On Spectral Learning of Mixtures of Distributions

Dimitris Achlioptas[1] and Frank McSherry[2]

[1] Microsoft Research, One Microsoft Way, Redmond WA 98052, USA
[2] Microsoft Research, 1065 La Avenida, Mountain View CA 94043, USA
{optas, mcsherry}@microsoft.com

Abstract. We consider the problem of learning mixtures of distributions via spectral methods and derive a characterization of when such methods are useful. Specifically, given a mixture-sample, let $\overline{\mu}_i, \overline{C}_i, \overline{w}_i$ denote the empirical mean, covariance matrix, and mixing weight of the samples from the i-th component. We prove that a very simple algorithm, namely spectral projection followed by single-linkage clustering, properly classifies every point in the sample provided that each pair of means $\overline{\mu}_i, \overline{\mu}_j$ is well separated, in the sense that $\|\overline{\mu}_i - \overline{\mu}_j\|^2$ is at least $\|\overline{C}_i\|_2(1/\overline{w}_i + 1/\overline{w}_j)$ plus a term that depends on the concentration properties of the distributions in the mixture. This second term is very small for many distributions, including Gaussians, Log-concave, and many others. As a result, we get the best known bounds for learning mixtures of arbitrary Gaussians in terms of the required mean separation. At the same time, we prove that there are many Gaussian mixtures $\{(\mu_i, C_i, w_i)\}$ such that each pair of means is separated by $\|C_i\|_2(1/w_i + 1/w_j)$, yet upon spectral projection the mixture collapses completely, i.e., all means and covariance matrices in the projected mixture are identical.

Keywords: learning mixtures of distributions, spectral methods, singular value decomposition, gaussians mixtures, log-concave and concentrated distributions.

1 Introduction

A mixture of k distributions $D_1, \ldots, D_k$ with mixing weights $w_1, \ldots, w_k$, where $\sum_i w_1 = 1$, is the distribution in which each sample is drawn from D_i with probability w_i. Learning mixtures of distributions is a classical problem in statistics and learning theory (see [4, 5]). Perhaps the most studied case is that of learning Gaussian mixtures. In such a mixture, each constituent distribution is a multivariate Gaussian, characterized by a mean vector $\mu_i \in \mathbb{R}^d$ and an arbitrary covariance matrix $C_i = R_i R_i^T \in \mathbb{R}^{d \times d}$. That is, a sample from the i-th Gaussian is a vector $\mu_i + R_i x$, where $x \in \mathbb{R}^d$ is a vector whose components are i.i.d. $N(0,1)$ random variables. We let $\sigma_i^2 = \|C_i\|$ denote the maximum directional variance of each Gaussian, where $\|\cdot\|$ denotes the matrix spectral norm.

P. Auer and R. Meir (Eds.): COLT 2005, LNAI 3559, pp. 458–469, 2005.

We begin with discussing some earlier works on learning Gaussian mixtures, which serve as the motivation (and canonical model) for our work. A generally fruitful approach to learning mixtures of Gaussians is to start by projecting the samples onto a low dimensional space. This idea, originated in non-parametric statistics in the 60s, is motivated by the fact that reducing the dimensionality of the host space, dramatically reduces the number of potential component separators, thus affording a more complete search among them. Moreover, it is well-known that the projection of a Gaussian mixture onto a fixed subspace is also a Gaussian mixture, one in which the means and mixing weights behave in the obvious way, while the covariance matrices get transformed to new matrices of no greater maximum directional variance.

Dasgupta [2] pioneered the idea of projecting Gaussian mixtures onto *random* low-dimensional subspaces. For a typical subspace, the separation of each mean μ_i from the other means shrinks at the same rate as $\mathbf{E}[\|R_i x\|^2]$, i.e., in proportion to the reduction in dimension. Thus, the random projection's main feature is to aid clustering algorithms that are exponential in the dimension. But, in order for a mixture to not collapse under a typical projection the separation between means μ_i, μ_j needs to grow as $(\sigma_i + \sigma_j) \times d^{1/2}$, i.e., not only must the Gaussians not touch but, in fact, they must be pulled further and further apart as their dimensionality grows.

In [3], Dasgupta and Schulman reduced this requirement to $(\sigma_i+\sigma_j)\times d^{1/4}$ for spherical Gaussians by showing that, in fact, under this conditions the EM algorithm can be initialized so as to learn the μ_i in only two rounds. Arora and Kannan [1] combined random projections with sophisticated distance-concentration arguments in the context of learning mixtures of general Gaussians. In their work, the separation of means is not the only relevant parameter and their results apply to many cases where a worst-case mixture with the given separation characteristics is not learnable by any algorithm. That said, the worst case separation required by the results in [1] is also $(\sigma_i + \sigma_j) \times d^{1/4}$.

