Efficiently Learning Mixtures of Two Arbitrary Gaussians

Adam Tauman Kalai [*] Ankur Moitra [†] Gregory Valiant [‡]

April 7, 2010

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. The main 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).

[*]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 paper is not eligible for best student paper.

[†]Massachusetts Institute of Technology. Supported in part by a Fannie and John Hertz Foundation Fellowship. Part of this work done while at Microsoft Research New England.

[‡]University of California, Berkeley. Supported in part by an NSF Graduate Research Fellowship. Part of this work done while at Microsoft Research New England.

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). We start with a brief history, then give 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 on 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 among them whose sixth moment most closely agreed with the empirical estimate.

Later work showed that "identifiability" is theoretically possible – every two different mixtures of different Gaussians (up to a permutation on the Gaussian labels, of course) have different probability distributions [26]. However, this work shed little light on convergence *rates* as they were based on differences in the density tails which would require enormous amounts of data to distinguish. In particular, to ϵ-approximate the Gaussian parameters in the sense that we will soon describe, previous work left open the possibility that it 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. 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]. This work generally required the Gaussians to have little *overlap* (statistical distance near 1); in many such cases they were able to find computationally efficient algorithms for GMMs of more than two Gaussians. Recently, a polynomial-time *density estimation*[1] algorithm was given for *axis-aligned* GMMs, without any nonoverlap assumption [10].

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 $F = w_1 F_1 + w_2 F_2$, 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 for estimation.
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.

[1] Density estimation refers to the easier problem of approximating the overall density without necessarily well-approximating individual Gaussians, and axis-aligned Gaussians are those whose principal axes are parallel to the coordinate axes.

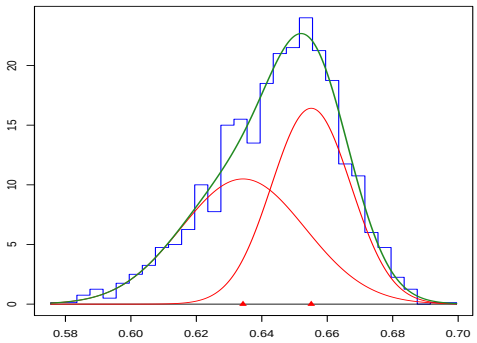

Figure 1: A fit of a mixture of two univariate Gaussians to the Pearson's data on Naples crabs [23]. The hypothesis was that the data was in fact a mixture of two different species of crabs. Although the empirical data histogram is single-peaked, the two constituent Gaussian parameters may be estimated.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

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. Second, we would like the algorithm to succeed in this goal using polynomially many samples. Lastly, we would 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 n-dimensional GMM $F = w_1F_1 + w_2F_2$, using m independent samples from F, Algorithm 5 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), 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 1 is at most $\text{poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{w_1}, \frac{1}{w_2}, \frac{1}{D(F_1,F_2)})$

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, 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 n-dimensional GMM $F = w_1F_1 + w_2F_2$, 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

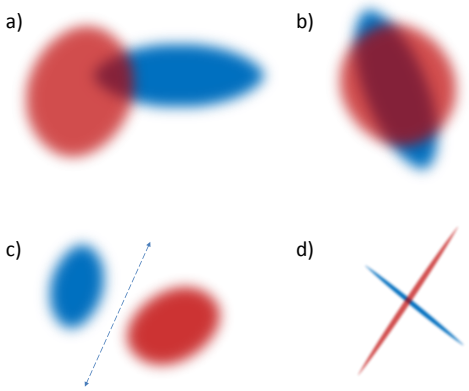

Figure 2: Mixtures of two multinormal distributions, with varying amounts of overlap. In (c), the Gaussians are nearly separable by a hyperplane, and the algorithm of Brubaker and Vempala [4] can cluster and learn them. In (d), the Gaussians are nearly separable but not by any hyperplane. Our algorithm will learn the parameters in all cases, and hence be able to cluster when possible.

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 random 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 the Gaussian with larger posterior probability. We are able to achieve near optimal clustering without *a priori* knowledge of the distribution parameters. See Appendix C for precise details.

Corollary 3. *For any $n \geq 1, \epsilon, \delta > 0$ and any n-dimensional GMM $F = w_1 F_1 + w_2 F_2$, 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 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 apply a one-dimensional algorithm.

When the data is projected down to a line, each $(mean, variance)$ pair recovered in this direction gives direct information about the corresponding $(mean, variance)$ pair in n dimensions. Lemma 13 states that for a suitably chosen *random* direction[2], 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, the projected mean and variance in $O(n^2)$ different directions is enough to approximate the Gaussian. The remaining challenge is identifying which Gaussian in one projection corresponds to which Gaussian in another projection; one must correctly *match up* the many pairs of Gaussians yielded by each one-dimensional problem. In practice, the mixing weights may be somewhat different, i.e., $|w_1 - w_2|$ is bounded from 0. Then 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].

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

We now proceed to describe how to identify univariate GMMs. Like many one-dimensional problems, it is algorithmically *easy* as simple brute-force algorithms (like that of [10]) will work. The surprising difficulty is proving that an algorithm approximates the constituent Gaussians well. 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 the bulk of our work.

We appeal to the old method of moments. In particular, the key fact is that 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. Formally, $\left|\mathrm{E}_{x\sim F}[x^i] - \mathrm{E}_{x\sim F'}[x^i]\right|$ 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, one algorithm which then works is the following. 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. We now describe the proof of Theorem 4. Two ideas are relating the statistical distance of two mixtures to the discrepancy in the moments, and *deconvolving*.

1.3.1 Relating statistical distance and discrepancy in moments

If two (bounded or almost bounded) distributions are statistically close, then there 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 "go up and down" many times, which cannot happen for mixtures of two univariate Gaussians. Lemma 9 shows that if two univariate GMMs have non-negligible 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 9. 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 address the issues of bounding the coefficients of $p(x)$, and making sure that the mass of $f(x)$ is not too concentrated near any zero). This finally 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 the two Gaussians. 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 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 will have large statistical distance. A more elaborate case analysis shows that the two GMMs have nonnegligible statistical distance when one of the Gaussians is skinny. (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 9); 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 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 One-Dimensional (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 10 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, showing that one of the first six raw moments will differ by at least $\mathrm{poly}(\epsilon)$. 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.

Throughout this section, we will make use of a variety of inequalities and concentration bounds for Gaussians which are included in Appendix K.

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. $\leq w_i, w_i' \in [\epsilon, 1]$
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 ϵ-standard F, F' and any $\epsilon < c$,*

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

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 $\text{poly}(\epsilon)$. 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 will be by 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 $\Omega(\epsilon^4)$ ℓ_1 norm 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 $\Omega(\epsilon^4)$ ℓ_1 norm 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 Appendix D.

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 $\text{poly}(\epsilon)$ 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 $\text{poly}(\epsilon)$ 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*

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

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 Appendix D.

To complete the proof of the theorem, we must show that the $\text{poly}(\epsilon)$ statistical distance between $\mathcal{F}_\alpha(F)$ and $\mathcal{F}_\alpha(F')$ gives rise to a $\text{poly}(\epsilon)$ 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 construct a degree six polynomial $p(x)$ that always has the same sign as $f(x)$, and when integrated against $f(x)$ is at least $\text{poly}(\epsilon)$. 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 7. *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 function is at most one more than the number of zeros of its derivative, one can prove that linear combinations 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 7 will hinge upon the following Theorem, due to Hummel and Gidas [13], and we defer the details to Appendix D.

Theorem 8 (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.*

We are now equipped to complete our proof of Theorem 4. 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 9. *There is some $i \leq 6$ such that*

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

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 7, the number of such zeros is at most the total number of zeros of $f(x)$ which is bounded by 6. (Although Proposition 7 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. The full proof is in Appendix D.

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 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 Appendix D.1.

Theorem 10. *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 3 will use $\mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\delta})$ 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 $\pi : \{1, 2\} \to \{1, 2\}$ and*

$$|w_i - \hat{w}_{\pi(i)}| \leq \epsilon, \quad |\mu_i - \hat{\mu}_{\pi(i)}| \leq \epsilon, \quad |\sigma_i^2 - \hat{\sigma}_{\pi(i)}^2| \leq \epsilon \text{ for each } i = 1, 2$$

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 parameter-learning 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 Appendix L we explain the general case which involves first using a number of samples to put the distribution in (approximately) isotropic position.

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 $F = w_1\mathcal{N}(0, 2-\epsilon') + w_2\mathcal{N}(0, \epsilon')$, where $\epsilon' \ll \epsilon$ is arbitrarily small (even possibly 0 – the Gaussian is a point mass). Note that an additive error of ϵ, say $\sigma_2 = \epsilon' + \epsilon$ leads to a variation distance near w_2. In high dimensions, this problem can occur in any direction in which one Gaussian has small variance. 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.[3] The solution is to use the additive approximation to the Gaussians to then cluster the data. From 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 3 and 4.

Lemma 11. *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 the samples and randomization of the algorithm), Algorithm 1 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,*

$$\|\hat{\mu}_i - \mu_{\pi(i)}\| \leq \epsilon, \|\hat{\Sigma}_i - \Sigma_{\pi(i)}\|_F \leq \epsilon, \text{ and } |\hat{w}_i - w_{\pi(i)}| \leq \epsilon, \text{for each } i = 1, 2.$$

And the runtime and number of samples drawn by Algorithm 1 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 12 and 13) that are independent of the algorithm.

Lemma 12. *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 $r^T\mu_2$

[3] 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. 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 3: 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 2. 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 4: 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.

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

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

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). 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 Appendix E.

Next, suppose that $r^T\mu_1 - r^T\mu_2 \geq poly(\epsilon, \frac{1}{n})$. 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 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 14. *For r, r^{ij} of Algorithm 1, (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 13, is given in Appendix F. 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 Appendices G and H.

Lemma 15. *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

In this section, we argue that we can guarantee, 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 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 1 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 16. *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 (Algorithm 4), and proof of the above lemma are given in Appendix I.

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

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, personal communication, 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), 244248.

[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.

A Conclusion

In conclusion, we have given polynomial rate bounds for the problem of estimating mixtures of two Gaussians in n dimensions, under provably minimal assumptions. No such efficient algorithms or rate bounds were known even for the problem in one dimension. The notion of accuracy we use is affine invariant, and our guarantees imply accurate density estimation and clustering as well. Questions: What is the optimal rate of convergence? How can one extend this to mixtures of more than two Gaussians?

B 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 $F = w_1 F_1 + w_2 F_2$ 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 is given in Appendix M, along with a proof of correctness.

C 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,

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

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 $C_F(x_1, \ldots, x_m)$ to be the classifier that outputs whichever Gaussian has a greater posterior: $C(x) = 1$ if $w_1 F_1(x) \geq w_2 F_2(x)$, 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 Appendix N, along with a proof for correctness.

D Proofs from Section 3

Proof of Lemma 6. Let X be a random variable with distribution $\mathcal{F}_\alpha\left(\mathcal{N}(\mu,\sigma^2)\right)$, and Y a random variable with distribution $\mathcal{N}(\mu,\sigma^2)$. From definition 2 and the fact that the sum of two independent Gaussian random variables is also a Gaussian random variable, it follows that $M_i(Y) = M_i(X+Z)$, where Z is a random variable, independent from X with distribution $\mathcal{N}(0,\alpha)$. From the independence of X and Z we have that

$$M_i(Y) = \sum_{j=0}^{i} \binom{i}{j} M_{i-j}(X) M_j(Z).$$

Since each moment $M_i(\mathcal{N}(\mu,\sigma^2))$ is some polynomial of μ,σ^2, which we shall denote by $m_i(\mu,\sigma^2)$, and the above equality holds for some interval of parameters, the above equation relating the moments of Y to those of X and Z is simply a polynomial identity:

$$m_i(\mu,\sigma^2) = \sum_{j=0}^{i} \binom{i}{j} m_{i-j}(\mu,\sigma^2-\beta) m_j(0,\beta).$$

Given this polynomial identity, if we set $\beta = -\alpha$, we can interpret this identity as

$$M_i(X) = \sum_{j=0}^{i} \binom{i}{j} M_{i-j}(Y)\left(c_j M_j(Z)\right),$$

where $c_j = \pm 1$ according to whether j is a multiple of 4 or not.

Let $d = \sum_{i=1}^{k} |M_i\left(\mathcal{F}_\alpha(D)\right) - M_i\left(\mathcal{F}_\alpha(D')\right)|$, and chose $j \le k$ such that $|M_j\left(\mathcal{F}_\alpha(D)\right) - M_j\left(\mathcal{F}_\alpha(D')\right)| \ge d/k$. From above, and by linearity of expectation, we get

$$\begin{aligned}
\frac{d}{k} &\le |M_j\left(\mathcal{F}_\alpha(D)\right) - M_j\left(\mathcal{F}_\alpha(D')\right)| \\
&= \sum_{i=0}^{j} \binom{j}{i} \left(M_{j-i}(D) - M_{j-i}(D')\right) c_i M_i(\mathcal{N}(0,\alpha)) \\
&\le \left(\sum_{i=0}^{j} \binom{j}{i} |M_{j-i}(D) - M_{j-i}(D')|\right) \max_{i\in\{0,1,\ldots,k-1\}} |M_i(\mathcal{N}(0,\alpha))|.
\end{aligned}$$

In the above we have used the fact that $M_k(\mathcal{N}(0,\alpha)))$ can only appear in the above sum along with $|M_0(D) - M_0(D')| = 0$. Finally, using the facts that $\binom{j}{i} < 2^k$, and expressions for the raw moments of $\mathcal{N}(0,\alpha)$ given by Equation (17), the above sum is at most $\frac{(k-1)!}{\lfloor k/2 \rfloor!} \sum_{i=0}^{k-1} |M_{j-i}(D) - M_{j-i}(D')|$, which completes the proof. □

The following claim will be useful in the proof of Lemma 5.

Claim 17. *Let $f(x^*) \ge M$ for $x^* \in (0,r)$ and suppose that $f(x) \ge 0$ on $(0,r)$ and $f(0) = f(r) = 0$. Suppose also that $f'(x) \le m$ everywhere. Then $\int_0^r f(x)dx \ge \frac{M^2}{m}$*

Proof. Note that for any $p \ge 0$, $f(x^* - p) \ge M - pm$, otherwise if $f(x^* - p) < M - pm$ then there must be a point $x \in (x^* - p, x^*)$ for which $f'(x) > \frac{M-(M-pm)}{x^*-(x^*-p)} = m$ which yields a contradiction.

So

$$\int_0^r f(x)dx \ge \int_{x^*-\frac{M}{m}}^{x^*} f(x)dx \ge \int_{x^*-\frac{M}{m}}^{x^*} M - m(x^* - x)dx$$

$$\int_{x^*-\frac{M}{m}}^{x^*} M - m(x^* - x)dx = \frac{M^2}{m} - m(\frac{M}{m})x^* + \frac{m}{2}[(x^*)^2 - (x^*)^2 + 2\frac{M}{m}x^* - \frac{M^2}{m}]$$

$$\int_{x^*-\frac{M}{m}}^{x^*} M - m(x^* - x)dx = \frac{M^2}{m} - Mx^* + \frac{m}{2}[2\frac{M}{m}x^* - \frac{M^2}{m}] = \frac{M^2}{2m}$$

And an identical argument on the interval (x^*, r) yields the inequality. □

Additionally, we use the following fact:

Fact 18.

$$\begin{aligned} \|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\mu, \sigma^2(1+\delta))\|_1 &\leq 10\delta \\ \|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\mu + \sigma\delta, \sigma^2)\|_1 &\leq 10\delta \end{aligned}$$

Proof. Let $F_1 = \mathcal{N}(\mu, \sigma^2(1+\delta))$ and let $F_2 = \mathcal{N}(\mu, \sigma^2)$. Then

$$\begin{aligned} KL(F_1\|F_2) &= \ln\frac{\sigma_2}{\sigma_1} + \frac{(\mu_1-\mu_2)^2 + \sigma_1^2 - \sigma_2^2}{2\sigma_2^2} \\ &= -\frac{1}{2}\ln(1+\delta) + \frac{\delta\sigma^2}{2\sigma^2} \\ &\leq -\frac{\delta}{2} + \frac{\delta^2}{4} + \frac{\delta}{2} = \frac{\delta^2}{4} \end{aligned}$$

where in the last line we have used the Taylor series expansion for $\ln 1 + x = x - \frac{x^2}{2} + \frac{x^3}{3} \ldots$ and the fact that $\ln 1 + x \geq x - \frac{x^2}{2}$. Then because $\|F_1 - F_2\| \leq \sqrt{2KL(F_1\|F_2)}$, we get that

$$\|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\mu, \sigma^2(1+\delta))\|_1 \leq 10\delta$$

Next, consider $F_1 = \mathcal{N}(\mu, \sigma^2)$ and $F_2 = \mathcal{N}(\mu + \sigma\delta, \sigma^2)$. In this case

$$\begin{aligned} KL(F_1\|F_2) &= \ln\frac{\sigma_2}{\sigma_1} + \frac{(\mu_1-\mu_2)^2 + \sigma_1^2 - \sigma_2^2}{2\sigma_2^2} \\ &= \ln\frac{\sigma}{\sigma} + \frac{(\mu + \sigma\sigma - \mu)^2 + \sigma^2 - \sigma^2}{2\sigma^2} \\ &= \frac{\delta^2}{2} \end{aligned}$$

And again because $\|F_1 - F_2\| \leq \sqrt{2KL(F_1\|F_2)}$, we get that

$$\|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\mu + \sigma\delta, \sigma^2)\|_1 \leq 10\delta$$

□

Proof of lemma 5: The restrictions that F, F' be ϵ-standard are symmetric w.r.t. F and F' so we will assume without loss of generality that the first constituent Gaussian of mixture F has the minimal variance among all Gaussians in F and F'. That is $\sigma_1^2 \leq \sigma_i^2, \sigma_i'^2$. We employ a case analysis:

Case 1: For both $i = 1$ and $i = 2$, either $\sigma_i'^2 - \sigma_1^2 \geq 16\epsilon^{2a}$, or $|\mu_i' - \mu_1| \geq 6\epsilon^a$

We choose $\alpha = \sigma_1^2 - \epsilon^{2a+2}$, which may be negative, and apply $\mathcal{F}_\alpha$ to F and F'. Note that $\mathcal{F}_\alpha$ transforms the first Gaussian component of F into a Gaussian of variance ϵ^{2a+2}, and thus $\mathcal{F}_\alpha(F)(\mu_1) \geq w_1\frac{1}{\epsilon^{a+1}\sqrt{2\pi}} \geq \frac{1}{\epsilon^a\sqrt{2\pi}}$.

