac{\sigma^4 m^4 n^2 \log^2 n \log^2(\sigma_{\max}/\delta)}{\delta^2}, \frac{\sigma^2\sigma_{\max}^2 m^3 n \log n \log(\sigma_{\max}/\delta)}{\delta^2}\right)\right)$$

In both cases, the lemma now follows by applying a Union Bound over $(\frac{8\sigma_{\max}}{\delta})^{2mn}$ choices of U and Y. $\square$

Lemma 11 *Let F and G be the matrices generated by sampling in Step 2 of the algorithm, and let $v_1, \ldots, v_m$ be the vectors output by the algorithm in Step 4. If the number of samples $|S|$ is greater than $\Omega(\frac{m^3 n^2 \log n(\log\Lambda + \log\frac{1}{\epsilon})}{\tau^2\epsilon^4})$ (for Binary Product Distributions), and $\max\left(\frac{\sigma^4 m^4 n^2 \log^2 n \log^2(\Lambda/\epsilon)}{\tau^2\epsilon^4}, \frac{\sigma^2\sigma_{\max}^2 m^3 n\log n\log(\Lambda/\epsilon)}{\tau^2\epsilon^4}\right)$ (for axis-aligned Gaussians), then, for each k, and any x in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$, $\langle v_k, x\rangle \leq \epsilon$.*

PROOF: Let $\{\bar{v}_1, \ldots, \bar{v}_m\}$ and $\{\bar{y}_1, \ldots, \bar{y}_m\}$ be the top m left and right singular vectors, and let $\{\lambda_1, \ldots, \lambda_m\}$ be the top m singular values of $\mathbf{E}[F^{\mathbf{T}}G]$. Let $\{y_1, \ldots, y_m\}$ be the top right singular vectors of $F^{\mathbf{T}}G$. If, for any k and any x in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}}_{\mathcal{F}})$, $\langle v_k, x\rangle \leq \epsilon$, then, for any set of orthonormal vectors $\{y_1, \ldots, y_m\}$,

$$\begin{aligned}\sum_{k=1}^m v_k^{\mathbf{T}}\mathbf{E}[F^{\mathbf{T}}G]y_k &= \sum_{k\neq l} v_k^{\mathbf{T}}\mathbf{E}[F^{\mathbf{T}}G]y_k + (v_l - \langle v_l, x\rangle x)^{\mathbf{T}}\mathbf{E}[F^{\mathbf{T}}G]y_k \\ &\leq \sum_{k=1}^m \lambda_k - (1-\sqrt{1-\epsilon^2})\lambda_m\end{aligned}$$

If ϵ is smaller than $1/4$, the left-hand side can be bounded as

$$\sum_{k=1}^m v_k^{\mathbf{T}}\mathbf{E}[F^{\mathbf{T}}G]y_k \leq \sum_{k=1}^m \lambda_k - \frac{\epsilon^2}{4}\lambda_m$$

Next we apply Lemma 10, which states that if the number of samples $|S|$ is large enough,

$$|\sum_{k=1}^{m} v_k^{\mathbf{T}}(F^{\mathbf{T}}G - \mathbf{E}[F^{\mathbf{T}}G])y_k| \leq \delta$$

Therefore, if $\delta \leq \epsilon^2\lambda_m/16$,

$$\sum_{k=1}^{m} v_k^{\mathbf{T}} F^{\mathbf{T}} G y_k \leq \sum_{k=1}^{m} \lambda_k - \frac{3\epsilon^2}{16}\lambda_m$$

On the other hand, using Lemma 10,

$$\sum_{k=1}^{m} \bar{v}_k^{\mathbf{T}} F^{\mathbf{T}} G \bar{y}_k \geq \sum_{k=1}^{m} \lambda_k - \frac{\epsilon^2}{16}\lambda_m$$

This contradicts the fact that $\{v_1, \ldots, v_m\}$ and $\{y_1, \ldots, y_m\}$ are the top singular vectors of $F^{\mathbf{T}}G$, and hence the lemma follows. □

Lemma 12 *Let F and G be the matrices generated by sampling in Step 2 of Algorithm* CORRELATION-SUBSPACE. *If the number of samples $|S|$ is greater than $\Omega\left(\frac{T^3n^2\log n\log\Lambda}{\tau^2}\right)$ (for binary product distributions) and $\Omega\left(\max\left(\frac{\sigma^4T^4n^2\log^2\log\Lambda}{\tau^2}, \frac{\sigma_{\max}^2\sigma^2T^3n\log n\log\Lambda}{\tau^2}\right)\right)$ for axis-aligned Gaussians, then, λ_{T+1}, the $T+1$-th singular value of the matrix $F^{\mathbf{T}}G$ is at most $\tau/8$.*

PROOF: For any k, let λ_k denote the k-th top singular value of the matrix $F^{\mathbf{T}}G$ and let $\hat{\lambda}_k$ denote the k-th top singular value of the matrix $\mathbf{E}[F^{\mathbf{T}}G]$. From Lemma 8, $\mathbf{E}[F^{\mathbf{T}}G] = \hat{F}^{\mathbf{T}}\hat{G}$ has rank at most T. Thus $\hat{\lambda}_{T+1} = 0$.

Let $\{v_1, \ldots, v_m\}$ (resp. $\{\hat{v}_1, \ldots, \hat{v}_m\}$) and $\{y_1, \ldots, y_m\}$ (resp. $\{\hat{y}_1, \ldots, \hat{y}_m\}$) denote the top m left and right singular vectors of $F^{\mathbf{T}}G$ (resp. $\mathbf{E}[F^{\mathbf{T}}G]$). From Lemma 10, if $|S|$ is greater than $\Omega\left(\frac{T^3n^2\log n\log\Lambda}{\tau^2}\right)$ (for binary product distributions) and $|S|$ is greater than $\Omega\left(\max\left(\frac{\sigma^4T^4n^2\log^2\log\Lambda}{\tau^2}, \frac{\sigma_{\max}^2\sigma^2T^3n\log n\log\Lambda}{\tau^2}\right)\right)$ (for axis-aligned Gaussians), then,

$$\lambda_1 + \ldots + \lambda_T \geq \sum_{k=1}^{T} \hat{v}_k^{\mathbf{T}} F^{\mathbf{T}} G \hat{y}_k \geq \hat{\lambda}_1 + \ldots \hat{\lambda}_T - \frac{\tau}{16}$$

Moreover, from Lemma 10,

$$\hat{\lambda}_1 + \ldots + \hat{\lambda}_{T+1} \geq \sum_{k=1}^{T} v_k^{\mathbf{T}} \mathbf{E}[F^{\mathbf{T}} G] y_k \geq \lambda_1 + \ldots + \lambda_{T+1} - \frac{\tau}{16}$$

Combining the above two equations, and the fact that $\hat{\lambda}_{T+1} = 0$, $\lambda_{T+1} \leq 2 \cdot \frac{\tau}{16} \leq \frac{\tau}{8}$. □

3.4.3 The Combined Analysis

In this section, we combine the lemmas proved in Sections 3.4.1 and 3.4.2 to prove Theorem 1.

We begin with a lemma which shows that if every vector in $\mathcal{C}$ has spread $32T \log \frac{\sigma}{\sigma_*}$, then the maximum directional variance in $\mathcal{K}$, the space output by Algorithm Correlation-Subspace, is at most $11\sigma_*^2$.

Lemma 13 *Let $\mathcal{K}$ be the subspace output by the algorithm, and let v be any vector in $\mathcal{K}$. If every vector in $\mathcal{C}$ has spread $32T \log \frac{\sigma}{\sigma_*}$, and the number of samples $|S|$ is greater than $\Omega\left(\max\left(\frac{\sigma^6 T^4 n^2 \log^2 \log \Lambda}{\tau^2 \sigma_*^4}, \frac{\sigma_{\max}^2 \sigma^4 T^3 n \log n \log \Lambda}{\tau^2 \sigma_*^4}\right)\right)$ then for any distribution i the maximum variance of D_i along v is at most $11\sigma_*^2$.*

Proof: We first show that for any v_k(or y_k) in the set $\mathcal{K}$ output by Algorithm Correlation-Subspace, and for any distribution D_i in the mixture, the maximum variance of D_i along v_k (or y_k) is at most $11\sigma_*^2$.

Let $v_k = u_k + x_k$ where u is in $\mathcal{C}_{\mathcal{F}}$ and x is in $\mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}_{\mathcal{F}}})$. From Lemma 11, we deduce that $||x_k|| \leq \frac{\sigma_*}{4\sigma}$. Let M_i be the matrix in which each row is a sample from the distribution D_i. Then the variance of distribution D_i along the direction v is the square of the norm of the vector $(M_i - \mathbf{E}[M_i])v$. This norm can be written as:

$$\begin{aligned}
||(M_i - \mathbf{E}[M_i])v_k||^2 &= ||(M_i - \mathbf{E}[M_i])u_k||^2 + ||(M_i - \mathbf{E}[M_i])x_k||^2 \\
&\quad +2\langle (M_i - \mathbf{E}[M_i])u_k, (M_i - \mathbf{E}[M_i])x_k \rangle \\
&\leq 5\sigma_*^2 ||u_k||^2 + \sigma^2 ||x_k||^2 + 4\sigma\sigma_* ||x_k|| ||u_k|| \\
&\leq 2(5\sigma_*^2 + ||x_k||^2 \sigma^2) \leq 11\sigma_*^2
\end{aligned}$$

The third line follows from Lemma 9, and the last step follows from the bound on $||x_k||$ from Lemma 11.

Now, let $v = \sum_l \alpha_l v_l + \bar{\alpha}_l y_l$ be any unit vector in the space $\mathcal{K}$. Then, $\sum_l \alpha_l^2 + \bar{\alpha}_l^2 = 1$, and for any i, the variance of D_i along v is $\sum_l \alpha_l^2 ||(M_i - \mathbf{E}[M_i])v_l||^2 + \bar{\alpha}_l^2 ||(M_i - \mathbf{E}[M_i])y_l||^2 \leq 11\sigma_*^2$. The lemma follows. $\square$

We are now ready to prove Theorem 1.

PROOF:(Of Theorem 1)

Suppose $\mathcal{K} = \mathcal{K}_L \cup \mathcal{K}_R$, where $\mathcal{K}_L = \{v_1, \ldots, v_m\}$, the top m left singular vectors of $F^{\mathbf{T}}G$ and $\mathcal{K}_R = \{y_1, \ldots, y_m\}$ are the corresponding right singular vectors. We abuse notation and use v_k to denote the vector v_k concatenated with a vector consisting of $n/2$ zeros, and use y_k to denote the vector consisting of $n/2$ zeros concatenated with y_k. Moreover, we use $\mathcal{K}$, $\mathcal{K}_L$, and $\mathcal{K}_R$ interchangeably to denote sets of vectors and the subspace spanned by those sets of vectors.

We show that with probability at least $1 - \frac{1}{T}$ over the splitting step, there exists no vector $v \in \mathcal{C}_{\mathcal{F} \cup \mathcal{G}}$ such that (1) v is orthogonal to the space spanned by the vectors $\mathcal{K}$ and (2) there exists some pair of centers i and j such that $\bar{d}_v(\mu_i, \mu_j) > 49Tc_{ij}^2 \log \Lambda$. For contradiction, suppose there exists such a vector v.

Then, if $v_{\mathcal{F}}$ and $v_{\mathcal{G}}$ denote the normalized projections of v onto $\mathcal{F}$-space and $\mathcal{G}$-space respectively, from Lemma 2, $v_{\mathcal{F}}^{\mathbf{T}} \hat{F}^{\mathbf{T}} G v_{\mathcal{G}} \geq \tau$ with probability at least $1 - \frac{1}{T}$ over the splitting step. From Lemma 10, if the number of samples $|S|$ is greater than $\Omega\left(\frac{T^3 n^2 \log n \log \Lambda}{\tau^2}\right)$ for binary product distributions, and if $|S|$ is greater than $\Omega\left(\max\left(\frac{\sigma^4 n^2 \log^2 \log \Lambda}{\tau^2}, \frac{\sigma^2 \sigma_{\max}^2 n \log n \log \Lambda}{\tau^2}\right)\right)$ for axis-aligned Gaussians, $v_{\mathcal{F}}^{\mathbf{T}} F^{\mathbf{T}} G v_{\mathcal{G}} \geq \frac{\tau}{2}$ with at least constant probability. Since v is orthogonal to the space spanned by $\mathcal{K}$, $v_{\mathcal{F}}$ is orthogonal to $\mathcal{K}_L$ and $v_{\mathcal{G}}$ is orthogonal to $\mathcal{K}_R$. As λ_{m+1} is the maximum value of $x^{\mathbf{T}} F^{\mathbf{T}} G y$ over all vectors x orthogonal to $\mathcal{K}_L$ and y orthogonal to $\mathcal{K}_R$, $\lambda_{m+1} \geq \frac{\tau}{2}$, which is a contradiction.

Moreover, from Lemma 12, $\lambda_{T+1} < \frac{\tau}{8}$, and hence $m \leq T$.

Let us construct an orthonormal series of vectors $v_1, \ldots, v_m, \ldots$ which are *almost*

in $\mathcal{C}_{\mathcal{F}}$ as follows. $v_1, \ldots, v_m$ are the vectors output by Algorithm CORRELATION-SUBSPACE. We inductively define v_l as follows. Suppose for each k, $v_k = u_k + x_k$, where $u_k \in \mathcal{C}_{\mathcal{F}}$ and $x_k \in \mathbf{P}_{\mathcal{F}}(\bar{\mathcal{C}_{\mathcal{F}}})$. Let u_l be a unit vector in $\mathcal{C}_{\mathcal{F}}$ which is perpendicular to $u_1, \ldots, u_{l-1}$. Then, $v_l = u_l$. By definition, this vector is orthogonal to $u_1, \ldots, u_{l-1}$. In addition, for any $k \neq l$,

$$\langle v_l, v_k \rangle = \langle u_l, u_k \rangle + \langle u_l, x_k \rangle = 0$$

and v_l is also orthogonal to $v_1, \ldots, v_{l-1}$. Moreover, if $\epsilon < \frac{1}{100T}$, $u_1, \ldots, u_m$ are linearly independent, and we can always find $\dim(\mathcal{C}_{\mathcal{F}})$ such vectors. Similarly, we construct a set of vectors $y_1, y_2, \ldots$. Let us call the combined set of vectors $\mathcal{C}^*$.

