Efficiently Learning Mixtures of Two Gaussians

Adam Tauman Kalai*
Microsoft Research, New England
Cambridge, MA 02139
adum@microsoft.com

Ankur Moitra†
Massachusetts Institute of Technology
Cambridge, MA 02139
moitra@mit.edu

Gregory Valiant‡
University of California, Berkeley
Berkeley, CA 94720
gvaliant@eecs.berkeley.edu

ABSTRACT

Given data drawn from a mixture of multivariate Gaussians, a basic problem is to accurately estimate the mixture parameters. We provide a polynomial-time algorithm for this problem for the case of two Gaussians in n dimensions (even if they overlap), with provably minimal assumptions on the Gaussians, and polynomial data requirements. In statistical terms, our estimator converges at an inverse polynomial rate, and no such estimator (even exponential time) was known for this problem (even in one dimension). Our algorithm reduces the n-dimensional problem to the one-dimensional problem, where the *method of moments* is applied. One technical challenge is proving that noisy estimates of the first six moments of a univariate mixture suffice to recover accurate estimates of the mixture parameters, as conjectured by Pearson (1894), and in fact these estimates converge at an inverse polynomial rate.

As a corollary, we can efficiently perform near-optimal clustering: in the case where the overlap between the Gaussians is small, one can accurately cluster the data, and when the Gaussians have partial overlap, one can still accurately cluster those data points which are not in the overlap region. A second consequence is a polynomial-time density estimation algorithm for arbitrary mixtures of two Gaussians, generalizing previous work on axis-aligned Gaussians (Feldman *et al*, 2006).

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: multivariate statistics

General Terms

Algorithms, Theory

*Microsoft Research New England. Part of this work was done while the author was at Georgia Institute of Technology, supported in part by NSF CAREER-0746550, SES-0734780, and a Sloan Fellowship.

†This research was supported in part by a Fannie and John Hertz Foundation Fellowship. Part of this work was done while the author was an intern at Microsoft Research New England.

‡This research was supported in part by an NSF Graduate Research Fellowship. Part of this work done while at Microsoft Research New England.

STOC'10, June 5–8, 2010, Cambridge, Massachusetts, USA.
ACM 978-1-4503-0050-6/10/06.

Keywords

Gaussians, Finite Mixture Models, method of moments

1. INTRODUCTION

The problem of estimating the parameters of a mixture of Gaussians has a rich history of study in statistics and more recently, computer science. This natural problem has applications across a number of fields, including agriculture, economics, medicine, and genetics [27, 21]. Consider a mixture of two *different* multinormal distributions, each with *mean* $\mu_i \in \mathbf{R}^n$, *covariance matrix* $\Sigma_i \in \mathbf{R}^{n\times n}$, and *weight* $w_i > 0$. With probability w_1 a sample is chosen from $\mathcal{N}(\mu_1, \Sigma_1)$, and with probability $w_2 = 1 - w_1$, a sample is chosen from $\mathcal{N}(\mu_2, \Sigma_2)$. The mixture is referred to as a Gaussian Mixture Model (GMM), and if the two multinormal densities are F_1, F_2, then the GMM density is,

$$F = w_1 F_1 + w_2 F_2.$$

The problem of *identifying* the mixture is that of estimating $\hat{w}_i$, $\hat{\mu}_i$, and $\hat{\Sigma}_i$ from m independent random samples drawn from the GMM.

In this paper, we prove that the parameters can be estimated at an inverse polynomial rate. In particular, we give an algorithm and polynomial bounds on the number of samples and runtime required under provably minimal assumptions, namely that w_1, w_2 and the statistical distance between the Gaussians are all bounded away from 0 (Theorem 1). No such bounds were previously known, even in one dimension. Our algorithm for accurately identifying the mixture parameters can also be leveraged to yield the first provably efficient algorithms for near-optimal clustering and density estimation (Theorems 3 and 2) for mixtures of two Gaussians. We start with a brief history, and then give the formal definition of the learning problem we consider and our main results and approach.

1.1 Brief history

In one of the earliest GMM studies, Pearson [23] fit a mixture of two univariate Gaussians to data (see Figure 1) using the *method of moments*. In particular, he computed empirical estimates of the first six (raw) moments $\mathrm{E}[x^i] \approx \frac{1}{m}\sum_{j=1}^{m} x_j^i$, for $i = 1, 2, \ldots, 6$ from sample points $x_1, \ldots, x_m \in \mathbf{R}$. Using only the first five moments, he solved a cleverly constructed ninth-degree polynomial *by hand* from which he derived a set of candidate mixture parameters. Finally, he heuristically chose the candidate set of parameters among them whose sixth moment most closely agreed with the empirical estimate.

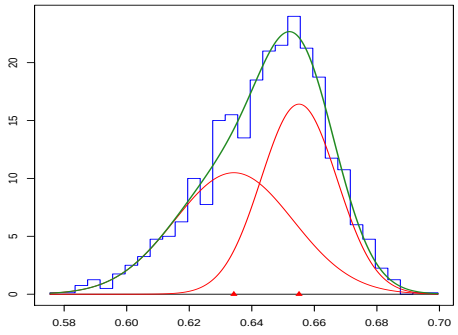

Figure 1: A fit of a mixture of two univariate Gaussians to the Pearson's data on Naples crabs [23]. The hypothesis was that the data was in fact a mixture of two different species of crabs. Although the empirical data histogram is single-peaked, the two constituent Gaussian parameters may be estimated. This density plot was created by Peter Macdonald using *R* [20].

Later work showed that "identifiability" is theoretically possible – every two distinct mixtures of Gaussians (i.e. the mixtures are not equivalent after some permutation of the labels) have different probability distributions [26]. However, this work shed little light on convergence *rates*: this result is based on demonstrating that distinct mixtures of Gaussians exhibit different behavior in the density tails, and even obtaining a single sample from the density tails could require an enormous number of random samples. In fact, previous work left open the possibility that distinguishing between GMMs that are ϵ-different (see Theorem 1 for our definition of ϵ-close) might require an amount of data that grows exponentially in $1/\epsilon$.

The problem of *clustering* is that of partitioning the points into two sets, with the hope that the points in each set are drawn from different Gaussians. Given an accurate clustering of sufficiently many points, one can recover good estimates of the mixture parameters. Starting with Dasgupta [5], a line of computer scientists designed *polynomial time* algorithms for identifying and clustering in high dimensions [2, 7, 30, 14, 1, 4, 31]. However, even if we were given the parameters of the mixture, we could not hope to cluster many points accurately unless the Gaussians have little *overlap* (statistical distance near 1). Thus this line of work must make such an assumption in order to cluster, and learn good estimates for the mixture from such a clustering. Here we are able to learn good estimates of the mixture parameters for a GMM of two Gaussians without clustering, and we do so using provably minimal assumptions on the GMM.

There is a vast literature that we have not touched upon (see, e.g., [27, 21]), including the popular EM and K-means algorithms.

1.2 Main results

In identifying a GMM $F = w_1F_1 + w_2F_2$, three limitations are immediately apparent:

1. Since permuting the two Gaussians does not change the resulting density, one cannot distinguish permuted mixtures. Hence, at best one hopes to estimate the parameter set, $\{(w_1, \mu_1, \Sigma_1), (w_2, \mu_2, \Sigma_2)\}$.
2. If $w_i = 0$, then one cannot hope to estimate F_i because no samples will be drawn from it. And, in general, at least $\Omega(1/\min\{w_1, w_2\})$ samples will be required in order to obtain any reasonable estimate.
3. If $F_1 = F_2$ (i.e., $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$) then it is impossible to estimate w_i. If the statistical distance between the two Gaussians is Δ, then at least $\Omega(1/\Delta)$ samples will be required.

Hence, the number of examples required will depend on the smallest of w_1, w_2, and the statistical distance between F_1 and F_2 denoted by $D(F_1, F_2)$ (see Section 2 for a precise definition).