Next we bound $\mathcal{F}_\alpha(F')(\mu_1)$, and we do this by considering the contribution of each component of $\mathcal{F}_\alpha(F')$ to $\mathcal{F}_\alpha(F')(\mu_1)$. Each component either has large variance, or the mean is far from μ_1. Consider the case in which a component has large variance - i.e. $\sigma^2 \geq 16\epsilon^{2a} + \epsilon^{2a+2} > 16\epsilon^{2a}$. Then for all x,

$$\mathcal{N}(0, \sigma^2, x) \leq \frac{1}{\sigma\sqrt{2\pi}} \leq \frac{1}{4\epsilon^a\sqrt{2\pi}}$$

Consider the case in which a component has a mean that is far from μ_1. Then from Corollary 24 for $|x| \geq 6\epsilon^a$:

$$\max_{\sigma^2>0}\mathcal{N}(0, \sigma^2, x) \leq \frac{1}{6\epsilon^a\sqrt{2\pi}}$$

So this implies that the contribution of each component of $\mathcal{F}_\alpha(F')$ to $\mathcal{F}_\alpha(F')(\mu_1)$ is at most $w_i \frac{1}{4\epsilon^a\sqrt{2\pi}}$. So we get

$$\mathcal{F}_\alpha(F)(\mu_1) - \mathcal{F}_\alpha(F')(\mu_1) \geq \frac{1}{\epsilon^a\sqrt{2\pi}}\Big[1 - \frac{1}{4}\Big] \geq \frac{3}{4\epsilon^a\sqrt{2\pi}}$$

We note that $|\frac{d\mathcal{N}(0,\sigma^2,x)}{dx}| \leq 1/\sigma^2$, and so the derivative of $\mathcal{F}_\alpha(F)(x) - \mathcal{F}_\alpha(F')(x)$ is at most $\frac{4}{\epsilon^{2a+2}}$ in magnitude since there are four Gaussians each with variance at least ϵ^{2a+2}. We can apply Claim 17 and this implies that

$$D(\mathcal{F}_\alpha(F), \mathcal{F}_\alpha(F')) = \Omega(\epsilon^2)$$

Case 2: Both $\sigma_1'^2 - \sigma_1^2 < 16\epsilon^{2a}$ and $|\mu_1' - \mu_1| < 6\epsilon^a$

We choose $\alpha = \sigma_1^2 - \epsilon^{2b}$. We have that either $\sigma_2^2 - \sigma_1^2 \geq \frac{\epsilon}{2}$, or $|\mu_1 - \mu_2| \geq \frac{\epsilon}{2}$ from the definition of ϵ-standard. We are interested in the maximum value of $\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F)$ over the interval $I = [\mu_1 - 6\epsilon^a, \mu_1 + 6\epsilon^a]$. Let $F_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and let $F_2 = \mathcal{N}(\mu_2, \sigma_2^2)$. We know that

$$\max_{x\in I} \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F) \geq \max_{x\in I} w_1\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1)$$

$\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1)$ is a Gaussian of variance exactly ϵ^{2b} so this implies that the value of $w_1\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F)$ at μ_1 is exactly $\frac{w_1}{\epsilon^b\sqrt{2\pi}}$. So

$$\max_{x\in I} \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F) \geq \frac{w_1}{\epsilon^b\sqrt{2\pi}}$$

We are interested an upper bound for $\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F)$ on the interval I. We achieve such a bound by bounding

$$\max_{x\in I} w_2\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_2)$$

Recall that either $\sigma_2^2 - \sigma_1^2 \geq \frac{\epsilon}{2}$, or $|\mu_1 - \mu_2| \geq \frac{\epsilon}{2}$. So consider the case in which the variance of F_2 is larger than the variance of F_1 by at least $\frac{\epsilon}{2}$. Because the variance of $\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1)$ is $\epsilon^{2b} > 0$, this implies that the variance of $\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_2)$ is at least $\frac{\epsilon}{2}$. In this case, the value of $\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_2)$ anywhere is at most

$$\frac{1}{\sqrt{2\pi(\sigma_2^2 - \alpha)}} \leq \frac{1}{\sqrt{2\pi\frac{\epsilon}{2}}} \leq \frac{2}{\epsilon}$$

Consider the case in which the mean μ_2 is at least $\frac{\epsilon}{2}$ far from μ_1. So any point x in the interval I is at least $\frac{\epsilon}{2} - 6\epsilon^a$ away from μ_2. So for ϵ sufficiently small, any such point is at least $\frac{\epsilon}{4}$ away from μ_2. In this case we can apply Corollary 24 to get

$$\max_{|x|\geq\frac{\epsilon}{4},\sigma^2} w_2\mathcal{N}(0,\sigma^2,x) \leq \frac{4}{\epsilon\sqrt{2\pi}} \leq \frac{2}{\epsilon}$$

So this implies that in either case (provided that $\sigma_2^2 - \sigma_1^2 \geq \frac{\epsilon}{2}$, or $|\mu_1 - \mu_2| \geq \frac{\epsilon}{2}$):

$$\max_{x\in I} w_2\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_2)(x) \leq \frac{2}{\epsilon}$$

So we get

$$\frac{w_1}{\epsilon^b\sqrt{2\pi}} \leq \max_{|x|\leq 6\epsilon^a} \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F)(\mu_1 + x) \leq \frac{w_1}{\epsilon^b\sqrt{2\pi}} + \frac{2}{\epsilon}$$

Again, from the definition of ϵ-standard either $|\sigma_1'^2 - \sigma_2'^2| \geq \frac{\epsilon}{2}$ or $|\mu_1' - \mu_2'| \geq \frac{\epsilon}{2}$. We note that $|\sigma_2'^2 - \sigma_1^2| \geq |\sigma_2'^2 - \sigma_1'^2| - |\sigma_1'^2 - \sigma_1^2| \geq |\sigma_2'^2 - \sigma_1'^2| - 16\epsilon^{2a}$. Also $|\mu_2' - \mu_1| \geq |\mu_2' - \mu_1'| - |\mu_1 - \mu_1'| \geq |\mu_2' - \mu_1'| - 6\epsilon^a$. And so for sufficiently small ϵ either $|\sigma_2'^2 - \sigma_1^2| \geq \frac{\epsilon}{4}$ or $|\mu_1 - \mu_2'| \geq \frac{\epsilon}{4}$. Then an almost identical argument as that used above to bound $\max_{x\in I} w_2\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_2)(x)$ by $\frac{2}{\epsilon}$ can be used to bound $\max_{x\in I} w_2'\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_2')(x)$ by $\frac{4}{\epsilon}$.

We now need only to bound $\max_{x\in I} w_1'\mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1')(x)$. And

$$\max_{x\in I} \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1')(x) \leq \max_x \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1')(x) \leq \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1')(\mu_1')$$

Because σ_1^2 is the smallest variance among all Gaussians components of F, F', we can conclude that $\sigma_1'^2 \geq \epsilon^{2b}$ and we can use this to get an upper bound:

$$\max_{x\in I} \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1')(x) \leq \mathcal{F}_{\sigma_1^2-\epsilon^{2b}}(F_1')(\mu_1') = \frac{1}{\sqrt{2\pi(\sigma_1'^2 - \alpha)}} \leq \frac{1}{\epsilon^b\sqrt{2\pi}}$$

We note that $\mu_1' \in I$ because $I = [\mu_1 - 6\epsilon^a, \mu_1 + 6\epsilon^a]$ and $|\mu_1' - \mu_1| < 6\epsilon^a$ in this case. So we conclude that

$$\max_{x \in I} \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F_1')(x) \geq \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F_1')(\mu_1') = \frac{1}{\sqrt{2\pi(\sigma_1'^2 - \alpha)}} \geq \frac{1}{\sqrt{2\pi(\epsilon^{2b} + 16\epsilon^{2a})}}$$

Consider the term $\sqrt{\epsilon^{2b} + 16\epsilon^{2a}} = \epsilon^b\sqrt{1 + 16\epsilon^{2a-2b}}$. So for $a > b$, we can bound $\epsilon^b\sqrt{1 + 16\epsilon^{2a-2b}} \leq \epsilon^b(1 + 16\epsilon^{2a-2b})$. We have already assumed that $a > b$, so we can bound $\epsilon^{2a-b} \leq \epsilon^a$. We can combine these bounds to get

$$\frac{w_1'}{(\epsilon^b + 16\epsilon^{b+a})\sqrt{2\pi}} \leq \max_{|x| \leq 6\epsilon^a} \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(\mu_1 + x) \leq \left(\frac{w_1'}{\epsilon^b\sqrt{2\pi}}\right) + \frac{4}{\epsilon}$$

These inequalities yield a bound on $\max_x |\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F)(x) - \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(x)|$ in terms of $|w_1 - w_1'|$: Suppose $w_1 > w_1'$, then

$$\begin{aligned}
\max_x |\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F)(x) - \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(x)| &\geq \frac{w_1}{\epsilon^b\sqrt{2\pi}} - \frac{w_1'}{\epsilon^b\sqrt{2\pi}} - \frac{4}{\epsilon} \\
&\geq \frac{|w_1 - w_1'|}{\epsilon^b\sqrt{2\pi}} - \frac{4}{\epsilon}
\end{aligned}$$

And if $w_1' > w_1$ then

$$\begin{aligned}
\max_x |\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F)(x) - \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(x)| &\geq \frac{w_1'}{(\epsilon^b + 16\epsilon^{b+a})\sqrt{2\pi}} - \frac{w_1}{\epsilon^b\sqrt{2\pi}} - \frac{2}{\epsilon} \\
&\geq \frac{w_1'}{\epsilon^b\sqrt{2\pi}}(1 - 32\epsilon^a) - \frac{w_1}{\epsilon^b\sqrt{2\pi}} - \frac{2}{\epsilon} \\
&\geq \frac{|w_1 - w_1'|}{\epsilon^b\sqrt{2\pi}} - 32\epsilon^{a-b} - \frac{2}{\epsilon}
\end{aligned}$$

So this implies that

$$\max_x |\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F)(x) - \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(x)| \geq \frac{|w_1 - w_1'|}{\epsilon^b\sqrt{2\pi}} - 32\epsilon^{a-b} - \frac{4}{\epsilon}$$

Case 2a: $|w_1 - w_1'| \geq \epsilon^c$

We have already assumed that $a > b$, and let us also assume that $b > c + 1$ in which case
Then for sufficiently small ϵ

$$\max_x |\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F)(x) - \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(x)| \geq \Omega(\epsilon^{-b+c})$$

We can bound the magnitude of the derivative of $\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F)(x) - \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')(x)$ by $\frac{4}{\epsilon^{2b}}$. Then we can use Claim 17 to get that $D(\mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F), \mathcal{F}_{\sigma_1^2 - \epsilon^{2b}}(F')) \geq \Omega(\epsilon^{2c})$

Case 2b: $|w_1 - w_1'| < \epsilon^c$

Suppose $\sigma_1'^2 - \sigma_1^2 < 16\epsilon^{2a}$ and $|\mu_1' - \mu_1| < 6\epsilon^a$ and $|w_1 - w_1'| < \epsilon^c$,
Let

$$\begin{aligned}
T_1 &= |w_1 - w_1'| \\
T_2 &= D(\mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2, \sigma_2^2)), \mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2', \sigma_2^2))) \\
T_3 &= D(\mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2', \sigma_2^2)), \mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2', \sigma_2'^2)))
\end{aligned}$$

And using Fact 18 and because the variance of each Gaussian after applying the operator $\mathcal{F}_{\sigma_1^2 - 1/2}$ is at least $\frac{1}{2}$

$$T_2 \leq O(\epsilon^a), T_3 \leq O(\epsilon^{2a})$$

Using the triangle inequality (and because we have already assumed $a > b$ and $b > c$ so $a > c$)

$$\begin{aligned}
D(\mathcal{F}_{\sigma_1^2 - 1/2}(F), \mathcal{F}_{\sigma_1^2 - 1/2}(F')) &\geq D(\mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2, \sigma_2^2)), \mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2', \sigma_2'^2))) - T_1 - T_2 - T_3 \\
&\geq D(\mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2, \sigma_2^2)), \mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2', \sigma_2'^2)) - O(\epsilon^a)
\end{aligned}$$

For this case, we have that $\sigma_1'^2 - \sigma_1^2 < 16\epsilon^{2a}$ and $|\mu_1' - \mu_1| < 6\epsilon^a$, and because F, F' are ϵ-standard we must have that $|w_2' - w_2| + |\sigma_2'^2 - \sigma_2^2| + |\mu_2' - \mu_2| \geq \epsilon - |w_1' - w_1| - |\sigma_1'^2 - \sigma_1^2| - |\mu_1' - \mu_1|$. So for sufficiently small ϵ, $|w_2' - w_2| + |\sigma_2'^2 - \sigma_2^2| + |\mu_2' - \mu_2| \geq \frac{\epsilon}{2}$. We can apply Lemma 38 and this yields

$$D(\mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2, \sigma_2^2)), \mathcal{F}_{\sigma_1^2 - 1/2}(\mathcal{N}(\mu_2', \sigma_2'^2)) \geq \Omega(\epsilon^3)$$

If we additionally require $a > 3$ then

$$D(\mathcal{F}_{\sigma_1^2 - 1/2}(F), \mathcal{F}_{\sigma_1^2 - 1/2}(F')) \geq \Omega(\epsilon^3) - O(\epsilon^a) = \Omega(\epsilon^3)$$

So if we set $a = 5, b = 4, c = 2$ these parameters satisfy all the restrictions we placed on a, b and c during this case analysis: i.e. $a > b$, $b > c + 1$ and $a > 3$. These parameters yield $D(\mathcal{F}_\alpha(F), \mathcal{F}_\alpha(F')) \geq \Omega(\epsilon^4)$ for some $\alpha \leq \sigma_1^2 - \Omega(\epsilon^{2a+2})$. So all variances after applying the operator $\mathcal{F}_\alpha$ are sill at least ϵ^{12}. $\square$

Proof of Proposition 7: First note that, assuming the number of zeros of any mixture of k Gaussians is finite, the maximum number of zeros, for any fixed k must occur in some distribution for which all the zeros have multiplicity 1. If this is not the case, then, assuming without loss of generality that there are at least as many zeros tangent to the axis on the positive half-plane, if one replaces a_1 by $a_1 - \epsilon$, for any sufficiently small ϵ, in the resulting mixture all zeros of odd multiplicity will remain and will have multiplicity 1, and each zeros of even multiplicity that had positive concavity will become 2 zeros of multiplicity 1, and thus the number of zeros will not have decreased.

We proceed by induction on the number of Gaussians in the linear combination. The base case, where $f(x) = a\mathcal{N}(\mu, \sigma^2)$ clearly holds. The intuition behind the induction step is that we will add the Gaussians in order of decreasing variance; each new Gaussian will be added as something very close to a Delta function, so as to increase the number of zeros by at most 2, adding only simple zeros. We then convolve by a Gaussian with width roughly $\sigma_i^2 - \sigma_{i+1}^2$ (the gap in variances of the Gaussian just added, and the one to be added at the next step).

Formalizing the induction step, assume that the proposition holds for linear combinations of $k-1$ Gaussians. Consider $f(x) = \sum_{i=1}^k a_k \mathcal{N}(\mu_i, \sigma_i^2, x)$, and assume that $\sigma_{i+1}^2 \geq \sigma_i^2$, for all i. Chose $\epsilon > 0$ such that for either any $a_1' \in [a_1 - \epsilon, a_1]$ or $a_1' \in [a_1, a_1 + \epsilon]$, the mixture with a_1' replacing a_1 has at least as many zeros as $f(x)$. (Such an $\epsilon > 0$ exists by the argument in the first paragraph; for the reminder of the proof we will assume $a_1' \in [a_1, a_1 + \epsilon]$, though identical arguments hold in the other case.) Let $g_c(x) = a_1'\mathcal{N}(\mu_1, \sigma_1^2 - \sigma_k^2 + c) + \sum_{i=2}^{k-1} a_i \mathcal{N}(\mu_i, \sigma_i^2 - \sigma_k^2 + c, x)$. By our induction assumption, $g_0(x)$ has at most $2(k-2)$ zeros, and we choose a_1' such that all of the zeros are simple zeros, and so that μ_k is not a zero of $g_0(x)$.

Let $\delta > 0$ be chosen such that $\delta < \sigma_k^2$, and so that for any $\delta' > 0$ such that $\delta' \leq \delta$, the function $g_{\delta'}(x)$ has the same number of zeros as $g_0(x)$ all of multiplicity one, and $g_{\delta'}(x)$ does not have any zero within a distance δ of μ_k. (Such a δ must exist by continuity of the evolution of the heat equation, which corresponds to convolution by Gaussian.)