Rather than projecting the mixture onto a random subspace, we could dream of projecting it onto the subspace spanned by the mean vectors. This would greatly enhance the "contrast" in the projected mixture since $\mathbf{E}[\|R_i x\|^2]$ is reduced as before, but the projected means remain fixed and, thus, at the same distance. Recently, Vempala and Wang [6] did just this, by exploiting the fact that in the case of *spherical* Gaussians, as the number of samples grows, the subspace spanned by the top singular vectors of the data set converges to the subspace spanned by the mean vectors. This allowed them to give a very simple and elegant algorithm for learning spherical Gaussians which works as long as each pair of means μ_i, μ_j is separated by $(\sigma_i + \sigma_j) \times k^{1/4}$, i.e., a length independent of the original dimensionality.

Unfortunately, for non-spherical Gaussians the singular vector subspace does not in general convergence to the subspace spanned by the means. Vempala and Wang [6] observed this and asked if spectral projections can be useful for distributions that are not weakly isotropic, e.g. non-spherical Gaussians. In recent related work, Kannan, Salmasian, and Vempala [8] show how to use spectral

projections to learn mixtures of Log-concave distributions in which each pair of means μ_i, μ_j is separated by, roughly, $k^{3/2}(\sigma_i + \sigma_j)/w_{\min}^2$.

Here, we show that combining spectral projection with single-linkage clustering gives a method for recursively dissecting mixtures of concentrated distributions, e.g., Gaussian mixtures, when each pair of means μ_i, μ_j in the mixture is separated by $(\sigma_i+\sigma_j)(1/w_i+1/w_j)^{1/2}$, plus a term describing the concentration of the constituent distributions. For example, for Gaussian mixtures this second term is of order $(\sigma_i + \sigma_j)(k + (k \log n)^{1/2})$.

At the same time, we also provide a lower bound that demonstrate that for spectral projection, separation in excess of $(\sigma_i + \sigma_j)(1/w_i + 1/w_j)^{1/2}$ is mandatory. That is, we prove that for any set of mixing weights $w_1, \ldots, w_k$, there is an arrangement of *identical* Gaussians, with maximal directional variance σ where every pair of means μ_i, μ_j is separated by $\sigma(1/w_i + 1/w_j)^{1/2}$, yet upon spectral projection the mixture collapses completely, i.e., all means and covariance matrices in the projected mixture are identical. Thus, with the exception of the concentration term, our upper and lower bounds coincide.

We should mention briefly an important difference of our approach as compared to much previous work. Given as input some $k' \geq k$, our algorithm recursively subdivides the data set using cuts that respect mixture boundaries whenever applied to mixtures with at least two components present. Unlike previous work, our algorithm does not terminate with a partition of the input samples into k sets, but instead continues subdivision, returning a hierarchy that describes many legitimate k-partitions for varying values of k. This subdivision tree admits simple dynamic programming algorithms that can reconstruct k-partitions minimizing a variety of loss functions, which we discuss later in further detail. Importantly, it gives the flexibility to explore several values of k in a uniform manner. Computing each cut in this tree reduces to computing the top k singular vectors of a sample submatrix followed by a Minimum Spanning Tree computation on the corresponding projected sample. As a result, a naive implementation of our algorithm runs in time $O(kd^3n^2)$. If one is a bit more careful and performs the MST computations on appropriately large subsamples, the running time becomes linear in n, specifically $O(n(k^2d^2 + d^3)/w_{\min})$.

2 Our Techniques and Results

From this point on, we adopt the convention of viewing the data set as a collection of samples with hidden labels, rather than samples from a pre-specified mixture of distributions. This will let us describe sufficient conditions for correct clustering that are independent of properties of the distributions. Of course, we must eventually determine the probability that a specific distribution yields samples with the requisite properties, but deferring this discussion clarifies the results, and aids in their generality.

Our exposition uses sample statistics: $\overline{\mu}_i$, $\overline{\sigma}_i$, and $\overline{w}_i$. These are the empirical analogues of μ_i, σ_i, and w_i, computed from a $d \times n$ matrix of labeled samples A. We also use $\overline{n}_i$ to denote the number of samples in the mixture with label i.

The advantages of sample statistics are twofold: i) they allow for more concise and accurate proofs, and ii) they yield pointwise bounds that may be applied to arbitrary sets of samples. We will later discuss the convergence of the sample statistics to their distributional equivalents, but for now the reader may think of them as equivalent.

2.1 Spectral Projection and Perturbation

We start our analysis with an important tool from linear algebra: the optimal rank k column projection. For every matrix A and integer k, there exists a rank k projection matrix P_A such that for any other matrix X of rank at most k,

$$\|A - P_A A\| \leq \|A - X\| \ . \tag{1}$$

The matrix P_A is spanned by the top k left singular vectors of A, read from A's singular value decomposition.