Our goal is, given m independently drawn samples from a GMM F, to construct an estimate GMM $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$. We will say that $\hat{F}$ is accurate to within ϵ if $|\hat{w}_i - w_i| \leq \epsilon$ and $D(F_i, \hat{F}_i) \leq \epsilon$ for each $i = 1, 2$. This latter condition is affine invariant and more appealing than bounds on the difference between the estimated and true parameters. In fact for arbitrary Gaussians, estimating parameters, such as the mean μ, to any given additive error ϵ is impossible without further assumptions since scaling the data by a factor of s will scale the error $\|\mu - \hat{\mu}\|$ by s. We would like the algorithm to succeed in this goal using polynomially many samples. And we would also like the algorithm itself to be computationally efficient, i.e., a polynomial-time algorithm.

Our main theorem is the following.

THEOREM 1. *For any $n \geq 1$, $\epsilon, \delta > 0$, and any GMM $F = w_1F_1 + w_2F_2$ in n dimensions, using m independent samples from F, there is an algorithm that outputs GMM $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$ such that, with probability $\geq 1 - \delta$ (over the samples and randomization of the algorithm), is ϵ-close - i.e. there is a permutation $\pi : \{1, 2\} \to \{1, 2\}$ such that,*

$$D(\hat{F}_i, F_{\pi(i)}) \leq \epsilon \text{ and } |\hat{w}_i - w_{\pi(i)}| \leq \epsilon, \text{for each } i = 1, 2.$$

And the runtime (in the Real RAM model) and number of samples drawn by Algorithm 2 is at most
$\text{poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{w_1}, \frac{1}{w_2}, \frac{1}{D(F_1,F_2)})$

Due to space considerations, we do not give the algorithm here but we give a detailed algorithm for the case in which F is in isotropic position in Algorithm 4, and the algorithm in the above theorem begins with a step where samples are used to put the distribution (nearly) in isotropic position, which suffices for the analysis.

Our primary goal is to understand the statistical and computational complexities of this basic problem, and the distinction between polynomial and exponential is a natural step. While the order of the polynomial in our analysis is quite large, to the best of our knowledge these are the first bounds on the convergence rate for the problem in this general context. In some cases, we have favored clarity of presentation over optimality of bounds. The challenge of achieving optimal bounds (optimal rate) is very interesting, and will most likely require further insights and understanding.

As mentioned, our approximation bounds are in terms of the statistical distance between the estimated and true Gaussians. To demonstrate the utility of this type of bound,

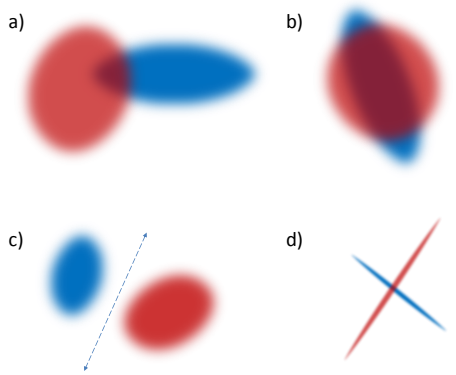

Figure 2: Mixtures of two multinormal distributions, with varying amounts of overlap. Our algorithm will learn the parameters in all cases, and hence be able to cluster when possible.

we note the following corollaries. For both problems, no assumptions are necessary on the underlying mixture. The first problem is simply that of approximating the density F itself.

COROLLARY 2. *For any $n \geq 1, \epsilon, \delta > 0$ and any GMM $F = w_1F_1 + w_2F_2$ in n dimensions, using m independent samples from F, there is an algorithm that outputs a GMM $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$ such that with probability $\geq 1 - \delta$ (over the samples and randomization of the algorithm)*

$$D(F, \hat{F}) \leq \epsilon$$

And the runtime (in the Real RAM model) and number of samples drawn from the oracle is at most $\text{poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta})$.

The second problem is that of clustering the m data points. In particular, suppose that during the data generation process, for each point $x \in \mathbf{R}^n$, a secret label $y_i \in \{1, 2\}$ (called *ground truth*) is generated based upon which Gaussian was used for sampling. A *clustering algorithm* takes as input m points and outputs a *classifier* $C : \mathbf{R}^n \to \{1, 2\}$. The *error* of a classifier is minimum, over all label permutations, of the probability that the the label of the classifier agrees with ground truth. Of course, achieving a low error is impossible in general. For example, suppose the Gaussians have equal weight and statistical distance $1/2$. Then, even armed with the correct mixture parameters, one could not identify with average accuracy greater than $3/4$, the label of a randomly chosen point. However, it is not difficult to show that given the correct mixture parameters, the optimal clustering algorithm (minimizing expected errors) simply clusters points based on which Gaussian has a larger posterior probability. We are able to approach the error rate of this classifier and achieve near optimal clustering without *a priori* knowledge of the distribution parameters. See Section 6 for precise details.

COROLLARY 3. *For any $n \geq 1, \epsilon, \delta > 0$ and any GMM $F = w_1F_1 + w_2F_2$ in n dimensions, using m independent samples from F, there is an algorithm that outputs a classifier $C_{\hat{F}}$ such that with probability $\geq 1 - \delta$ (over the samples and randomization of the algorithm), the error of $C_{\hat{F}}$ is at most ϵ larger than the error of any classifier, $C' : \mathbf{R}^n \to \{1, 2\}$. And the runtime (in the Real RAM model) and number of samples drawn from the oracle is at most* $\text{poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta})$

In a recent extension of Principal Component Analysis, Brubaker and Vempala [4] give a polynomial-time clustering algorithm that will succeed, with high probability, whenever the Gaussians are nearly separated by any hyperplane. (See 2c for an example.) This algorithm inspired the present work, and our algorithm follows theirs in that both are invariant to affine transformations of the data. Figure 2d illustrates a mixture where clustering is possible although the two Gaussians are not separable by a hyperplane.

1.3 Outline of Algorithm and Analysis

The problem of identifying Gaussians in high dimensions is surprising in that much of the difficulty seems to be present in the one-dimensional problem. We first briefly explain our reduction from n to 1 dimensions, based upon the fact that the projection of a multivariate GMM is a univariate GMM to which we can apply a one-dimensional algorithm.

When the data is projected down onto a line, each pair $(mean, variance)$ recovered in this direction gives some direct information about the corresponding $(mean, variance)$ pair in n dimensions. Lemma 12 states that for a suitably chosen *random* direction[1], two different Gaussians (statistical distance bounded away from 0) will project down to two reasonably distinct one-dimensional Gaussians, with high probability. For a single Gaussian, knowing the approximate value of the projected mean and variance in $O(n^2)$ linearly independent directions is enough to recover a good approximation for the Gaussian. The remaining challenge is identifying which univariate Gaussian in one projection corresponds to which univariate Gaussian in another projection; one must correctly *match up* the many pairs of univariate Gaussians in each one-dimensional problem. In practice, the mixing weights may be somewhat different, i.e., $|w_1 - w_2|$ is bounded from 0. In such cases, matching would be quite easy because each one-dimensional problem should have one Gaussian with weight close to the true w_1. In the general case, however, we must do something more sophisticated. The solution we employ is simple but certainly not the most efficient – we project to $O(n^2)$ directions which are all very close to each other, so that with high probability the means and variances change very little and are easy to match up. The idea of using random projection for this problem has been used in a variety of theoretical and practical contexts. Independently, Belkin and Sinha considered using random projections to one dimension for the problem of learning a mixture of multiple identical spherical Gaussians [3].

We now proceed to describe how to identify univariate GMMs. Like many one-dimensional problems, it is algorithmically *easy* because simple brute-force algorithms (like that of [10]) will work. The surprising difficulty is proving that such brute force algorithms cannot return wrong (or spurious) estimates. What if there were two mixtures where all four Gaussians were at least ϵ-different in statistical distance, yet the resulting mixtures were exponentially close in statistical distance? Ruling out this possibility is, in fact, a central hurdle in this work.

[1]The random direction is not uniform but is chosen in accordance with shape (covariance matrix) of the data, making the algorithm affine invariant.