There exists constants a, a', b, b', s, s' with $a, a' \neq 0$, $s, s' > 0$ and $b < \mu_k < b'$, such that for $x < b$, either $g_0(x) > 0, a > 0$ and $\frac{g_0(x)}{a\mathcal{N}(0,s,x)} > 1$, or $g_0(x) < 0$, $a < 0$ and $\frac{g_0(x)}{a\mathcal{N}(0,s,x)} > 1$. Correspondingly for $x > b'$, the left tail is dominated in magnitude by some Gaussian of variance s'. Next, let $m = \max_{x \in \mathbb{R}} |g_0'(x)|$ and $w = \max_{x \in \mathbb{R}} |g_0(x)|$. Let $\epsilon_1 = \min_{x : g_0(x) = 0} |g_0'(x)|$. Pick ϵ_2, with $0 < \epsilon_2 < \delta$ such that for any x s.t. $g_0(x) = 0$, $|g_0'(x + y)| > \epsilon_1/2$, for any $y \in [-\epsilon_2, \epsilon_2]$. Such an $\epsilon_2 > 0$ exists since $g_0(x)$ is analytic. Finally, consider the set $\mathcal{A} \subset \mathbb{R}$ defined by $\mathcal{A} = [b, b'] \setminus \bigcup_{x : g_0(x) = 0} [x - \epsilon_2, x + \epsilon_2]$. Set $\epsilon_3 = \min_{x \in \mathcal{A}} |g_0(x)|$, which exists since $\mathcal{A}$ is compact and $g_0(x)$ is analytic.

Consider the scaled Gaussian $a_k\mathcal{N}(\mu_k, v, x)$, where $v > 0$ is chosen sufficiently small so as to satisfy the following conditions:

- For all $x < b$, $|a\mathcal{N}(0, s, x)| > |a_k\mathcal{N}(\mu_k, v, x)|$,
- For all $x > b'$, $|a'\mathcal{N}(0, s', x)| > |a_k\mathcal{N}(\mu_k, v, x)|$,
- For all $x \in \mathbb{R} \setminus \mathcal{A}$, $|a_k \frac{d\mathcal{N}(\mu_k, v, x)}{dx}| < \epsilon_1/2$,.
- For all $x \in \mathcal{A}$, at least one of the following holds:
  - $|a_k\mathcal{N}(\mu_k, v, x)| < \epsilon_3$,
  - $|a_k\mathcal{N}(\mu_k, v, x)| > w$,
  - $|a_k \frac{d\mathcal{N}(\mu_k, v, x)}{dx}| > m$.

The above conditions guarantee that $g_0(x) + a_k\mathcal{N}(\mu_k, v', x)$ will have at most two more zeros than $g_0(x)$, for any v' with $0 < v' < v$. Since g_c is uniformly continuous as a function of $c > 0$, fixing $v'' < v$, there exists some $d > 0$ such that for any $d' \leq d$, $g_{d'}(x) + a_k\mathcal{N}(\mu_k, v', x)$ will also have at most two more zeros than $g_0(x)$, for any $v' \leq v''$. Let $\alpha = \min(d, v'')$, and define $h(x) = g_\alpha(x) + a_k\mathcal{N}(\mu_k, \alpha, x)$. To complete the proof, note that by Theorem 8 the function obtained by convolving $h(x)$ by $\mathcal{N}(0, \sigma_k^2 - \alpha, x)$ has at most $2(k-1)$ zeros, and, by assumption, at least as many zeros as $f(x)$.

To see that this bound is tight, consider $f_k(x) = k\mathcal{N}(0, k^2, x) - \sum_{i=1}^{k-1}\mathcal{N}(i, 1/25, x)$, which is easily seen to have $2(k-1)$ zeros. $\square$

Proof of lemma 9: Let $x_1, x_2, \ldots, x_k$ be the zeros of $f(x)$ which have $|x_i| \leq \frac{2}{\epsilon}$. Using Proposition 7, the number of such zeros is at most the total number of zeros of $f(x)$ which is bounded by 6. (Although Proposition 7 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

$$\Big|\int_x f(x)p(x)dx\Big| = \Omega(poly(\epsilon))$$

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 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 implies that $\int_{-\infty}^{\infty}|f(x)|dx \geq \Omega(\epsilon^4)$.

The polynomial $p(x)$ will only match the sign of $f(x)$ on the interval I so we will need to get tail estimates for the total contribution of the tails $(-\infty, \frac{-2}{\epsilon})$ and $[\frac{2}{\epsilon}, \infty)$ to various integrals. We first prove that these tails cannot contribute much to the total ℓ_1 norm of $f(x)$. This is true because each Gaussian component has mean at most $\frac{1}{\epsilon}$ (because F, F' are ϵ-standard) and variance at most 1 so each Gaussian component only has an exponentially small amount of mass distributed along the tails. Let $T = (-\infty, \frac{-2}{\epsilon}) \cup [\frac{2}{\epsilon}, \infty)$. We can use Lemma 26 to bound the contribution of each Gaussian component to the integral $\int_T |f(x)|dx$ and this implies that

$$\int_I |f(x)|dx \quad \geq \quad \int_x |f(x)|dx - \int_T |f(x)|dx \geq \Omega(\epsilon^4) - O(e^{-\frac{1}{4\epsilon^2}}) \geq \Omega(\epsilon^4)$$

Then there must be an interval $J = (a, b) \subset I$ and for which $\int_J |f(x)|dx \geq \frac{1}{6}\Omega(\epsilon^4)$ and so that $f(x)$ does not change signs on J. The derivative of $f(x)$ is bounded by $\frac{4}{\epsilon^{12}}$ because, from Lemma 5, the smallest variance of any of the transformed Gaussians is at least ϵ^{12}. We extend the interval $J = (a, b)$ to $J' = (a', b')$ so that $f(a') = f(b') = 0$. In particular, $f(x)$ does not change sign on J'. Note that J' need not be contained in the interval $I = [\frac{-2}{\epsilon}, \frac{2}{\epsilon}]$ anymore. Let $I' = J' \cup I$. Note that $p(x)$ matches the sign of $f(x)$ on I, and $p(x)$ only changes sign on the interval I so this implies that $p(x)$ matches the sign of $f(x)$ on the entire interval I'.

We now need to lower bound $\int_x f(x)p(x)dx$. We write

$$\begin{aligned}\int_x f(x)p(x)dx \quad &\geq \quad \int_{I'} f(x)p(x)dx - \Big|\int_{\Re - I'} f(x)p(x)dx\Big| \\ &\geq \quad \int_{J'} f(x)p(x)dx - \Big|\int_{\Re - I'} f(x)p(x)dx\Big|\end{aligned}$$

And the last line follows because $p(x)$ matches the sign of $f(x)$ on I', and $J' \subset I'$. The polynomial $p(x)$ can be arbitrarily small for some values of $x \in J'$. However, we consider the reduced interval $J'' = [a' + \gamma, b' - \gamma]$. There are no zeros of $f(x)$ on this interval so there are also no zeros of $p(x)$ on this interval. In fact the closest zero of $p(x)$ must be at least distance γ from this interval J''. So this implies that on the interval J'', $p(x) \geq \gamma^7$. Note that $\int_{J'}|f(x)|dx \geq \int_J |f(x)|dx = \Omega(\epsilon^4)$ so

$$\begin{aligned}\int_{J''}|f(x)|dx \quad &\geq \quad \int_{J'}|f(x)|dx - \int_{x=a'}^{a'+\gamma}|f(x)|dx - \int_{x=b'-\gamma}^{b'}|f(x)|dx \\ &\geq \quad \Omega(\epsilon^4) - \int_{x=a'}^{a'+\gamma}|f(x)|dx - \int_{x=b'-\gamma}^{b'}|f(x)|dx \\ &\geq \quad \Omega(\epsilon^4) - O(\frac{\gamma^2}{\epsilon^{12}})\end{aligned}$$

This last line follows because $f(x)$ is zero at a', b' and the derivative of $f(x)$ is bounded by $O(\frac{1}{\epsilon^{12}})$. So this yields a bound on $\int_{x=a'}^{a'+\gamma} |f(x)|dx$ of $\frac{\gamma^2}{2\max_x |f'(x)|}$. So if we choose $\gamma = O(\epsilon^8)$ then we conclude

$$\int_{J'} f(x)p(x)dx \geq \int_{J''} f(x)p(x)dx \geq \int_{J''} |f(x)|\gamma^7 dx = \Omega(\epsilon^4)\gamma^7$$

And this yields

$$\int_x f(x)p(x)dx \geq \Omega(\epsilon^{60}) - \Big|\int_{\Re - I'} f(x)p(x)dx\Big|$$

The largest coefficient in $p(x)$ is at most $O(\frac{1}{\epsilon^6})$ and the degree of $p(x)$ is at most 6 by construction. Note that $I \subset I'$ so we can use Lemma 29 to conclude

$$\begin{aligned}
\Big|\int_{\Re - I'} f(x)p(x)dx\Big| &\leq \sum_i w_i \int_{\Re - I'} |p(x)|\mathcal{N}(\mu_i, \sigma_i^2)(x)dx + \sum_i w_i' \int_{\Re - I'} |p(x)|\mathcal{N}(\mu_i', \sigma_i'^2)(x)dx \\
&\leq \sum_i w_i \int_{\Re - I} |p(x)|\mathcal{N}(\mu_i, \sigma_i^2)(x)dx + \sum_i w_i' \int_{\Re - I} |p(x)|\mathcal{N}(\mu_i', \sigma_i'^2)(x)dx \\
&\leq O(\frac{1}{\epsilon^{12}})e^{-\frac{1}{2\epsilon^2}}
\end{aligned}$$

So for sufficiently small ϵ

$$\int_x f(x)p(x)dx \geq \Omega(\epsilon^{60}) - O(\frac{1}{\epsilon^{12}})e^{-\frac{1}{2\epsilon^2}} = \Omega(\epsilon^{60})$$

And again because each coefficient in $p(x)$ is at most $O(\frac{1}{\epsilon^6})$ by construction, this implies that there is some $i \leq 6$ such that

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

□

D.1 The Univariate Algorithm

The algorithm for reconstructing the parameters of the mixture is given in Figure 5. For simplicity, the algorithm requires the mixture to be in standard (isotropic) position – meaning that the distribution's mean (expectation) is 0 and its variance is 1. Achieving this is fairly straightforward by a linear transformation, which we assume has already been done in our reduction from the n-dimensional problem to the one-dimensional problem.

Proof of theorem 10: There are three parts to the proof. In the first part, we argue that the estimates of the moments $\hat{M}_i$ are all within an additive α of their true values, with high probability. In the second part, we argue that there will be at least one candidate set of parameters $(\hat{w}_1, \hat{\mu}_1, \hat{\sigma_1}, \hat{\mu}_2, \hat{\sigma}_2) \in \mathcal{A} \times \mathcal{B}^4$ in our brute-force search whose moments (which we can compute analytically as a function of $\hat{w}_1, \hat{\mu}_1, \hat{\sigma_1}, \hat{\mu}_2$, and $\hat{\sigma}_2$) are all within α of the true values (and which satisfy $|\hat{\mu}_1 - \hat{\mu}_2| + |\hat{\sigma}_1^2 - \hat{\sigma}_2^2| > \epsilon$). This means that with probability at least $1 - \delta$, the discrepancy between the estimated moments and those of the output mixture will be at most 2α, and that the discrepancy between the true moments and those of the output mixture will be at most 3α. In the final part, we apply Theorem 4 to show that if the discrepancy in moments is less than 3α, then all parameters are accurate to within an additive ϵ.

The variance of the mixture can be written as $w_1\sigma_1^2 + w_2\sigma_2^2 + w_1w_2(\mu_1 - \mu_2)^2$. Combining this with $w_1\mu_1 + w_2\mu_2 = 0$, gives,

$$1 = w_1\sigma_1^2 + w_2\sigma_2^2 + w_1w_2\left(\mu_1 + \frac{w_1}{w_2}\mu_1\right)^2 \tag{4}$$

$$= w_1\sigma_1^2 + w_2\sigma_2^2 + w_1w_2\mu_1^2\left(1 + \frac{w_1}{w_2}\right)^2 \tag{5}$$

$$= w_1\sigma_1^2 + w_2\sigma_2^2 + w_1w_2\mu_1^2\frac{1}{w_2^2}. \tag{6}$$

From this, we have that $\sigma_i \leq \sqrt{1/w_i} \leq \epsilon^{-1/2}$ and $\mu_1 \leq \sqrt{w_2/w_1} \leq \epsilon^{-1/2}$, and similarly, $\mu_2 \leq \epsilon^{-1/2}$.

Algorithm 3. 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:

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

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

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

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

Figure 5: 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 O.

Part 1. In this part, we argue that the empirical moment estimates are all accurate to within an additive α, with probability $\geq 1 - \delta$. By Lemma 30 of Appendix K.3, using the union bound, with probability $\geq 1 - \delta$, the 4 estimated moments will all be close to their expectations,

$$\left|\frac{1}{m} \sum_{i=1}^{m} x_i^k - \mathrm{E}[x^k]\right| \leq \sqrt{\frac{2^k k!}{(\delta/4)m}} \leq \frac{500}{\sqrt{\delta m}}, \text{ for } k = 3, 4, 5, 6.$$

This is at most α, for the specified m.

Part 2. By our choice of N, the true means will be in $[-N, N]$ and the true standard deviations will be within $[0, N]$. Hence, there will be some candidate mixture where the weights, means, and standard deviations, are all within $1/N$ of the truth. For the rest of this part, we refer to these nearby parameters as $\hat{w}_i, \hat{\mu}_i, \hat{\sigma}_i$. Now, consider $M_k(\hat{\mu}_i, \hat{\sigma}_i^2) - M_k(\mu_i, \sigma_i^2)$. For $k \leq 6$, this is a polynomial in $\hat{\mu}, \hat{\sigma}_i$ of total degree at most 6. Furthermore, as can be verified from Appendix O, the sum of the magnitudes of the coefficients (excluding the constant term) is at most 76. Therefore, by changing the mean and standard deviation by at most $1/N$, this can change each such moment by at most an additive,

$$\begin{aligned} 76\left(\left(\epsilon^{-1/2} + N^{-1}\right)^6 - \left(\epsilon^{-1/2}\right)^6\right) &= 76\epsilon^{-3}\left(\left(1 + N^{-1}\epsilon^{1/2}\right)^6 - 1\right) \\ &\leq 76\epsilon^{-3}\left(7N^{-1}\epsilon^{1/2}\right) = 76 \cdot 7 \cdot \epsilon^{-2.5} N^{-1}. \end{aligned}$$

Hence, if we used the true mixing weights with the candidate Gaussian, the moments of the mixture $w_1 \hat{F}_1 + w_2 \hat{F}_2$ would be off of the true moments by at most $152 \cdot 7 \cdot \epsilon^{-2.5} N^{-1}$. Now, note that the moments of each candidate Gaussian are at most $76\epsilon^{-3}$. Therefore,using the candidate weights, which are off by at most $1/N$ can cause an additional additive error of at most $\frac{152}{\epsilon^3 N}$. Hence, each of the moments will be off by at most $152\left(\frac{7}{\epsilon^{2.5} N} + \frac{1}{\epsilon^3 N}\right) \leq \alpha$ from the truth.

Putting these two parts together, the algorithm will find some candidate Gaussian all of whose moments are within 2α of the estimated moments. Hence, the output mixture will have all moments within 3α of correct.

It remains to show that if all of the first 6 moments are correct to within $3\alpha \leq 3\epsilon^{150}$, then the mixture parameters are within ϵ of correct. Now, we would like to apply Theorem 4. However, that theorem requires that the variances be bounded by at most 1. Imagine shrinking F by rescaling it by a factor of $\sqrt{\epsilon}$ so that the variances are at most 1. After such a transformation, we would have $|\mu_1 - \mu_2| + |\sigma_1^2 - \sigma_2^2| \geq \epsilon^2$, and similarly for the scaled candidate parameters. Thus the transformed mixture is ϵ^2-standard, so Theorem 4 implies that, if the mixture parameters of the (rescaled) candidate are not within ϵ^2 of correct, then the (rescaled) moments will be off by at least ϵ^{134}. This implies that if the parameters of a candidate are not to within $\epsilon^{1.5} < \epsilon$, then the moments would be off by a factor of at least ϵ^{134}, using the fact that the rescaled kth moment is simply the kth moment times a factor of exactly $\epsilon^{k/2}$. □

E Anisotropy Preservation Lemma

Before we prove Lemma 13, we state a simple fact about averaging.