Our key technical result is that the sample means $\overline{\mu}_i$ are only slightly perturbed when projected through P_A. We use the notation $\overline{\sigma}^2 = \sum_i \overline{w}_i \overline{\sigma}_i^2$ for the weighted maximum directional variance.

Theorem 1. *For any set A of labeled samples, for all i, $\|\overline{\mu}_i - P_A\overline{\mu}_i\| \ \leq \ \overline{\sigma}/\overline{w}_i^{1/2}$.*

Proof. Let $x_i \in \{0, 1/\overline{n}_i\}^n$ be the scaled characteristic vector of samples in A with label i, i.e., $x_i^q = 1/\overline{n}_i$ iff the q-th sample has label i. Thus, $\overline{\mu}_i = Ax_i$ and

$$\|\overline{\mu}_i - P_A\overline{\mu}_i\| \ = \ \|(A - P_A A)x_i\| \ \leq \ \|A - P_A A\| \, \|x_i\| \ \leq \ \|A - P_A A\|/\overline{n}_i^{1/2} \ . \tag{2}$$

Let B be the $d \times n$ matrix that results by replacing each sample (column) in A by the empirical mean of its component. B has rank at most k, and so by (1)

$$\|A - P_A A\| \ \leq \ \|A - B\| \ . \tag{3}$$

Write $D = A - B$ and let D_j be the $d \times \overline{n}_j$ submatrix of samples with label j, so that $\|D_j D_j^T/\overline{n}_j\| = \overline{\sigma}_j^2$. Then

$$\|D\|^2 \ = \ \|DD^T\| \ = \ \|\sum_j D_j D_j^T\| \ \leq \ \sum_j \|D_j D_j^T\| \ = \ \sum_j \overline{\sigma}_j^2 \overline{n}_j \ = \ \overline{\sigma}^2 n \ . \tag{4}$$

Combining (2),(3) and (4) we get $\|\overline{\mu}_i - P_A\overline{\mu}_i\| \leq \overline{\sigma}(n/\overline{n}_i)^{1/2} = \overline{\sigma}/\overline{w}_i^{1/2}$. □

Theorem 1 and the triangle inequality immediately imply that for every i, j the separation of $\overline{\mu}_i, \overline{\mu}_j$ is reduced by the projection onto P_A by no more than

$$\|(\overline{\mu}_i - \overline{\mu}_j) - P_A(\overline{\mu}_i - \overline{\mu}_j)\| \leq \overline{\sigma}(1/\overline{w}_i^{1/2} + 1/\overline{w}_j^{1/2}) \ . \tag{5}$$

In Theorem 2 below we sharpen (5) slightly (representing an improvement of no more than a factor of $\sqrt{2}$). As we will prove in Section 4, the result of Theorem 2 is *tight*.

Theorem 2. *For any set A of labeled samples, for all i, j, $\|(\overline{\mu}_i - \overline{\mu}_j) - P_A(\overline{\mu}_i - \overline{\mu}_j)\| \leq \overline{\sigma}(1/\overline{w}_i + 1/\overline{w}_j)^{1/2}$.*

Proof. Analogously to Theorem 1, we now choose $x_{ij} \in \{0, 1/\overline{n}_i, -1/\overline{n}_j\}^n$ so that $\overline{\mu}_i - \overline{\mu}_j = Ax_{ij}$ and $\|x_{ij}\| = (1/\overline{n}_i + 1/\overline{n}_j)^{1/2}$. Recall that by (3) and (4) we have $\|(A - P_A A)\| \leq \overline{\sigma}n^{1/2}$. Thus,

$$\begin{aligned}\|(\overline{\mu}_i - \overline{\mu}_j) - P_A(\overline{\mu}_i - \overline{\mu}_j)\| &= \|(A - P_A A)x_{ij}\| \\ &\leq \|A - P_A A\|(1/\overline{n}_i + 1/\overline{n}_j)^{1/2} \\ &= \overline{\sigma}(1/\overline{w}_i + 1/\overline{w}_j)^{1/2} \ .\end{aligned}$$

□

2.2 Combining Spectral Projection and Single-Linkage

We now describe a simple partitioning algorithm combining spectral projection and single-linkage that takes as input a training set A, a set to separate B, and a parameter k. The algorithm computes an optimal rank k projection for the samples in A which it applies to the samples in B. Then, it applies single-linkage to the projected samples, i.e., it computes their minimum spanning tree and removes the longest edge from it.

Separate(A, B, k):

1. Construct the Minimum Spanning Tree on $P_A B$ with respect to the 2-norm.
2. Cut the longest edge, and return the connected components.

Separate will be the core primitive we build upon in the following sections, and so it is important to understand the conditions under which it is guaranteed to return a proper cut.