We appeal to the old method of moments. In particular, the key fact is that univariate mixtures of two Gaussians are *polynomially robustly identifiable*–that is, if two mixtures have parameter sets differing by ϵ then one of the low-order moments will differ: i.e. $\left|\mathrm{E}_{x\sim F}[x^i] - \mathrm{E}_{x\sim F'}[x^i]\right|$ will be at least $\mathrm{poly}(\epsilon)$ for some $i \leq 6$.

> **Polynomially Robust Identifiability (Informal version of Theorem 4):** Consider two one-dimensional mixtures of two Gaussians, F, F', where F's mean is 0 and variance is 1. If the parameter sets differ by ϵ, then at least one of the first six raw moments of F will differ from that of F' by $\mathrm{poly}(\epsilon)$.

Using this theorem, a brute force search will work correctly: First normalize the data so that it has mean 0 and variance 1 (called *isotropic position*). Then perform a brute-force search over mixture parameters, choosing the one whose moments best fit the empirical moments and this will necessarily be a good estimate for the parameters. We now describe the proof of Theorem 4. The two ideas are to relate the statistical distance of two mixtures to the discrepancy in the moments, and *deconvolution*.

1.3.1 Relating statistical distance and discrepancy in moments

If two (bounded or almost bounded) distributions are statistically close, then their low-order moments must be close. However, the converse is not true in general. For example, consider the uniform distribution over $[0, 1]$ and the distribution whose density is proportional to $|\sin(Nx)|$ over $x \in [0, 1]$, for very large N. Crucial to this example is that the difference in the two densities oscillate many times, which cannot happen for mixtures of two univariate Gaussians. Lemma 8 shows that if two univariate GMMs have nonnegligible statistical distance, then they must have a nonnegligible difference in one of the first six moments. Hence statistical distance and moment discrepancy are closely related.

We very briefly describe the proof of Lemma 8. Denote the difference in the two probability density functions by $f(x)$; by assumption, $\int |f(x)|dx$ is nonnegligible. We first argue that $f(x)$ has at most six zero-crossings (using a general fact about the effect of convolution by a Gaussian on the number of zeros of a function), from which it follows that there is a degree-six polynomial whose sign always matches that of $f(x)$. Call this polynomial p. Intuitively, $\mathrm{E}[p(x)]$ should be different under the two distributions; namely $\int_{\mathbf{R}} p(x)f(x)dx$ should be bounded from 0 (provided we make sure that the mass of $f(x)$ is not too concentrated near any zero). Then if the coefficients of $p(x)$ are bounded, this implies $\mathrm{E}[x^i]$ differs under the two distributions, for some $i \leq 6$.

1.3.2 Deconvolving Gaussians

The convolution of two Gaussians is a Gaussian, just as the sum of two normal random variables is normal. Hence, we can also consider the "deconvolution" of the mixture by a Gaussian of variance, say, α – this is a simple operation which subtracts α from the variance of each Gaussian in the mixture. In fact, it affects all the moments in a simple, predictable fashion, and we show that a discrepancy in the low-order moments of two mixtures is roughly preserved by convolution. (See Lemma 6).

If we choose α close to the smallest variance of the four Gaussians that comprise the two mixtures, then after deconvolving, one of the mixtures has a Gaussian component that is very skinny – nearly a Dirac Delta function. When one of the four Gaussians is very skinny, it is intuitively clear that unless this skinny Gaussian is closely matched by a similar skinny Gaussian in the other mixture, the two mixtures will have large statistical distance. And if instead there are two closely matched Gaussians, these can be removed from the respective mixtures and we can compare the remaining Gaussians directly and obtain a large statistical distance directly (see Lemma 5).

The proof of Theorem 4 then follows: (1) after deconvolution, at least one of the four Gaussians is very skinny; (2) combining this with the fact that the parameters of the two GMMs are slightly different, the deconvolved GMMs have nonnegligible statistical distance; (Lemma 5) (3) nonnegligible statistical distance implies nonnegligible moment discrepancy (Lemma 8); and (4) if there is a discrepancy in one the low-order moments of two GMMs, then after convolution by a Gaussian, there will still be a discrepancy in some low-order moment (Lemma 6).

2. NOTATION AND PRELIMINARIES

Let $\mathcal{N}(\mu, \Sigma)$ denote the multinormal distribution with mean $\mu \in \mathbf{R}^n$ and $n \times n$ covariance matrix Σ, with density

$$\mathcal{N}(\mu, \Sigma, x) = (2\pi)^{-n/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

For probability distribution F, define the *mean* $\mu(F) = \mathrm{E}_{x\sim F}[x]$ and *covariance matrix*

$$\mathrm{var}(F) = \mathrm{E}_{x\sim F}[xx^T] - \mu(F)(\mu(F))^T$$

A distribution is *isotropic* or in *isotropic position* if the mean is zero and the covariance matrix is the identity matrix.

For distributions F and G with densities f and g, define the ℓ_1 distance $\|F - G\|_1 = \int_{\mathbf{R}^n} |f(x) - g(x)|dx$. Define the *statistical distance* or *variation distance* by $D(F, G) = \frac{1}{2}\|F - G\|_1 = F(S) - G(S)$, where $S = \{x|f(x) \geq g(x)\}$.

For vector $v \in \mathbf{R}^n$, Let P_v be the projection onto v, i.e., $P_v(w) = v \cdot w$, for vector $w \in \mathbf{R}^n$. For probability distribution F over $\mathbf{R}^n$, $P_v(F)$ denotes the marginal probability distribution over $\mathbf{R}$, i.e., the distribution of $x \cdot v$, where x is drawn from F. For Gaussian G, we have that $\mu(P_v(G)) = v \cdot \mu(G)$ and $\mathrm{var}(P_v(G)) = v^T\mathrm{var}(G)v$.

Let $\mathbb{S}_{n-1} = \{x \in \mathbf{R}^n : \|x\| = 1\}$. We write $\Pr_{u\in\mathbb{S}_{n-1}}$ over u chosen uniformly at random from the unit sphere. For probability distribution F, we define an *sample oracle* $\mathrm{SA}(F)$ to be an oracle that, each time invoked, returns an independent sample drawn according to F. Note that given $\mathrm{SA}(F)$ and a vector $v \in \mathbf{R}^n$, we can efficiently simulate $\mathrm{SA}(P_v(F))$ by invoking $\mathrm{SA}(F)$ to get sample x, and then returning $v \cdot x$.

For probability distribution F over $\mathbf{R}$, define $M_i(F) = \mathrm{E}_{x\sim F}[x^i]$ to be the ith (raw) moment.

3. THE UNIVARIATE PROBLEM

In this section, we will show that one can efficiently learn one-dimensional mixtures of two Gaussians. To be most useful in the reduction from n to 1 dimensions, Theorem 9 will be stated in terms of achieving estimated parameters that are off by a small additive error (and will assume the true mixture is in isotropic position).

The main technical hurdle in this result is showing the *polynomially robust identifiability* of these mixtures: that is, given two such mixtures with parameter sets that differ by ϵ, we show that one of the first six raw moments will differ by at least poly(ϵ). Given this result, it will be relatively easy to show that by performing essentially a brute-force search over a sufficiently fine (but still polynomial-sized) mesh of the set of possible parameters, one will be able to efficiently learn the 1-d mixture.