Lemma 19. *Suppose $w_1(1+\alpha) + w_2(1-\beta) \leq 1$, $w_1, w_2 \geq \epsilon \geq 0$, $w_1 + w_2 = 1$ and $\alpha > 0$. Then, $\beta \geq \epsilon\alpha$.*

Proof. Clearly $1 \geq \epsilon(1+\alpha) + (1-\epsilon)(1-\beta)$ since this weighting puts the greatest amount of weight on $1-\alpha$. Rearranging terms gives $\beta \geq \epsilon\alpha/(1-\epsilon) \geq \epsilon\beta$. □

Proof of Lemma 13. We first argue that one of the covariance matrices is far from the identity matrix. In particular, we will show that,

$$\max\{\ \|\Sigma_1^{-1}\|_2, \|\Sigma_2^{-1}\|_2\ \} \geq 1 + a, \quad a = \frac{\epsilon^3 - t^2}{3n}. \tag{7}$$

This means that it has an eigenvalue which is bounded from 1. The covariance matrix of F is the identity matrix, but as a mixture it is also helpful to write it as,

$$I_n = w_1\Sigma_1 + w_2\Sigma_2 + w_1w_2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T. \tag{8}$$

By Lemma 22, the squared variation distance between F_1 and F_2 is,

$$\epsilon^2 \leq (D(F_1, F_2))^2 \leq \frac{1}{2}\sum_{i=1}^{n}(\lambda_i + \frac{1}{\lambda_i} - 2) + (\mu_1 - \mu_2)^T\Sigma_1^{-1}(\mu_1 - \mu_2).$$

In the above, $\lambda_1, \ldots, \lambda_n > 0$ are the eigenvalues of $\Sigma_1^{-1}\Sigma_2$. By (8), we also have,

$$\|\mu_1 - \mu_2\|^2 = (\mu_1 - \mu_2)^T I_n(\mu_1 - \mu_2) \geq w_1(\mu_1 - \mu_2)^T\Sigma_1(\mu_1 - \mu_2).$$

Together with $w_1 \geq \epsilon$, and $\|\mu_1 - \mu_2\| < t$, this implies that $(\mu_1 - \mu_2)^T\Sigma_1(\mu_1 - \mu_2) \leq t^2/\epsilon$. Hence,

$$\epsilon^2 \leq \frac{1}{2}\sum_{i=1}^{n}(\lambda_i + \frac{1}{\lambda_i} - 2) + \frac{t^2}{\epsilon}.$$

In particular, there must be some eigenvalue λ_j, such that,

$$\lambda_j + 1/\lambda_j - 2 \geq \frac{2}{n}\left(\epsilon^2 - \frac{t^2}{\epsilon}\right) = \frac{6a}{\epsilon^2}.$$

Let v_j be a unit (eigen)vector corresponding to λ_j, i.e., $v_j = \lambda_j\Sigma_1^{-1}\Sigma_2 v_j$. Then we have that,

$$v_j^T\Sigma_1 v_j = \lambda_j v_j^T\Sigma_2 v_j$$
$$\left(\frac{v_j^T\Sigma_1 v_j}{v_j^T\Sigma_2 v_j} - 1\right) + \left(\frac{v_j^T\Sigma_2 v_j}{v_j^T\Sigma_1 v_j} - 1\right) = \lambda_j + \frac{1}{\lambda_j} - 2 \geq \frac{6a}{\epsilon^2}.$$

Since one of the two terms in parentheses above must be at least $3a/\epsilon^2$, WLOG, we can take $\frac{v_j^T\Sigma_1 v_j}{v_j^T\Sigma_2 v_j} \geq 1 + 3a/\epsilon^2$. This means that the numerator or denominator is bounded from 1. We can break this into two cases.

Case 1: $v_j^T \Sigma_2 v_j < 1/(1+a)$. This establishes (7) immediately.

Case 2: $v_j^T \Sigma_1 v_j \geq (1+3a/\epsilon^2)/(1+a) = 1 + (3/\epsilon^2 - 1)a/(1+a) \geq 1 + (3/\epsilon^2 - 1)a/2$. By Lemma 19, since $w_1 v_j^T \Sigma_1 v_j + w_2 v_j^T \Sigma_2 v_j \leq 1$, we have

$$v_j^T \Sigma_2 v_j \leq 1 - \frac{\epsilon}{2}\left(\frac{3}{\epsilon^2} - 1\right) a \leq 1 - a.$$

This means that $\|\Sigma_2^{-1}\|_2 \geq 1/(1-a) \geq 1 + a$.

(end of cases).

We have now established (7) by a case argument. WLOG, suppose $\|\Sigma_1^{-1}\|_2 \geq 1 + a$. As discussed in the preliminaries, $P_u(F_1) = \mathcal{N}(\mu_1 \cdot u, u^T \Sigma_1 u)$. We claim that this implies that,

$$\Pr_{u \in \mathbb{S}_{n-1}}\left[\ u^T \Sigma_1 u \in [1-c, 1+c]\ \right] < \delta, \quad c = \frac{\delta^2 a}{4n} = \frac{\delta^2(\epsilon^3 - t^2)}{12n^2}. \tag{9}$$

To see that this is sufficient for the lemma, note that if $u^T \Sigma_1 u \leq 1 - c$, then we have the lemma directly. If $u^T \Sigma_1 u > 1 + c$, then again by Lemma 19, we have $u^T \Sigma_2 u < 1 - \epsilon c$, which gives the lemma. It remains to show (9).

We know there is some eigenvalue $\lambda < 1/(1+a)$ of Σ_1. Let unit vector $v \in \mathbf{R}^n$ be a corresponding eigenvector of Σ_1. WLOG we may assume that the random u satisfies $u \cdot v \geq 0$, since it doesn't matter whether we project onto u or $-u$. Let $u = \sqrt{p}v + \sqrt{1-p}v'$, where v' is another unit vector that is orthogonal to v, i.e., $\sqrt{p} = u \cdot v$, and $v' = (u - \sqrt{p}v)/\|u - \sqrt{p}v\|$. One would like to say that as long as p is bounded from 0, then $u^T \Sigma_1 u$ is bounded from 1, but this is not true because $v'^T \Sigma_1 v'$ contributes as well. More precisely,

$$u^T \Sigma_1 u = (\sqrt{p}v + \sqrt{1-p}v')^T \Sigma_1 (\sqrt{p}v + \sqrt{1-p}v') = pv^T \Sigma_1 v + (1-p)v'^T \Sigma_1 v' = p\lambda + (1-p)v'^T \Sigma_1 v'.$$

(The cross-terms in the above product are 0.) Let $v'^T \Sigma_1 v' = \gamma$. So $u^T \Sigma_1 u = p\lambda + (1-p)\gamma$, where $p \in [0,1]$. It does not suffice to argue that p is bounded from 0, even though $\lambda < 1/(1+a)$, because we haven't made any assumption about γ. However, we only have to consider the case where $\gamma > \lambda$. In the other case, (9) holds trivially.

Instead, imagine picking unit vector u uniformly at random by first picking v' uniformly at random from the set of unit vectors orthogonal to v, and then choosing p, conditional on v'. By symmetry, the distribution on $\sqrt{p}$ will be the same, regardless of the choice v'. $\sqrt{p}$ will be distributed exactly like the absolute value of the first coordinate of a random unit vector. Let P be the set of $p \in [0,1]$ that satisfy (9),

$$P = \left\{p \in [0,1] \mid p\lambda + (1-p)\gamma \in [1-c, 1+c]\right\}.$$

Since P is convex, it is an interval. We claim that its width is at most $4c/a$. Geometrically, the width of P is the fraction that the interval $[1-c, 1+c]$ covers of the interval $[\lambda, \gamma]$. This width is maximized when $\gamma = 1+c$, in which case it is $2c/(\gamma - \lambda) \leq \frac{2c}{1-1/(1+a)} \leq 4c/a$. Finally, the distribution over p has a density that is monotonically decreasing. Hence, the probability that $p \in P$ is at most the probability that $p \in [0, 4c/a]$. From Lemma 12, we know

$$\Pr_{u \in \mathbf{R}^n : \|u\| = 1}\left[\sqrt{p} = |u \cdot v| \in [0, 2\sqrt{c/d}]\right] \leq 2\sqrt{cn/a} = \delta.$$

Now that we have established (9), we are done. □

F Continuity Lemma

Proof of lemma 14:

First note that the formula for the covariance of a mixture is:

$$\mathrm{var}(F) = w_1 \Sigma_1 + w_2 \Sigma_2 + w_1 w_2 (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T. \tag{10}$$

Let $t = 2\epsilon_3 \sqrt{n}/\delta$.

Case 1: $\|\mu_1 - \mu_2\| > t$. In this case, by Lemma 12, with probability $\geq 1 - \delta$, $|r \cdot (\mu_1 - \mu_2)| \geq \delta t/\sqrt{n} = 2\epsilon_3$, which is what we wanted.

Case 2: $\|\mu_1 - \mu_2\| < t$. By Lemma 13, with probability $\geq 1-\delta$, for some ℓ,

$$r^T \Sigma_\ell r \leq 1 - \frac{\epsilon\delta^2(\epsilon^3 - t^2)}{12n^2} \leq 1 - \frac{\epsilon\delta^2(\epsilon^3/2)}{12n^2} \leq 1 - \epsilon_4 \tag{11}$$

WLOG, we can take $\ell = 1$. Now, if $|r \cdot (\mu_1 - \mu_2)| \leq 2\epsilon_3$, then,

$$w_1 r^T \Sigma_1 r + w_2 r^T \Sigma_2 r + 2\epsilon_3 \geq 1$$

If $r^T\Sigma_1 r \leq 1 - \epsilon_4 < 1 - 4\epsilon_3$, then $r^T\Sigma_2 r \geq 1 - 2\epsilon_3$. This gives, $r^T(\Sigma_2 - \Sigma_1)r \geq \epsilon_4 - 2\epsilon_3 \geq 2\epsilon_3$.

(end of cases) Hence we have established part (a) of the lemma. For (b), first notice that $\|\mu_1 - \mu_2\| \leq \sqrt{2/\epsilon}$. The reason is because, by (10), the variance of F projected in the direction $v = (\mu_1 - \mu_2)/\|\mu_1 - \mu_2\|$ is 1 and is at least $w_1 w_2 (\mu_1 - \mu_2)^2 \geq \frac{1}{2}\epsilon(\mu_1 - \mu_2)^2$. Furthermore, since the origin lies on the segment joining the two, we have $\|\mu_\ell\| \leq \sqrt{2/\epsilon}$, for each $\ell \in [2]$. Further,

$$|r\mu_\ell - r^{ij}\mu_\ell| \leq \|\epsilon_2 b_i + \epsilon_2 b_j\| \cdot |\mu_\ell| \leq \frac{4\epsilon_2}{\sqrt{\epsilon}} \leq \epsilon_3/3.$$

This gives (b).

For (c), also by (10), we have that $\|\Sigma_\ell\|_2 \leq 1/\epsilon$. Similarly,

$$\begin{aligned}
|(r^{ij})^T\Sigma_\ell r^{ij} - r^T\Sigma_\ell r| &= |\epsilon_2^2(b_i + b_j)^T\Sigma_\ell(b_i + b_j) + 2\epsilon_2(b_i + b_j)^T\Sigma_\ell r| \\
&\leq (4\epsilon_2^2 + 2\sqrt{2}\epsilon_2)\|\Sigma_\ell\|_2 \\
&\leq \frac{10\epsilon_2}{\epsilon} \leq \epsilon_3/3
\end{aligned}$$

□

G Recovery Lemma

Proof of lemma 15: Let *idealizations* $\bar{m}^0 = \mu \cdot r$, $\bar{m}^{ij} = \mu \cdot r^{ij}$, $\bar{v}^0 = r^T\Sigma r$, $\bar{v}^{ij} = (r^{ij})^T\Sigma r^{ij}$, $\bar{v}^i = \frac{1}{n}\sum_j \bar{v}^{ij}$, and $\bar{v} = \frac{1}{n^2}\sum_{i,j}\bar{v}^{ij}$.

Since $\bar{m}^0 = r \cdot \mu$ and $\bar{m}^{ii} = (r + 2\epsilon_2 b_i)\mu$, we get that,

$$b_i \cdot \mu = \frac{\bar{m}^{ii} - \bar{m}}{2\epsilon_2}.$$

Hence, since we have assumed that each variable is within ϵ_1 of its idealization,

$$\|\mu - \hat{\mu}\|^2 = \sum_{i=1}^n (b_i \cdot \mu - b_i \cdot \hat{\mu})^2 = \sum_{i=1}^n \left(b_i \cdot \mu - \frac{m^{ii} - m}{2\epsilon_2}\right)^2 \leq n\left(\frac{2\epsilon_1}{2\epsilon_2}\right)^2.$$

This gives a bound of $\|\mu - \hat{\mu}\| \leq \epsilon_1\sqrt{n}/\epsilon_2$. The variance calculation is also straightforward, but a bit more involved. For vectors $u, v \in \mathbf{R}^n$, let $u \diamond v = u^T\Sigma v$. Note that $\diamond$ is symmetric since Σ is. So, $\bar{v}^0 = r \diamond r$, and

$$\begin{aligned}
\bar{v}^{ij} &= (r + \epsilon_2(b_i + b_j)) \diamond (r + \epsilon_2(b_i + b_j)) \\
&= r \diamond r + \epsilon_2^2(b_i \diamond b_i + b_j \diamond b_j + 2b_i \diamond b_j) + 2\epsilon_2(b_i + b_j) \diamond r \\
\bar{v}^i &= r \diamond r + \epsilon_2^2\left(b_i \diamond b_i + \frac{1}{n}\sum_\ell b_\ell \diamond b_\ell + \frac{2}{\sqrt{n}} b_i \diamond r\right) + 2\epsilon_2 b_i \diamond r + \frac{2\epsilon_2}{n}\sum_\ell b_\ell \diamond r \\
&= \left(1 + \frac{2\epsilon_2}{\sqrt{n}}\right) r \diamond r + \epsilon_2^2\left(b_i \diamond b_i + \frac{1}{n}\sum_\ell b_\ell \diamond b_\ell + \frac{2}{\sqrt{n}} b_i \diamond r\right) + 2\epsilon_2 b_i \diamond r \\
\bar{v} &= \left(1 + \frac{4\epsilon_2}{\sqrt{n}} + \frac{2\epsilon_2^2}{n}\right) r \diamond r + \frac{2\epsilon_2^2}{n}\sum_\ell b_\ell \diamond b_\ell \\
\bar{v} - \bar{v}^i - \bar{v}^j &= \left(\frac{2\epsilon_2^2}{n} - 1\right) r \diamond r - \epsilon_2^2(b_i \diamond b_i + b_j \diamond b_j) - \left(\frac{2\epsilon_2^2}{\sqrt{n}} + 2\epsilon_2\right)(b_i + b_j) \diamond r \\
\frac{\bar{v}^{ii} + \bar{v}^{jj}}{2} &= r \diamond r + 2\epsilon_2^2(b_i \diamond b_i + b_j \diamond b_j) + 2\epsilon_2 r \diamond (b_i + b_j).
\end{aligned}$$

A straightforward calculation (by hand or computer) verifies that,

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

Now, since each variable is within ϵ_1 of its idealization, the matrix V of the algorithm satisfies,

$$|V_{ij} - b_i \diamond b_j| \leq \epsilon_1 \left(\frac{3\sqrt{n}}{(2\epsilon_2 + \sqrt{n})2\epsilon_2^2} + \frac{2}{(2\epsilon_2 + \sqrt{n})4\epsilon_2} + \frac{1}{2\epsilon_2\sqrt{n}} + \frac{1}{2\epsilon_2^2} \right) \leq \frac{3\epsilon_1}{\epsilon_2^2}.$$

Also note that $b_i \diamond b_j = Q_{ij}$ where $Q = B^T \Sigma B$. Hence, as matrices, the difference obeys $\|V - Q\|_F \leq 3n\epsilon_1/\epsilon_2^2$. Now, let $R = \arg\min_{M \succeq 0} \|M - V\|$. Note that, since R is the closest positive semidefinite matrix in Frobenius norm, and Q is positive semidefinite, we have $\|V - R\|_F \leq \|V - Q\|_F \leq 3n\epsilon_1/\epsilon_2^2$. By the triangle inequality, $\|Q - R\|_F \leq 6n\epsilon_1/\epsilon_2^2$. Since a change of basis does not alter the Frobenius norm, $\|B(Q - R)B^T\|_F = \|\Sigma - \hat{\Sigma}\|_F \leq 6n\epsilon_1/\epsilon_2^2$. Finally, observe that R is symmetric since V is symmetric by construction, and the set of positive semidefinite matrices is symmetric. A change of basis does not alter a matrices symmetry, so $\Sigma = BRB^T$ is also symmetric. □

H Additive Approximation Lemma

Proof of lemma 11: We show a failure probability of $\leq 2\delta n^2$. Since our goal is to prove polynomial bounds, this suffices.

By Lemma 14(a), with probability $\geq 1 - \delta$, we will have $|r \cdot (\mu_1 - \mu_2)| > 2\epsilon_3$ or $|r^T(\Sigma_1 - \Sigma_2)r| > 2\epsilon_3$.

Case 1: $|r \cdot (\mu_1 - \mu_2)| > 2\epsilon_3$. WLOG, we can assume $r \cdot \mu(F_1) < r \cdot \mu(F_2)$. By Lemma , with probability $\geq 1 - \delta$ we will correctly estimate all the means of the one-dimensional mixture to within ϵ_1 using $m_1 = \text{poly}(1/\epsilon_1)$ samples. Hence, the parameters of G_ℓ^0 will be within ϵ_1 of $P_r(F_\ell)$. Then $|\mu(G_1^0) - \mu(G_2^0)| > 2\epsilon_3 - 2\epsilon_1 > \epsilon_3$ and the algorithm will correctly permute so that $\mu(G_2^0) > \mu(G_1^0)$. By Lemma 14(b), the means in all the projected directions will satisfy $r^{ij} \cdot (\mu_2 - \mu_1) > \epsilon_3 - 2\epsilon_3/3 > \epsilon_3/3$ for all $i, j \in [n]$. Again by Lemma , with probability $1 - \delta n^2$, we will correctly estimate all the means of the projection onto each r^{ij} within ϵ_1 using $n^2 m_1 = \text{poly}(1/\epsilon_1)$ samples.[4] Since $\epsilon_3/3 > 2\epsilon_1$, we will correctly surmise the Gaussian of greater (projected) mean. Hence, with probability $\geq 1 - 2\delta n^2$, all Gaussians will be correctly "matched up." By Lemma 15, this will result in a reconstructed mean with $|\hat{\mu} - \mu| < \epsilon$.