We now show that if there are sufficient samples, $d_{\bar{C^*}}(\mu_i, \mu_j) \leq c_{ij}^2$. Note that for any unit vector v^* in $\mathcal{C}^*$, and any unit $x \in \bar{C}_{\mathcal{F}\cup\mathcal{G}}$, $\langle v, x \rangle \leq m\epsilon$. Also, note that for any u_k and u_l, $k \neq l$, $|\langle u_k, u_l \rangle| \leq \epsilon^2$, and $||u_k||^2 \geq 1 - \epsilon^2$. Let $v = \sum_k \alpha_k u_k$ be any unit vector in $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$. Then,

$$1 = ||v||^2 = \sum_{k,k'} \alpha_k \alpha_{k'} \langle u_k, u_{k'} \rangle \geq \sum_k \alpha_k^2 ||u_k||^2 - \Omega(T^2\epsilon^2)$$

The projection of v on C^* can be written as:

$$\begin{aligned}
\sum_k \langle v, v_k \rangle^2 &= \sum_k \langle v, u_k \rangle^2 \\
&= \sum_k \sum_l \alpha_l^2 \langle u_k, u_l \rangle^2 + 2 \sum_{l,l'} \alpha_l \alpha_{l'} \langle u_k, u_l \rangle \langle u_k, u_{l'} \rangle \\
&\geq \sum_k \alpha_k^2 ||u_k||^4 - T^3\epsilon^4 \geq 1 - \Omega(T^2\epsilon^2)
\end{aligned}$$

The last step follows because for each k, $||u_k||^2 \geq 1 - \epsilon^2$. If the number of samples $|S|$ is greater than $\Omega(\frac{m^3 n^2 \log n (\log \Lambda + \log 100T)}{\tau^2 T^4})$ (for Binary Product Distributions), and $\max\left(\frac{\sigma^4 m^4 n^2 \log^2 n \log^2(100T\Lambda)}{\tau^2 T^4}, \frac{\sigma_{\max}^2 \sigma^2 m^3 n \log n \log(100T\Lambda)}{\tau^2 T^4}\right)$ (for axis-aligned Gaussians), , then, $\epsilon < 1/100T$. Therefore, $d_{\bar{C^*}}(\mu_i, \mu_j) \leq \frac{1}{100} d(\mu_i, \mu_j)$.

For any i and j, $d(\mu_i, \mu_j) = d_{\mathcal{K}}(\mu_i, \mu_j) + d_{\mathcal{C}^* \setminus \mathcal{K}}(\mu_i, \mu_j) + d_{\bar{\mathcal{C}^*}}(\mu_i, \mu_j)$. Since vectors $v_{m+1}, \ldots$ and $y_{m+1}, \ldots$, all belong to $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}$ (as well as $\mathcal{C}^* \setminus \mathcal{K}$, there exists no $v \in \mathcal{C}^* \setminus \mathcal{K}$

with the Conditions (1) and (2) in the previous paragraph, and $\bar{d}_{\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}\setminus\mathcal{K}}(\mu_i,\mu_j) \leq 49Tc_{ij}^2\log\Lambda$. That is, the actual distance between μ_i and μ_j in $\mathcal{C}_{\mathcal{F}\cup\mathcal{G}}\setminus\mathcal{K}$ (as well as $\mathcal{C}^*\setminus\mathcal{K}$) is at most the contribution to $d(\mu_i,\mu_j)$ from the top $49Tc_{ij}^2\log\Lambda$ coordinates, and the contribution to $d(\mu_i,\mu_j)$ from $\mathcal{K}$ and $\bar{C^*}$ is at least the contribution from the rest of the coordinates. Since $d_{\bar{C^*}}(\mu_i,\mu_j) \leq \frac{1}{100}d(\mu_i,\mu_j)$, the distance between μ_i and μ_j in $\mathcal{K}$ is at least $\frac{99}{100}\bar{d}(\mu_i,\mu_j) - 49T\log\Lambda c_{ij}^2)$. The first part of the theorem follows.

The second part of the theorem follows directly from Lemma 13. □

3.4.4 Distance Concentration

In this section, we show how to prove Theorem 2 by combining Theorem 1 and distance-concentration methods. We begin with the following distance-concentration lemmas of [1] and [19], which we prove for the sake of completeness.

Lemma 14 *Let $\mathcal{K}$ be a d-dimensional subspace of $\mathbf{R}^n$, and x be a point drawn from axis-aligned Gaussian. Let $\sigma_{\mathcal{K}}^2$ be the maximum variance of x along any direction in the subspace $\mathcal{K}$. Then,*

$$\Pr[||\mathbf{P}_{\mathcal{K}}(x-\mathbf{E}[x])|| > \sigma_{\mathcal{K}}\sqrt{2d\log(d/\delta)}] \leq \delta$$

PROOF: Let $v_1,\ldots,v_d$ be an orthonormal basis of $\mathcal{K}$. Since the projection of a Gaussian is a Gaussian, the projection of the distribution of x along any v_k is a Gaussian with variance at most $\sigma_{\mathcal{K}}^2$. By the properties of the normal distribution,

$$\Pr[|\langle x-\mathbf{E}[x], v_k\rangle| > \sigma_{\mathcal{K}}\sqrt{2\log(d/\delta)}] \leq \frac{\delta}{d}$$

Since $||\mathbf{P}_{\mathcal{K}}(x-\mathbf{E}[x])||^2 = \sum_k\langle x-\mathbf{E}[x], v_k\rangle^2$, the lemma follows by a Union Bound over $v_1,\ldots,v_d$. □

Lemma 15 *Let $\mathcal{K}$ be a d-dimensional subspace of $\mathbf{R}^n$ and x be a point drawn from a binary product distribution. Then,* $\Pr[||\mathbf{P}_{\mathcal{K}}(x-\mathbf{E}[x])|| > \sqrt{2d\log(d/\delta)}] \leq \delta$.

PROOF: Let $v_1, \ldots, v_d$ be an orthonormal basis of $\mathcal{K}$. For a fixed v_k, we bound $\langle v_k, x - \mathbf{E}[x] \rangle$, where x is generated by a binary product distribution by applying the Method of Bounded Differences (Theorem 19 in the Appendix). Let γ_f be the change in $\langle v_k, x - \mathbf{E}[x] \rangle$, when the value of coordinate f in x changes. Then, $\gamma_f = v_k^f$, and $\gamma = \sum_f \gamma_f^2 = ||v_k||^2 = 1$. From the Method of Bounded Differences,

$$\Pr[|\langle x - \mathbf{E}[x], v_k \rangle| > \sqrt{2 \log(d/\delta)}] \leq \frac{\delta}{d}$$

Since $||\mathbf{P}_{\mathcal{K}}(x - \mathbf{E}[x])||^2 = \sum_k \langle v_k, x - \mathbf{E}[x] \rangle^2$, the lemma follows by a union bound over $v_1, \ldots, v_d$. $\square$

Now we are ready to prove Theorem 2.

PROOF:(Of Theorem 2) From Theorem 1, if for all i and j, $\bar{d}(\mu_i, \mu_j) \geq 49 T c_{ij}^2 \log \Lambda$, then, for all i and j, with constant probability,

$$d_{\mathcal{K}}(\mu_i, \mu_j) \geq \frac{99}{100} (\bar{d}(\mu_i, \mu_j) - 49 T c_{ij}^2 \log \Lambda)$$

From the separation conditions in Theorem 2, this means that for all i and j, we have that $d_{\mathcal{K}}(\mu_i, \mu_j) \geq 9T(\log \Lambda + \log n)$ for binary product distributions, and $d_{\mathcal{K}}(\mu_i, \mu_j) \geq 9\sigma^2 T(\log \Lambda + \log n)$ for axis-aligned Gaussians.

Applying Lemma 15, for binary product distributions, any two samples from a fixed distribution D_i in the mixture are at a distance of at most $4\sqrt{T(\log T + \log n)}$ in $\mathcal{K}$, with probability $1 - \frac{1}{n}$. On the other hand, two points from different distributions are at distance $5\sqrt{T(\log T + \log n)}$. Therefore, with probability $1 - \frac{1}{n}$, the distance concentration algorithm succeeds.

Similarly, for axis-aligned Gaussians, from Lemma 14, any two samples from a fixed distribution D_i in the mixture are at a distance of at most $4\sigma^2\sqrt{T(\log T + \log n)}$ in $\mathcal{K}$, with probability $1 - \frac{1}{n}$. On the other hand, two points from different distributions are at distance $5\sigma^2\sqrt{T(\log T + \log n)}$. Distance concentration therefore works, and the first part of the theorem follows.

If every vector in $\mathcal{C}$ has spread at least $49T \log \Lambda$, from Theorem 1, the maximum variance of any D_i in $\mathcal{K}$ is at most $11\sigma_*^2$. The Theorem now follows by the same

arguments as above. □

3.5 Discussions

The Spreading Condition. Unlike previous algorithms, our method requires a *spreading condition* – namely, a condition that the centers of any pair of distributions be separated along $\Omega(T \log \Lambda)$ coordinates. This condition is specific to our methods: we show in Section 3.6.2 that unless every pair of centers are separated by $\Omega(T)$ coordinates, there might be no correlation between any pair of coordinates that our algorithm can take advantage of. On the other hand, if such correlation is present, and the distributions in question are product distributions, even if the separation between the centers is too small for classification, our algorithm can find a subspace such that the centers are far apart when projected onto this subspace. Previous techniques, including SVD-based methods fail to do so for general Gaussians and product distributions. ([22], however, still works in this case, but requires the distributions to be exactly spherical.) We also note that for a mixture of binary product distributions, the separation condition implies (a weaker version of) the spreading condition.

Learning with Separation proportional to σ_*. Theorem 2 shows that our algorithm learns mixtures of Gaussians with separation proportional to σ_*, when every vector in the subspace containing the center has high spread. In general, we do not require such a strong constraint on the subspace containing the centers.

To successfully cluster the samples, our algorithm requires that the variance of any distribution in the mixture along the subspaces $\mathcal{C}_\mathcal{F}$ and $\mathcal{C}_\mathcal{G}$ is low. This does not immediately follow from the fact that the variance of any distribution is low in $\mathcal{C}$. For example, suppose the space containing the centers is spanned by the vectors $(0.1, 0.1, 1, 1)$ and $(0.1, 0.1, -1, 1)$, the directional variances of some distribution D_i along the coordinate directions are $(10, 10, 1, 1)$, and $\mathcal{F}$ is the set $\{1, 2\}$. Then, σ_*^2 is about 2.8, whereas, the variance along $\mathcal{C}_\mathcal{F}$ is 10.

However, in general, the maximum directional variance of any distribution in the

mixture along $\mathcal{C}_{\mathcal{F}}$ and $\mathcal{C}_{\mathcal{G}}$ may still be low, even though the condition in Theorem 2 is not satisfied. As an example, if the centers lie in the space spanned by the first T coordinate vectors $e_1, \ldots, e_T$, and the maximum variance of any distribution in the mixture is low in $\mathcal{C}$, then for any division of the coordinates, the maximum variance along $\mathcal{C}_{\mathcal{F}}$ and $\mathcal{C}_{\mathcal{G}}$ are also low.

Finally, in Chapter 5, we provide an algorithm which can learn mixtures of axis-aligned Gaussians with separation proportional to σ_*, without the condition that the maximum directional variance in both $\mathcal{C}_{\mathcal{F}}$ and $\mathcal{C}_{\mathcal{G}}$ is low. However, the algorithm is the current chapter is simple and easy to implement.

3.6 Lower Bounds

In this section, we show that there exists a mixture of axis-aligned Gaussians, such that any algorithm which learns the mixture well enough to classify a $1-\delta$ fraction of the samples correctly, requires a separation of $\Omega(\sigma_* \sqrt{\log(1/\delta)})$.

3.6.1 Information Theoretic Lower Bounds

We demonstrate by an example, that the Bayes-optimal algorithm, which knows the parameters of the distributions comprising the mixture, but not the labels of the individual points, may require a separation of as much as $\Omega(\sigma_* \log(\frac{1}{\delta}))$ to classify $1-\delta$ fraction of the points correctly. Since no algorithm can achieve a better classification guarantee than the Bayes-optimal algorithm, this is an information theoretic lower bound on the separation required to learn mixtures of distributions.

Given the parameters of the distributions comprising a mixture with equal mixing weights, the Bayes-optimal algorithm, assigns a sample x to distribution D_i if the probability density function of D_i evaluated at x, $\varphi_i(x)$ is maximum out of all i in $\{1, \ldots, T\}$. This classification is incorrect with probability $\frac{1-\varphi_i(x)}{\sum_{j=1}^{T} \varphi_j(x)}$. The following lemma shows a lower bound on the separation required by the Bayes-optimal algorithm for correct classification of $1-\delta$ fraction of the points.

Lemma 16 *Let $\mathcal{D} = \{D_1, D_2\}$ be a mixture of axis-aligned Gaussians where D_1 has center $(0, \ldots, 0)$ and standard deviation $(\sigma_*, \sigma, \ldots, \sigma)$ and D_2 has center $(\sigma_*\sqrt{2\log(\frac{1}{\delta})}, 0, \ldots, 0)$ and standard deviation $(\sigma_*, \sigma, \ldots, \sigma)$. Then, the Bayes-optimal algorithm has mis-classification probability at least $\frac{\delta}{\sqrt{2\pi}}$.*

PROOF: For a point x,

$$\begin{aligned}
\varphi_1(x) &= \frac{1}{\sigma_*\sigma^{n-1}(2\pi)^{n/2}} e^{-(x_1^2/2\sigma_*^2 + \sum_{j\neq 1} x_j^2/2\sigma^2)} \\
\varphi_2(x) &= \frac{1}{\sigma_*\sigma^{n-1}(2\pi)^{n/2}} e^{-((x_1 - \sigma_*\sqrt{2\log(\frac{1}{\delta})})^2/2\sigma_*^2 + \sum_{j\neq 1} x_j^2/2\sigma^2)}
\end{aligned}$$

Therefore, the Bayes-optimal algorithm will classify a sample x as D_1 if x_1 is less than $\frac{1}{2}\sigma_*\sqrt{2\log(\frac{1}{\delta})}$ and as D_2 otherwise. The probability of misclassification for the Bayes-optimal algorithm is thus

$$\frac{e^{-\left(x_1 - \sigma_*\sqrt{2\log(\frac{1}{\delta})}\right)^2/2\sigma_*^2}}{e^{-\left(x_1 - \sigma_*\sqrt{2\log(\frac{1}{\delta})}\right)^2/2\sigma_*^2} + e^{-x_1^2/2\sigma_*^2}}$$

, when $x_1 > \frac{1}{2}\sigma_*\sqrt{2\log(\frac{1}{\delta})}$ and

$$\frac{e^{-x_1^2/2\sigma_*^2}}{e^{-\left(x_1 - \sigma_*\sqrt{2\log(\frac{1}{\delta})}\right)^2/2\sigma_*^2} + e^{-x_1^2/2\sigma_*^2}}$$

otherwise. Therefore, integrating over all values of x_1, the total probability of misclassification is at least $\frac{\delta}{\sqrt{2\pi}}$. The lemma follows. □

On the other hand, for any two axis-aligned Gaussians, the Bayes-optimal algorithm will successfully classify at least $1-\delta$ fraction of the samples when the separation is $\Theta(\sigma_*\sqrt{\log(\frac{1}{\delta})})$ regardless of the value of σ. When projected on the line joining the centers, the two distributions become one-dimensional Gaussians with standard deviation at most σ_* and means at the distance of $\Theta(\sigma_*\sqrt{\log(\frac{1}{\delta})})$. Calculations similar to Lemma 16 show that the Bayes-optimal algorithm can correctly classify at least $1 - \delta$ fraction of points from such distributions.