3.1 Polynomially Robust Identifiability

Throughout this section, we will consider two mixtures of one-dimensional Gaussians:

$$F(x) = \sum_{i=1}^{2} w_i \mathcal{N}(\mu_i, \sigma_i^2, x), \text{ and } F'(x) = \sum_{i=1}^{2} w_i' \mathcal{N}(\mu_i', \sigma_i'^2, x).$$

DEFINITION 1. *We will call the pair F, F' ϵ-standard if $\sigma_i^2, \sigma_i'^2 \leq 1$ and if ϵ satisfies:*

1. $w_i, w_i' \in [\epsilon, 1]$
2. $|\mu_i|, |\mu_i'| \leq \frac{1}{\epsilon}$
3. $|\mu_1 - \mu_2| + |\sigma_1^2 - \sigma_2^2| \geq \epsilon$ *and* $|\mu_1' - \mu_2'| + |\sigma_1'^2 - \sigma_2'^2| \geq \epsilon$
4. $\epsilon \leq \min_\pi \sum_i \left(|w_i - w_{\pi(i)}'| + |\mu_i - \mu_{\pi(i)}'| + |\sigma_i^2 - \sigma_{\pi(i)}'^2|\right)$, *where the minimization is taken over all permutations π of $\{1, 2\}$.*

THEOREM 4. *There is a constant $c > 0$ such that, for any $\epsilon < c$ and any ϵ-standard F, F',*

$$\max_{i \leq 6} |M_i(F) - M_i(F')| \geq \epsilon^{67}$$

In order to prove this theorem, we rely on "deconvolving" by a Gaussian with an appropriately chosen variance (this corresponds to running the heat equation in reverse for a suitable amount of time). We define the operation of deconvolving by a Gaussian of variance α as $\mathcal{F}_\alpha$; applying this operator to a mixture of Gaussians has a particularly simple effect: subtract α from the variance of each Gaussian in the mixture (assuming that each constituent Gaussian has variance at least α).

DEFINITION 2. *Let $F(x) = \sum_{i=1}^{n} w_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ be the probability density function of a mixture of Gaussian distributions, and for any $\alpha < \min_i \sigma_i^2$, define*

$$\mathcal{F}_\alpha(F)(x) = \sum_{i=1}^{n} w_i \mathcal{N}(\mu_i, \sigma_i^2 - \alpha, x).$$

Consider any two mixtures of Gaussians that are ϵ-standard. Ideally, we would like to prove that these two mixtures have statistical distance at least poly(ϵ). We settle instead for proving that there is some α for which the resulting mixtures (after applying the operation $\mathcal{F}_\alpha$) have large statistical distance. Intuitively, this deconvolution operation allows us to isolate Gaussians in each mixture and then we can reason about the statistical distance between the two mixtures locally, without worrying about the other Gaussian in the mixture. We now show that we can always choose an α so as to yield a large ℓ_1 distance between $\mathcal{F}_\alpha(F)$ and $\mathcal{F}_\alpha(F')$.

LEMMA 5. *Suppose F, F' are ϵ-standard. There is some α such that*

$$D(\mathcal{F}_\alpha(F), \mathcal{F}_\alpha(F')) \geq \Omega(\epsilon^4),$$

and such an α can be chosen so that the smallest variance of any constituent Gaussian in $\mathcal{F}_\alpha(F)$ and $\mathcal{F}_\alpha(F')$ is at least ϵ^{12}.

The proof of the above lemma is through an analysis of several cases. Assume without loss of generality that the first constituent Gaussian of mixture F has the minimal variance among all Gaussians in F and F'. Consider the difference between the two density functions. We lower-bound the ℓ_1 norm of this function on $\mathbf{R}$. The first case to consider is when both Gaussians in F' either have variance significantly larger than σ_1^2, or means far from μ_1. In this case, we can pick α so as to show that there is $\Omega(\epsilon^4)$ statistical distance in a small interval around μ_1 in $\mathcal{F}_\alpha(F) - \mathcal{F}_\alpha(F')$. In the second case, if one Gaussians in F' has parameters that very closely match σ_1, μ_1, then if the weights do not match very closely, we can use a similar approach as to the previous case. If the weights do match, then we choose an α very, very close to σ_1^2, to essentially make one of the Gaussians in each mixture nearly vanish, except on some tiny interval. We conclude that the parameters σ_2, μ_2 must not be closely matched by parameters of F', and demonstrate an $\Omega(\epsilon^4)$ statistical distance coming from the mismatch in the second Gaussian components in $\mathcal{F}_\alpha(F)$ and $\mathcal{F}_\alpha(F')$. The details are laborious, and are deferred to the the full version of our paper.

Unfortunately, the transformation $\mathcal{F}_\alpha$ does not preserve the statistical distance between two distributions. However, we show that it, at least roughly, preserves the disparity in low-order moments of the distributions. Specifically, we show that if there is an $i \leq 6$ such that the i^{th} raw moment of $\mathcal{F}_\alpha(F)$ is at least poly(ϵ) different than the i^{th} raw moment of $\mathcal{F}_\alpha(F')$ then there is a $j \leq 6$ such that the j^{th} raw moment of F is at least poly(ϵ) different than the j^{th} raw moment of F'.

LEMMA 6. *Suppose that each constituent Gaussian in F or F' has variances in the interval $[\alpha, 1]$. Then*

$$\frac{\sum_{i=1}^{k} |M_i(\mathcal{F}_\alpha(F)) - M_i(\mathcal{F}_\alpha(F'))|}{\sum_{i=1}^{k} |M_i(F) - M_i(F')|} \leq \frac{(k+1)!}{\lfloor k/2 \rfloor !},$$

The key observation here is that the moments of F and $\mathcal{F}_\alpha(F)$ are related by a simple linear transformation; and this can also be viewed as a recurrence relation for Hermite polynomials. We defer a proof to the full version of our paper.

To complete the proof of the theorem, we must show that the poly(ϵ) statistical distance between $\mathcal{F}_\alpha(F)$ and $\mathcal{F}_\alpha(F')$ gives rise to a poly(ϵ) disparity in one of the first six raw moments of the distributions. To accomplish this, we show that there are at most 6 zero-crossings of the difference in densities, $f = \mathcal{F}_\alpha(F) - \mathcal{F}_\alpha(F')$, using properties of the evolution of the heat equation, and then we construct a degree six polynomial $p(x)$ that always has the same sign as $f(x)$, so that when $p(x)$ is integrated against $f(x)$ the result is at least poly(ϵ). We construct this polynomial so that the coefficients are bounded, and this implies that there is some raw moment i (at most the degree of the polynomial) for which the difference between the i^{th} raw moment of $\mathcal{F}_\alpha(F)$ and of $\mathcal{F}_\alpha(F')$ is large.

Our first step is to show that $\mathcal{F}_\alpha(D)(x) - \mathcal{F}_\alpha(D')(x)$ has a constant number of zeros.

PROPOSITION 1. *Given $f(x) = \sum_{i=1}^{k} a_i \mathcal{N}(\mu_i, \sigma_i^2, x)$, the linear combination of k one-dimensional Gaussian probability density functions, such that $\sigma_i^2 \neq \sigma_j^2$ for $i \neq j$, assuming that not all the a_i's are zero, the number of solutions to $f(x) = 0$ is at most $2(k-1)$. Furthermore, this bound is tight.*

Using only the facts that quotients of Gaussians are Gaussian and that the number of zeros of a differentiable function is at most one more than the number of zeros of its derivative, one can prove that linear combinations of k Gaussians have at most 2^k zeros. However, since the number of zeros dictates the number of moments that we must match in our univariate estimation problem, we will use more powerful machinery to prove the tighter bound of $2(k-1)$ zeros. Our proof of Proposition 1 will hinge upon the following Theorem, due to Hummel and Gidas [13], and we defer the details to the full version of our paper.

THEOREM 7 (THM 2.1 IN [13]). *Given $f(x) : \mathbb{R} \to \mathbb{R}$, that is analytic and has n zeros, then for any $\sigma^2 > 0$, the function $g(x) = f(x) \circ \mathcal{N}(0, \sigma^2, x)$ has at most n zeros.*

Let $f(x) = \mathcal{F}_\alpha(F)(x) - \mathcal{F}_\alpha(F')(x)$, where α is chosen according to Lemma 5 so that $\int_x |f(x)|dx = \Omega(\epsilon^4)$.