Case 2: $|r^T(\Sigma_1 - \Sigma_2)r| > 2\epsilon_3$. It is still possible that $|\mu(G_1^0) - \mu(G_2^0)| > \epsilon_3$. If this is the case, then the argument in the previous paragraph, starting with the word "then" still applies. If $|\mu(G_1^0) - \mu(G_2^0)| < \epsilon_3$, then the completely analogous argument to the previous paragraph, for variances rather than means, still applies.□

I Statistical Approximation Lemma

In this section, we state and prove the correctness of an algorithm that takes a polynomial number of samples from a GMM, and efficiently outputs, with high probability, approximations to the two constituent Gaussians that are close in terms of variation distance.

Proof of lemma 16: Consider two cases.

Case 1: The smallest eigenvalues of Σ_1, Σ_2 are both greater than $2\varepsilon_2$. Now, with probability $\geq 1 - \delta/3$, for $\hat{F}_i$ output by Algorithm 1, we have $\|\hat{\mu}_i - \mu_i\| \leq \varepsilon_1, \|\hat{\Sigma}_i - \Sigma_i\|_F \leq \varepsilon_1$, and $|\hat{w}_i - w_i| \leq \varepsilon_1$, for each i. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ be the eigenvalues of $\Sigma_i^{-1}\hat{\Sigma}_i$. By Lemma 22:

$$D^2(\hat{F}_i, F_i) \leq \sum_{j=1}^{n} (\lambda_j + \frac{1}{\lambda_j} - 2) + (\mu_i - \hat{\mu}_i)^T \Sigma_i^{-1} (\mu_i - \hat{\mu}_i). \tag{12}$$

By assumption, we know $\|\Sigma_i^{-1}\|_2 \leq \frac{1}{2\varepsilon_2}$. Hence,

$$(\mu_i - \hat{\mu}_i)^T \Sigma_i^{-1} (\mu_i - \hat{\mu}_i) \leq \|\mu_i - \hat{\mu}_i\|^2 \cdot \|\Sigma_i^{-1}\|_2 \leq \frac{\varepsilon_1^2}{2\varepsilon_2}.$$

[4]There is a slight mismatch in parameters here since r^{ij} is not a unit vector, and hence $P_{r^{ij}}(F)$ is not isotropic, as required by Lemma H. However, we can use the unit vector $r^{ij}/\|r^{ij}\|$ for the projection, request accuracy $\epsilon_1/2$, and then rescale the parameters. Since $\|r^{ij}\| > 1/2$, this will result in a final accuracy of ϵ_1, with probability δ. The number of samples required is still $\text{poly}(1/\epsilon)$.

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 1$(n, \varepsilon_1, \delta/3, \mathrm{SA}(F))$ 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 the 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.

Let $\Delta = \hat{\Sigma}_i - \Sigma_i$, so $\|\Delta\|_F \leq \varepsilon_1$. Then we have that,

$$\lambda_1 = \|\Sigma_i^{-1}\hat{\Sigma}_i\|_2 = \|\Sigma_i^{-1}(\Sigma_i + \Delta)\|_2 \leq \|I_n\|_2 + \|\Sigma_i^{-1}\|_2\|\Delta\|_2 \leq 1 + \frac{\varepsilon_1}{2\varepsilon_2}. \tag{13}$$

Similarly,

$$\frac{1}{\lambda_n} = \|\hat{\Sigma}_i^{-1}\Sigma_i\|_2 = \|\hat{\Sigma}_i^{-1}(\hat{\Sigma}_i - \Delta)\|_2 \leq \|I_n\|_2 + \|\hat{\Sigma}_i^{-1}\|_2\|\Delta\|_2.$$

The quantity $\|\hat{\Sigma}_i^{-1}\|_2^{-1}$ is the smallest eigenvalue of $\hat{\Sigma}_i = \Sigma_i - \Delta$, which is at least the smallest eigenvalue of Σ_i minus the largest eigenvalue of Δ. Hence, $\|\hat{\Sigma}_i^{-1}\|_2^{-1} \geq 2\varepsilon_2 - \|\Delta\|_F \geq 2\varepsilon_2 - \varepsilon_1$. This gives,

$$\frac{1}{\lambda_n} \leq 1 + \frac{\varepsilon_1}{2\varepsilon_2 - \varepsilon_1} \leq 1 + \frac{\varepsilon_1}{\varepsilon_2}. \tag{14}$$

Combining these with (12) gives,

$$D^2(\hat{F}_i, F_i) \leq \frac{2n\varepsilon_1}{\varepsilon_2} + \frac{\varepsilon_1^2}{2\varepsilon_2} \leq \epsilon^2.$$

As argued, the smallest eigenvalue of $\hat{\Sigma}_i$ is at least $2\varepsilon_2 - \varepsilon_1 \geq \varepsilon_2$. Since this applies to both $i = 1, 2$, the algorithm will halt and output $\hat{F}$, which meets the conditions of the lemma.

Case 2: Σ_1 or Σ_2 has an eigenvalue $\lambda < 2\varepsilon_2$. We further divide into two subcases.

Case 2a: Both $\hat{\Sigma}_i$ have all eigenvalues greater than ε_2, and the algorithm will output $\hat{F}$. This is not a problem, and the argument above (with small modification to the parameters) will guarantee that $D(F_i, \hat{F}_i) \leq \epsilon$ in this case. The only change to the above argument needed is that we can guarantee only that the smallest eigenvalue of $\hat{\Sigma}_i$ is at least $\varepsilon_2 - \varepsilon_1$, so the bound in (13) becomes $1 + \frac{\varepsilon_1}{\varepsilon_2 - \varepsilon_1} \leq 1 + 2\frac{\varepsilon_1}{\varepsilon_2}$, which is still sufficient for the argument. (The bound of (14) remains valid.)

Case 2b: $\hat{\Sigma}_i$ has an eigenvalue smaller than ε_2, for some i. By possibly renumbering, the algorithm chooses eigenvalue $\lambda < \varepsilon_2$ of $\hat{\Sigma}_1$ along with corresponding unit eigenvector, v. Suppose again that $|\hat{\mu}_i - \mu_i| \leq \varepsilon_1, \|\hat{\Sigma}_i - \Sigma_i\|_F \leq \varepsilon_1,$ and $|\hat{w}_i - w_i| \leq \varepsilon_1$, for each i, which happens with probability at least $1 - \delta/3$.

We will now argue that,

$$\Pr_{x\sim F_1}\left[\left|P_v(x) - P_v(\mu(\hat{F}_1))\right| > \frac{\sqrt{\varepsilon_2}}{\varepsilon_3}\right] \le \varepsilon_3 \tag{15}$$

$$\Pr_{x\sim F_2}\left[\left|P_v(x) - P_v(\mu(\hat{F}_1))\right| \le \frac{\sqrt{\varepsilon_2}}{\varepsilon_3}\right] \le \varepsilon_3 \tag{16}$$

Hence, by the union bound, with probability at least $1-m\varepsilon_3 \ge 1-\delta/3$ we make no clustering mistakes. Given $(10n\epsilon^{-1}\delta^{-1})^4$ samples (actually many fewer), it is known that one can estimate the parameters of a Gaussian to within variation distance ϵ, with probability $\ge 1-\delta/6$. **[Adam: Get reference!]** With probability $\ge 1-\delta/6$, the number of samples from each Gaussian is at least $w_i m\delta/3 \ge \epsilon m\delta/3 > (10n\epsilon^{-1}\delta^{-1})^4$. Hence, hypothetically speaking, if our algorithm chose the set D_i based on the true labels of which Gaussian each sample came from, then our Gaussians would have variation distance ϵ from the truth. By the union bound, all of this happens with probability $\ge 1-\delta/3-\delta/3-2\delta/6 = 1-\delta$.

It now remains to show (15) and (16). For (15), by the fact that $|v\cdot(\mu_1-\hat{\mu}_1)| \le \varepsilon_1$ it suffices to upper bound the probability that x drawn from F_1 satisfies,

$$\left|v\cdot(x-\mu_1)\right| > \frac{\sqrt{\varepsilon_2}}{\varepsilon_3} - \varepsilon_1 \ge \frac{\sqrt{\varepsilon_2}}{2\varepsilon_3}.$$

Since the standard deviation of F_1 in the direction of v is at most $\sqrt{2\varepsilon_2}$, points satisfying the above are at least $1/(2\varepsilon_3)$ standard deviations from their true mean. Using the fact that, for a one-dimensional Gaussian random variable, the probability of being at least s standard deviations from the mean is at most $2e^{-s^2/2}/(\sqrt{2\pi}s) \le 1/s$, we get a bound of $2\varepsilon_3$ which is sufficient for (15).

Next, since $v^T\hat{\Sigma}_1 v = \lambda$, $v\Sigma_1 v^T \le \lambda + \varepsilon_1 \le 2\varepsilon_2$. By isotropy and (10),

$$w_1 v\Sigma_1 v^T + w_2 v\Sigma_2 v^T + w_1 w_2 (v\cdot(\mu_1-\mu_2))^2 = 1.$$

At a high level, we can conclude that either the projected means or variances are far apart. In particular, since $w_1 v\Sigma_1 v^T \le 1/2$, a crude statement is that $v\Sigma_2 v^T \ge 1/2$ or $|v\cdot(\mu_1-\mu_2)| \ge 1/2$ (or both).

To establish (16), we break into two further subsubcases.

Case 2bi: $|v\cdot(\mu_1-\mu_2)| \ge 1/2$.

Next consider a random sample from F_2. Note that $|v\cdot(\hat{\mu}_1-\mu_2)| > 1/4$. A crude bound on the probability that it is put in D_1 is,

$$\frac{2\sqrt{\varepsilon_2}/\varepsilon_3}{1/8} = 16\sqrt{\varepsilon_2}/\varepsilon_3.$$

To see this, note simply that the (marginal) density of points in the interval of width $1/8$ around μ_2 are all larger than those in the interval tested by the algorithm, which has width $2\sqrt{\varepsilon_2}/\varepsilon_3$. This establishes (16) in this case.

Case 2bii: $|v\cdot(\mu_1-\mu_2)| < 1/2$ and $v\Sigma_2 v^T > 1/2$. In this case, the density of F_2 is never larger than $1/\sqrt{\pi} \le 1$. Hence, the probability of falling in the specified interval is at most its width, $2\sqrt{\varepsilon_2}/\varepsilon_3$. This establishes (16). □

J Total Variance Estimates

J.1 Kullback-Leibler Divergence for Gaussians

Fact 20. *Let F_1, F_2 be one-dimensional normal distributions with means and variances μ_1, μ_2 and σ_1, σ_2, respectively. Then*

$$KL(F_1\|F_2) = \ln\frac{\sigma_2}{\sigma_1} + \frac{(\mu_1-\mu_2)^2 + \sigma_1^2 - \sigma_2^2}{2\sigma_2^2}$$

The KL between two n-dimensional Gaussians is:

Fact 21. *Let F_1, F_2 be n-dimensional normal distributions with means and variances μ_1, μ_2 and Σ_1, Σ_2, respectively. Then*

$$KL(F_1\|F_2) = \frac{1}{2}\left(\log\frac{det(\Sigma_2)}{det(\Sigma_1)} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) - n\right).$$

KL divergence is convex

J.2 Total Variance via Kullback-Leibler Divergence

Lemma 22. *Let $F_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $F_2 = \mathcal{N}(\mu_2, \Sigma_2)$ be two n-dimensional Gaussian distributions. Let $\lambda_1, \ldots, \lambda_n > 0$ be the eigenvalues of $\Sigma_1^{-1}\Sigma_2$. Then the variation distance between them satisfies,*

$$(D(F_1, F_2))^2 \leq \sum_{i=1}^{n}(\lambda_i + \frac{1}{\lambda_i} - 2) + (\mu_1 - \mu_2)^T\Sigma_1^{-1}(\mu_1 - \mu_2).$$

Proof. The Kullback-Leibler divergence (KL) between two Gaussians is well-known to be,

$$\mathrm{KL}(F_1\|F_2) = \frac{1}{2}\left(Tr(\Sigma_1^{-1}\Sigma_2) + \ln\frac{det(\Sigma_1)}{det(\Sigma_2)} - n + (\mu_1 - \mu_2)^T\Sigma_1^{-1}(\mu_1 - \mu_2)\right).$$

Note that $det(\Sigma_1^{-1}\Sigma_2) = \frac{det(\Sigma_2)}{det(\Sigma_1)} = \lambda_1 \ldots \lambda_n$, and hence $\ln\frac{det(\Sigma_1)}{det(\Sigma_2)} = \sum \ln\frac{1}{\lambda_i}$. Also, $Tr(\Sigma_1^{-1}\Sigma_2) = \lambda_1 + \ldots + \lambda_n$. It is also well-known that $D^2(F_1, F_2) \leq 2\mathrm{KL}(F_1\|F_2)$. This gives,

$$(D(F_1, F_2))^2 \leq \sum_{i=1}^{n}(\lambda_i + \ln\frac{1}{\lambda_i} - 1) + (\mu_1 - \mu_2)^T\Sigma_1^{-1}(\mu_1 - \mu_2).$$

Using the fact that $\log x \leq x - 1$, we are done. □

K Gaussian Inequalities

K.1 Maximization Inequalities

Lemma 23. *Let $f(x, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}$. Then*

$$\arg\max_{r\geq\sigma\geq 0} f(x, \sigma) = \begin{cases} x & r \geq |x| \\ r & r < |x| \end{cases}$$

Proof.

$$\frac{\partial}{\partial\sigma}f(x, \sigma) = \frac{1}{\sqrt{2\pi}}\Big[\frac{-1}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}} + \frac{x^2}{\sigma^4}e^{-\frac{x^2}{2\sigma^2}}\Big] = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma^2\sqrt{2\pi}}\Big[-1 + \frac{x^2}{\sigma^2}\Big]$$

For $x^2 > \sigma^2$, $\frac{\partial}{\partial\sigma}f(x, \sigma) > 0$ and for $x^2 < \sigma^2$, $\frac{\partial}{\partial\sigma}f(x, \sigma) < 0$. □

As a simple corollary:

Corollary 24.

$$\max_{|x|\geq r, \sigma^2 > 0}\mathcal{N}(0, \sigma^2, x) \leq \frac{1}{r\sqrt{2\pi}}$$

Proof.

$$\max_{|x|\geq r, \sigma^2>0}\mathcal{N}(0, \sigma^2, x) = \max_{|x|\geq r}\max_{\sigma^2>0}\mathcal{N}(0, \sigma^2, x) = \max_{|x|\geq r}\mathcal{N}(0, x^2, x) \leq \max_{|x|\leq r}\frac{1}{x\sqrt{2\pi}} \leq \frac{1}{r\sqrt{2\pi}}$$

□

Lemma 25. *Let $f(x, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}$. Then*

$$\frac{\partial^2}{\partial x^2}f(x, \sigma) = \begin{cases} < 0 & |x| < \sigma \\ 0 & |x| = \sigma \\ > 0 & |x| > \sigma \end{cases}$$

Proof.

$$\frac{\partial^2}{\partial x^2}f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\Big[\frac{-1}{\sigma^2} + \frac{x^2}{\sigma^4}\Big]e^{-\frac{x^2}{2\sigma^2}}$$

□

K.2 Central Moment Integrals

Lemma 26. *Let $\sigma^2 \leq 1$. Then*

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\frac{1}{\epsilon}}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx \leq \frac{1}{2} e^{-\frac{1}{4\epsilon^2}}$$

Proof. Let $I(x) = \int_{\frac{1}{\epsilon}}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx$. Let $B = [0, \frac{1}{\epsilon}] \times [0, \frac{1}{\epsilon}]$. Let $C = \{(x,y) | \sqrt{x^2+y^2} \leq \frac{1}{\epsilon} \text{ and } x, y \geq 0\}$. Let $B' = \Re^2 - B$, and $C' = \Re^2 - C$.