Theorem 3. *Assume that A, B are sets of samples containing the same set of labels and that the sample statistics of A satisfy, with $i = \arg\max_i \overline{\sigma}_i$,*

$$\forall j \neq i : \quad \|\overline{\mu}_i - \overline{\mu}_j\| > \overline{\sigma}_i(1/\overline{w}_i + 1/\overline{w}_j)^{1/2} + 4 \max_{x_u \in B} \|P_A(x_u - \overline{\mu}_u)\| \ . \quad (6)$$

*If B contains at least two labels, then **Separate**(A, B, k) does not separate samples of the same label.*

Proof. The proof idea is that after projecting B on P_A, the samples in B with label i will be sufficiently distant from all other samples so that the following is true: all intra-label distances are shorter than the shortest inter-label distance involving label i. As a result, by the time an inter-label edge involving label i is added to the Minimum Spanning Tree, the samples of each label already form a connected component.

By the triangle inequality, the largest intra-label distance is at most

$$\|P_A(x_i - x_j)\| \leq 2 \max_{x_v \in B} \|P_A(x_v - \overline{\mu}_v)\| \ . \quad (7)$$

On the other hand, also by the triangle inequality, all inter-label distances are at least

$$\|P_A(x_i - x_j)\| \geq \|P_A(\overline{\mu}_i - \overline{\mu}_j)\| - 2 \max_{x_v \in B} \|P_A(x_v - \overline{\mu}_v)\| \,. \tag{8}$$

To bound $\|P_A(\overline{\mu}_i - \overline{\mu}_j)\|$ from below we first apply the triangle inequality one more time to get (9). We then bound the first term in (9) from below using (6) and the second term using Theorem 2, thus getting

$$\|P_A(\overline{\mu}_i - \overline{\mu}_j)\| \geq \|\overline{\mu}_i - \overline{\mu}_j\| - \|(I - P_A)(\overline{\mu}_i - \overline{\mu}_j)\| \tag{9}$$

$$> (\overline{\sigma}_i - \overline{\sigma})(1/\overline{w}_i + 1/\overline{w}_j)^{1/2} + 4 \max_{x_v \in B} \|P_A(x_v - \overline{\mu}_v)\| \,. \tag{10}$$

As $\overline{\sigma}_i \geq \overline{\sigma}$, combining (8) and (10) we see, by (7), that all inter-label distances involving label i have length exceeding the upper bound on intra-label distances. □

2.3 k-Partitioning the Full Sample Set

Given two sets of samples A, B and a parameter k, **Separate** bisects B by projecting it onto the optimal rank k subspace of A and applying single-linkage clustering. Now, we show how to use **Separate** recursively and build an algorithm **Segment** which on input A, B, k outputs a full k-partition of B. To classify n sample points from a mixture of distributions we simply partition them at random into two sets X, Y and invoke **Segment** twice, with each set being used once as the training set and once as the set to be partitioned.

Applying **Separate** recursively is non-trivial. Imagine that we are given sets A, B meeting the conditions of Theorem 3 and by running **Separate**(A, B, k) we now have a valid bipartition $B = B_1 \cup B_2$. Recall that one of the conditions in Theorem 3 is that the two sets given as input to **Separate** contain the same set of labels. Therefore, if we try to apply **Separate** to either B_1 or B_2 using A as the training set we are guaranteed to not meet that condition! Another, more technical, problem is that we would like each recursive invocation to succeed or fail independently of the rest. Using the same training set for all invocations introduces probabilistic dependencies among them that are very difficult to deal with.

To address these two problems we will need to be a bit more sophisticated in our use of recursion: given sets A, B rather than naively running **Separate**(A, B, k), we will instead first subsample A to get a training set A_1 and then invoke **Separate**$(A_1, A \cup B - A_1, k)$. The idea is that if A_1 is big enough it will have all the good statistical properties of A (as demanded by Theorem 3) and **Separate** will return a valid bipartition of $A \cup B - A_1$. The benefit, of course, is that each part of B will now be accompanied by the subset of $A - A_1$ of same labels. Therefore, we can now simply discard A_1 and proceed to apply the same idea to each of the two returned parts, as we know which points in each part came from A and which came from B.

Our algorithm **Segment** will very much follow the above idea, the only difference being that rather than doing subsampling with each recursive call we will fix a partition of $A = A_1 \cup \cdots \cup A_k$ at the outset and use it throughout the recursion. More specifically, we will think of the execution of **Segment** as a full binary tree with $2^k - 1$ nodes, each of which will correspond to an invocation of **Separate**. In each level $1 \leq \ell \leq k$ of the tree, all invocations will use A_ℓ as the training set and they will partition some subset of $A_{\ell+1} \cup \cdots \cup A_k \cup B$. So, for example, at the second level of the tree, there will be two calls to **Separate**, both using A_2 as the training set and each one partitioning the subset of $A \cup B - A_1$ that resulted by the split at level 1. Clearly, one of these two parts can already consist of samples from only one label, in which case the invocation at level 2 will produce a bipartition which is arbitrary (and useless). Nevertheless, as long as these are the only invocations in which samples with the same label are split, there exists a subset of k nodes in the tree which corresponds exactly to the labels in B. As we will see, we will be able to identify this subset in time $O(k^2 \min(n, 2^k))$ by dynamic programming.