LEMMA 8. *There is some $i \leq 6$ such that*

$$\left| \int_x x^i f(x)dx \right| = |M_i(\mathcal{F}_\alpha(F)) - M_i(\mathcal{F}_\alpha(F'))| = \Omega(\epsilon^{66})$$

A sketch of the proof of the above lemma is as follows: Let $x_1, x_2, \ldots, x_k$ be the zeros of $f(x)$ which have $|x_i| \leq \frac{2}{\epsilon}$. Using Proposition 1, the number of such zeros is at most the total number of zeros of $f(x)$ which is bounded by 6. (Although Proposition 1 only applies to linear combinations of Gaussians in which each Gaussian has a distinct variance, we can always perturb the Gaussians of $f(x)$ by negligibly small amounts so as to be able to apply the proposition.) We prove that there is some $i \leq 6$ for which $|M_i(\mathcal{F}_\alpha(F)) - M_i(\mathcal{F}_\alpha(F'))| = \Omega(poly(\epsilon))$ by constructing a degree 6 polynomial (with bounded coefficients) $p(x)$ for which $|\int_x f(x)p(x)dx| = \Omega(poly(\epsilon))$. Then if the coefficients of $p(x)$ can be bounded by some polynomial in $\frac{1}{\epsilon}$ we can conclude that there is some $i \leq 6$ for which the i^{th} moment of F is different from the i^{th} moment of $\hat{F}$ by at least $\Omega(poly(\epsilon))$. So we choose $p(x) = \pm \prod_{i=1}^{k}(x - x_i)$ and we choose the sign of $p(x)$ so that $p(x)$ has the same sign as $f(x)$ on the interval $I = [\frac{-2}{\epsilon}, \frac{2}{\epsilon}]$. Lemma 5 together with tail bounds imply that $\int_I |f(x)|dx \geq \Omega(\epsilon^4)$. To finish the proof, we show that $\int_I p(x)f(x)dx$ is large, and that $\int_{\mathbb{R}\setminus I} p(x)f(x)dx$ is negligibly small. We defer a full proof to the full version of our paper.

These tools are enough to yield a proof of Theorem 4.

3.2 The Univariate Algorithm

We now leverage the robust identifiability shown in Theorem 4 to prove that we can efficiently learn the parameters of 1-d GMM via a brute-force search over a set of candidate parameter sets. Roughly, the algorithm will take a polynomial number of samples, compute the first 6 sample moments, and compare those with the first 6 (analytic) moments of each of the candidate parameter sets. The algorithm then returns the parameter set whose moments most closely match the sample moments. Theorem 4 guarantees that if the first 6 sample moments closely match those of the chosen parameter set, then the parameter set must be nearly accurate. To conclude the proof, we argue that a polynomial-sized set of candidate parameters suffices to guarantee that at least one set of parameters will yield moments sufficiently close to the sample moments. We state the theorem below, and defer the details of the algorithm, and the proof of its correctness to the full version of our paper.

THEOREM 9. *Suppose we are given access to independent samples from any* **isotropic** *mixture $F = w_1F_1 + w_2F_2$, where $w_1 + w_2 = 1$, $w_i \geq \epsilon$, and each F_i is a univariate Gaussian with mean μ_i and variance σ_i^2, satisfying $|\mu_1 - \mu_2| + |\sigma_1^2 - \sigma_2^2| \geq \epsilon$,. Then Algorithm 1 will use $\text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta})$ samples and with probability at least $1-\delta$ will output mixture parameters $\hat{w}_1, \hat{w}_2, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma_1}^2, \hat{\sigma}_2^2$, so that there is a permutation $\pi : \{1,2\} \to \{1,2\}$ so that for each $i = 1, 2$*

$$|w_i - \hat{w}_{\pi(i)}| \leq \epsilon, \quad |\mu_i - \hat{\mu}_{\pi(i)}| \leq \epsilon, \quad |\sigma_i^2 - \hat{\sigma}_{\pi(i)}^2| \leq \epsilon$$

The brute-force search in the univariate algorithm is rather inefficient – we presented it for clarity of intuition, and ease of description and proof. Alternatively, we could have proceeded along the lines of Pearson's work [23]: using the first five sample moments, one generates a ninth degree polynomial whose solutions yield a small set of candidate parameter sets (which, one can argue, includes one set whose sixth moment closely matches the sixth sample moment). After picking the parameters whose sixth moment most closely matches the sample moment, we can use Theorem 4 to prove that the parameters have the desired accuracy.

4. THE N-DIMENSIONAL ALGORITHM

In this section, via a series of projections and applications of the univariate parameter learning algorithm of the previous section, we show how to efficiently learn the mixture parameters of an n-dimensional GMM. Let $\epsilon > 0$ be our target error accuracy. Let $\delta > 0$ be our target failure probability. For this section, we will suppose further that $w_1, w_2 \geq \epsilon$ and $D(F_1, F_2) \geq \epsilon$.

We first analyze our algorithm in the case where the GMM F is in *isotropic position*. This means that $\mathrm{E}_{x\sim F}[x] = 0$ and, $\mathrm{E}_{x\sim F}[xx^T] = I_n$. The above condition on the co-variance matrix is equivalent to $\forall u \in \mathbb{S}_{n-1}$ $\mathrm{E}_{x\sim F}[(u \cdot x)^2] = 1$. In the full version of our paper we explain the general case which involves first using a number of samples to put the distribution in (approximately) isotropic position, and then running the isotropic algorithm.

Given a mixture in isotropic position, we first argue that we can get ϵ-close additive approximations to the weights, means and variances of the Gaussians. This does not suffice to upper-bound $D(F_i, \hat{F}_i)$ in the case where F_i has small variance along one dimension. For example, consider a univariate GMM $F = \frac{1}{2}\mathcal{N}(0, 2-\epsilon') + \frac{1}{2}\mathcal{N}(0, \epsilon')$, where $\epsilon' \ll \epsilon$ is arbitrarily small (even possibly 0 – the Gaussian is a point mass). Note that an additive error of ϵ, say $\hat{\sigma}_2 = \epsilon' + \epsilon$ leads to a variation distance near $\frac{1}{2}$. In this case, however, $D(\hat{F}_1, \hat{F}_2)$ must be very close to 1, i.e., the Gaussians nearly do not overlap.[2] The solution is to use the additive approximation to the Gaussians to then cluster the data. From

[2]We are indebted to Santosh Vempala for suggesting this idea, namely, that if one of the Gaussians is very thin, then they must be almost non-overlapping and therefore clustering may be applied.

ALGORITHM 1. UNIVARIATE ESTIMATION

Input: $\epsilon > 0, \delta > 0$, sample oracle $\mathrm{SA}(F)$.

Output: For $i = 1, 2$, $(\hat{w}_i, \hat{\mu}_i, \hat{\sigma}_i) \in \mathbf{R}^3$.

(* Note: this algorithm assumes that the mixture is isotropic position, i.e., $\mathrm{E}[x] = 0, \mathrm{E}[x^2] = 1$. *)

1. Let $\hat{M}_1 = 0$, $\hat{M}_2 = 1$.
2. Let $\alpha = \epsilon^{150}$, and choose $m, N \in \mathbb{N}$ such that $m \geq \frac{500^2}{\alpha^2 \delta}$, and $N \geq \frac{1}{\alpha \epsilon^3}$.
3. For $i = 3, 4, 5, 6$:
   - Let $\hat{M}_i = \frac{1}{m} \sum_{j=1}^{m} x_i^j$, where $x_1, x_2, \ldots, x_m$ are m samples drawn from $\mathrm{SA}(F)$.