Then

$$I(x)^2 = \int_{\frac{1}{\epsilon}}^{\infty} \int_{\frac{1}{\epsilon}}^{\infty} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy = \int_{B'} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy$$

$$I(x)^2 \leq \int_{C'} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy = \frac{1}{4} \int_0^{2\pi} \int_{\frac{1}{\epsilon}}^{\infty} e^{-\frac{r^2}{2\sigma^2}} r dr d\theta$$

because $B' \subset C'$ and $e^y > 0$ for all real y. Thus

$$I(x)^2 \leq 2\pi \frac{-\sigma^2}{4} e^{-\frac{r^2}{2\sigma^2}} \Big|_{\frac{1}{\epsilon}}^{\infty} = 2\pi \frac{\sigma^2}{4} e^{-\frac{1}{2\epsilon^2\sigma^2}} \Rightarrow \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\frac{1}{\epsilon}}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx \leq \frac{1}{2} e^{-\frac{1}{4\epsilon^2\sigma^2}}$$

□

Claim 27. *For i odd:*

$$\begin{aligned} H_i(x,\sigma) := \int x^i e^{-\frac{x^2}{2\sigma^2}} dx &= -x^{i-1}\sigma^2 e^{-\frac{x^2}{2\sigma^2}} - (i-1)x^{i-3}\sigma^4 e^{-\frac{x^2}{2\sigma^2}} \\ &\quad -(i-1)(i-3)x^{i-5}\sigma^6 e^{-\frac{x^2}{2\sigma^2}} \ldots - (i-1)!!\sigma^{i+1} e^{-\frac{x^2}{2\sigma^2}} \end{aligned}$$

Proof. We can check

$$\begin{aligned} \frac{\partial}{\partial x} H_i(x,\sigma) &= x^i e^{-\frac{x^2}{2\sigma^2}} + (i-1)x^{i-2}\sigma^2 e^{-\frac{x^2}{2\sigma^2}} \ldots + (i-1)!! x\sigma^{i-1} e^{-\frac{x^2}{2\sigma^2}} \\ &\quad -(i-1)x^{i-2}\sigma^2 e^{-\frac{x^2}{2\sigma^2}} - (i-1)(i-3)x^{i-4}\sigma^4 e^{-\frac{x^2}{2\sigma^2}} \ldots - (i-1)!! x e^{-\frac{x^2}{2\sigma^2}} \\ &= x^i e^{-\frac{x^2}{2\sigma^2}} \end{aligned}$$

□

Similarly

Claim 28. *For i even:*

$$\begin{aligned} H_i(x,\sigma) := \int x^i e^{-\frac{x^2}{2\sigma^2}} dx &= -x^{i-1}\sigma^2 e^{-\frac{x^2}{2\sigma^2}} - (i-1)x^{i-3}\sigma^4 e^{-\frac{x^2}{2\sigma^2}} \\ &\quad -(i-1)(i-3)x^{i-5}\sigma^6 e^{-\frac{x^2}{2\sigma^2}} \ldots + (i-1)!!\sigma^i \int e^{-\frac{x^2}{2\sigma^2}} dx \end{aligned}$$

We can apply these identities to get bounds on the contribution of the tails - i.e. $|x| \geq \frac{2}{\epsilon}$ for all finite moments $i \geq 0$.

Lemma 29. *Let $\sigma^2 \leq 1$ and $|\mu| \leq \frac{1}{\epsilon}$.*

$$\int_{|x| \geq \frac{2}{\epsilon}} x^i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \leq O(\frac{1}{\epsilon^i} e^{-\frac{1}{2\epsilon^2}})$$

Proof. Let

$$T_i = \int_{|x| \geq \frac{2}{\epsilon}} x^i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Assume that $\mu \leq 0$. Then

$$T_i \leq 2 \int_{\frac{1}{\epsilon}}^{\infty} (x-\mu)^i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

We can compute the binomial expansion of $(x-\mu)^i$ and using this expansion, a bound on the contribution of the tails to the central moments of order $j \le i$ will yield a bound on the contribution of the tails to the raw moment T_i.

Using Claim 27 and Claim 28 (and the fact that $H_i(x,\sigma)$ for $\sigma^2 \le 1$ is maximized for $\sigma = 1$):

$$\int_{\frac{1}{\epsilon}}^{\infty} x^j \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = O(\frac{1}{\epsilon^{j-1}} e^{-\frac{1}{4\epsilon^2}})$$

Then using the bound $|\mu| \le \frac{1}{\epsilon}$

$$M_i \le O(\frac{1}{\epsilon^i} e^{-\frac{1}{4\epsilon^2}})$$

Note that here the hidden constant depends on i. □

K.3 Moment Concentration

Lemma 30. *Let $x_1, x_2, \ldots, x_m$ be independent draws from a Normal distribution of mean 0 and variance 1, and let $k \ge 1$ be an integer. Then, with probability $\ge 1-\delta$,*

$$\left| \frac{1}{m} \sum_{i=1}^{m} x_i^k - \mathrm{E}_{x \sim \mathcal{N}(0,1)}[x^k] \right| \le \sqrt{\frac{2^k k!}{\delta m}}.$$

Proof. Let $a = \mathrm{E}_x[x^k]$, the kth moment of a standard normal distribution. By Chebyshev's inequality, with probability at most δ,

$$\left(\frac{1}{m} \sum_{i=1}^{m} x_i^k - a \right)^2 \le \frac{1}{\delta} \mathrm{E} \left[\left(\frac{1}{m} \sum_{i=1}^{m} x_i^k - a \right)^2 \right].$$

Hence it suffices to show that $\mathrm{E}[\left(\frac{1}{m}\sum_{i=1}^m x_i^k - a\right)^2] \le \frac{2^k k!}{m}$. Clearly, $\mathrm{E}_{x_1,\ldots,x_m}\left[\frac{1}{m}\sum_{i=1}^m x_i^k - a\right] = 0$. Using the fact that the variance of a sum of independent random variables is the sum of the variances,

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{1}{m} \sum_{i=1}^{m} x_i^k - a \right)^2 \right] &= \frac{1}{m} \mathrm{E}_x \left[\left(x^k - a \right)^2 \right] \\
&= \frac{1}{m} \left(\mathrm{E}_x[x^{2k}] - a^2 \right) \\
&\le \frac{1}{m} \mathrm{E}[x^{2k}] \\
&= \frac{1}{m} \frac{(2k)!}{2^k k!}
\end{aligned}$$

In the last equality we have used (17). Using the fact that $\binom{2k}{k} < 2^{2k}$, we get that the above is at most $\frac{2^k k!}{m}$. □

K.4 Moment Estimates via Bernstein Inequalities

Let X_j be independent random variables, and $E[X_j] = 0$. Then the Bernstein Inequality states:

Theorem 31. *Assume that $E[|X_j^k| \le \frac{E[X_j^2]}{2} L^{k-2} k!$. Then*

$$Pr[\sum_j X_j > 2t \sqrt{\sum_j E[X_j^2]}] \le e^{-t^2}$$

for $0 < t \le \frac{\sqrt{\sum_j E[X_j^2]}}{2L}$

We will apply this inequality to get estimates for how many samples we need to estimate the central moments of a Gaussian. Suppose X_j is a Gaussian random variable that has mean 0 and variance 1. We need to bound the k^{th} central moments of the Gaussian, which we can use Claim 27 and Claim 28 to explicitly calculate.

$$E[|X_j^k|] = \begin{cases} (k-1)!! & k \text{ is even} \\ \sqrt{\frac{2}{\pi}}(k-1)!! & k \text{ is odd} \end{cases}$$

So in particular the criteria in Bernstein's Inequality holds for $L = \Theta(1)$. Then we can sample m points from the Gaussian, $X_1, X_2, \ldots X_m$.

Fact 32. *Let $t = \Theta(\sqrt{\ln \frac{1}{\delta}})$, and $m = \frac{4t^2}{\gamma^2}$. Then*

$$Pr[\frac{|\sum_j X_j|}{m} > \gamma] \leq \delta^a$$

We also need to bound how quickly an estimate of the 2^{nd} moment converges: So let $Y_j = X_j^2 - 1$, be the square of a sample from the Gaussian of mean 0 and variance 1. Then

$$E[|Y_j^k|] = \left|(2k-1)!! - k(2k-3)!! + \binom{k}{2}(2k-5)!! \ldots \right| = \Theta((2k-1)!!)$$

Again, we can choose $L = \Theta(1)$ to get the criteria in Bernstein's Inequality to hold. This implies

Fact 33. *Let $t = \Theta(\sqrt{\ln \frac{1}{\delta}})$, and $m = \frac{4t^2}{\gamma^2}$. Then*

$$Pr[\frac{|\sum_j X_j^2 - m|}{m} > \gamma] \leq \delta^a$$

Finally, we need to bound how quickly correlation estimates will converge. So suppose that X_j, Y_j are independent samples from a Gaussian of mean 0 and variance 1. Then let $Z_j = X_j Y_j$.

$$E[|Z_j^k|] = \begin{cases} (k-1)!!(k-1)!! & k \text{ is even} \\ \frac{2}{\pi}(k-1)!!(k-1)!! & k \text{ is odd} \end{cases}$$

Here also, $L = \Theta(1)$ suffices, and we get

Fact 34. *Let $t = \Theta(\sqrt{\ln \frac{1}{\delta}})$, and $m = \frac{4t^2}{\gamma^2}$. Then*

$$Pr[\frac{|\sum_j X_j Y_j|}{m} > \gamma] \leq \delta^a$$

K.5 Total Variation Distance Lower Bounds

Lemma 35. *Assume $\epsilon \leq \frac{1}{2}$. Then*

$$\|\mathcal{N}(0,1,x) - \mathcal{N}(0,1+\epsilon,x)\|_1 \geq \frac{\epsilon}{100}$$

Proof. Consider

$$\frac{\mathcal{N}(0,1,x)}{\mathcal{N}(0,1+\epsilon,x)} = \sqrt{1+\epsilon} e^{-\frac{\epsilon x^2}{2(1+\epsilon)}} = f(x)$$

We can use the Taylor series expansion for $\sqrt{1+x} = (1 + \frac{1}{2}x - \frac{1}{8}x^2 \ldots)$ and for $|x| \leq \frac{1}{2}$, $\sqrt{1+x} \geq 1 + \frac{x}{4}$. So using the assumption that $\epsilon \leq \frac{1}{2}$, we get $\sqrt{1+\epsilon} \geq 1 + \frac{\epsilon}{4}$. We can also use $e^x \geq 1 + x$ to get

$$f(x) \geq (1 + \frac{\epsilon}{4})(1 - \frac{\epsilon x^2}{2(1+\epsilon)}) \geq (1 + \frac{\epsilon}{4})(1 - \frac{\epsilon x^2}{2})$$

So for $x = \frac{1}{3}$:

$$f(\frac{1}{3}) \geq (1 + \frac{\epsilon}{4})(1 - \frac{\epsilon}{18}) \geq (1 + \frac{\epsilon}{4})(1 - \frac{\epsilon}{9}) > 1$$

We can use Lemma 25 to get that $f(x)$ is concave on $[\frac{-1}{3}, \frac{1}{3}]$. Note that $f(0) \geq 1 + \frac{\epsilon}{4}$. Then the concavity, and the value at 0 together imply that for $x \in [\frac{-1}{6}, \frac{1}{6}]$, $f(x) > 1 + \frac{\epsilon}{8}$.

$$\|\mathcal{N}(0,\sigma^2,x) - \mathcal{N}(0,\sigma^2+\epsilon,x)\|_1 \geq (1 - \frac{1}{1+\frac{\epsilon}{8}}) \int_{\frac{-1}{6}}^{\frac{1}{6}} \mathcal{N}(0,1,x)dx \geq \frac{1}{3}\mathcal{N}(0,1,\frac{1}{6}) \geq \frac{\epsilon}{100}$$

□

Lemma 36.

$$\|\mathcal{N}(\mu,\sigma^2,x)-\mathcal{N}(0,\sigma^2+\epsilon,x)\|_1 \geq \frac{1}{4}\|\mathcal{N}(0,\sigma^2,x)-\mathcal{N}(0,\sigma^2+\epsilon,x)\|_1$$

Proof. Let $F_1=\mathcal{N}(\mu,\sigma^2)$ and $F_2=\mathcal{N}(0,\sigma^2+\epsilon)$. Assume without loss of generality that $\mu\leq 0$.

$$\|\mathcal{N}(\mu,\sigma^2,x)-\mathcal{N}(0,\sigma^2+\epsilon,x)\|_1 \geq \int_0^\infty |F_1(x)-F_2(x)|1_{F_1(x)\leq F_2(x)}dx$$

So we are integrating the ℓ_1 difference over only the range $[0,\infty]$ and only where $F_1(x)\leq F_2(x)$. For every point $x\in[0,\infty]$, $F_1(x)$ is an increasing function of μ - and this just follows from the fact that a Gaussian is decreasing for all x larger than the mean. So if we let $F_3=\mathcal{N}(0,\sigma^2)$ then

$$\int_0^\infty |F_2(x)-F_1(x)|1_{F_1(x)\leq F_2(x)}dx \geq \int_0^\infty |F_2(x)-F_1(x)|1_{F_3(x)\leq F_2(x)}dx$$

because on $[0,\infty]$, $F_1(x)\leq F_3(x)$ so on the range $[0,\infty]$, $1_{F_1(x)\leq F_2(x)}\geq 1_{F_3(x)\leq F_2(c)}$. Also

$$\int_0^\infty |F_2(x)-F_1(x)|1_{F_3(x)\leq F_2(x)}dx \geq \int_0^\infty |F_2(x)-F_3(x)|1_{F_1(x)\leq F_3(x)}dx$$

because whenever $F_3(x)\leq F_2(x)$, we must have $F_1(x)\geq F_2(x)$, so we can replace $|F_2(x)-F_1(x)|$ by $F_2(x)-F_1(x)\geq F_2(x)-F_3(x)$ and lastly whenever $F_3(x)\leq F_2(x)$ we have $F_2(x)-F_3(x)=|F_2(x)-F_3(x)|$. And using symmetry properties:

$$\int_0^\infty |F_2(x)-F_3(x)|1_{F_1(x)\leq F_3(x)}dx = \frac{1}{2}\int_{-\infty}^\infty |F_2(x)-F_3(x)|1_{F_1(x)\leq F_3(x)}dx$$

And because $F_1(x),F_3(x)$ are distributions, the contribution from $F_1(x)\leq F_3(x)$ to the ℓ_1 difference is exactly $\frac{1}{2}$ of the total ℓ_1 difference. So

$$\frac{1}{2}\int_{-\infty}^\infty |F_2(x)-F_3(x)|1_{F_1(x)\leq F_3(x)}dx \geq \frac{1}{4}\|\mathcal{N}(0,\sigma^2,x)-\mathcal{N}(0,\sigma^2+\epsilon,x)\|_1$$

□

Lemma 37. *Let $\epsilon\leq\frac{1}{4}$.*

$$\|\mathcal{N}(\epsilon,1)-\mathcal{N}(0,1)\|_1 \geq \frac{\epsilon}{20}$$

Proof. First observe that the derivative of $\mathcal{N}(0,1,x)$ is nondecreasing on $(-\infty,-1]$, since the only solutions to $\frac{d^2\mathcal{N}(0,1,x)}{dx^2}=0$ are at ±1. Thus for $x\in[-3/2,-1]$ we have that $\mathcal{N}(0,1,x)\geq\mathcal{N}(\epsilon,1,x)+\epsilon\frac{d\mathcal{N}(\epsilon,1,x)}{dx}(x)$. For $x\in[-3/2,-1]$, and $\epsilon\leq 1/4$, we have

$$1/10\leq\frac{d\mathcal{N}(0,1,x)}{dx}(-3/2-1/4)\leq\frac{d\mathcal{N}(\epsilon,1,x)}{dx}(-3/2)\leq\frac{d\mathcal{N}(\epsilon,1,x)}{dx}(x).$$

Putting everything together,

$$\|\mathcal{N}(\epsilon,1,x),\mathcal{N}(0,1,x)\|_1\geq\int_{-3/2}^{-1}\left(\mathcal{N}(0,1,x)-\mathcal{N}(\epsilon,1,x)\right)dx\geq\int_{-3/2}^{-1}\epsilon\frac{d\mathcal{N}(\epsilon,1,x)}{dx}(x)dx\geq\epsilon\frac{1}{2}\frac{1}{10}.$$

□

Lemma 38. *Consider two weighted Gaussian distributions, $w_1\mathcal{N}(\mu_1,\sigma_1^2)$ and $w_2\mathcal{N}(\mu_2,\sigma_2^2)$ and suppose that $|w_1-w_2|+|\mu_1-\mu_2|+|\sigma_1^2-\sigma_2^2|\geq\epsilon$, and that $\frac{1}{2}\leq\sigma_1^2,\sigma_2^2\leq\frac{3}{2}$ and $w_1,w_2\geq\epsilon$: Then*

$$\|w_1\mathcal{N}(\mu_1,\sigma_1^2)-w_2\mathcal{N}(\mu_2,\sigma_2^2)\|_1=\Omega(\epsilon^3)$$

Proof. We will use the previous lemmas to break this into a case analysis.

Case 1: Suppose that $|w_1 - w_1| \geq c_1\epsilon^3$
Then $\|w_1\mathcal{N}(\mu_1, \sigma_1^2) - w_2\mathcal{N}(\mu_2, \sigma_2^2)\|_1 \geq |w_1 - w_2| \geq c_1\epsilon^3$

Case 2: Suppose that $|\sigma_1^2 - \sigma_2^2| \geq c_2\epsilon^2$. Then using the triangle inequality

$$\begin{aligned}\|w_1\mathcal{N}(\mu_1, \sigma_1^2) - w_2\mathcal{N}(\mu_2, \sigma_2^2)\|_1 &\geq \|w_1\mathcal{N}(\mu_1, \sigma_1^2) - w_1\mathcal{N}(\mu_2, \sigma_2^2)\|_1 - |w_1 - w_2| \\ &\geq w_1\|\mathcal{N}(\mu_1, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_2^2)\|_1 - c_1\epsilon^3 \\ &\geq \frac{w_1}{4}\|\mathcal{N}(0, \sigma_1^2) - \mathcal{N}(0, \sigma_2^2)\|_1 - c_1\epsilon^3\end{aligned}$$

where the last line follows from Lemma 36. And applying Lemma 35 and that $w_1 \geq \epsilon$

$$\|w_1\mathcal{N}(\mu_1, \sigma_1^2) - w_2\mathcal{N}(\mu_2, \sigma_2^2)\|_1 \geq \frac{c_2\epsilon^2}{800} - c_1\epsilon^3$$

So if $c_2 \geq 1600c_1$, then this case is complete.