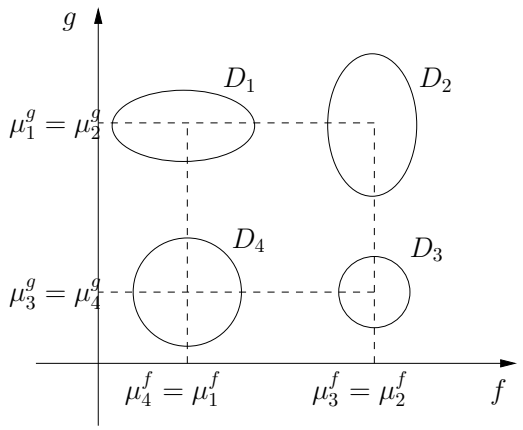

Figure 3.4: An Example where All Covariances are 0

3.6.2 Limitations of Linear Correlations

We demonstrate by an example, that a spread of $\Omega(T)$, is a natural limit for all methods that use linear correlations between coordinates, such as our methods and SVD based methods [22; 17; 1]. This example is based on the Hadamard code, in which a codeword for a k-bit message is 2^k bits long, and includes a parity bit for each subset of the bits of the message. The input distributions are defined as follows. Each of the $T = 2^k$ centers is a codeword for a $k-$bit string appended by a string of length $n - k$ in which each coordinate has value $1/2$. Notice that the last $n - k$ bits are noise. Thus, the centers are separated by $T/2$ coordinates. As there are no linear correlations between any two bits in the Hadamard code, all directions will look the same to our algorithm and any SVD based technique, and thus, the algorithms will fail. Another example in two dimensions, where $T = 4$ is illustrated in Figure 3.4.

We also note that learning binary product distributions with minimum separation 2 and average separation $1+\frac{1}{2}\log T$ would allow one to learn parities of $\log T$ variables with noise.

3.7 Conclusions and Open Problems

In summary, in this chapter, we provide a correlation-based algorithm for learning mixtures of product distributions and axis-aligned Gaussians, which work with lower separation than previous methods. There are two main open questions in this chapter. First, can we provide an efficient algorithm which works with smaller separation? This seems to be information-theoretically possible in some cases.

Problem 1 *Can we provide efficient algorithms for learning mixtures of product distributions and axis-aligned Gaussians, which require a lower separation than* $\Theta(\sigma_* \sqrt{T \log \Lambda})$*?*

A major restriction of our algorithm is that it requires the distributions in the mixture to be product distributions. A second question is whether we can extend it to mixtures of distributions which are not product distributions. A starting point would be extending it to mixtures of general Gaussians, where different Gaussians in the mixture have different axes.

Problem 2 *Can we apply our algorithm to mixtures of distributions which are not product distributions ?*

Chapter 4

The Sample Complexity of Learning Mixtures

4.1 Overview

In this chapter, we study the sample complexity of learning mixtures of binary product distributions. Our study is motivated by applications in population stratification. Recall from Chapter 1 that in such applications, we are given data corresponding to genetic factors from individuals belonging to different populations, and the goal is to cluster the individuals according to source population. The sampling process here involves determining the genetic factors present in an individual, and collecting a large number of samples is expensive. Hence it is very important to have an algorithm which works reliably with a small number of samples.

The state of the art algorithms for learning mixtures of binary product distributions are singular-value decomposition-based approaches, which require $\Omega(\frac{nT}{w_{\min}})$ samples when working with distributions in n-dimensional space. In the special case when the separation is $\Omega(n^{1/4})$, one can also apply distance-based approaches [2], which require a number of samples polylogarithmic in the number of dimensions. Typical population stratification data consists of genetic data on a few thousand factors for a few hundred individuals; the guarantee on the sample requirement of SVD-based algorithms is therefore insufficient for such applications. In addition, in population stratification data, the separation between the centers is fairly high, and therefore we

can use this fact to trade off separation against a lower sample complexity.

In this chapter, we take a step towards designing such an algorithm by providing an algorithm which learns a mixture of two binary product distributions with uniform mixing weights and low sample complexity. Given samples from such a mixture, our algorithm clusters all the samples correctly with high probability, so long as $d(\mu_1, \mu_2)$, the square of the Euclidean distance between the centers of distributions is at least polylogarithmic in s, the number of samples and the following trade-off holds between the separation and the number of samples:

$$s \cdot d^2(\mu_1, \mu_2) \geq a \cdot n \log s \log(ns)$$

for some constant a. As shown in Section 4.4, the requirement on $d(\mu_1, \mu_2)$ is essentially optimal, except for the logarithmic factors. We note that in the worst case for our algorithm, when the separation $d(\mu_1, \mu_2)$ is logarithmic in s, the number of samples required is $\tilde{O}(n)$, which matches the sample complexity bounds of SVD-based approaches.

4.1.1 Notation

We use indices i, j, ... to index over distributions, and f, g, ... to index over coordinates. For a vector $v \in \mathbf{R}^n$, we use $||v||$ to denote the L_2 norm of v. For two vectors x, y in $\mathbf{R}^n$, we use $\langle x, y \rangle$ to denote the dot-product of x and y. We use μ_1 and μ_2 to denote the means of the distributions in the mixture and $d(\mu_1, \mu_2)$ to denote the square of the Euclidean distance between centers μ_1 and μ_2.

Partitions. For the rest of the chapter, we work with partitions of sets of sample points into two components. For disjoint sets of sample points S_1 and S_2, we write (S_1, S_2) to denote the partition of $S_1 \cup S_2$ into S_1 and S_2. We call a partition (S_1, S_2) *uniform* if $|S_1| = |S_2|$. All partitions we work with in this chapter are uniform unless otherwise specified. A correct partition of a set of samples is a partition of the samples according to source distribution.

Imbalance. Let S be a set of sample points generated from a mixture of distributions $\mathcal{D} = \{D_1, D_2\}$ with uniform mixing weights. Given a partition (S_1, S_2) of S, we assume without loss of generality that S_1 has at least as many samples from D_1 as S_2. Then, the imbalance of (S_1, S_2) is the number of samples in S_1 from D_1, divided by the total number of samples S_1. The imbalance of a partition is a measure of how close it is to the correct partition. For example, the imbalance of the correct partition is 1, and on expectation, the imbalance of a random partition on s samples is $1/2 + \Omega(1/\sqrt{s})$.

4.1.2 A Summary of Our Results

The main contribution of this chapter is an algorithm which successfully learns mixtures of two binary product distributions with uniform mixing weights using a small number of samples. The guarantees provided by our algorithm are summarized in the theorem below.

Theorem 3 *Suppose we are given s samples from a mixture of two product distributions over binary vectors, each with a mixing weight of $1/2$. If the following conditions hold:*

1. $d(\mu_1, \mu_2) > a_1 \log s \log(ns)$ *and*
2. $s \cdot d^2(\mu_1, \mu_2) > a_2 n \log s \log(ns)$, *for some fixed constants a_1 and a_2,*

then, our algorithm outputs a partition of $1 - \frac{2\sqrt{\log s}}{\sqrt{s}}$ fraction of the samples according to source distribution with probability $1 - \frac{1}{s}$ (over the samples) and constant (over the random choices made by the algorithm).

Theorem 3 provides a trade-off between $d(\mu_1, \mu_2)$, the square of the Euclidean distance between the means, and the number of samples required to learn the mixture. As expected, the smaller the separation, the higher is the number of samples required. In the worst case for our algorithm, when the separation between the means

Cluster (S)

1. Partition the set of samples into A and B uniformly at random.
2. Partition the set of coordinates into $F_1, \ldots, F_{\log s}$ uniformly at random.
3. Let $(A_{0,1}, A_{0,2})$ be a random partition of A.
4. For $k = 1, \ldots, \log s$:

   $$(B_{k,1}, B_{k,2}) = \textbf{Repartition}(B, (A_{k-1,1}, A_{k-1,2}), F_k)$$

   $$(A_{k,1}, A_{k,2}) = \textbf{Repartition}(A, (B_{k-1,1}, B_{k-1,2}), F_k)$$

5. Output the partition: $(A_{\log s,1} \cup B_{\log s,1}, A_{\log s,2} \cup B_{\log s,2})$

Figure 4.1: Main Algorithm

is $\Theta(\log^2 s)$, the number of samples required by our algorithm can be as much as $\Theta(n \log n)$, which agrees with the number of samples required by SVD-based methods within a logarithmic factor of the dimension. We also show that a somewhat weaker version of Condition (1) in Theorem 3 is necessary for any algorithm which classifies $1 - \frac{1}{s}$ fraction of the samples correctly.

4.2 Our Algorithm

Our algorithm takes as input a set S of samples from a mixture of binary product distributions with uniform mixing weights, and produces as output a partition of the samples according to source distribution. This clustering can be further used to estimate the centers of the source distribution. Our algorithm is summarized in Figure 4.1.

Our algorithm begins by splitting the set of samples into sets A and B uniformly at random and the set of coordinates into $F_1, \ldots, F_{\log s}$, also uniformly at random.

Repartition$(V, (S_1, S_2), F_k)$

1. Compute c_1, c_2, the centers of the components S_1 and S_2.
2. For each $x \in V$,
$$\textbf{Score}_k(x) = \sum_{f \in F_k} (c_1^f - c_2^f) \cdot x^f$$
3. Let m_k be the median score among all $x \in V$.
4. Assign:
$$\begin{aligned} V_1 &= \{x | x \in V, \textbf{Score}_k(x) \leq m_k\} \\ V_2 &= \{x | x \in V, \textbf{Score}_k(x) > m_k\} \end{aligned}$$
5. Output the partition (V_1, V_2).

Figure 4.2: The Repartitioning Procedure

Our algorithm then proceeds in $\log s$ rounds. Each round yields partitions which are progressively closer to the correct partitions of A and B, and in $\log s$ rounds the algorithm converges to the correct partitions of A and B with high probability.

In the sequel, we use $(A_{k,1}, A_{k,2})$ (respectively $(B_{k,1}, B_{k,2})$) to denote the partition of A (respectively B) at the end of round k. $(A_{0,1}, A_{0,2})$ is a random partition of A. In round k, $(B_{k,1}, B_{k,2})$ (respectively $(A_{k,1}, A_{k,2})$) is produced by partitioning the samples in B (resp. A) using $(A_{k-1,1}, A_{k-1,2})$ (resp. $(B_{k-1,1}, B_{k-1,2})$) and the set of coordinates F_k. We note that the coordinates used in each round are disjoint from those used in any other round; we see in Lemma 18 that this results in some useful properties.

The repartitioning procedure is described in Figure 4.2. It takes as input a set V of nodes to partition, a partition (S_1, S_2) and a set of coordinates. A proximity score is then computed for each sample $x \in V$, which determines whether x is closer

to S_1 or S_2. Finally, a partition is computed by placing samples with scores above the median score on one side and samples with scores below the median score on the other side of the partition.

4.3 Analysis

Notation The following notation is used in the analysis of the algorithm.In round k, we call the partitions $(A_{k-1,1}, A_{k-1,2})$ and $(B_{k-1,1}, B_{k-1,2})$ the current partitions, and $(A_{k,1}, A_{k,2})$ and $(B_{k,1}, B_{k,2})$ the output partitions. We use b as the extra imbalance of the current partition over $\frac{1}{2}$. For a real number t, we write $\mathcal{N}(t)$ as the probability that a standard normal variable is less than or equal to t.

The main components of the proof of Theorem 3 are Lemmas 17 and 18.