Formally, **Segment** takes as input a sample set $S \subseteq A \cup B$ and a parameter ℓ indicating the level. Its output is the hierarchical partition of S as captured by the binary tree mentioned above. To simplify notation below, we assume that the division of A into $A_1, \ldots, A_k$ is known to the algorithm.

Segment(S, ℓ)
1. Let $[L, R] =$ **Separate**$(A_\ell \cap S, S \setminus A_\ell, k)$.
2. If $\ell < k$ invoke **Segment**$(L, \ell + 1)$ and **Segment**$(R, \ell + 1)$.

To state the conditions that guarantee the success of **Segment** we need to introduce some notation. For each i, ℓ, let $\overline{\mu}_i^\ell$, $\overline{\sigma}_i^\ell$, and $\overline{w}_i^\ell$ be the sample statistics associated with label i in A_ℓ. For each vector $\mathbf{v} \subseteq \{1, \ldots, k\}$ let $A_\ell^{\mathbf{v}}$ denote the set of samples from A_ℓ with labels from $\mathbf{v}$, and let $B_\ell^{\mathbf{v}}$ denote the set of samples from $\bigcup_{m>\ell} A_m \cup B$ with labels from $\mathbf{v}$. Finally, we say that a hierarchical clustering is *label-respecting* if for any set of at least two labels, the clustering does not separate samples of the same label.

Theorem 4. *Assume that $A_1, \ldots, A_k$ and B each contain the same set of labels and that for every pair $(\ell, \mathbf{v})$, with $i = \arg\max_{i \in \mathbf{v}} \overline{\sigma}_i^\ell$, we have:*

$$\forall\, j \in \mathbf{v} - i : \quad \|\overline{\mu}_i^\ell - \overline{\mu}_j^\ell\| \geq \overline{\sigma}_i^\ell (1/\overline{w}_i^\ell + 1/\overline{w}_j^\ell)^{1/2} + 4 \max_{x_u \in B_\ell^v} \|P_{A_\ell^{\mathbf{v}}}(x_u - \overline{\mu}_u^\ell)\| \,.$$

*The hierarchical clustering **Segment**$(A \cup B, 1)$ produces will be label-respecting.*

Proof. The proof is inductive, starting with the inductive hypothesis that in any invocation of **Segment**(S, ℓ) where S contains at least two labels, the set S equals $B_{\ell-1}^{\mathbf{v}}$ for some $\mathbf{v}$. Therefore, we need to prove that **Separate**$(A_\ell \cap S, S \setminus A_\ell, k) =$ **Separate**$(A_\ell^{\mathbf{v}}, B_\ell^{\mathbf{v}}, k)$ will produce sets L and R that do not share labels.

For every $(\mathbf{v}, \ell)$, if $i = \arg\max_{i \in \mathbf{v}} \overline{\sigma}_i^\ell$, our assumed separation guarantees that label i satisfies

$$\forall\, j \in \mathbf{v} - i : \quad \|\overline{\mu}_i^\ell - \overline{\mu}_j^\ell\| \geq \overline{\sigma}_i^\ell (1/\overline{w}_i^\ell + 1/\overline{w}_j^\ell)^{1/2} + 4 \max_{x_u \in B_\ell^{\mathbf{v}}} \|P_{A_\ell^{\mathbf{v}}}(x_u - \overline{\mu}_u^\ell)\| \,.$$

While the above separation condition refers to the sample statistics of A_ℓ, when we restrict our attention to $A_\ell^{\mathbf{v}}$, the samples means and standard deviations do not change and the sample mixing weights only increase. Therefore, the requirements of Theorem 3 hold for $A_\ell^{\mathbf{v}}, B_\ell^{\mathbf{v}}$ concluding the proof. □

Given the hierarchical clustering generated by **Segment** we must still determine which set of $k-1$ splits is correct. We will, in fact, solve a slightly more general problem. Given an arbitrary function scoring subsets of B, $score : 2^B \to \mathbb{R}$, we will find the k-partition of the samples with highest total (sum) score. For many distributions, such as Gaussians, there are efficient estimators of the likelihood that a set of data was generated from the distribution and such estimators can be used as the score function. For example, in cross training log-likelihood estimators, the subset under consideration is randomly partitioned into two parts. First, the parameters of the distribution are learned using one part and then the likelihood of the other part given these parameters is computed.