Case 3: In the remaining case, $|\mu_1 - \mu_2| \geq \epsilon - c_2\epsilon^2 - c_1\epsilon^3 \geq \frac{\epsilon}{2}$ for sufficiently small ϵ. Then using the triangle inequality

$$\|w_1\mathcal{N}(\mu_1, \sigma_1^2) - w_2\mathcal{N}(\mu_2, \sigma_2^2)\|_1 \geq w_1\|\mathcal{N}(\mu_1, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_2^2)\|_1 - |w_1 - w_2|$$

And using Fact 18 and that $\sigma^2 \geq \frac{1}{2}$ and the triangle inequality:

$$\begin{aligned}w_1\|\mathcal{N}(\mu_1, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_2^2)\|_1 &\geq w_1\|\mathcal{N}(\mu_1, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_1^2)\|_1 - w_1\|\mathcal{N}(\mu_2, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_2^2)\|_1 \\ &\geq w_1\|\mathcal{N}(\mu_1, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_1^2)\|_1 - 20w_1c_2\epsilon^2\end{aligned}$$

Then we can use Lemma 37 and that $w_1 \geq \epsilon$, $\sigma_i^2 \leq 3/2$ to get:

$$w_1\|\mathcal{N}(\mu_1, \sigma_1^2) - \mathcal{N}(\mu_2, \sigma_1^2)\|_1 \geq w_1\frac{\epsilon}{20 \cdot 2\sqrt{3/2}} \geq w_1\frac{\epsilon}{50}$$

and this then implies that

$$\|w_1\mathcal{N}(\mu_1, \sigma_1^2) - w_2\mathcal{N}(\mu_2, \sigma_2^2)\|_1 \geq w_1\Big[\frac{\epsilon}{50} - 20c_2\epsilon^2\Big] - c_1\epsilon^3$$

and the lemma is complete. □

L The General Anisotropic Case

The basic idea of the general anisotropic algorithm, Algorithm 5 given in Figure 7, is simple: first use a polynomial number of samples to put the data very nearly into isotropic position, then run Algorithm 4 to get an approximation to the transformed mixture model, and then transform the resulting model back.

Other than for clarity of exposition, there does not seem to be a good reason to use separate data for estimating the covariance matrix of the data – one may very well be able to use the same data for both putting the data in isotropic position and estimating the mixture. The present approach we adopt is modular. It enables the analysis in the previous sections to deal with the case where the data is exactly in isotropic position and postpone discussion of isotropy completely to this section.

The formal argument here goes as follows. We first argue that, with high probability, the estimates of the covariance matrices will be sufficiently accurate that the transformed distribution will be nearly isotropic. Formally, we argue that it is in fact *statistically very close* to a GMM F' which is exactly in isotropic position. In fact, they are statistically so close that Algorithm 4 would behave identically on data from F and F'. To be precise, suppose Algorithm 4 used at most m samples. For the sake of analysis, imagine running the algorithm on samples from F and samples from F'. Then, if F and F' are at statistical distance at most $\delta/(2m)$, then we can *couple* samples from the two distributions (the sets of samples from F and F' would each be i.i.d., but the two sets would be dependent) such that with probability $1 - \delta/2$, the algorithm would give identical outputs. If we choose the parameters so that Algorithm 4 succeeds with probability $1 - \delta/2$ in approximating F' well, which in turn implies that we approximate F well.

The analysis we present here can be improved by applying results in [25, 19], but here we favor clarity of exposition over optimality in our bounds.

Algorithm 5. The general anisotropic algorithm

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

Output: n-dimensional GMM

1. Let p be the number of samples needed by Algorithm 4 (with input parameters $\epsilon, \frac{\delta}{2}, n$. Let $\epsilon_1 = \frac{\delta}{2p}$, and let $m = \Omega(\frac{\log \frac{1}{\delta} n^4}{\epsilon_1^5})$
2. Draw samples $x_1, x_2, \ldots, x_m$ from $\mathrm{SA}(F)$.
3. Let μ_s and Σ_s be the mean and covariance matrices of the sample.
4. Let $T(x) = \Sigma_s^{-1/2}(x - \mu_s)$ and $T^{-1}(x) = \mu_s + \Sigma_s^{1/2}x$.
5. Let $F' = T(F)$ be the GMM F under the transformation T.
6. Run Algorithm $4(n, \epsilon/2, \delta/2, \mathrm{SA}(F'))$ to get approximation $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$, where $\hat{F}_i = \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)$. (* Note that one can efficiently simulate $\mathrm{SA}(F')$ using $\mathrm{SA}(F)$. *)
7. Output $T^{-1}(\hat{F})$.

Figure 7: The general parameter learning algorithm.

Lemma 39. *Let $x_1, x_2, \ldots x_m \in \Re^n$, and let μ_1, Σ_1 be the mean and co-variance of these points. Let μ_2, Σ_2 be the mean and covariance of an n-dimensional Gaussian. Let $AA^T = \Sigma_2$. And let F_1, F_2 be the Gaussian distributions defined by μ_1, Σ_1 and μ_2, Σ_2 respectively. Let $y_i = A^{-1}(x_i - \mu_2)$*

$$KL(F_1 \| F_2) = \frac{1}{2}\Big[- \log det(\frac{1}{m}\sum_i y_i y_i^T - \frac{1}{m^2}\sum_{i,j} y_i y_j^T) + \frac{1}{m}\sum_i y_i^T y_i - n\Big]$$

Proof. Using Fact 21, we can write:

$$KL(F_1 \| F_2) = \frac{1}{2}(\log \frac{det(\Sigma_2)}{det(\Sigma_1)} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T\Sigma_2^{-1}(\mu_1 - \mu_2) - n)$$

Consider the term

$$Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T\Sigma_2^{-1}(\mu_1 - \mu_2)$$

Using the properties of the trace operator, we can re-write this as

$$Tr(A^{-1}\Sigma_1 A^{-T}) + Tr(A^{-1}(\mu_1 - \mu_2)(A^{-1}(\mu_1 - \mu_2))^T) =$$

$$Tr(A^{-1}(\frac{1}{m}\sum_i x_i x_i^T - \mu_1\mu_1^T)A^{-T}) + Tr(A^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T A^{-T}) =$$

$$Tr(A^{-1}(\frac{1}{m}\sum_i x_i x_i^T - \mu_1\mu_1^T + \mu_1\mu_1^T - \mu_1\mu_2^T - \mu_2\mu_1^T + \mu_2\mu_2^T)A^{-T})$$

We can apply the identity $\frac{1}{m}\sum_i x_i\mu_2^T = \mu_1\mu_2^T$ (and a similar identity for $\mu_2\mu_1^T$ to rewrite this as

$$Tr(A^{-1}(\frac{1}{m}\sum_i (x_i - \mu_2)(x_i - \mu_2)^T)A^{-T}) =$$

$$\frac{1}{m}\sum_i (A^{-1}(x_i - \mu_2))^T(A^{-1}(x_i - \mu_2))$$

Next consider the term

$$\log \frac{det(\Sigma_2)}{det(\Sigma_1)} = \log \frac{det(AA^T)}{det(\Sigma_1)} = -\log det(A^{-1}\Sigma_1 A^{-T})$$

We can write

$$\Sigma_1 = \frac{1}{m}\sum_i x_i x_i^T - \mu_1\mu_1^T = \frac{1}{m}\sum_i (x_i-\mu_2)(x_i-\mu_2)^T - \mu_1\mu_1^T + \mu_1\mu_2^T + \mu_2\mu_1^T - \mu_2\mu_2^T$$

$$\Sigma_1 = \frac{1}{m}\sum_i (x_i-\mu_2)(x_i-\mu_2)^T - (\mu_1-\mu_2)(\mu_1-\mu_2)^T$$

This implies that

$$A^{-1}\Sigma_1 A^{-T} = \frac{1}{m}\sum_i (A^{-1}(x_i-\mu_2))(A^{-1}(x_i-\mu_2))^T - \frac{1}{m^2}\sum_{i,j}(A^{-1}(x_i-\mu_2))(A^{-1}(x_j-\mu_2))^T$$

□

We will think of x_i as being sampled from $\mathcal{N}(\mu_2, \Sigma_2)$, and we will use this expression for the Kullback-Leibler divergence to bound how quickly sampling points from $\mathcal{N}(\mu_2, \Sigma_2)$ will converge to a Gaussian that is close in statistical distance to the original Gaussian. If we sample x_i from $\mathcal{N}(\mu_2, \Sigma_2)$, and then apply the transformation $y_i = A^{-1}(x_i - \mu_2)$, this is equivalent to sampling from the distribution $\mathcal{N}(0, I)$. We can then use bounds on how quickly moment estimates from the distribution $\mathcal{N}(0, I)$ converge to the actual moments to bound how quickly the Kullback-Leibler divergence converges to zero, which will in turn give us a bound on how quickly the total variance of the empirical distribution $\mathcal{N}(\mu_2, \Sigma_1)$ converges in total variation distances to the distribution $\mathcal{N}(\mu_2, \Sigma_2)$.

Lemma 40. *Let $I + A$ be a real, symmetric matrix. Suppose that for all i, j, $|A_{i,j}| \leq \gamma$. Then*

$$(1-\gamma n)^n \leq det(I+A) \leq (1+\gamma n)^n$$

Proof. Consider the Rayleigh quotient:

$$\frac{x^T I x + x^T A x}{x^T x} = 1 + \frac{x^T A x}{x^T x}$$

Let $u = \frac{x}{\sqrt{x^T x}}$. Then

$$u^T A u = Tr(u^T A u) = Tr(A u u^T) \leq \|A\|_F \|u u^T\|_F$$

where $\|B\|_F$ denotes the Frobenius norm of the matrix B. $\|uu^T\|_F = 1$ and $\|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2} \leq \gamma n$. So

$$u^T A u \leq \gamma n$$

and this implies that the Rayleigh quotient is always in the range $[1-\gamma n, 1+\gamma n]$. Because $I+A$ is a symmetric matrix, we can compute an eigen-decomposition of $I + A = \sum_i \lambda_i v_i v_i^T$ and then $det(I+A) = \prod_i \lambda_i$. Each eigenvalue λ_i is in the range $[1-\gamma n, 1+\gamma n]$ because the Rayleigh quotient is always in this range. So

$$(1-\gamma n)^n \leq det(I+A) \leq (1+\gamma n)^n$$

□

Let y_i be a sample from $\mathcal{N}(0, I)$. Note that $E[y_i^p] = 0$, $E[(y_i^p)^2] = 1$ and $E[y_i^p y_i^q] = 0$.

Lemma 41. *Suppose that $|\frac{\sum_i y_i^p}{m}| \leq \gamma$, $|\frac{\sum_i (y_i^p)^2}{m} - 1| \leq \gamma$, and $|\frac{\sum_i y_i^p y_i^q}{m}| \leq \gamma$ for all $p, q \in [n]$. Then*

$$KL(F_1 \| F_2) = \frac{1}{2}\Big[-\log det\Big(\frac{1}{m}\sum_i y_i y_i^T - \frac{1}{m^2}\sum_{i,j} y_i y_j^T\Big) + \frac{1}{m}\sum_i y_i^T y_i - n\Big] \leq 6\gamma n^2$$

Proof. Consider the term

$$\Big|\frac{1}{m}\sum_i y_i^T y_i - n\Big| = \Big|\sum_p \Big(\frac{1}{m}\sum_i (y_i^p)^2 - 1\Big)\Big| \leq n\gamma$$

Let $B' = \frac{1}{m}\sum_i y_i y_i^T$ and $B'' = \mu_1\mu_1^T$ where $\mu_1 = \frac{1}{m}\sum_i y_i$.

$$B'_{p,q} = \frac{1}{m}\sum_i y_i^p y_i^q$$

$$\text{if } p = q \text{ then } 1-\gamma \leq B'_{p,q} \leq 1+\gamma \text{else } p \neq q \text{ and } |B'_{p,q}| \leq \gamma$$

Also consider the vector μ_1: $|\frac{1}{m}\sum_i y_i^p| \leq \gamma$ So this implies that $|B''_{p,q}| \leq \gamma^2$. This implies that we can write $B' + B''$ as $B' + B'' = I + A$ where $|A_{p,q}| \leq \gamma + \gamma^2 \leq 2\gamma$ for sufficiently small γ. So we can apply Lemma 40 and get

$$n\log(1-2\gamma n) \leq \log det(I+A) \leq n\log(1+2\gamma n)$$

Using the inequality $\log 1 + x \leq x$, and that for $|x| \leq \frac{1}{2}$ $1 - 2x \leq \log 1 - x$:

$$-4\gamma n^2 \leq det(I+A) \leq 2\gamma n^2$$

and this implies the lemma. □

We can use Fact 32, Fact 33 and Fact 34 and set $\gamma = \frac{\epsilon^{2c}}{3n^2}$ to get that with probability $1-\delta$, and $m = O(\frac{n^4 \ln \frac{1}{\delta}}{\epsilon^{4c}})$ samples, we get an empirical Gaussian $\mathcal{N}(\mu_1, \Sigma_1)$ (i.e. we sample points $x_1, x_2, \ldots x_m \in \Re^n$, and let μ_1, Σ_1 be the mean and co-variance of these points) for which

$$D(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)) \leq \sqrt{2KL(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2))} \leq O(\epsilon^c)$$

Proposition 42. *Let $m = \Omega(\frac{\log \frac{1}{\delta} n^4}{\epsilon^{4c+1}})$. Let $x_1, x_2 \ldots x_m$ denote m samples from a GMM $F = w_1F_1 + w_2F_2$. Let μ_s and Σ_s be the mean and co-variance matrix of the m sample points, and let $AA^T = \Sigma_s$. Then with probability at least $1-\delta$ there is a GMM $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$ such that, $D(\hat{F}, F) \leq \epsilon^c$ and*

$$D(\hat{F}_i, F_i) \leq \epsilon^c \text{ and } |\hat{w}_i - w_i| \leq \epsilon^c, \text{ for each } i = 1, 2$$

and the transformation $A^{-1}(x - \mu_s)$ puts the GMM $\hat{F}$ in isotropic position.

Proof. We will construct the GMM $\hat{F}$ explicitly as follows: partition the sample set $x_1, x_2, \ldots x_m$ into X_1, X_2, the sample points that came from F_1, F_2 respectively. Let $m_1 = |X_1|$, i.e. the number of points (among the m samples) sampled from F_1 (and let $m_2 = |X_2|$). Then applying Hoeffding's bound:

$$Pr[|\frac{m_1}{m} - w_1| \geq \frac{\epsilon^c}{4}] \leq 2e^{-2m\frac{\epsilon^{2c}}{16}} \leq \frac{\delta}{4}$$

because $m \geq \Omega(\frac{\log \frac{1}{\delta}}{\epsilon^{2c}})$. Denote the event that $|\frac{m_1}{m} - w_1| \leq \frac{\epsilon^c}{4}$ as $(E1)$. If $(E1)$ does occur, then $m_1, m_2 = \Omega(\frac{n^4 \ln \frac{1}{\delta}}{\epsilon^{4c}})$

Then given the partition X_1, X_2 let $\hat{\mu}_1, \hat{\Sigma}_1$ be the mean and co-variance of X_1, and similarly let $\hat{\mu}_2, \hat{\Sigma}_2$ be the mean and co-variance of X_2. We construct $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$ with $\hat{w}_1 = \frac{m_1}{m}, \hat{w}_2 = \frac{m_2}{m}$ and let $\hat{F}_1 = \mathcal{N}(\hat{\mu}_1, \hat{\Sigma}_1), \hat{F}_2 = \mathcal{N}(\hat{\mu}_2, \hat{\Sigma}_2)$.

We can use Fact 32, Fact 33 and Fact 34 and set $\gamma = \frac{\epsilon^{2c}}{3n^2}$ to get that with probability $1 - \frac{\delta}{3}$, and $m_i = O(\frac{n^4 \ln \frac{1}{\delta}}{\epsilon^{4c}})$ samples, we get an empirical Gaussian $N(\hat{\mu}_i, \hat{\Sigma}_i)$ for which

$$D(\mathcal{N}(\hat{\mu}_i,\hat{\Sigma}_i),\mathcal{N}(\mu_i,\Sigma_i)) \leq \sqrt{2KL(\mathcal{N}(\hat{\mu}_i\|\hat{\Sigma}_i)\|\mathcal{N}(\mu_i,\Sigma_i))} \leq O(\epsilon^c)$$

Let the event that this happens be $(E2), (E3)$ respectively.

So with probability at least $1-\delta$, all events $(E1), (E2)$ and $(E3)$ occur. This implies that

$$D(\hat{F}, F) \leq D(\hat{w}_1\hat{F}_1, w_1F_1) + D(\hat{w}_2\hat{F}_2, w_2F_2)$$

and using the triangle inequality and events $(E1)$ and $(E2)$

$$D(\hat{w}_1\hat{F}_1, w_1F_1) \leq D(\hat{w}_1\hat{F}_1, w_1\hat{F}_1) + D(w_1\hat{F}_1, w_1F_1) \leq |\hat{w}_1 - w_1| + D(\hat{F}_1, F_1) \leq \frac{\epsilon^c}{2}$$

Algorithm 6. DENSITY ESTIMATION
Input: Integer $n \geq 1$, sample oracle $\mathrm{SA}(F)$, $\epsilon, \delta > 0$
Output: For $i = 1, 2$, $\hat{\mu}_i \in \mathbf{R}^n, \hat{\Sigma}_i \in \mathbf{R}^{n \times n}$.