   (* Brute-force search: *)
4. Let $\mathcal{A} = \left\{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}\right\}$, $\mathcal{B} = \left\{-N, -N + \frac{1}{N}, -N + \frac{2}{N}, \ldots, N\right\}$.
5. For each $(\hat{w}_1, \hat{\mu}_1, \hat{\sigma_1}, \hat{\mu}_2, \hat{\sigma}_2) \in \mathcal{A} \times \mathcal{B}^4$, let:

$$\hat{w}_2 = 1 - \hat{w}_1 \tag{1}$$

$$\text{for } i = 1, 2, \ldots, 6: \quad \tilde{M}_i = M_i\left(\hat{w}_1 \mathcal{N}(\hat{\mu}_1, \hat{\sigma}_1^2) + \hat{w}_2 \mathcal{N}(\hat{\mu}_2, \hat{\sigma}_2^2)\right) \tag{2}$$

$$\mathrm{disc}_{\hat{w}_1, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{w}_2, \hat{\mu}_2, \hat{\sigma}_2^2} = \max_{i \in \{1,2,\ldots,6\}} \left|\hat{M}_i - \tilde{M}_i\right| \tag{3}$$

6. Output $(\hat{w}_1, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{w}_2, \hat{\mu}_2, \hat{\sigma}_2^2)$ of minimal $\mathrm{disc}_{\hat{w}_1, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{w}_2, \hat{\mu}_2, \hat{\sigma}_2^2}$ that satisfies $|\hat{\mu}_1 - \hat{\mu}_2| + |\hat{\sigma}_1^2 - \hat{\sigma}_2^2| > \epsilon$.

Figure 3: The one-dimensional estimation algorithm. For (2), evaluation of the moments of the distributions may be done exactly using explicit formulas for the first six moments of a Gaussian, given in Appendix ??.

clustered data, the problem is simply one of estimating a single Gaussian from random samples, which is easy to do in polynomial time.

4.1 Additive approximation

The algorithm for this case is given in Figures 4 and 5.

LEMMA 10. *For any $n \geq 1$, $\epsilon, \delta > 0$, for any any* isotropic *GMM mixture $F = w_1 F_1 + w_2 F_2$, where $w_1 + w_2 = 1$, $w_i \geq \epsilon$, and each F_i is a Gaussian in $\mathbf{R}^n$ with $D(F_1, F_2) \geq \epsilon$, with probability $\geq 1 - \delta$, (over the samples and randomization of the algorithm), Algorithm 2 will output GMM $\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ such that there exists a permutation $\pi : [2] \to [2]$ so that for each $i = 1, 2$*

$$\|\hat{\mu}_i - \mu_{\pi(i)}\| \leq \epsilon, \|\hat{\Sigma}_i - \Sigma_{\pi(i)}\|_F \leq \epsilon, \text{ and } |\hat{w}_i - w_{\pi(i)}| \leq \epsilon$$

And the runtime and number of samples drawn by Algorithm 2 is at most $\mathrm{poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta})$.

The rest of this section gives an outline of the proof of this lemma. We first state two geometric lemmas (Lemmas 11 and 12) that are independent of the algorithm.

LEMMA 11. *For any $\mu_1, \mu_2 \in \mathbf{R}^n, \delta > 0$, over uniformly random unit vectors u,*

$$\Pr_{u \in \mathbb{S}_{n-1}}\left[|u \cdot \mu_1 - u \cdot \mu_2| \leq \delta \|\mu_1 - \mu_2\| / \sqrt{n}\right] \leq \delta.$$

PROOF. If $\mu_1 = \mu_2$, the lemma is trivial. Otherwise, let $v = (\mu_1 - \mu_2)/\|\mu_1 - \mu_2\|$. The lemma is equivalent to claiming that

$$\Pr_{u \in \mathbb{S}_{n-1}}[|u \cdot v \leq t] \leq t\sqrt{n}.$$

This is a standard fact about random unit vectors (see, e.g., Lemma 1 of [6]). □

We next prove that, given a random unit vector r, with high probability either the projected means onto r or the projected variances onto r of F_1, F_2 must be different by at least $\mathrm{poly}(\epsilon, \frac{1}{n})$. A qualitative argument as to why this lemma is true is roughly: suppose that for most directions r, the projected means $r^T \mu_1$ and $r^T \mu_2$ are close, and the projected variances $r^T \Sigma_1 r$ and $r^T \Sigma_2 r$ are close, then the statistical distance $D(F_1, F_2)$ must be small too. So conversely, given $D(F_1, F_2) \geq \epsilon$ and $w_1, w_2 \geq \epsilon$ (and the distribution is in isotropic position), for most directions r either the projected means or the projected variances must be different.

LEMMA 12. *Let $\epsilon, \delta > 0$, $t \in (0, \epsilon^2)$. Suppose that $\|\mu_1 - \mu_2\| \leq t$. Then, for uniformly random r,*

$$\Pr_{r \in \mathbb{S}_{n-1}}\left[\min\{r^T \Sigma_1 r, r^T \Sigma_2 r\} > 1 - \frac{\epsilon \delta^2 (\epsilon^3 - t^2)}{12 n^2}\right] \leq \delta.$$

This lemma holds under the assumptions that we have already made about the mixture in this section (namely isotropy and lower bounds on the weights and statistical distance). While the above lemma is quite intuitive, the proof involves a probabilistic analysis based on the eigenvalues of the two covariance matrices, and is deferred to the full version.

Next, suppose that $r^T \mu_1 - r^T \mu_2 \geq \mathit{poly}(\epsilon, \frac{1}{n})$. Continuity arguments imply that if we choose a direction $r^{i,j}$ sufficiently close to r, then $(r^{i,j})^T \mu_1 - (r^{i,j})^T \mu_2$ will not change much from $r^T \mu_1 - r^T \mu_2$. So given a univariate algorithm that computes estimates for the mixture parameters in direction

ALGORITHM 2. HIGH-DIMENSIONAL ISOTROPIC ADDITIVE APPROXIMATION

Input: Integers $n \geq 1$, reals $\epsilon, \delta > 0$, sample oracle $\mathrm{SA}(F)$.

Output: For $i = 1, 2$, $(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i) \in \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^{n \times n}$.

1. Let $\epsilon_4 = \frac{\epsilon\delta}{100n}$, and $\epsilon_i = \epsilon_{i+1}^{10}$, for $i = 3, 2, 1$.
2. Choose uniformly random orthonormal basis $B = (b_1, \ldots, b_n) \in \mathbf{R}^{n \times n}$.
   Let $r = \sum_{i=1}^n b_i/\sqrt{n}$. Let $r^{ij} = r + \epsilon_2 b_i + \epsilon_2 b_j$ for each $i, j \in [n]$.
3. Run $\textsc{Univariate}\big(\epsilon_1, \frac{\delta}{3n^2}, \mathrm{SA}(P_r(F))\big)$ to get a univariate mixture of Gaussians $\hat{w}_1 G_1^0 + \hat{w}_2 G_2^0$.
4. If $\min\{\hat{w}_1, \hat{w}_2\} < \epsilon_3$ or $\max\big\{|\mu(G_1^0) - \mu(G_2^0)|, |\mathrm{var}(G_1^0) - \mathrm{var}(G_2^0)|\big\} < \epsilon_3$, then halt and output $FAIL$ ELSE:
5. If $|\mu(G_1^0) - \mu(G_2^0)| > \epsilon_3$, then:
   - Permute G_1^0, G_2^0 (and $\hat{w}_1, \hat{w}_2$) as necessary so that $\mu(G_1^0) < \mu(G_2^0)$. ELSE
   - Permute G_1^0, G_2^0 (and $\hat{w}_1, \hat{w}_2$) as necessary so $\mathrm{var}(G_1^0) < \mathrm{var}(G_2^0)$.
6. For each $i, j \in [n]$:
   - Run $\textsc{Univariate}\big(\epsilon_1, \frac{\delta}{3n^2}, \mathrm{SA}(P_{r^{ij}}(F))\big)$ to get estimates of a univariate mixture of two Gaussians, G_1^{ij}, G_2^{ij}. (* We ignore the weights returned by the algorithm. *)
   - If $\mu(G_2^0) - \mu(G_1^0) > \epsilon_3$, then:
     - Permute G_1^{ij}, G_2^{ij} as necessary so that $\mu(G_1^{ij}) < \mu(G_2^{ij})$. ELSE
     - Permute G_1^{ij}, G_2^{ij} as necessary so that $\mathrm{var}(G_1^{ij}) < \mathrm{var}(G_2^{ij})$.
7. Output $\hat{w}_1, \hat{w}_2$, and for $\ell \in \{1, 2\}$, and output

$$(\hat{\mu}_\ell, \hat{\Sigma}_\ell) = \textsc{Solve}\left(n, \epsilon_2, B, \mu(G_\ell^0), \mathrm{var}(G_\ell^0), \left\langle \mu(G_\ell^{ij}), \mathrm{var}(G_\ell^{ij}) \right\rangle_{i,j \in [n]}\right)$$

Figure 4: A dimension reduction algorithm. Although ϵ_4 is not used by the algorithm, it is helpful to define it for the analysis. We choose such ridiculously small parameters to make it clear that our efforts are placed on simplicity of presentation rather than tightness of parameters.