Lemma 17 *Let (S_1, S_2) be a random partition of s samples from a mixture of two binary product distributions, with uniform mixing weights. Then, with constant probability, (S_1, S_2) has imbalance at least $\frac{1}{2} + \frac{1}{\sqrt{s}}$.*

PROOF: For $l = 1, \ldots, n$, let Z_l be a 0/1 random variable which is 1 when sample k from distribution D_1 is in S_1 and 0 otherwise. If $Z = Z_1 + \ldots + Z_n$, then Z denotes the number of samples from distribution D_1 in S_1, and is distributed as a binomial variable with parameters s and $\frac{1}{2}$. Using the fact that the binomial distribution with parameters s and $\frac{1}{2}$ has mean $\frac{s}{2}$ and standard deviation $\frac{\sqrt{s}}{2}$, and the Berry-Esseen Central Limit Theorem 23,

$$\Pr[Z > \frac{s}{2} + \sqrt{s}] \geq a$$

for some constant a. The lemma follows. □

Lemma 18 *If $(A_{k-1,1}, A_{k-1,2})$ has imbalance at least $(\frac{1}{2} + b)$, and the following conditions hold:*

1. $b \geq \frac{1}{\sqrt{s}}$

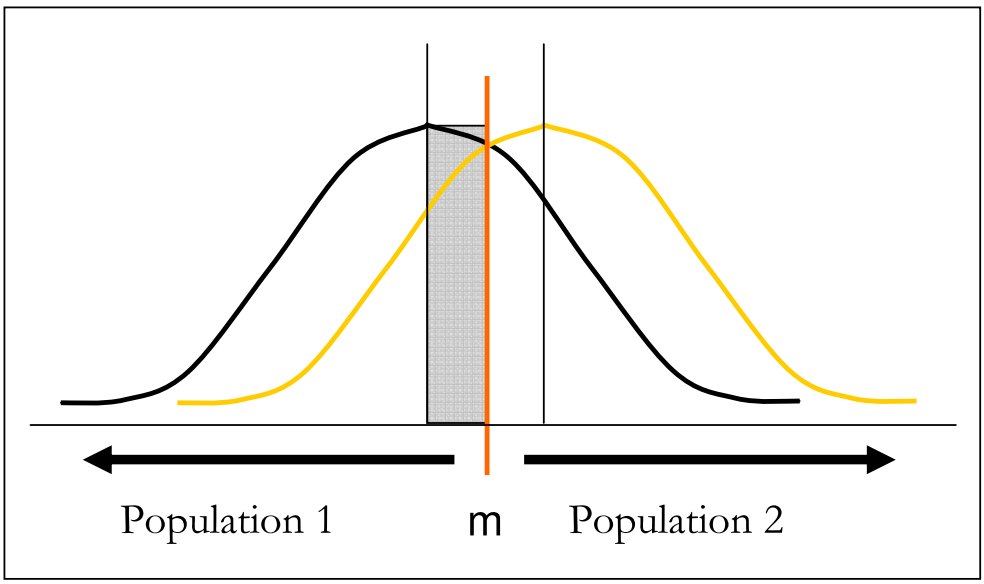

Figure 4.3: Score Distributions When b is Small

2. $d(\mu_1, \mu_2) > a_1 \log s \log(ns)$

3. $s \cdot d^2(\mu_1, \mu_2) > a_2 n \log s \log(ns)$,

then, $(A_{k,1}, A_{k,2})$ *and* $(B_{k,1}, B_{k,2})$ *have imbalance at least* $\min(\frac{1}{2} + 2b, 1 - \frac{2\sqrt{\log s}}{\sqrt{s}})$ *with probability at least* $1 - \frac{1}{s}$ *over the samples.*

The intuition behind the proof of Lemma 18 is to show that the distribution of the scores of the samples in B belonging to D_1 is sufficiently different from the distribution of the scores of the samples of B belonging to D_2. This is shown by Lemmas 19 and 20. Further, Lemma 21 shows that in each round, a higher fraction of the samples, which have scored below the median belong to one of the distributions, say D_1. As a result, in each round, the partition output by the repartitioning procedure has higher imbalance than the partition input to the procedure. When a partition (S_1, S_2) is provided as input to the repartitioning procedure, we define random variables $Y_1(S_1, S_2)$ and $Y_2(S_1, S_2)$ as scores of randomly chosen samples from D_1 and D_2 respectively. Then, $Y_1(S_1, S_2)$ and $Y_2(S_1, S_2)$ are random functions of the samples in S_1 and S_2.

Lemma 19 *Suppose in round k, the inputs to the repartitioning procedure is a par-*

tition $(A_{k-1,1}, A_{k-1,2})$ *with imbalance at least* $\frac{1}{2} + b$. *If the following conditions hold, (1)*$b \geq \frac{1}{\sqrt{s}}$ *(2)* $d(\mu_1, \mu_2) > a_1 \log s \log(ns)$, *for some fixed constant* $a_1 \geq 32$, *then,*

$$|\mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})] - \mathbf{E}[Y_2(A_{k-1,1}, A_{k-1,2})]| \geq \frac{b \cdot d(\mu_1, \mu_2)}{\log s}$$

Here, the expectation is taken over the distribution of the samples in $A_{k-1,1}$ *and* $A_{k-1,2}$ *as well as over* D_1 *and* D_2.

PROOF: Since the set of coordinates F_k involved in the score in round k is disjoint from those involved in the computation of $(A_{k-1,1}, A_{k-1,2})$, and D_1 and D_2 are product distributions, the distribution of any coordinate f in F_k in the partition $(A_{k-1,1}, A_{k-1,2})$ is the same as the distribution of f in a randomly chosen partition of $\frac{s}{2}$ samples with imbalance $\frac{1}{2} + b$. This follows because of the independence between the distributions of disjoint coordinates. Therefore, we can write:

$$\begin{aligned}
\mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2}) &- Y_2(A_{k-1,1}, A_{k-1,2})] \\
&= \sum_{f \in F_k} \mathbf{E}_{x \sim D_1}[(c_1^f - c_2^f) \cdot x^f] - \sum_{f \in F_k} \mathbf{E}_{x \sim D_2}[(c_1^f - c_2^f) \cdot x^f] \\
&= \sum_{f \in F_k} \mathbf{E}[c_1^f - c_2^f] \cdot (\mathbf{E}_{x \sim D_1}[x^f] - \mathbf{E}_{x \sim D_2}[x^f]) \\
&= \sum_{f \in F_k} 2b(\mu_1^f - \mu_2^f) \cdot (\mathbf{E}_{x \sim D_1}[x^f] - \mathbf{E}_{x \sim D_2}[x^f]) \\
&= 2b \sum_{f \in F_k} (\mu_1^f - \mu_2^f)^2 \\
&= \frac{2b \cdot d(\mu_1, \mu_2)}{\log s}
\end{aligned}$$

The third step follows because $\mathbf{E}[c_1^f] = \frac{1}{s}((\frac{1}{2} + b)s\mu_1^f + (\frac{1}{2} - b)s\mu_2^f)$ and $\mathbf{E}[c_2^f] = \frac{1}{s}((\frac{1}{2} + b)s\mu_2^f + (\frac{1}{2} - b)s\mu_1^f)$. □

Lemma 20 *Suppose in round* k, *the inputs to the repartitioning procedure is a partition* $(A_{k-1,1}, A_{k-1,2})$ *with imbalance at least* $\frac{1}{2} + b$. *If the following conditions hold,*

1. $b \geq \frac{1}{\sqrt{s}}$
2. $d(\mu_1, \mu_2) > a_1 \log s \log(ns)$,

then, with probability $1 - \frac{1}{s}$ *over the samples in* $A_{k-1,1}$ *and* $A_{k-1,2}$,

$$|\mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})] - \mathbf{E}[Y_2(A_{k-1,1}, A_{k-1,2})]| \geq \frac{b \cdot d(\mu_1, \mu_2)}{2 \log s}$$

Here, the expectation is taken over D_1 *and* D_2.

PROOF: Since the set of coordinates F_k involved in the score in round k is disjoint from those involved in the computation of $(A_{k-1,1}, A_{k-1,2})$, and D_1 and D_2 are product distributions, the distribution of any coordinate f in F_k in the partition $(A_{k-1,1}, A_{k-1,2})$ is the same as the distribution of f in a randomly chosen partition of $\frac{s}{2}$ samples with imbalance $\frac{1}{2} + b$. This follows because of the independence between the distribution of disjoint coordinates. Let $Z = \mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})] - \mathbf{E}[Y_2(A_{k-1,1}, A_{k-1,2})]$. Then, Z is a random variable, the value of which depends on c_1 and c_2, computed in Step 2 of the repartitioning procedure. We can write:

$$Z = \sum_{f \in F_k} (\mu_1^f - \mu_2^f) \cdot (c_1^f - c_2^f)$$

We can now apply the Method of Bounded Differences [20] (Theorem 20 in the Appendix). Each of c_1^f and c_2^f is a function of $\frac{s}{4}$ sample points. Changing coordinate f of any sample x in $A_{k-1,1}$ and $A_{k-1,2}$ changes Z by at most $\gamma_{x,f} = \frac{4(\mu_1^f - \mu_2^f)}{s}$. Then,

$$\gamma = \sum_{x,f} \gamma_{x,f}^2 = \frac{16s \cdot \sum_{f \in F_k} (\mu_1^f - \mu_2^f)^2}{s^2} = \frac{16 d(\mu_1, \mu_2)}{s \log s}$$

Using the Method of Bounded Differences,

$$\Pr[|Z - \mathbf{E}[Z]| > t] \leq e^{-t^2/2\gamma}$$

Plugging in $t = b \cdot d(\mu_1, \mu_2)$, this probability is at most $e^{-b^2 d(\mu_1, \mu_2) s \log s/32} \leq \frac{1}{s}$ using conditions (1) and (2). □

Lemmas 19 and 20 show that with high probability, the expected scores of the samples from distributions D_1 and D_2 are sufficiently different. Next, in Lemma 21 we show that the variance of the scores is low enough that in each round, a higher fraction

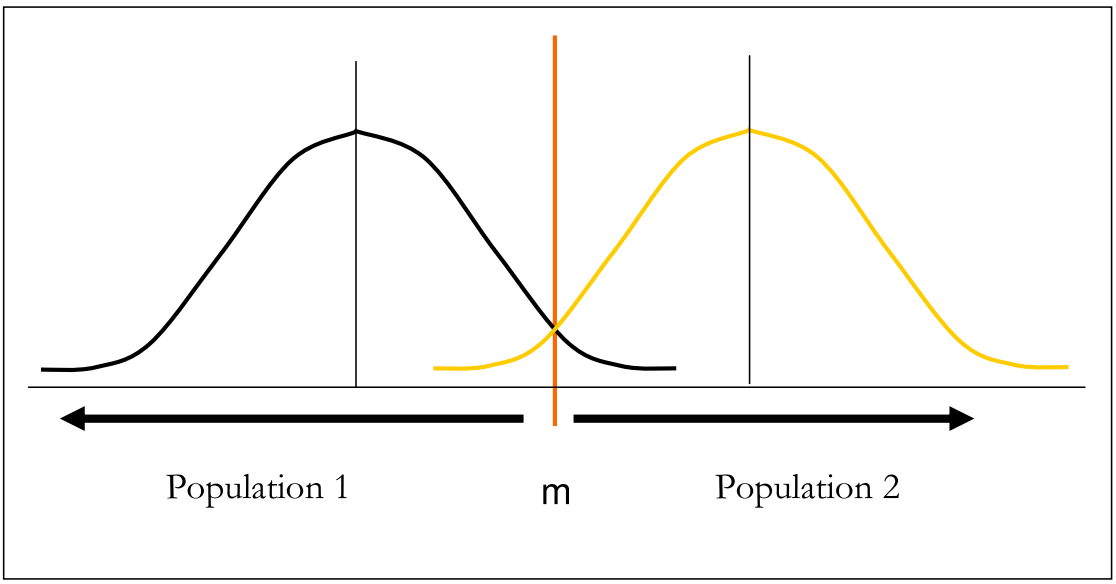

Figure 4.4: Score Distribution with Large b

of the points from D_1 have scores less than the combined median of the scores than from D_2. For the rest of this section, we assume that without loss of generality, $\mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})] < \mathbf{E}[Y_2(A_{k-1,1}, A_{k-1,2})]$.

Lemma 21 *Suppose that in round k, the input to the repartitioning procedure is a partition $(A_{k-1,1}, A_{k-1,2})$ with imbalance at least $\frac{1}{2}+b$, and let a_4 be a positive constant such that $\mathcal{N}(a_4) = \frac{1}{2} + \frac{3}{4}a_4$. If the following conditions are true,*

1. $b \geq \frac{1}{\sqrt{s}}$
2. $d(\mu_1, \mu_2) \geq a_1 \log s \log(ns)$
3. $s \cdot d^2(\mu_1, \mu_2) \geq a_2 n \log s \log(ns)$, *for some fixed constants a_1 and a_2.*

Then, if $b < a_4/4$,

$$\Pr[Y_1(A_{k-1,1}, A_{k-1,2}) \leq m_k] > \frac{1}{2} + 2b + \frac{2\sqrt{\log s}}{\sqrt{s}}$$

Otherwise,

$$\Pr[Y_1(A_{k-1,1}, A_{k-1,2}) \leq m_k] > 1 - \frac{1}{s^2}$$

The probabilities in both equations are computed over the distribution of the samples in $A_{k-1,1}$ and $A_{k-1,2}$.

PROOF: Since, $|\mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})] - \mathbf{E}[Y_2(A_{k-1,1}, A_{k-1,2})]| > \frac{b \cdot d(\mu_1,\mu_2)}{\log s}$, from Lemma 20, by the triangle inequality, there exists some $i \in \{1, 2\}$ such that

$$|\mathbf{E}[Y_i(A_{k-1,1}, A_{k-1,2})] - m_k| > \frac{b \cdot d(\mu_1, \mu_2)}{2 \log s}$$

Without loss of generality, let $i = 1$. The other case follows similarly.

Let, for a coordinate f, $\beta_f = c_1^f - c_2^f$. Recall that since the set of coordinates F_k involved in the score in round k is disjoint from those involved in the computation of $(A_{k-1,1}, A_{k-1,2})$, and D_1 and D_2 are product distributions, the distribution of any coordinate f in F_k in the partition $(A_{k-1,1}, A_{k-1,2})$ is the same as the distribution of f in a randomly chosen partition of $\frac{s}{2}$ samples with imbalance $\frac{1}{2} + b$. Then, $\mathbf{E}[\beta_f] = 2b(\mu_1^f - \mu_2^f)$, and using the Chernoff Bounds (Theorem 18 in the Appendix), for all f,

$$|\beta_f| \leq 2b(\mu_1^f - \mu_2^f) + \frac{2\sqrt{\log(ns)}}{\sqrt{s}}$$

with probability at most $1 - \frac{1}{s}$. We now apply the Berry-Esseen Central Limit Theorem (Theorem 23 in the Appendix). We define, for each coordinate f, a random variable W_f as the contribution from coordinate f to the score of an individual in distribution D_1. Then, $W_f = \beta_f$, with probability μ_1^f and $W_f = -\beta_f$ with probability $1 - \mu_1^f$, and $Y_1(A_{k-1,1}, A_{k-1,2}) = \sum_f W_f$. Let $Z_f = W_f - \mathbf{E}[W_f]$. Then,

$$\begin{aligned} Z_f &= 2\beta_f(1 - \mu_1^f) \quad \text{with probability} \mu_1^f \\ &= -2\beta_f \mu_1^f \quad \text{otherwise} \end{aligned}$$

Also, for any two coordinates f and f', Z_f is distributed independently of $Z_{f'}$, and $Y_1(A_{k-1,1}, A_{k-1,2}) - \mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})] = \sum_f Z_f$. We can compute that:

$$\begin{aligned} \mathbf{E}[Z_f] &= 0 \\ \mathbf{E}[{Z_f}^2] &= \mathbf{Var}(Z_f) = 4\beta_f^2 \mu_1^f (1 - \mu_1^f) \\ \mathbf{E}[|Z_f|^3] &= 16\beta_f^3 \mu_1^f (1 - \mu_1^f)((\mu_1^f)^2 + (1 - \mu_1^f)^2) \end{aligned}$$

Following the notation in the Berry-Esseen Central Limit Theorem,

$$\begin{aligned} r_n &= \sum_f \mathbf{E}[|Z_f|^3] = \sum_f 16\beta_f^3 \mu_1^f (1-\mu_1^f)((\mu_1^f)^2 + (1-\mu_1^f)^2) \\ s_n^2 &= \sum_f 4\beta_f^2 \mu_1^f (1-\mu_1^f) \end{aligned}$$

Before we can apply the Berry-Esseen Central Limit Theorem, we need the following observation.