We will use dynamic programming to efficiently determine which subset set of $k-1$ splits corresponds to a k-partition for which the sum of the scores of its parts is highest. As one of the options in the $k-1$ splits by labels, our result will score at least as high as the latent partition. The dynamic program computes, for every node S in the tree and integer $i \leq k$, the quantity $opt(S, i)$, the optimal score gained by budgeting i parts to the subset S. If we let $S = L \cup R$ be the cut associated with S, the dynamic program is defined by the rules

$$opt(S, 1) \;=\; score(S) \;\text{ and }\; opt(S, i) \;=\; \max_{j<i}[opt(L, j) + opt(R, i - j)] \; .$$

We are ultimately interested in $opt(B, k)$ which we can be computed efficiently in a bottom up fashion in time $O(k^2 \min(n, 2^k))$.

Finally, all of the techniques that we have used to partition B can be used to partition A. We can divide B into k sets $B_1, \ldots B_k$ to use as training in the classification of A. For all but the most obtuse sets of samples, a random partition into A and B will yield samples for which $\|\overline{\mu}_i^A - \overline{\mu}_j^B\|$ is minimized at $i = j$ allowing us to merge the partition of A with the partition of B. We avoid stating a theorem generally about the combination of these three steps, but do so in the next section with concrete distributions.

3 Results for Gaussian, Log-Concave, and Concentrated Mixtures

We now examine how our results apply to specific distributions, such as Gaussian and Log-concave distributions, as well as a more general class that we define below. In fact, we will start with the more general class, and instantiate the other two from it.

First, we say that a distribution x is f-concentrated for a function $f : \mathbb{R} \to \mathbb{R}$ if for every unit vector v

$$\Pr\left[|v^T(x - \mathbf{E}[x])| > f(\delta)\right] \leq \delta \; . \tag{11}$$

In words, when we project the distribution onto any fixed line, a random sample will be within $f(\delta)$ of the mean with probability $1-\delta$. Second, we say that a distribution is *g-convergent* for a function $g : \mathbb{R} \to \mathbb{R}$ if a sample of size $g(\delta)$ with probability $1-\delta$ satisfies

$$\|\overline{\mu} - \mu\| \;\leq\; \sigma/8 \quad \text{and} \quad \sigma/2 \;\leq\; \overline{\sigma} \;\leq\; 2\sigma \;, \tag{12}$$

where $\overline{\mu}$ and $\overline{\sigma}^2$ denote the sample mean and the sample maximum directional variance, respectively.

Before proceeding, we prove an extension of f-concentration to low dimensional projections:

Lemma 1. *Let x be a distribution that is f-concentrated. For any fixed k dimensional projection P,*

$$\Pr\left[\|P(x - \mathbf{E}[x])\| > k^{1/2} f(\delta/k)\right] \leq \delta \;.$$

Proof. Given any set of k orthogonal basis vectors $v_1, \ldots, v_k$ for the space associated with P, we can write $P = \sum_i v_i v_i^T$. As the v_i are orthonormal, we can use the Pythagorean equality

$$\|P(x - \mathbf{E}[x])\|^2 \;=\; \|\sum_i v_i v_i^T (x - \mathbf{E}[x])\|^2 \;=\; \sum_i |v_i^T (x - \mathbf{E}[x])|^2 \;. \tag{13}$$

Taking a union bound, the probability that any of the k terms in the last sum exceeds $f(\delta/k)^2$ is at most δ, giving a squared distance of at most $k f(\delta/k)^2$ and completing the proof. □

With these definitions in hand, we now state and prove a result about the classification of concentrated, convergent distributions.

Theorem 5. *Consider any mixture of k distributions where each distribution i is f_i-concentrated and g_i-convergent. Assume that A contains at least*

$$k \times \max_i \left(\left(g_i(\delta/k^2) + 8\log(k^2/\delta)\right) w_i^{-1} \right)$$

samples from the mixture and that B contains n samples. If

$$\forall\, i,\ \forall\, j \neq i : \quad \|\mu_i - \mu_j\| > 4\sigma_i (1/w_i + 1/w_j)^{1/2} + 4k^{1/2} \max_{\sigma_v < 4\sigma_i} f_v\left(\frac{\delta}{nk2^k}\right)$$

then with probability at least $1 - 3\delta$, the hierarchical clustering produced by ***Segment***$(A \cup B, 1)$ *will be label-respecting.*

Proof. We argue that with probability $1 - 3\delta$ the sets $A_1, \ldots, A_k, B$ meet the conditions of Theorem 4.

As A is broken uniformly into $A_1, \ldots, A_k$, each of these k sets will contain a number of samples that is at least $\max_i \left(g_i(\delta/k^2)/w_i + 8\log(k^2/\delta)/w_i\right)$. Importantly, the first term is sufficient to ensure that with probability $1 - \delta$ each of

the mixtures in each of A_ℓ have "converged", in the sense of (12). The second term ensures that with probability $1-\delta$ we have $\overline{w}_i^\ell \geq w_i/2$ for each i, ℓ.