ALGORITHM 3. SOLVE

Input: $n \geq 1$, $\epsilon_2 > 0$, basis $B = (b_1, \ldots, b_n) \in \mathbf{R}^{n \times n}$, means and variances m^0, v^0, and $m^{ij}, v^{ij} \in \mathbf{R}$ for each $i, j \in [n]$.

Output: $\hat{\mu} \in \mathbf{R}^n$, $\hat{\Sigma} \in \mathbf{R}^{n \times n}$.

1. Let $v^i = \frac{1}{n}\sum_{j=1}^n v^{ij}$ and $v = \frac{1}{n^2}\sum_{i=1}^n v^{ij}$.
2. For each $i \leq j \in [n]$, let

$$V_{ij} = \frac{\sqrt{n}(v - v^i - v^j)}{(2\epsilon_2 + \sqrt{n})2\epsilon_2^2} - \frac{v^{ii} + v^{jj}}{(2\epsilon_2 + \sqrt{n})4\epsilon_2} - \frac{v^0}{2\epsilon_2\sqrt{n}} + \frac{v^{ij}}{2\epsilon_2^2}.$$

3. For each $i > j \in [n]$, let $V_{ij} = V_{ji}$. (* So $V \in \mathbf{R}^{n \times n}$ *)
4. Output

$$\hat{\mu} = \sum_{i=1}^n \frac{m^{ii} - m^0}{2\epsilon_2} b_i, \quad \hat{\Sigma} = B\left(\arg\min_{M \succeq 0} \|M - V\|_F\right) B^T.$$

Figure 5: Solving the equations. In the last step, we project onto the set of positive semidefinite matrices, which can be done in polynomial time using semidefinite programming.

r and in direction $r^{i,j}$, we can determine a pairing of these parameters so that we now have estimates for the mean of F_1 projected on r and estimates for the mean of F_1 projected on $r^{i,j}$, and similarly we have estimates for the projected variances (on r and $r^{i,j}$) of F_1. From sufficiently many of these estimates in different directions $r^{i,j}$, we can hope to recover the mean and covariance matrix of F_1, and similarly for F_2. An analogous statement will also hold in the case that for direction r, the projected variances are different in which case choosing a direction $r^{i,j}$ sufficiently close to r will result in not much change in the projected variances, and we can similarly use these continuity arguments (and a univariate algorithm) to again recover many estimates in different directions.

LEMMA 13. *For r, r^{ij} of Algorithm 2, (a) With probability $\geq 1-\delta$ over the random unit vector r, $|r\cdot(\mu_1-\mu_2)| > 2\epsilon_3$ or $|r^T(\Sigma_1-\Sigma_2)r| > 2\epsilon_3$, (b) $|(r^{ij}-r)\cdot(\mu_1-\mu_2)| \leq \epsilon_3/3$, and (c) $|(r^{ij})^T(\Sigma_1-\Sigma_2)r^{ij} - r^T(\Sigma_1-\Sigma_2)r| \leq \epsilon_3/3$.*

The proof, based on Lemma 12, is given in the full version. We then argue that SOLVE outputs the desired parameters. Given estimates of the projected mean and projected variance of F_1 in n^2 directions $r^{i,j}$, each such estimate yields a linear constraint on the mean and covariance matrix. Provided that each estimate is close to the correct projected mean and projected variance, we can recover an accurate estimate of the parameters of F_1, and similarly for F_2. Thus, using the algorithm for estimating mixture parameters for univariate GMMs $F = w_1F_1 + w_2F_2$, we can get a polynomial time algorithm for estimating mixture parameters in n-dimensions for isotropic Gaussian mixtures. Further details are deferred to the full version.

LEMMA 14. *Let $\epsilon_2, \epsilon_1 > 0$. Suppose $|m^0 - \mu\cdot r|$,$|m^{ij} - \mu\cdot r^{ij}|$, $|v^0 - r^T\Sigma r|$,$|v^{ij} - (r^{ij})^T\Sigma r^{ij}|$ are all at most ϵ_1. Then* SOLVE *outputs $\hat{\mu} \in \mathbf{R}^n$ and $\hat{\Sigma} \in \mathbf{R}^{n\times n}$ such that $\|\hat{\mu}-\mu\| < \epsilon$, and $\|\hat{\Sigma}-\Sigma\|_F \leq \epsilon$. Furthermore, $\hat{\Sigma} \succeq 0$ and $\hat{\Sigma}$ is symmetric.*

4.2 Statistical approximation

It remains to achieve, with high probability, approximations to the Gaussians that are close in terms of variation distance. An additive bound on error yields bounded variation distance, only for Gaussians that are relatively "round," in the sense that their covariance matrix has a smallest eigenvalue that is bounded away from 0. However, if, for isotropic F, one of the Gaussians has a very small eigenvalue, then this means that they are practically nonoverlapping, i.e., $D(F_1, F_2)$ is close to 1. In this case, our estimates from Algorithm 2 are good enough, with high probability, to cluster a polynomial amount of data into two clusters based on whether it came from Gaussian F_1 or F_2. After that, we can easily estimate the parameters of the two Gaussians.

LEMMA 15. *There exists a polynomial p such that, for any $n \geq 1$, $\epsilon, \delta > 0$, for any any* isotropic *GMM mixture $F = w_1F_1 + w_2F_2$, where $w_1 + w_2 = 1$, $w_i \geq \epsilon$, and each F_i is a Gaussian in $\mathbf{R}^n$ with $D(F_1, F_2) \geq \epsilon$, with probability $\geq 1-\delta$, (over its own randomness and the samples), Algorithm 4 will output GMM $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$ such that there exists a permutation $\pi : [2] \to [2]$ with,*

$$D(\hat{F}_i, F_{\pi(i)}) \leq \epsilon, \text{ and } |\hat{w}_i - w_{\pi(i)}| \leq \epsilon, \text{for each } i = 1, 2.$$

The runtime and number of samples drawn by Algorithm 4 is at most $\text{poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta})$.

The statistical approximation algorithm is given in Algorithm 4, and proof of the above lemma is deferred to the full version.

5. DENSITY ESTIMATION

The problem of PAC learning a distribution (or density estimation) was introduced in [16]: Given parameters $\epsilon, \delta > 0$, and given oracle access to a distribution F (in n dimensions), the goal is to learn a distribution $\hat{F}$ so that with probability at least $1-\delta$, $D(F, \hat{F}) \leq \epsilon$ in time polynomial in $\frac{1}{\epsilon}$, n, and $\frac{1}{\delta}$. Here we apply our algorithm for learning mixtures of two arbitrary Gaussians to the problem of polynomial-time density estimation (aka PAC learning distributions) for arbitrary mixtures of two Gaussians without any assumptions. We show that given oracle access to a distribution $F = w_1F_1 + w_2F_2$ for $F_i = \mathcal{N}(\mu_i, \Sigma_i)$, we can efficiently construct a mixture of two Gaussians $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$ for which $D(F, \hat{F}) \leq \epsilon$. Previous work on this problem [10] required that the Gaussians be axis aligned.

The algorithm for density estimation and a proof of correctness is deferred to the full version.