Observation 1 *Let $a_3 \geq 24$ be a constant. If (1) $b \geq \frac{1}{\sqrt{s}}$ and (2) $d(\mu_1, \mu_2) > a_1 \log s \log(ns)$, for some fixed constant a_1, then,*

$$\frac{b \cdot d(\mu_1, \mu_2)}{2 s_n \log s} > a_3 \cdot \frac{r_n}{s_n^3}$$

PROOF: For a fixed coordinate f,

$$\frac{\mathbf{E}[|Z_f|^3]}{\mathbf{E}[Z_f^2]} \leq 8b + \frac{16\sqrt{\log(ns)}}{\sqrt{s}}$$

if the denominator and numerator are nonzero. The inequality follows because $(\mu_1^f)^2 + (1-\mu_1^f)^2 \leq 2$, $\mu_1^f \leq 1$, and $|\beta_f| \leq 2b(\mu_1^f - \mu_2^f) + \frac{2\sqrt{\log(ns)}}{\sqrt{s}}$. Since $\frac{r_n}{s_n^2} = \frac{\sum_f \mathbf{E}[|Z_f|^3]}{\sum_f \mathbf{E}[Z_f^2]}$, this ratio is at most $8b + \frac{16\sqrt{\log(ns)}}{\sqrt{s}}$. If $a_1 > 64a_3$, $d(\mu_1, \mu_2) > 64a_3 \log^2 s$. Since $b \geq \frac{1}{\sqrt{s}}$,

$$\frac{b \cdot d(\mu_1, \mu_2)}{2 \log s} > 64 a_3 \log(ns) \cdot b > a_3 \cdot \frac{r_n}{s_n^2}$$

The observation follows. □

We consider two cases, depending on the value of b.

Case 1: Low Imbalance $b < \frac{a_4}{4}$, where a_4 is a constant such that $\mathcal{N}(a_4) = \frac{3}{4}a_4 + \frac{1}{2}$.

In this case, we can apply the Berry-Esseen Central Limit Theorem on the

distribution of $Y_1(A_{k-1,1}, A_{k-1,2})$ to conclude that:

$$\begin{aligned}
& \Pr[Y_1(A_{k-1,1}, A_{k-1,2}) \leq m_k] \\
\geq\ & \Pr[Y_1(A_{k-1,1}, A_{k-1,2}) \leq \mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2}) + \frac{b \cdot d(\mu_1, \mu_2)}{4 \log s}] \\
\geq\ & \Pr[\sum_f Z_f < -\frac{b \cdot d(\mu_1, \mu_2)}{4 \log s} \\
\geq\ & \frac{1}{2} + \frac{3b \cdot d(\mu_1, \mu_2)}{16 s_n \log s} - \frac{6 r_n}{s_n^3} \\
\geq\ & \frac{1}{2} + \frac{b \cdot d(\mu_1, \mu_2)}{4 s_n \log s} \cdot (\frac{3}{4} - \frac{6}{a_3}) \\
\geq\ & \frac{1}{2} + \frac{b \cdot d(\mu_1, \mu_2)}{8 s_n \log s}
\end{aligned}$$

The third line follows because $a_3 \geq 24$. s_n can be bounded as:

$$\begin{aligned}
s_n^2 \leq \sum_f \beta_f^2 &\leq 2 \left(\sum_f b^2 (\mu_1^f - \mu_2^f)^2 + \frac{4 \log(ns)}{s} \right) \\
&\leq 4b^2 \cdot d(\mu_1, \mu_2) + \frac{8n \log(ns)}{s}
\end{aligned}$$

From Conditions (1), (2), and (3),

$$\frac{b \cdot d(\mu_1, \mu_2)}{8 s_n \log s} \geq \frac{b \cdot d(\mu_1, \mu_2)}{8 \log s (4b^2 d(\mu_1, \mu_2) + \frac{8n \log(ns)}{s})^{1/2}}$$

Since,

$$\begin{aligned}
\frac{b \cdot d(\mu_1, \mu_2)}{8 \log s \cdot 2b \sqrt{d(\mu_1, \mu_2)}} &\geq 4b + \frac{4\sqrt{\log s}}{\sqrt{s}} \\
\frac{b \cdot d(\mu_1, \mu_2)}{8 \log s (\frac{8n \log(ns)}{s})^{1/2}} &\geq \frac{b \cdot \sqrt{a_2 n \log(ns)}}{8 \log s (\frac{8n \log(ns)}{s})^{1/2}} \geq 4b + \frac{4\sqrt{\log s}}{\sqrt{s}}
\end{aligned}$$

Therefore,

$$\frac{b \cdot d(\mu_1, \mu_2)}{8 s_n \log s} > 2b + \frac{2\sqrt{\log s}}{\sqrt{s}}$$

from which the lemma follows.

Case 2: High Imbalance Otherwise, $b \geq \frac{a_4}{4}$. In this case,

$$|m_k - \mathbf{E}[Y_1(A_{k-1,1}, A_{k-1,2})]| \geq \frac{a_4 \cdot d(\mu_1, \mu_2)}{4 \log s}$$

and $s_n^2 \leq \frac{a_4^2}{4} \cdot d(\mu_1, \mu_2) + \frac{8n \log ns}{s}$. In addition, the parameter β in the statement of the Method of Bounded Variances is at most 1. We can therefore apply the Method Of Bounded Variances to conclude that

$$\Pr[Y_1(A_{k-1,1}, A_{k-1,2}) \geq m_k] \leq e^{-|m_k - Y_1(A_{k-1,1}, A_{k-1,2})|^2/4s_n^2} \leq e^{-\log s} \leq \frac{1}{s}$$

from Condition (3).

□

Finally we are ready to prove Lemma 18.

PROOF:(Of Lemma 18) Let for each sample $x \in B$, Z_x be a 0/1 random variable which is 1 if x is classified correctly, and 0 otherwise. Also let a_4 be a constant such that $\mathcal{N}(a_4) = \frac{1}{2} + \frac{3}{4}a_4$.

From Lemma 21, if $b < a_4$, $\mathbf{E}[\sum_{x \in B} Z_x] \geq \frac{s}{4} + bs + \sqrt{s \log s}$. We now apply the Method of Bounded Differences 20. Let γ_x be the change in $\sum_{x \in B} Z_x$ when we change the allocation of sample x. Then, $\gamma = \gamma_x^2 = s/2$. Therefore, from the Method of Bounded Differences,

$$\Pr[\sum_{x \in B} Z_x > \frac{s}{2} + sb] \leq e^{-\log s} \leq \frac{1}{s}$$

from which the lemma follows.

If $b \geq a_4/4$, from Lemma 21, $\mathbf{E}[\sum_{x \in B} Z_x] \geq \frac{s}{2} - \frac{1}{s}$. The lemma follows by an argument similar to the one above. □

4.4 Lower Bounds

In this section, we show that the separation condition required by our algorithm is almost optimal, in an information-theoretic sense.

To be more specific, we first demonstrate that for any $\delta \geq \Omega(\frac{1}{\sqrt{n}})$, a separation of $\Omega\left(\sqrt{\log(\frac{1}{\delta})}\right)$ is necessary for correctly learning a mixture of two binary product distributions with probability at least $1 - \delta$. This means that to be able to classify s

samples correctly, we need a separation of $\Omega(\sqrt{\log s})$.

To prove a lower bound on the separation condition, we examine the performance of the Bayes-Optimal algorithm, which knows the centers of the distributions in the mixture, but not the labels of the individual points. We show that this algorithm requires a separation of $\Omega\left(\sqrt{\log(\frac{1}{\delta})}\right)$ to classify $1-\delta$-fraction of the samples correctly. Since no algorithm can achieve a better classification probability than the Bayes-Optimal algorithm, this is an information theoretic lower bound on the separation requirement.

Lemma 22 *Let $\mathcal{D} = \{D_1, D_2\}$ be a mixture of two binary product distributions with uniform mixing weights. If $\mu_1 = (\frac{1}{2}, \ldots, \frac{1}{2})$ and $\mu_2 = \left(\frac{1}{2} + \frac{\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}, \ldots, \frac{1}{2} + \frac{\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right)$, then, the Bayes-Optimal algorithm has a misclassification probability at least $\Omega(\delta - \frac{1}{\sqrt{n}})$.*

PROOF: Given the parameters of D_1 and D_2, the Bayes-Optimal algorithm assigns a sample x to D_1 if

$$\sum_{f=1}^{n} x^f \leq \frac{n}{2} + \sqrt{2n\log(1/\delta)}$$

and assigns x to D_2 otherwise. We note that when x is generated from D_1, $\sum_f x^f$ is distributed as a binomial with parameters n and $1/2$, and when x is generated from D_2, $\sum_f x^f$ is distributed as a binomial with parameters n and $\frac{1}{2} + \frac{\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}$. The total probability of misclassification is therefore:

$$\frac{1}{2}\Pr\left[\mathbf{Bin}(n, \frac{1}{2}) > \frac{n}{2} + \sqrt{2n\log(1/\delta)}\right]$$
$$+\frac{1}{2}\Pr\left[\mathbf{Bin}(n, \frac{1}{2} + \frac{\sqrt{2n\log(1/\delta)}}{\sqrt{n}}) < \frac{n}{2} + \sqrt{2n\log(1/\delta)}\right]$$

We use the Berry-Esseen Central Limit Theorem to estimate each probability in the equation. As for each f, $\mathbf{Var}(x^f)$ and $\mathbf{E}[|x^f - \mathbf{E}[x^f]|^3]$ are constants, the term $\frac{r_n}{s_n^3}$ in the Berry-Esseen Central Limit Theorem is at most $\Omega(\frac{1}{\sqrt{n}})$. As the term $|t|$ in the Berry-Esseen Central Limit Theorem is $\Omega(\sqrt{2\log(\frac{1}{\delta})})$, $\mathcal{N}(t) = \Omega(\delta)$. The lemma therefore follows by the application of the Berry-Esseen Central Limit Theorem. □

4.5 Conclusions and Open Problems

In this chapter, we make partial progress in analyzing the sample complexity of learning mixtures of distributions. We present an algorithm for learning mixtures of two binary product distributions with uniform mixing weights with low sample complexity. We leave open the following, more general question:

Problem 3 *Given samples from a mixture of T binary product distributions or Gaussians, with separation $d(\mu_1, \mu_2)$, what is the optimal number of samples required to learn the mixture? Can we find an algorithm which learns such a mixture with optimal sample complexity?*

Another open question is to analyze the sample complexity of existing algorithms. From [1], an upper bound on the number of samples required by SVD-based algorithms is $\Theta(n)$. The question is, does there exist a better bound when the separation $d(\mu_1, \mu_2)$ is higher?

Problem 4 *How many samples are required by SVD-based algorithms to learn mixtures of T distributions?*

Chapter 5

Learning Mixtures of Heavy-Tailed Distributions

5.1 Overview

In this chapter, we study the problem of learning mixtures of product distributions over high-dimensional space, in which the individual coordinates are heavy-tailed. A distribution on $\mathbf{R}$ with median m is loosely called heavy-tailed, if the probability mass in an interval $[-\infty, m-t] \cup [m+t, \infty]$ decreases slowly (usually slower than an exponential) with increasing t. An example of a heavy-tailed distribution is the Cauchy distribution, with mean μ, which has the probability density function: $f(x) = \frac{2}{\pi((x-\mu)^2+1)}$. A canonical example of a distribution, which is not heavy-tailed, is a Gaussian, in which the probability density function decays exponentially with distance from the mean.

The common measure of separation used for learning mixtures of Gaussians and binary product distributions is the distance between the means, parameterized by the maximum directional standard deviation of any distribution in the mixture. However, the standard deviation may not be finite for heavy-tailed distributions. Therefore, for learning mixtures of heavy-tailed distributions, the analogous separation measure used is the distance between the medians of the distributions, as a function of the maximum β-radius of any distribution for some constant β. The β-radius of a distribution on $\mathbf{R}$ with median m is the minimum value r_β such that the interval

$[m - r_\beta, m + r_\beta]$ contains a β fraction of the probability mass. If X is a random variable with finite variance, then its β-radius and variance are related: for any β, $\mathbf{Var}(X) \geq r_\beta^2(1 - \beta)$.

One of the main challenges in learning mixtures of heavy-tailed product distributions is that *distance concentration*, a tool extensively used in the literature for learning mixtures of Gaussians and binary product distributions, does not yield good guarantees when the distributions have heavy tails or a high range. For heavy-tailed product distributions in high dimensions, even if the distribution of a single coordinate has constant probability mass very far from the median, a sample can be very far from the median of the distribution with as much as constant probability. In contrast, in a light-tailed distribution in high dimensions, the probability that a sample is far off from the median of the distribution is vanishingly small, which makes distance concentration possible.

In addition, if one of the coordinates has infinite variance, the covariance matrix of the samples does not converge. As a result, the techniques of [17; 1] and [5], which work with the sample covariance matrix, do not apply.

The main contribution of this chapter is to provide an embedding from $\mathbf{R}^n$ to $\{0, 1\}^{n'}$ where $n' > n$. The embedding has the property that samples from distributions in $\mathbf{R}^n$ which satisfy certain conditions and have medians that are far apart, map to samples from distributions in $\{0, 1\}^{n'}$ which have centers that are far apart. We then apply this embedding to design an algorithm for learning mixtures of heavy-tailed product distributions. Our algorithm learns mixtures of general product distributions, as long as the distribution of each coordinate satisfies the following two properties. First, the distribution of each coordinate is symmetric about its median, and second, 3/4 of the probability mass is present in an interval of length $2R$ around the median. The separation condition required to correctly classify a $1 - \delta$ fraction of the samples is that the distance between the medians of any two distributions in the mixture is $\Omega(R\sqrt{T \log \Lambda} + R\sqrt{T \log(T/\delta)})$, where Λ is polynomial in n and T. In

addition, we require a spreading constraint, which states that the distance between the medians of any two distributions should be spread across $\Omega(T \log \Lambda + T \log(T/\delta))$ coordinates. The number of samples required by the algorithm is polynomial in n, T, and $\frac{1}{w_{\min}}$ and our algorithm is based on the algorithm in [5].