Given these bounds relating the sample statistics to their limits, and letting $s = \frac{\delta}{nk2^k}$ to simplify notation, the assumed separation of $\|\mu_i - \mu_j\|$ ensures that for all ℓ, for all i, and for all $j \neq i$,

$$\|\overline{\mu}_i^\ell - \overline{\mu}_j^\ell\| > \overline{\sigma}_i^\ell (1/\overline{w}_i^\ell + 1/\overline{w}_j^\ell)^{1/2} + 4 \max_{\sigma_u < 4\sigma_i} \left(k^{1/2} f_u(s) + \|\mu_u - \overline{\mu}_u^\ell\| \right) \quad (14)$$

As there are at most $k2^k$ matrices $A_\ell^{\mathbf{v}}$, the f_i-concentration of the distributions ensures that with probability at least $1-\delta$, for all $A_\ell^{\mathbf{v}}, B_\ell^{\mathbf{v}}$

$$\max_{x_u \in B_\ell^{\mathbf{v}}} \|P_{A_\ell^{\mathbf{v}}}(x_u - \mu_u)\| \leq k^{1/2} \max_{j \in \mathbf{v}} f_j(s) \ . \quad (15)$$

By the triangle inequality and submultiplicativity,

$$\max_{x_u \in B_\ell^{\mathbf{v}}} \|P_{A_\ell^{\mathbf{v}}}(x_u - \overline{\mu}_u^\ell)\| \leq \max_{j \in \mathbf{v}} \left(k^{1/2} f_j(s) + \|\mu_j - \overline{\mu}_j^\ell\| \right) \ . \quad (16)$$

Now, for each $\ell, \mathbf{v}$, from (12) we have that for any $\overline{\sigma}_j^\ell \leq \overline{\sigma}_i^\ell$, it is the case that $\sigma_j \leq 4\sigma_i$. Specifically, considering $i = \arg\max_{i \in \mathbf{v}} \overline{\sigma}_i^\ell$ we have that

$$\max_{j \in \mathbf{v}} \left(k^{1/2} f_j(s) + \|\mu_j - \overline{\mu}_j^\ell\| \right) \leq \max_{\sigma_u \leq 4\sigma_i} \left(k^{1/2} f_u(s) + \|\mu_u - \overline{\mu}_u^\ell\| \right) \ . \quad (17)$$

Combining (15), (16), and (17) with (14), we see that for all $\ell, \mathbf{v}$, if we let $i = \arg\max_{i \in \mathbf{v}} \overline{\sigma}_i^\ell$, then

$$\forall j \in \mathbf{v} - i \quad \|\overline{\mu}_i^\ell - \overline{\mu}_j^\ell\| > \overline{\sigma}_i^\ell (1/\overline{w}_i^\ell + 1/\overline{w}_j^\ell)^{1/2} + 4 \max_{x_u \in B_\ell^{\mathbf{v}}} \|P_{A_\ell^{\mathbf{v}}}(x_u - \overline{\mu}_u^\ell)\| \ .$$

□

3.1 Gaussian and Log-Concave Mixtures

We now show that for mixtures of both Gaussian and Log-concave distributions, **Segment** produces a hierarchical clustering that is label-respecting, as desired. From this, using dynamic programming as discussed in Section 2.3, we can efficiently find the k-partition that maximizes any scoring function which scores each part independently of the others. For example, in the case of Gaussians, this allows us to find a k-partition with cross-training log-likelihood at least as high as the latent partition in time $O(k^2 \min(n, 2^k))$.

Theorem 6. *Consider any mixture of k Gaussian distributions with parameters $\{(\mu_i, \sigma_i, w_i)\}$ and assume that $n \gg k(d + \log k)/w_{\min}$ is such that*

$$\forall i \ \forall j : \quad \|\mu_i - \mu_j\| \geq 4\sigma_i (1/w_i + 1/w_i)^{1/2} + 4\sigma_i (k \log(nk) + k^2)^{1/2} \ .$$

*Let A and B each contain n samples from the mixture and partition $A = A_1 \cup \ldots \cup A_k$ randomly. With probability that tends to 1 as $n \to \infty$, the hierarchical clustering produced by **Segment**$(A \cup B, 1)$ will be label-respecting.*

Proof. Standard results show that any Gaussian is f-concentrated for $f(\delta) = \sigma(2\log(1/\delta))^{1/2}$. Using techniques from Soshnikov [7] describing concentration of the median of σ_i^ℓ for various sample counts, one can show that a d-dimensional Gaussian with maximum directional variance σ^2 is g-convergent for $g(\delta) = cd\log(1/\delta)$ for a universal constant c. □

A recent related paper of Kannan et al. [8] shows that Log-concave distributions, those for which the logarithm of the probability density function is concave, are also reasonably concentrated and convergent.