6. CLUSTERING

It makes sense that knowing the mixture parameters should imply that one can perform optimal clustering, and approximating the parameters should imply approximately optimal clustering. In this section, we formalize this intuition. For GMM F, it will be convenient to consider the *labeled distribution* $\ell(F)$ over $(x, y) \in \mathbf{R}^n \times \{1, 2\}$ in which a *label* $y \in \{1, 2\}$ is drawn with probability w_i of i, and then a sample x is chosen from F_i.

A clustering algorithm takes as input m examples

$$x_1, x_2, \ldots, x_m \in \mathbf{R}^n$$

and outputs a classifier $C : \mathbf{R}^n \to \{1, 2\}$ for future data (a similar analysis could be done in terms of partitioning data $x_1, \ldots, x_m$). The *error* of a classifier C is defined to be,

$$\text{err}(C) = \min_{\pi} \Pr_{(x,y)\sim\ell(F)}[C(x) \neq y],$$

where the minimum is over permutations $\pi : \{1, 2\} \to \{1, 2\}$. In other words, it is the fraction of points that must be relabeled so that they are partitioned correctly (actual label is irrelevant).

For any GMM F, define C_F to be the classifier that outputs whichever Gaussian has a greater posterior: $C_F(x) = 1$ if $w_1F_1(x) \geq w_2F_2(x)$, and $C(x) = 2$ otherwise. It is not difficult to see that this classifier has minimum error.

Corollary 3 implies that given a polynomial number of points, one can cluster *future samples* with near-optimal expected error. But using standard reductions, this also implies that we can learn and accurately cluster our training set as well. Namely, one could run the clustering algorithm on, say, $\sqrt{m}$ of the samples, and then use it to partition the data. The algorithm for near-optimal clustering is given in the full version, along with a proof for correctness.

Acknowledgments. We are grateful to Santosh Vempala, Charlie Brubaker, Yuval Peres, Daniel Stefankovic, and Paul Valiant for helpful discussions.

ALGORITHM 4. HIGH-DIMENSIONAL ISOTROPIC VARIATION-DISTANCE APPROXIMATION

Input: Integers $n \geq 1$, reals $\epsilon, \delta > 0$, sample oracle $\mathrm{SA}(F)$.
Output: n-dimensional GMM

1. Let $\varepsilon_1, \varepsilon_2, \varepsilon_3 =$
2. Run Algorithm 2$(n, \varepsilon_1, \delta/3, \mathrm{SA}(F))$ to get $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$.
3. Permute $\hat{w}_i, \hat{F}_i$ so that the *smallest* eigenvalue of $\mathrm{var}(\hat{F})_1$ is no larger than the smallest eigenvalue of $\mathrm{var}(\hat{F})_2$.
4. If the smallest eigenvalue of $\mathrm{var}(\hat{F}_1)$ is greater than ε_2, then halt and output the mixture $\hat{F}$. ELSE: (* Clustering step *)
   (a) Let λ, v be a smallest eigenvalue and corresponding unit eigenvector of $\hat{F}_1$.
   (b) Draw $m = \varepsilon_4^{-1}$ samples $x_1, \ldots, x_m$ from $\mathrm{SA}(F)$.
   (c) Partition the data into two sets, $D_1 \cup D_2 = \{x_1, \ldots, x_m\}$, where,

$$D_1 = \left\{ x_i : \left| P_v(x_i) - P_v(\mu(\hat{F}_1)) \right| \leq \frac{\sqrt{\varepsilon_2}}{\varepsilon_3} \right\}.$$

   (d) Output GMM $\hat{G} = \hat{w}_1\hat{G}_1 + \hat{w}_2\hat{G}_2$, where $\hat{G}_i$ is the Gaussian with mean and covariance matrix that matches the empirical mean and covariance on set D_i, and $\hat{w}_i$ are those from Step 2.

Figure 6: The algorithm that guarantees low variation distance.

7. REFERENCES

[1] D. Achlioptas and F. McSherry: On Spectral Learning of Mixtures of Distributions. *Proc. of COLT*, 2005.

[2] S. Arora and R. Kannan: Learning mixtures of arbitrary Gaussians. *Ann. Appl. Probab.* 15 (2005), no. 1A, 69–92.

[3] M. Belkin and K. Sinha, technical report arXiv:0907.1054v1, July 2009.

[4] C. Brubaker and S. Vempala: Isotropic PCA and Affine-Invariant Clustering. *Proc. of FOCS*, 2008.

[5] S. Dasgupta: Learning mixtures of Gaussians. *Proc. of FOCS*, 1999.

[6] S. Dasgupta, A. Kalai, and C. Monteleoni: Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281-299, 2009

[7] S. Dasgupta and L. Schulman: A two-round variant of EM for Gaussian mixtures. *Uncertainty in Artificial Intelligence*, 2000.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. With discussion. *J. Roy. Statist. Soc. Ser. B* 39 (1977), no. 1, 1–38.

[9] A. Dinghas: Über eine Klasse superadditiver Mengenfunktionale von Brunn–Minkowski–Lusternik-schem Typus, *Math. Zeitschr.* **68**, 111–125, 1957.

[10] J. Feldman, R. Servedio and R. O'Donnell: PAC Learning Axis-Aligned Mixtures of Gaussians with No Separation Assumption. *Proc. of COLT*, 2006.

[11] A. A. Giannopoulos and V. D. Milman: Concentration property on probability spaces. Adv. Math. 156(1), 77–106, 2000.

[12] G. Golub and C. Van Loan: *Matrix Computations*, Johns Hopkins University Press, 1989.

[13] R. A. Hummel and B. C. Gidas, "Zero Crossings and the Heat Equation", Technical Report number 111, Courant Insitute of Mathematical Sciences at NYU, 1984.

[14] R. Kannan, H. Salmasian and S. Vempala: The Spectral Method for Mixture Models. *Proc. of COLT*, 2005.

[15] M. Kearns and U. Vazirani: An Introduction to Computational Learning Theory, MIT Press, 1994.

[16] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire and L. Sellie: On the learnability of discrete distributions. *Proc of STOC*, 1994

[17] L. Leindler: On a certain converse of Hölder's Inequality II, *Acta Sci. Math. Szeged* 33 (1972), 217–223.

[18] B. Lindsay: *Mixture models: theory, geometry and applications*. American Statistical Association, Virginia 1995.

[19] L. Lovász and S. Vempala: The Geometry of Logconcave functions and sampling algorithms. *Random Strucures and Algorithms*, 30(3) (2007), 307–358.

[20] P.D.M. Macdonald, personal communication, November 2009.

[21] G.J. McLachlan and D. Peel, *Finite Mixture Models* (2009), Wiley.

[22] R. Motwani and P. Raghavan: Randomized Algorithms, Cambridge University Press, 1995.

[23] K. Pearson: Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London. A*, 1894.

[24] A. Prékopa: Logarithmic concave measures and functions, *Acta Sci. Math. Szeged* 34 (1973), 335–343.

[25] M. Rudelson: Random vectors in the isotropic position, *J. Funct. Anal.* **164** (1999), 60–72.

[26] H. Teicher. Identifiability of mixtures, *Ann. Math. Stat.* 32 (1961), 244Ű248.

[27] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions* (1985), Wiley.

[28] L. Valiant: A theory of the learnable. *Communications of the ACM*, 1984.

[29] S. Vempala: On the Spectral Method for Mixture Models, IMA workshop on Data Analysis and Optimization, 2003 http://www.ima.umn.edu/talks/workshops/5-6-9.2003/vempala/vempala.html

[30] S. Vempala and G. Wang: A spectral algorithm for learning mixtures of distributions, *Proc. of FOCS*, 2002; J. Comput. System Sci. 68(4), 841–860, 2004.

[31] S. Vempala and G. Wang: The benefit of spectral projection for document clustering. *Proc. of the 3rd Workshop on Clustering High Dimensional Data and its Applications*, SIAM International Conference on Data Mining (2005).

[32] C.F.J. Wu: On the Convergence Properties of the EM Algorithm, *The Annals of Statistics* (1983), Vol.11, No.1, 95–103.