A second application of our embedding is in designing an algorithm for learning mixtures of product distributions with finite variance, but with separation proportional to σ_*, the maximum directional standard deviation of any distribution in the mixture along the subspace containing the centers. Given samples from a mixture of distributions with finite variance, our algorithm can learn the mixture provided the following conditions hold. The distance between every pair of centers is at least $\Omega(\sigma_* \sqrt{T \log \Lambda})$ and this distance is spread along $\Omega(T \log \Lambda)$ coordinates, with no coordinate contributing more than a $\frac{1}{T \log \Lambda}$ fraction of the distance. This condition is implied by the strong spreading condition in Chapter 3, which states that every vector in the space spanned by the centers have high spread.

The main idea behind our embedding is to use many random *shifts* or *cutting points*, and map points to the left of the cutting point to 0 and point to the right of the cutting point to 1. This succeeds in separating two distributions, with medians which are far apart, only when the medians of both distributions lie in some fixed interval of length $\Omega(R)$ known in advance. To address this limitation, we divide the real line into intervals of length $\Omega(R)$ and map point in alternate segments to 0 and 1. This results in an embedding which separates two distributions with medians which are far apart, no matter where the medians are located. These methods are perhaps related to techniques in metric embeddings [16]. Combining this embedding with spectral methods for clustering results in an efficient algorithm for learning mixtures of heavy-tailed product distributions with possibly infinite variances.

5.1.1 Notation

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

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

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

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 $\mathcal{H}$.

5.1.2 A Summary of Our Results

The main contribution of this chapter 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 4 *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), and the following conditions hold. For some $Q > 1$,*

1. *For any i and j,*

$$||m(D_i) - m(D_j)|| \geq RQ$$

2. *For any i and j,*

$$||m(D_i) - m(D_j)|| \geq \Omega(Q) \cdot \max_f |m(D_i^f) - m(D_j^f)|$$

Then, there exists an embedding $\Phi : \mathbf{R}^n \to \{0,1\}^{nq}$ such that after the transformation, for any i and j,

$$\begin{aligned} ||\tilde{\mu}_i - \tilde{\mu}_j|| &\geq \Omega(\sqrt{q}Q) \\ ||\tilde{\mu}_i - \tilde{\mu}_j|| &\geq \Omega(Q) \cdot \max_f ||\tilde{\mu}_i^f - \tilde{\mu}_j^f|| \end{aligned}$$

More precisely, the properties of our embedding Φ can be described as follows.

Lemma 23 *Let $R_1 \geq 26R$, $R_2 \geq 8R$, and $q = 4\sqrt{n}\log n \log T$. Then, for all i and j, the embedding $\Phi = \oplus_f \Phi_f$ defined in Equation 5.3 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,*

$$||\mathbf{E}_{x \sim D_i^f}[\Phi_f(x)] - \mathbf{E}_{x \sim D_j^f}[\Phi_f(x)]|| \geq \Omega(\sqrt{q})$$

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

$$||\mathbf{E}_{x \sim D_i^f}[\Phi_f(x)] - \mathbf{E}_{x \sim D_j^f}[\Phi_f(x)]|| \geq \Omega(\sqrt{q}) \cdot \Omega\left(\frac{|m(D_i^f) - m(D_j^f)|}{R}\right)$$

In this chapter, we provide two applications of our embedding in learning mixtures of distributions. First, we present an algorithm that learns mixtures of heavy-tailed distributions and uses our embedding as a preprocessing step. The properties of this algorithm are summarized as follows.

Theorem 5 *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), and the following conditions hold.*

1. *For any i and j,*

$$||m(D_i) - m(D_j)|| \geq \Omega\left(R\sqrt{T\log\Lambda} + R\sqrt{T\log(T/\delta)}\right)$$

2. *For any i and j,*

$$||m(D_i) - m(D_j)|| \geq \Omega\left(\sqrt{T\log\Lambda} + \sqrt{T\log(T/\delta)}\right) \cdot \max_f |m(D_i^f) - m(D_j^f)|$$

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

The first condition is a separation condition, which states that every pair of centers is sufficiently far apart in space. The second condition is the spreading condition, which states that the distance between every pair of centers is spread across $\Omega\left(T\log\Lambda + T\log(T/\delta)\right)$ coordinates.

Next we present an algorithm which uses our algorithm to learn mixtures of product distributions with finite variance, with separation $\Omega(\sigma_*\sqrt{T\log\Lambda})$, where σ_* is the maximum directional standard deviation of any distribution in the mixture along the space containing the centers. The guarantees provided by this algorithm are formally stated below.

Theorem 6 *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 *has finite variance, and the following conditions hold.*

1. *For any* i *and* j,

$$||m(D_i) - m(D_j)|| \geq \Omega\left(\sigma_*\sqrt{T\log\Lambda} + \sigma_*\sqrt{T\log(T/\delta)}\right)$$

2. *For any* i *and* j,

$$||m(D_i) - m(D_j)|| \geq \Omega\left(\sqrt{T\log\Lambda} + \sqrt{T\log(T/\delta)}\right)\cdot\max_f |m(D_i^f) - m(D_j^f)|$$

where Λ *is at more* $\Theta(\frac{Tn\log^2 n}{w_{\min}})$. *Then, there exists an algorithm, which runs in time polynomial in* n *and clusters* $1-\delta$ *fraction of the samples correctly. The number of samples required by our algorithm is polynomial in* n*,* T*, and* $\frac{1}{w_{\min}}$.

The algorithms in Theorem 5 and 6 are modified versions of Algorithm CORRELATION-CLUSTER in Chapter 3. After the set of coordinates is randomly split into two halves,

Φ is applied to each half, and then the remaining steps of Algorithm Correlation-Cluster are carried out.

We note that for both applications, we can use our embedding as a preprocessing step along with SVD-based algorithms for clustering. The details of these algorithms, along with the bounds obtained are stated in Section 5.3.

5.2 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 5.2.3; 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 $\Omega\big(\frac{|m(D_i^f)-m(D_j^f)|}{R}\big)$. 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.

5.2.1 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 the embedding can be summarized as follows.

Theorem 7 *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

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

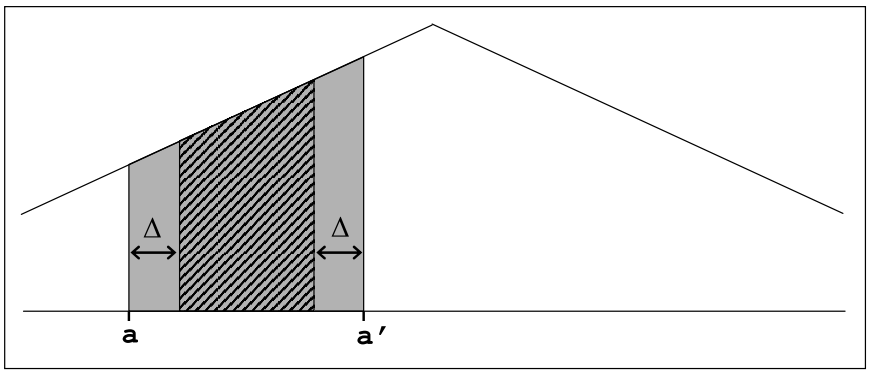

Figure 5.1: Proof of Lemma 25

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. This area 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 7 can be written as follows.

$$\mathbf{E}[|\Pr_{x\sim D_i^f}[\psi_r(x)=0] - \Pr_{x\sim D_j^f}[\psi_r(x)=0]|] = \frac{1}{R_1}\int_{r=-R_1/2}^{R_1/2} |\alpha_j^f(r+\Delta) - \alpha_i^f(r)|\mathbf{dr} \quad (5.1)$$

The proof of Theorem 7 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 7 would hold. This is shown in Lemma 24. Next,Lemma 27 shows that even if D_i^f is not a shifted version of D_j^f, the statements in Theorem 7 still hold.

Lemma 24 *For any Δ, if $R_1 > 3\Delta + 2R$, then, for any i,*

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

Note that the difference between the statement of Theorem 7 and Lemma 24 is that the left-hand side of the equation in Theorem 7 has an absolute value, and hence Lemma 24 makes a stronger statement (under stronger assumptions).

Before we prove Lemma 24, we need the following lemma.

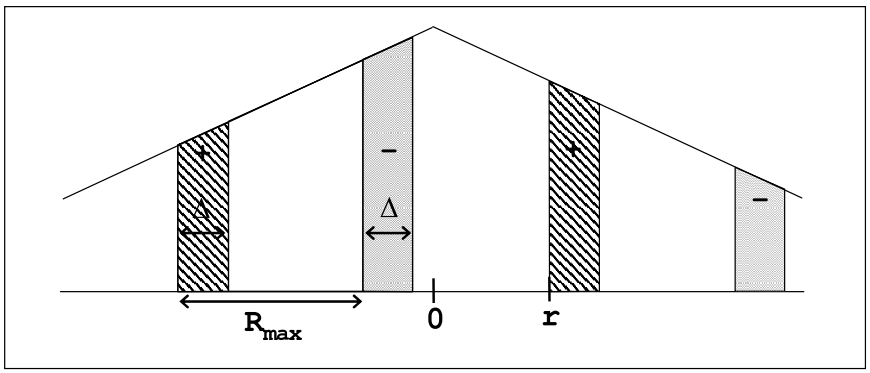

Figure 5.2: Proof of Lemma 24

Lemma 25 *Let $[a, a']$ be any interval of length more than 2Δ. Then, for any i,*

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

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 5.1. Since $\varphi_i^f(t) \geq 0$ for all t, the lemma follows in the limit when $\bar{\delta} \to 0$. □

PROOF:(Of Lemma 24) The shaded area in Figure 5.2 shows the value of $\alpha_i^f(r) -$

$\alpha_i^f(r+\Delta)$ for a distribution D_o. From Lemma 25,

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

From Lemma 25, the first term is at least

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

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

From Lemma 25, the second term is at most

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

This is the total probability mass of D_i in the intervals $[(2\lambda+1)R_1 - R - \Delta, (2\lambda+1)R_1 + R + \Delta]$, 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. $\square$

Next we show that Theorem 7 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 26 and 27, which are both consequences of the symmetry of the distributions D_i^f and D_j^f.

Lemma 26 *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: For $i = 1, 2$, 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 27 *For any i and j, if D_i^f and D_j^f have properties (1) and (2), then,*

$$\int_{r=-\Delta-R}^{\Delta+R} |\alpha_j^f(r + \Delta) - \alpha_i^f(r)| \mathbf{dr} \geq \int_{r=-\Delta-R}^{\Delta+R} (\alpha_j^f(r) - \alpha_j^f(r + \Delta)) \mathbf{dr}$$

PROOF: By Lemma 26, for every $r \in [-\Delta - R, \Delta + R]$, 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'),

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

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

PROOF: (Of Theorem 7) From Equation 5.1 and Lemma 24,

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

The second step follows because $R_1 \geq 2R+3\Delta$, for the interval $[-R_1/2, R_1/2]$ includes the interval $[-R-\Delta, R+\Delta]$ and the last step follows from Lemma 24. □

5.2.2 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.

$$\phi_\zeta(x) = \varepsilon_{k(x)}, \text{where} \quad k(x) = \lfloor \frac{x-\rho}{R_2} \rfloor \tag{5.2}$$

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 8 *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 $2R$,

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

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. □

5.2.3 Combining the Embeddings

In this section, we show how to combine the embeddings of Sections 5.2.1 and 5.2.2 to provide a map Φ which obeys the guarantees of Theorem 4. Given parameters R_1, R_2, and q, we define Φ_f for a coordinate f as follows.

$$\Phi_f(x) = (\phi_{\zeta_1}(x^f), \ldots, \phi_{\zeta_q}(x^f), \psi_{r_1}(x^f), \ldots, \psi_{r_q}(x^f)) \tag{5.3}$$

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:

$$\Phi(x) = \Phi_1(x) \oplus \ldots \oplus \Phi_n(x) \tag{5.4}$$

The properties of the embedding Φ are summarized in Theorem 4. Next, we prove Theorem 4. We begin with the following lemma, which demonstrates the properties of each Φ_f.

PROOF:(Of Lemma 23) The first part of the lemma follows by Theorem 8, 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 7. □

PROOF:(Of Theorem 4) We call a coordinate f *low* for distributions i and j if $|m(D_i^f) - m(D_j^f)| < 8R$, otherwise we call it a *high* coordinate. Let $L_{i,j}$ and $H_{i,j}$ denote the set of low and high coordinates for distributions D_i and D_j. We consider two cases. First, suppose $H_{i,j} = \emptyset$. Then,

$$\begin{aligned} ||\tilde{\mu}_i - \tilde{\mu}_j||^2 &= \sum_{f \in L_{i,j}} ||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2 \\ &\geq \sum_{f \in L_{i,j}} \Omega(q) \cdot \Omega\left(\frac{|m(D_i^f) - m(D_j^f)|^2}{R^2}\right) \end{aligned}$$

The second line follows from Lemma 23. Statements (1) and (2) now follow by applying the Separation Condition and the Spreading Condition respectively. Otherwise, $H_{i,j} \neq \emptyset$. Then,

$$\begin{aligned} |H_{i,j}| \cdot \max_f |m(D_i^f) - m(D_j^f)|^2 &\geq \sum_{f \in H_{i,j}} |m(D_i^f) - m(D_j^f)|^2 \\ &\geq \Omega(Q^2) \cdot \max_f |m(D_i^f) - m(D_j^f)|^2 \\ &\quad - \sum_{f \in L_{i,j}} |m(D_i^f) - m(D_j^f)|^2 \end{aligned}$$

By Lemma 23, there exists a constant a_2 such that for all f in $H_{i,j}$, $||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2 \geq a_2 q$. Therefore,

$$\begin{aligned} ||\tilde{\mu}_i - \tilde{\mu}_j||^2 &\geq |H_{i,j}| \cdot a_2 q + \sum_{f \in L_{i,j}} ||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2 \\ &\geq a_2 q \cdot \left(\Omega(Q^2) - \sum_{f \in L_{i,j}} \frac{|m(D_i^f) - m(D_j^f)|^2}{\max_f |m(D_i^f) - m(D_j^f)|^2} \right) \\ &\quad + a_3 q \cdot \sum_{f \in L_{i,j}} \frac{|m(D_i^f) - m(D_j^f)|^2}{R^2} \\ &\geq a_4 q \cdot \Omega(Q^2) \end{aligned}$$

Here, a_3 is the constant in Lemma 23 and $a_4 = \min(a_2, 64a_3)$. Since for $f \in L_{ij}$, $\max_f |m(D_i^f) - m(D_j^f)|^2 \leq 64R^2$, we can write:

$$||\tilde{\mu}_i - \tilde{\mu}_j||^2 \geq \Omega(Q^2) \cdot \max_f ||\tilde{\mu}_i^f - \tilde{\mu}_j^f||^2$$

The first condition follows from the second condition. □

5.3 Applications to Learning Mixtures

5.3.1 Clustering Distributions with Heavy Tails

In this section, we present Algorithm HT-Correlations –a modified version of Algorithm Correlation-Cluster in [5], which can successfully learn mixtures of

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 $q = 4\sqrt{n}\log n \log T$. 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 [2] to partition the points in $\tilde{S}_A$ and $\tilde{S}_B$ after projection.