Theorem 7. *Given a mixture of k Log-concave distributions with parameters $\{(\mu_i, \sigma_i, w_i)\}$ assume that for some fixed $n \gg k(d(\log d)^5 + \log k)/w_{\min}$ the following holds:*

$$\forall i\ \forall j: \quad \|\mu_i - \mu_j\| \geq 4\sigma_i(1/w_i + 1/w_i)^{1/2} + 4\sigma_i k^{1/2}(\log(nk) + k) \ .$$

*Let A and B each contain n samples from the mixture and partition $A = A_1 \cup \ldots \cup A_k$ randomly. With probability that tends to 1 as $n \to \infty$, the hierarchical clustering produced by **Segment**$(A \cup B, 1)$ will be label-respecting.*

Proof. Lemma 2 of [8] shows that any Log-concave distribution is f-concentrated for $f(\delta) = \sigma\log(1/\delta)$. Lemma 4 of [8] shows that for any Log-concave distribution there is a constant c such that the distribution is g-convergent for $g(\delta) = cd(\log(d/\delta))^5$. □

4 Lower Bounds

We now argue that for any set of mixing weights $w_1, \ldots, w_k$, there is an arrangement of *identical* Gaussians for which spectral projection is not an option. This also demonstrates that the bound in Theorem 1 is tight.

Theorem 8. *For any $\sum_i w_i = 1$, there exists a mixture of Gaussians with $\|C_i\| = \sigma^2$ satisfying*

$$\|\mu_i - \mu_j\| = \sigma(1/w_i + 1/w_j)^{1/2} \tag{18}$$

for which the optimal rank k subspace for the distribution is arbitrary.

Proof. We choose the μ_i to be mutually orthogonal and of norm $\sigma/w_i^{1/2}$. To each we assign the common covariance matrix $C = \sigma^2 I - \sum_i w_i \mu_i \mu_i^T$. The optimal rank k subspace for the distribution is the optimal rank k subspace for the expected outer product of a random sample x from the mixture which is

$$\mathbf{E}[xx^T] \ = \ \sum_i w_i \mu_i \mu_i^T + \sum_i w_i C \ = \ \sigma^2 I \ .$$

Since the identity matrix favors no dimensions for its optimal approximation, the proof is complete. □

Remark: The theorem above only describes a mixture for which there is no preference for a particular subspace. By diminishing the norms of the μ_i ever so slightly, we can set the optimal rank k subspace arbitrarily and ensure that it does not intersect the span of the means.

Remark: One can construct counterexamples with covariance matrices of great generality, so long as they discount the span of the means $\sum_i w_i \mu_i \mu_i^T$, and promote some other k dimensions. In particular, the $d - 2k$ additional dimensions can have 0 variance, demonstrating that the maximum variance $\sigma^2 = \|C\|_2^2$ is the parameter of interest, as opposed to the average variance $\|C\|_F^2/d$, or any other function that depends on more than the first k singular values of C.

Remark: If one is willing to weaken Theorem 8 slightly by dividing the RHS of (18) by 2, then we can take as the common covariance matrix $C = 2\sigma^2 I - \sum_i w_i \mu_i \mu_i^T$, which has eccentricity bounded by 2. Bounded eccentricity was an important assumption of Dasgupta [2], who used random projections, but we see here that it does not substantially change the lower bound.

References

1. S. Arora and R. Kannan, Learning mixtures of arbitrary Gaussians, In Proc. *33rd ACM Symposium on Theory of Computation*, 247–257, 2001.
2. S. Dasgupta, Learning mixtures of Gaussians, In Proc. *40th IEEE Symposium on Foundations of Computer Science*, 634–644, 1999.
3. S. Dasgupta, L. Schulman, A 2-round variant of EM for Gaussian mixtures, In Proc. *16th Conference on Uncertainty in Artificial Intelligence*, 152–159, 2000.
4. B. Lindsay, Mixture models: theory, geometry and applications, *American Statistical Association*, Virginia, 2002.
5. D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical analysis of finite mixture distributions*, Wiley, 1985.
6. S. Vempala and G. Wang, A Spectral Algorithm of Learning Mixtures of Distributions, In Proc. *43rd IEEE Symposium on Foundations of Computer Science*, 113–123, 2002.
7. A. Soshnikov, A Note on Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices, *J. Stat. Phys.*, v.108, Nos. 5/6, pp. 1033-1056, (2002)
8. H. Salmasian, R. Kannan, S. Vempala, The Spectral Method for Mixture Models, In *Electronic Colloquium on Computational Complexity (ECCC)* (067), 2004.