Figure 5.3: Algorithm Using Correlations

heavy-tailed product distributions. The input to the algorithm is a set S of s samples, and the output is a partitioning of the samples. The algorithm is described in Figure 5.3.

The properties of Algorithm HT-Correlations are described in Theorem 5. This section is devoted to proving Theorem 5. The proof proceeds in three steps. First, we deduce from Theorem 4 that if the distributions satisfy the conditions in Theorem 5, then the transformed distributions satisfy the separation and spreading requirements of Theorem 1 in Chapter 3. We can now 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 28 and 29 to show that distance concentration algorithms work in these

output subspaces.

Lemma 28 *For any i, the maximum variance of $\tilde{D}_i$ in any direction is at most q.*

PROOF: Let v be any unit vector in the transformed space. The variance of $\tilde{D}_i$ along v can be written as:

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

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 q coordinates corresponding to a single f. □

The following lemma shows that the distance concentration algorithm works correctly as long as the separation between the transformed centers is high enough, for every i and j.

Lemma 29 *Let $\mathcal{H}$ be a d-dimensional subspace of $\{0,1\}^{nq}$. Then for any i,*

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

PROOF: Let $v_1, \ldots, v_d$ 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. □

PROOF:(Of Theorem 5) From Theorem 4 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 $\Omega\left(\sqrt{qT \log \Lambda} + \sqrt{qT \log(T/\delta)}\right)$. Note that the proof of Theorem 1 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 to conclude that for each i and j, there exists some constant a such that:

$$d_{\mathcal{K}_B}(\tilde{\mu}_i, \tilde{\mu}_j) \geq \Omega(d(\tilde{\mu}_i, \tilde{\mu}_j)) \geq a\left(\sqrt{qT \log \Lambda} + \sqrt{qT \log(T/\delta)}\right)$$

As $\mathcal{K}_B$ has dimension at most $2T$, from Lemma 29, any two points drawn from a distribution D_i have distance at most $2\sqrt{qT \log(T/\delta)}$ in the subspace $\mathcal{K}_B$. On the other hand, a point drawn from D_i and a point drawn from D_j are at least $(a_1 - 2)\sqrt{2qT \log(T/\delta)}$ apart in $\mathcal{K}_B$. Algorithm HT-CORRELATIONS therefore works. □

We can also use our embedding along with SVD-based algorithms for clustering mixtures of heavy-tailed distributions. We present Algorithm HT-SVD –a simple singular value decomposition based algorithm which succeeds in correctly clustering mixtures of heavy-tailed distributions. The algorithm, which is very similar to the algorithm of [1; 22; 17], is described in Figure 5.4. The input to the algorithm is a

HT-SVD(S)

1. Let $R_1 = 26R$, $R_2 = R$, and $q = 4\sqrt{n}\log n \log T$. 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 $\{v_{1,A}, \ldots, v_{T,A}\}$ (resp. $\{v_{1,B}, \ldots, v_{T,B}\}$) 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 $v_{1,A}, \ldots, v_{T,A}$ (resp. $v_{1,B}, \ldots, v_{T,B}$).

4. Use a distance-based clustering algorithm as in [2] to partition the points in $\tilde{S}_A$ and $\tilde{S}_B$ after projection.

Figure 5.4: Algorithm Using SVDs

set S of s samples, and the output is a partitioning of the samples.

The properties of Algorithm HT-SVD are summarized by Theorem 9, which is stated and proved below. The main components of the proof are Theorem 10, which states that if the transformed distributions are sufficiently far apart in space, then the algorithm of [1] successfully clusters the samples, and Theorem 4, which states that if the original distributions are far apart in space, then so are the transformed distributions.

Theorem 9 *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), and the following conditions hold.*

1. *For any i and j,*

$$||m(D_i) - m(D_j)|| \geq \Omega\left(R \cdot (\sqrt{w_i} + \sqrt{w_j})^{-1} + R \cdot \sqrt{T\log(T/\delta)}\right)$$

2. For any i and j,

$$||m(D_i) - m(D_j)|| \geq \Omega((\sqrt{w_i} + \sqrt{w_j})^{-1} + \sqrt{T\log(T/\delta)}) \max_f |m(D_i^f) - m(D_j^f)|$$

Then, there exists an algorithm, which runs in time polynomial in n and clusters $1-\delta$ fraction of the samples correctly. The number of samples required by our algorithm is $O(\frac{n^{3/2}T\log T}{w_{\min}})$.

Theorem 10 *Let $q = 4\sqrt{n}\log n \log T$, and let $\tilde{\mathcal{D}} = \{\tilde{D}_1, \ldots, \tilde{D}_T\}$ be a mixture of distributions on $\{0,1\}^{\Theta(nq)}$ obtained by applying the embedding Φ on a mixture of product distributions. If, for all i, j,*

$$||\tilde{\mu}_i - \tilde{\mu}_j|| \quad \geq \quad \Omega\left(\sqrt{q}\cdot(\sqrt{w_i} + \sqrt{w_j})^{-1} + \sqrt{qT\log(T/\delta)}\right)$$

Then, the algorithm in Figure 5.4 produces a correct partitioning of $1-\delta$ fraction of the samples.

The following theorem, inspired by [1], 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.

Theorem 11 *Let, for each i, $c_{i,A}$ be the empirical centers of $\tilde{D}_i$ computed from the points in $\tilde{S}_A$. Then,*

$$||\mathbf{P}_{\mathcal{K}_B}(c_{i,A} - c_{j,A})|| \geq ||c_{i,A} - c_{j,A}|| - \sqrt{q}\cdot(\sqrt{w_i} + \sqrt{w_j})^{-1}$$

The proof follows from the proof of Theorem 1 in [1] and Lemma 28.

PROOF:(Of Theorem 10) From Theorem 4 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

$$\Omega\left(\sqrt{q}(\sqrt{w_i} + \sqrt{w_j})^{-1} + \sqrt{qT\log(T/\delta)}\right)$$

Therefore, from Theorem 11,

$$||\mathbf{P}_{\mathcal{K}_B}(c_{i,A} - c_{j,A})|| \geq ||c_{i,A} - c_{j,A}|| - \sqrt{q} \cdot (\sqrt{w_i} + \sqrt{w_j})^{-1}$$

where $c_{i,A}$ and $c_{j,A}$ are the empirical centers of the transformed distributions. As $\mathcal{K}_B$ has dimension at most T, from Lemma 29, any two points drawn from a distribution D_i have distance at most $2\sqrt{qT\log(T/\delta)}$ in the subspace $\mathcal{K}_B$. On the other hand, for some constant a', a point drawn from D_i and a point drawn from D_j are at least $a'\sqrt{qT\log(T/\delta)}$ apart in $\mathcal{K}_B$. Algorithm HT-SVD therefore works for $a' > 2$. □

5.3.2 Clustering Distributions with Finite Variance

In this section, we present Algorithm Low Separation Cluster –a small variant of Algorithm HT-Correlations, which learns mixtures of product distributions with finite variance. Unlike Algorithm Correlation-Cluster, Algorithm Low Separation Cluster requires the centers of the distributions to be separated only by $\Omega(\sigma_*\sqrt{T\log\Lambda})$, where σ_* is the maximum directional variance of any distribution in the mixture in the space containing the centers. Unlike Algorithm Low-Variance Cluster, Algorithm Low Separation Cluster only requires time polynomial in n, T, and the number of samples. The algorithm is described below.

The properties of Algorithm Low Separation Cluster are summarized in Theorem 6, which is proved below.

Proof:(Of Theorem 6) For a fixed pair i and j, let $\nu_{ij} = \mu_i - \mu_j$. If the standard deviation of distribution D_i along coordinate f is σ_i^f, then, the variance of D_i along ν_{ij} is $\frac{\sum_f (\nu_{ij}^f)^2(\sigma_i^f)^2}{||\nu_{ij}||^2}$. Since σ_*^2 is the maximum variance of any distribution in the mixture along the space containing the centers,

$$\frac{\sum_f (\nu_{ij}^f)^2(\sigma_i^f)^2}{||\nu_{ij}||^2} \leq \sigma_*^2$$

Let $\mathcal{F}_i^-$ (resp. $\mathcal{F}_j^-$) be the set of coordinates for which $\sigma_i^f \leq 2\sigma_*$ (resp. $\sigma_j^f \leq 2\sigma_*$).

LOW-SEPARATION CLUSTER(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 $R_1 = 52\sigma_*$, $R_2 = 16\sigma_*$, and $q = 4\sqrt{n}\log n \log T$. 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 [2] to partition the points in $\tilde{S}_A$ and $\tilde{S}_B$ after projection.

Figure 5.5: Algorithm LOW SEPARATION CLUSTER

Then,

$$\begin{aligned}\frac{\sum_{f\in\mathcal{F}_i^-}(\nu_{ij}^f)^2}{||\nu_{ij}||^2} &\geq \frac{3}{4}\\ \frac{\sum_{f\in\mathcal{F}_j^-}(\nu_{ij}^f)^2}{||\nu_{ij}||^2} &\geq \frac{3}{4}\end{aligned}$$

Since $\sum_f(\nu_{ij}^f)^2 = ||\nu_{ij}||^2$,

$$\frac{\sum_{f\in\mathcal{F}_i^-\cap\mathcal{F}_j^-}(\nu_{ij}^f)^2}{||\nu_{ij}||^2} \geq \frac{1}{2}$$

Therefore, from Condition (1),

$$\sum_{f\in\mathcal{F}_i^-\cap\mathcal{F}_j^-}(\mu_i^f-\mu_j^f)^2 \geq \frac{1}{2}\sum_f(\mu_i^f-\mu_j^f)^2 \geq \Omega\left(\sigma_*\sqrt{T\log\Lambda}+\sigma_*\sqrt{T\log(T/\delta)}\right)$$

From Condition (2),

$$\begin{aligned}\sum_{f\in\mathcal{F}_i^-\cap\mathcal{F}_j^-}(\mu_i^f-\mu_j^f)^2 &\geq \frac{1}{2}\sum_f(\mu_i^f-\mu_j^f)^2\\ &\geq \Omega\left(\sqrt{T\log\Lambda}+\sqrt{T\log(T/\delta)}\right)\cdot\max_f(\mu_i^f-\mu_j^f)^2\end{aligned}$$

Since the $\frac{3}{4}$-radius of any distribution is at most 4 times its standard deviation, we can apply Theorem 4 with parameter $4\sigma_*$ to conclude that:

$$||\tilde{\mu}_i-\tilde{\mu}_j|| \geq \Omega(\sqrt{qT\log\Lambda}+\sqrt{qT\log(T/\delta)})$$

The rest of the proof is similar to the proof of Theorem 5. □

5.4 Discussions

Spreading Condition. To show Theorems 5, 6 and 9, we need a spreading condition, which states that for all i and j, $\sum_f|\mu_i^f-\mu_j^f|^2 \geq \Omega(T\log\Lambda)\max_f|\mu_i^f-\mu_j^f|^2$. This essentially means that the distance between any pair of centers is spread out along $T\log\Lambda$ coordinates, and no one coordinate contributes too much to the distance. This spreading condition is stronger than the spreading condition we use in Chapter 3,

which allows a few coordinates to have a high contribution to the distance between some pair of centers.

We also observe that Theorems 5 and 9 work with a spreading condition analogous to the condition in Chapter 3. The algorithms corresponding to these theorems work correctly so long as there is sufficient distance between the transformed centers to allow distance concentration. This is achieved only if there is sufficient distance between the medians of the original distributions, and this distance comes from $\Omega(T \log \Lambda)$ coordinates, regardless of whether a few coordinates has very high contribution.

Theorem 6 however requires the stronger spreading condition, which implies that the *variance* of any distribution in the subspace containing the centers is also spread out along several coordinates. This ensures that there are about $\Theta(T \log \Lambda)$ coordinates along which some pair of distributions have variance comparable to σ_*, and the embedding provides the desired results.

5.5 Related Work

Dasgupta *et. al* [6] 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 two classes of heavy-tailed distributions. The first class of distributions, which they call $\mathcal{F}_0(R')$, is the class of all product distributions in which the distribution of each coordinate has the following properties:

1. Symmetry around the median.
2. Decreasing probability density with distance from the median.
3. Half-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. We do not require the second

assumption of [6], that is the assumptions of decreasing probability density from the center. However, the condition that the $\frac{3}{4}$-radius is upper bounded by R, is stronger than assumption (3) of [6].

[6] provide two algorithms for learning mixtures of distributions in $\mathcal{F}_0(R')$.

1. An algorithm which works with a separation of $\Omega(R'\sqrt{T})$ and a spreading condition that the distance between the medians of any two distributions in the mixture should be spread over $\Theta(T)$ coordinates. 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 $\Theta(\frac{n\log(nT)}{w_{\min}})$.

2. An algorithm which works with a separation 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 $\Theta(T)$ coordinates. The 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 $\Theta(\frac{\log(nT)}{w_{\min}})$, which may be polynomial in n but 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 [6].

In addition, [6] works with a second class of distributions, called $\mathcal{F}_1(R')$. $\mathcal{F}_1(R')$ includes all distributions D in $\mathcal{F}_0(R')$ in which

$$\forall \alpha \geq 1, \Pr_{x\sim D}[|x-\mu| > \alpha R'] \leq \frac{1}{2\alpha R'}$$

They provide an algorithm which clusters correctly a $1-\delta$ fraction of the samples from a mixture of distributions in $\mathcal{F}_1(R')$ so long as the separation between any two centers is $\Omega(\frac{RT^{5/2}}{\delta^2})$.

5.6 Conclusions and Open Problems

In summary, in this chapter, we provide an embedding from $\mathbf{R}^n$ to $\{0,1\}^{n'}$ where $n > n'$. This embedding has the property that samples from two distributions on $\mathbf{R}^n$ which have medians that are far apart, map to samples from distributions on $\{0,1\}^{n'}$ with centers that are far apart. We show applications of our embedding in designing fast algorithms for learning mixtures of heavy-tailed product distributions, as well as for learning mixtures of distributions with finite variance and low separation.

We leave open two major questions. First, how can we learn mixtures of heavy-tailed product distributions better? From Lemmas 16 and 22, the separation condition seems necessary. However, do we really need the spreading condition?

Problem 5 *Can we design an algorithm for learning mixtures of heavy-tailed product distributions without the spreading condition? Can the spreading condition be relaxed?*

Secondly, here we use an embedding to convert a learning problem on $\mathbf{R}^n$ to an easier learning problem on $\{0,1\}^{n'}$. Can our embedding be used in other such learning problems?

Problem 6 *Can our embedding be applied to reduce other learning problems on* $\mathbf{R}^n$ *to problems on* $\{0,1\}^{n'}$*?*

Bibliography

[1] 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. 7, 12, 13, 21, 22, 42, 47, 64, 66, 84, 85, 86

[2] S. Arora and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of 33rd ACM Symposium on Theory of Computing*, pages 247–257, 2001. 10, 22, 49, 82, 85, 88

[3] K. Chaudhuri, E. Halperin, S. Rao, and S. Zhou. A rigorous analysis of population stratification with limited data. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2007. 8, 17

[4] K. Chaudhuri and S. Rao. Beyond gaussians: Spectral methods for learning mixtures of heavy-tailed product distributions. unpublished manuscript, 2007. 8

[5] K. Chaudhuri and S. Rao. Learning mixtures of distributions using correlations and independence. unpublished manuscript, 2007. 7, 8, 66, 67, 81

[6] 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. 6, 90, 91

[7] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer S cience*, pages 634–644, 1999. 3, 9, 10, 14

[8] S. Dasgupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000. 10

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society B, 39*, pages 1–38, 1977. 10

[10] J. Pritchard et al. Structure 2.1 software. http://pritch.bsd.uchicago.edu/structure.html. 5, 13, 16

[11] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proceedings of FOCS*, 2005. 13, 14

[12] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of gaussians with no separation assumptions. In *Proceedings of COLT*, 2006. 13, 14

[13] W. Feller. *An Introduction to Probability Theory and Its Application: Volumes I and II.* New York: John Wiley and Sons, 1971. 100

[14] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1999. 13, 14

[15] G. Golub and C. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1996. 97

[16] Piotr Indyk. Algorithmic applications of low-distortion geometric embeddings. In *FOCS*, pages 10–33, 2001. 67

[17] 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. 7, 12, 13, 22, 47, 66, 84

[18] M. Ledoux. *The Concentration of Measure Phenomenon*. Americal Mathematical Society, 2000. 103

[19] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer S cience*, pages 529–537, 2001. 42

[20] A. Panconesi and D. Dubhashi. Concentration of measure for the analysis of randomised algorithms. Draft, 2005. 57, 100

[21] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:954–959, June 2000. 5, 13, 16

[22] 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. 12, 22, 44, 47, 84

Appendix A

Linear Algebra and Statistics

We introduce some basic facts from linear algebra and statistics, which are used throughout the thesis.

A.1 Singular Value Decomposition

Singular Value Decomposition. For a $m \times n$ matrix A, a *singular value decomposition* or SVD of A is a decomposition of A into three matrices U, Σ, and V such that:

$$A = U^{\mathbf{T}} \Sigma V$$

Here, U is a $m \times m$ matrix, Σ is $m \times n$, and V is a $n \times n$ matrix. Σ is a diagonal matrix, and U and V are unitary matrices, that is the rows and columns of U and V form orthonormal bases of $\mathbf{R}^m$ and $\mathbf{R}^n$ respectively.

The diagonal entries of Σ are called *singular values* and the columns of U and V are called the *left singular vectors* and *right singular vectors* respectively.

A basic fact of linear algebra is that the SVD of every matrix exists and is unique, except for permutations of the singular value and singular vectors.

A notion closely related to singular value decompositions is an eigen-decomposition.

Eigen-Decomposition. For a $n \times n$ symmetric matrix A, an *eigen-decomposition* is a decomposition of A into two matrices U and D such that:

$$A = U^{\mathbf{T}} D U$$

Here D is a diagonal matrix and U is unitary. The diagonal entries of D are called the *eigenvalues* of A, and the columns of U are called the *eigenvectors*.

Now we present two theorems of linear algebra connecting eigen-decompositions and singular value decompositions.

Theorem 12 *For a matrix A, the eigenvalues of $AA^{\mathbf{T}}$ are the squares of the corresponding singular values of A, and the eigenvectors of $AA^{\mathbf{T}}$ are the corresponding left singular vectors of A.*

PROOF: If $A = U^{\mathbf{T}}\Sigma V$, then, $AA^{\mathbf{T}} = U^{\mathbf{T}}\Sigma VV^{\mathbf{T}}\Sigma^{\mathbf{T}}U$. As V is unitary, $VV^{\mathbf{T}} = I$. Therefore,

$$AA^{\mathbf{T}} = U^{\mathbf{T}}\Sigma\Sigma^{\mathbf{T}}U = U^{\mathbf{T}}\Sigma^2 U$$

where $\Sigma^2 = \Sigma\Sigma^{\mathbf{T}}$ is a diagonal matrix such that $\Sigma^2_{ii} = (\Sigma_{ii})^2$. The theorem follows from the uniqueness of the eigen-decompositions and singular value decompositions. □

Theorem 13 *If A is a symmetric matrix, then the eigenvectors of A are the corresponding singular vectors of A.*

The theorem follows from the uniqueness of the singular value and eigen-decompositions.

A.2 Matrix Norms

L_2 Norm. For a symmetric matrix A, the L_2 norm is defined as its top eigenvalue. We use the notation $||A||$ to denote the L_2 norm of a matrix A.

Frobenius Norms. For any matrix M, the *Frobenius norm* written $|M|_{\mathbf{F}}$ is defined as $\sqrt{\sum_{i,j} M^2_{ij}}$. We write $|A|_{\mathbf{F}}$ to denote the Frobenius norm of a matrix A. The following theorem relates the Frobenius norm, rank and top singular value of a matrix.

Theorem 14 ([15]) *For a matrix M of rank r, the top singular value is greater than or equal to $|M|_{\mathbf{F}}/\sqrt{r}$.*

Rank k Approximation. The rank k approximation to a matrix A is the projection of the matrix A onto the span of its top k left singular vectors.

Theorem 15 *If A_k is the rank k approximation to a matrix A, and B is any rank k matrix, then,*

$$||A - A_k|| \leq ||A - B||$$

A.3 Basic Statistics

Variance. Let X be a random variable on $\mathbf{R}$. Then, the variance of X, denoted by $\mathbf{Var}(X)$, is defined as:

$$\mathbf{Var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

Covariance. Let X and Y be two random variables on $\mathbf{R}$. Then, the covariance of X and Y, denoted by $\mathbf{Cov}(X, Y)$ is defined as:

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])]$$

Covariance Matrix. Let $X = (X_1, \ldots, X_n)$ be a vector on $\mathbf{R}^n$ distributed according to some distribution D. Then, the covariance matrix of X is a matrix M with entries as follows:

$$\begin{aligned} M_{ij} &= \mathbf{Cov}(X_i, X_j), \text{ if} i \neq j \\ M_{ii} &= \mathbf{Var}(X_i) \end{aligned}$$

Note that the covariance matrix is symmetric.

Directional Variance. Let D be a distribution on $\mathbf{R}^n$. For a unit vector v, the directional variance of D along v is the variance of the distribution D_v induced by the projection of D onto v.

Theorem 16 *Let M be the covariance matrix of a distribution D on $\mathbf{R}^n$ and v be a unit vector in $\mathbf{R}^n$. Then, the directional variance of D along v is $v^{\mathbf{T}} M v$.*

PROOF: Let X be a vector in $\mathbf{R}^n$ distributed according to D and let $\mathbf{E}[X] = (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n])$ be the center of D. Then the directional variance of D along v can be written as:

$$\begin{aligned}
\mathbf{E}[\langle v, X - \mathbf{E}[X]\rangle^2] &= \mathbf{E}[\sum_{i,j} v_i v_j (X_i - \mathbf{E}[X_i]) \cdot (X_j - \mathbf{E}[X_j]) \\
&= \sum_{i,j} v_i v_j \mathbf{E}[(X_i - \mathbf{E}[X_i]) \cdot (X_j - \mathbf{E}[X_j])] \\
&= \sum_{i,j,i\neq j} v_i v_j \mathbf{Cov}(X_i, X_j) + \sum_i v_i^2 \mathbf{Var}(X_i) \\
&= \sum_{i,j} v_i v_j M_{ij} \\
&= v^{\mathbf{T}} M v
\end{aligned}$$

□

Theorem 17 *For a distribution D with covariance matrix M, the top k eigenvectors of M form a basis for the k-dimensional subspace along which D has the highest variance.*

This theorem follows from Theorem 16.

Appendix B

Concentration of Measure Inequalities

In this section, we briefly survey a few concentration of measure inequalities which are used extensively throughout our thesis. The proofs can be found in standard texts for probability- see, for example [13; 20].

B.1 Chernoff and Hoeffding Bounds

The most popular concentration of measure inequality used in computer science literature is the Chernoff Bound, which relates to sums of independent 0/1 random variables.

Theorem 18 (Chernoff Bounds) *Let* $X = X_1 + \ldots + X_n$ *where* $X_1, \ldots, X_n$ *are independent* 0/1 *random variables. If* $\mathbf{E}[X] = \mu$*, then,*

$$\begin{aligned} \Pr[X < (1-\delta)\mu] &\leq e^{-\delta^2\mu/2} \\ \Pr[X > (1+\delta)\mu] &\leq e^{-\delta^2\mu/3}, \textit{for } \delta \leq 1 \\ &\leq e^{-\delta^2\mu/(1+\delta)}, \textit{for } \delta > 1 \end{aligned}$$

The Hoeffding Bound extends the inequality to random variables with a range of $[0, 1]$, which may not necessarily be discrete.

Theorem 19 (Hoeffding Bounds) *Let $X = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are independent random variables with range $[0, 1]$. If $\mathbf{E}[X] = \mu$, then,*

$$\begin{aligned}
\Pr[X < (1-\delta)\mu] &\leq e^{-\delta^2\mu/2} \\
\Pr[X > (1+\delta)\mu] &\leq e^{-\delta^2\mu/3}, \textit{for } \delta \leq 1 \\
&\leq e^{-\delta^2\mu/(1+\delta)}, \textit{for } \delta > 1
\end{aligned}$$

B.2 Method of Bounded Differences

The Method of Bounded Differences is a concentration inequality similar to the Chernoff Bounds, but which applies to any arbitrary functions of random variables and not necessarily sums.

Theorem 20 (Method of Bounded Differences) *Let $X_1, \ldots, X_n$ be arbitrary independent random variables and let f be any function of $X_1, \ldots, X_n$. For each i, if for any a and a',*

$$|f(X_1, \ldots, X_i = a, \ldots, X_n) - f(X_1, \ldots, X_i = a', \ldots, X_n)| \leq \gamma_i$$

and $\gamma = \sum_i \gamma_i^2$, then,

$$\Pr[|f(X_1, \ldots, X_n) - \mathbf{E}[f(X_1, \ldots, X_n)]| > t] \leq 2e^{-t^2/2\gamma}$$

The Method of Averaged Bounded Differences applies to situations where the worst case difference when we change the value of a certain X_i can be high, but the average difference is low.

Theorem 21 (Method of Averaged Bounded Differences) *Let $X_1, \ldots, X_n$ be arbitrary independent random variables and let f be any function of $X_1, \ldots, X_n$. For each i, if for any a and a',*

$$|\mathbf{E}[f(X_1, \ldots, X_i = a, \ldots)|X_1, \ldots, X_{i-1}] - \mathbf{E}[f(X_1, \ldots, X_i = a', \ldots)|X_1, \ldots, X_{i-1}]| \leq \gamma_i$$

and $\gamma = \sum_i \gamma_i^2$, then,

$$\Pr[|f(X_1, \ldots, X_n) - \mathbf{E}[f(X_1, \ldots, X_n)]| > t] \leq 2e^{-t^2/2\gamma}$$

Note that the expectation in Theorem 21 is taken over $X_{i+1}, \ldots, X_n$ for any worst case configuration of $X_1, \ldots, X_{i-1}$.

B.3 Method of Bounded Variances

The method of Bounded Variances applies to situations when the worst case difference when we change the value of a certain variable X_i is high, but the worst case happens rarely.

Theorem 22 (Method of Bounded Variances) *Let $X_1, \ldots, X_n$ be arbitrary independent random variables and let f be any function of $X_1, \ldots, X_n$. For each i, if for any a and a',*

$$|\mathbf{Var}(f(X_1, \ldots, X_i = a, \ldots, X_n) - f(X_1, \ldots, X_i = a', \ldots, X_n))| \leq \sigma_i^2$$

and $s_n^2 = \sum_i \gamma_i^2$, and for any i,

$$|f(X_1, \ldots, X_i = a, \ldots, X_n) - f(X_1, \ldots, X_i = a', \ldots, X_n)| \leq \beta$$

then,

$$\Pr[|f(X_1, \ldots, X_n) - \mathbf{E}[f(X_1, \ldots, X_n)]| > t] \leq 2e^{-t^2/2s_n^2(1+\beta t/3s_n^2)}$$

B.4 The Berry-Esseen Central Limit Theorem

A related inequality is the Berry-Esseen Central Limit Theorem.

Theorem 23 (Berry-Esseen Central Limit Theorem) *Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbf{E}[X_i] = 0$, $\mathbf{E}[X_i^2] = \sigma_i^2$ and $\mathbf{E}[|X_i|^3] = \rho_i$. Also let $s_n^2 = \sum_i \sigma_i^2$ and $r_n = \sum_i \rho_i$, and then $F_n(t)$ denote the distribution function of $\sum_i X_i/s_n$. Then, for any t,*

$$|F_n(t) - \mathcal{N}(t)| \leq \frac{6r_n}{s_n^3}$$

where $\mathcal{N}(t)$ is the distribution function of the standard normal distribution.

B.5 Gaussian Concentration of Measure

A concentration inequality for Gaussians, related to the Method of Bounded Variances, is the following theorem.

Theorem 24 *[18] Let $F(X_1, \ldots, X_n)$ be any function of independent Gaussian variables $X_1, \ldots, X_n$, such that $\nabla F \leq \gamma$, for any value of $X_1, \ldots, X_n$. Then,*

$$\Pr[|F(t) - \mathbf{E}[F(t)]| > t] \leq e^{-t^2/\gamma}$$