1. Let $m_1 = O(\frac{n^4 \ln \frac{1}{\delta}}{\epsilon^4}), m_2 = O(\frac{\ln \frac{1}{\delta}}{\epsilon}), \epsilon_1 = \frac{\delta}{4m_1}$
2. Run Algorithm 5$(n, \epsilon_1, \frac{\delta}{4}, \mathrm{SA}(F))$ to get approximation $\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$
3. Draw m_1 samples S_1 from $\mathrm{SA}(F)$, and set $\hat{S} = \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)$ where $\hat{\mu}_s$ and $\hat{\Sigma}_s$ are the mean and co-variance of the set S_1
4. Draw m_2 samples S_2 from $\mathrm{SA}(F)$. Let $U = \{x | x \in \mathbf{R}^n \text{ and } \hat{F}(x) > \hat{S}(x)\}$
5. Let p be the fraction of samples from S_2 in U
6. Construct a sample oracle $\mathrm{SA}(\hat{F})$ for $\hat{F}$. Draw m_2 samples from $\mathrm{SA}(\hat{F})$ and let $\hat{p}$ be the fraction of samples in U.
7. If $|p - \hat{p}| \leq \frac{\epsilon}{4}$, Output $\hat{F}$, and otherwise Output $\hat{S}$

Figure 8: The density estimation algorithm.

An identical argument for $i = 2$ yields the inequality

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

We note that the means for the set of points X and for the distribution $\hat{F}$ are equal:

$$E_{X=X_1 \cup X_2}[x_i] = \mu_s = E_{\hat{F}}[x]$$

and the co-variance plus the squared mean are also equal:

$$E_{X=X_1 \cup X_2}[x_i x_i^T] = \Sigma_s + \mu_s \mu_s^T = E_{\hat{F}}[xx^T]$$

This implies that the means and co-variance matrices are equal, and as a result the same transformation $A^{-1}(x - \mu_s)$ that puts the point set X into isotropic position will also put the mixture $\hat{F}$ into isotropic position. □

We now conclude the proof of our main theorem:
Proof of theorem 1: Let p and ϵ_1 be as in Algorithm 5. Then $D(\hat{F}, F) \leq \frac{\delta \epsilon_1}{2}$ so with probability at least $1 - \frac{\delta}{2}$, we can assume that the $p = \frac{1}{\epsilon_1}$ random samples needed by Algorithm 4 all originate from $SA(\hat{F})$. $\hat{F}$ is exactly in isotropic position after the transformation T, so Algorithm 4 will return an $\frac{\epsilon}{2}$-accurate estimate with probability at least $1 - \frac{\delta}{2}$. Note that the notion of $\frac{\epsilon}{2}$-accurate is affine-invariant, so an ϵ-accurate estimate for F' is also an $\frac{\epsilon}{2}$-accurate estimate for $\hat{F}$. And again using Proposition 42, any estimate that is $\frac{\epsilon}{2}$-accurate compared to $\hat{F}$ must also be ϵ-accurate compared to F. □

M Density Estimation

In this section we state the efficient density estimation algorithm, and prove Corollary 2.

Proof of 2. We first prove that one of the two estimates $\hat{F}$, or $\hat{S}$ will be close in statistical distance to the actual distribution F:

Let $\nu = \min(D(F_1, F_2), w_1, w_2)$. If $\nu \geq \epsilon_1$, then the distribution $F = w_1 F_1 + w_2 F_2$ satisfies the requirements of Theorem 1 and with probability at least $1 - \frac{\delta}{4}$ Algorithm 5 will output a distribution $\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ satisfies $D(\hat{F}_i, F_{\pi(i)}) \leq \epsilon_1$ and $|\hat{w}_i - w_{\pi(i)}| \leq \epsilon_1$ for some permutation $\pi : \{1, 2\} \to \{1, 2\}$. This implies that $D(F, \hat{F}) \leq 4\epsilon_1$. Also note that $32\epsilon_1 \leq \epsilon$ and this implies that $D(F, \hat{F}) \leq \frac{\epsilon}{8}$ with probability at least $1 - \frac{\delta}{4}$.

In the remaining case, $\nu \leq \epsilon_1$. In this case we can assume that all the samples come from one of the two Gaussians, either F_1 or F_2. Consider the case $w_1 = \nu$, and the remaining cases will follow an identical

argument. We sample m_1 points from SA(F) and so the probability that any sample comes from F_1 is at most $m_1 w_1 \leq \epsilon_1 m_1 \leq \frac{\delta}{4}$. So in this case, with probability at least $1 - \frac{\delta}{4}$ we can assume that all m_2 samples come from F_2. Given m_1 samples from F_2, we can apply Fact 32, Fact 33, Fact 34 and Lemma 41 and set $\gamma = \frac{\epsilon^2}{3n^2}$ to get that with probability $1 - \frac{\delta}{4}$, and $m_1 = O(\frac{n^4 \ln \frac{1}{\delta}}{\epsilon^4})$ samples, we get an empirical Gaussian $\hat{S} = \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)$ for which

$$D(\mathcal{N}(\mu_2, \Sigma_2), \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)) \leq \sqrt{2KL(\mathcal{N}(\mu_2, \Sigma_2) \| \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s))} \leq \epsilon_1$$

And because $\nu \leq \epsilon_1$, this implies that $D(F, \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)) \leq 2\epsilon_1 \leq \frac{\epsilon}{8}$ with probability at least $1 - \frac{\delta}{4}$.

So with probability at least $1 - \frac{\delta}{2}$, we have that either $D(F, \hat{F}) \leq \frac{\epsilon}{8}$ or $D(F, \hat{S}) \leq \frac{\epsilon}{8}$. All we need to do is prove that we output (close to) the correct estimate, either $\hat{F}$ or $\hat{S}$.

Suppose that $D(\hat{F}, \hat{S}) \leq \frac{\epsilon}{2}$. In this case, we can apply the triangle inequality and both estimates $\hat{F}$, $\hat{S}$ are statistical distance at most ϵ from F, so we can output either estimate.

If not, then $D(\hat{F}, \hat{S}) > \frac{\epsilon}{2}$. We assume that $D(\hat{F}, F) \leq \frac{\epsilon}{8}$ and the remaining case is identical.

Let $U = \{x | x \in \mathbf{R}^n \text{ and } \hat{F}(x) > \hat{S}(x)\}$. $\|\hat{F}, \hat{S}\|_1 \geq \epsilon$ and this implies

$$\int_U \hat{F}(x)dx - \int_U \hat{S}(x)dx = Pr_{x \leftarrow \hat{F}}[x \in U] - Pr_{x \leftarrow \hat{S}}[x \in U] \geq \frac{\epsilon}{2}$$

In particular, this implies that

$$\int_U \hat{F}(x)dx = Pr_{x \leftarrow \hat{F}}[x \in U] \geq \frac{\epsilon}{2}$$

Also, because $D(\hat{F}, F) \leq \frac{\epsilon}{8}$ this implies that

$$\|Pr_{x \leftarrow \hat{F}}[x \in U] - Pr_{x \leftarrow F}[x \in U]\|_1 \leq \frac{\epsilon}{8}$$

So we draw samples from F to estimate $Pr_{x \leftarrow F}[x \in U]$. We get an estimate p for this probability, and we can also estimate $Pr_{x \leftarrow F}[x \in U]$ by constructing a sample oracle for $\hat{F}$ and empirically measuring this probability from samples. Let this estimate be $\hat{p}$. If $p, \hat{p}$ are sufficiently close, then we output $\hat{F}$ and if not we output $\hat{S}$.

Using the standard Chernoff bound, choosing $m_2 = O(\frac{\ln \frac{1}{\delta}}{\epsilon})$ samples we can guarantee that with probability at least $1 - \frac{\delta}{4}$ both the estimate $\hat{p}$ for $Pr_{x \leftarrow \hat{F}}[x \in U]$ and the estimate p for $Pr_{x \leftarrow F}[x \in U]$ will be $\frac{\epsilon}{16}$ close to the actual probabilities. $Pr_{x \leftarrow \hat{F}}[x \in U]$ and $\|Pr_{x \leftarrow \hat{F}}[x \in U] - Pr_{x \leftarrow F}[x \in U]\|_1 \leq \frac{\epsilon}{8}$ and so with probability at least $1 - \frac{\delta}{4}$

$$|p - \hat{p}| < \frac{\epsilon}{4}$$

and the Algorithm 6 will output $\hat{F}$ correctly in this case. If instead $D(F, \hat{S}) \leq \frac{\epsilon}{8}$ then because p is within $\frac{\epsilon}{16}$ of $Pr_{x \leftarrow F}[x \in U]$ and $\hat{p}$ is within $\frac{\epsilon}{16}$ of $Pr_{x \leftarrow \hat{F}}[x \in U]$ and

$$\|Pr_{x \leftarrow \hat{S}}[x \in U] - Pr_{x \leftarrow F}[x \in U]\|_1 \leq \frac{\epsilon}{8}$$

and

$$\|Pr_{x \leftarrow \hat{F}}[x \in U] - Pr_{x \leftarrow \hat{S}}[x \in U]\|_1 > \frac{\epsilon}{2}$$

so this implies that $|p - \hat{p}| > \frac{\epsilon}{4}$ and in this case we will correctly output $\hat{S}$. □

N Proof of Optimal Clustering

Lemma 43. *Given a* univariate *GMM* $F = w_1 F_1 + w_2 F_2$ *such that* $w_i \geq \epsilon$, *and* $D(F_1, F_2) \geq \epsilon$, *then* $D(F, \mathcal{N}(0, 1)) > poly(\epsilon)$.

Proof. This lemma is essentially a much simpler version of the polynomially robust identifiability of mixtures (Theorem 4); the proof approach and intuition carry over. Throughout, let μ_i, σ_i^2 be the mean and variance of Gaussian F_i. Assume without loss of generality that $\sigma_1^2 \leq \sigma_2^2$. We first show that there will be a $poly(\epsilon)$ disparity in one of the low-order moments of F and $\mathcal{N}(0, 1)$.

Case 1: Assume $\sigma_1^2 > 1 - \epsilon^4$. If $\sigma_1^2 > 2$, then we immediately have a disparity in the variance of F and $\mathcal{N}(0,1)$ of at least 1. Otherwise, by Fact 20 and the triangle inequality, since $D(F_1, F_2) > \epsilon$, and $\sigma_1^2 \leq 2$, either $|\mu_1 - \mu_2| > \epsilon/40$ or $|\sigma_1^2 - \sigma_2^2| > \epsilon/40$. In the first case, we have:

$$\begin{aligned} var(F) &= w_1\sigma_1^2 + w_2\sigma_2^2 + w_1\mu_1^2 + w_2\mu_2^2 \\ &\geq \sigma_1^2 + \epsilon\left(\frac{|\mu_1 - \mu_2|}{2}\right)^2 \geq 1 - \epsilon^4 + \epsilon\frac{\epsilon^2}{40^2} \geq 1 + \epsilon^4, \end{aligned}$$

and we have a disparity in the variances of F and $\mathcal{N}(0,1)$.

Case 2: Assume $\sigma_1^2 \leq 1 - \epsilon^4$. Consider the deconvolved distributions $\mathcal{F}_{\sigma_1^2-\epsilon^8}(F), \mathcal{F}_{\sigma_1^2-\epsilon^8}(\mathcal{N}(0,1))$. From our assumption on σ_1^2, it follows that $\max_x \mathcal{F}_{\sigma_1^2-\epsilon^8}(\mathcal{N}(0,1))(x) \leq \frac{1}{.9\epsilon^2\sqrt{2\pi}}$. Additionally, $\mathcal{F}_{\sigma_1^2-\epsilon^8}(F)(\mu_1) \geq w_1\frac{1}{\epsilon^4\sqrt{2\pi}} \geq \frac{1}{\epsilon^3\sqrt{2\pi}}$. Together with a bound of $\frac{2}{\epsilon^8}$ for the derivative of the probability density function of $\mathcal{F}_{\sigma_1^2-\epsilon^8}(F)(x)$, this implies that $D(\mathcal{F}_{\sigma_1^2-\epsilon^8}(F), \mathcal{F}_{\sigma_1^2-\epsilon^8}(\mathcal{N}(0,1))) \geq \frac{1}{72\epsilon^2}$. From Proposition 7, the function $f(x) = \mathcal{F}_{\sigma_1^2-\epsilon^8}(F)(x) - \mathcal{F}_{\sigma_1^2-\epsilon^8}(\mathcal{N}(0,1))(x)$ has at most 4 (which is less than 6) zeros, and thus the proof of Lemma 9 applies without modification to yield that at least one of the first six moments of $\mathcal{F}_{\sigma_1^2-\epsilon^8}(F)$ differs from that of $\mathcal{F}_{\sigma_1^2-\epsilon^8}(\mathcal{N}(0,1))$ by $\Omega(\epsilon^6 6)$. By Lemma 6, it follows that one of the first six moments of F differs from that of $\mathcal{N}(0,1)$ by at least $c = \Omega(\epsilon^6 6)$.

To conclude our proof, we argue that a $poly(\epsilon)$ disparity in the first six moments implies a $poly(\epsilon)$ statistical distance. First note that if $\sigma_2^2 > 2$, we can directly show a statistical distance of at least ϵ^4. Assuming that $\sigma_2^2 \leq 2$, from Lemma 29 (bounds on the contribution of the tails of a Gaussian to the i^{th} moment), it follows that for some $i \in [6]$,

$$\int_{x \in [-\sqrt{2}/\epsilon, \sqrt{2}/\epsilon]} x^i(F(x) - \mathcal{N}(0,1,x))dx \geq c/2.$$

Finally, we have:

$$\begin{aligned} c/2 &\leq \int_{x \in [-\sqrt{2}/\epsilon, \sqrt{2}/\epsilon]} x^i(F(x) - \mathcal{N}(0,1,x))dx \\ &\leq \int_{x \in [-\sqrt{2}/\epsilon, \sqrt{2}/\epsilon]} |x^i(F(x) - \mathcal{N}(0,1,x))|dx \\ &\leq \int_{x \in [-\sqrt{2}/\epsilon, \sqrt{2}/\epsilon]} \left(\frac{\sqrt{2}}{\epsilon}\right)^i |(F(x) - \mathcal{N}(0,1,x))|dx \leq \frac{8}{\epsilon^6} 2D(F, \mathcal{N}(0,1)), \end{aligned}$$

from which the lemma follows. □

Lemma 44. *Consider a mixture of two* multivariate *Gaussians: $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2 + \hat{F}_2$ in isotropic position and a single Gaussian $\hat{S}$. Suppose that* $\min(\hat{w}_1, \hat{w}_2, D(\hat{F}_1, \hat{F}_2)) \geq \epsilon$. *Then for a random direction r with probability at least $1 - \delta$ the statistical distance between the two distributions $\hat{F}$ and $\hat{S}$ when projected onto r is at least $\epsilon_1 = poly(\epsilon, \frac{1}{n}, \frac{1}{\delta})$.*

Proof. Given a mixture of two Gaussians: $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2 + \hat{F}_2$ in isotropic position, we can apply Lemma 13 and the resulting univariate mixture, which we will denote by $\hat{S} = \hat{w}_1\hat{S}_1 + \hat{w}_2 + \hat{S}_2$ is in isotropic position, and has $\hat{w}_1, \hat{w}_2 \geq \epsilon$. We can apply Lemma 38 (setting $w_1, w_2 = 1$ when applying the lemma in order to get a lower bound on the statistical distance between $\hat{S}_1$ and $\hat{S}_2$). So we know that for some $\epsilon_2 = poly(\epsilon, \frac{1}{n}, \frac{1}{\delta})$, $\hat{S} = \hat{w}_1\hat{S}_1 + \hat{w}_2 + \hat{S}_2$ satisfies $\hat{w}_1, \hat{w}_2, D(\hat{S}_1 \| \hat{S}_2) \geq \epsilon_2$. Then we can apply Lemma 43 to get that the statistical distance between the two distributions $\hat{F}$ and $\hat{S}$ when projected onto r is at least $\epsilon_1 = poly(\epsilon, \frac{1}{n}, \frac{1}{\delta})$. □

Claim 45. *Suppose $F = w_1F_1 + w_2F_2$. Then if* $\min(w_1, w_2, D(F_1, F_2)) \leq \epsilon$ *then $\hat{p} = 1$ has error probability at most $OPT + 2\epsilon$.*

Proof. The error probability of a clustering scheme $\hat{p}$ is measured w.r.t. the best permutation $\pi : \{1,2\} \to \{1,2\}$, so we can assume without loss of generality that if either of the two mixing weights w_1, w_2 is at most ϵ, that w_2 is at most ϵ. So consider this case. The error probability of $\hat{p}$ is:

$$E_{x \leftarrow F}[|\hat{p}(x) - 1 - \frac{w_2F_2(x)}{w_1F_1(x) + w_2F_2(x)}|] = E_{x \leftarrow F}[|\frac{w_2F_2(x)}{w_1F_1(x) + w_2F_2(x)}|] \leq \epsilon$$

So the error probability of $\hat{p}$ is certainly at most $OPT+\epsilon$. Next, consider the case in which $D(F_1\|F_2) \leq \epsilon$. Again, the error probability of a clustering scheme $\hat{p}$ is measured w.r.t. the best permutation $\pi : \{1,2\} \to \{1,2\}$, so here we can assume that $w_1 \geq \frac{1}{2}$. We can re-write the distributions F_1, F_2 as $F_1 = f_c + f_1$ and $F_2 = f_c + f_2$ where $\|f_i\|_1 = D(F_1\|F_2)$ and f_i is non-negative. So we can express sampling from the distribution $w_1F_1 + w_2F_2$ as, with probability $1 - D(F_1\|F_2)$, sample from f_c and with probability $w_1D(F_1\|F_2)$ sample from f_1, and with the remaining probability sample from $w_2D(F_1\|F_2)$. We can assume that the clustering scheme $\hat{p}$ chooses the wrong cluster in the latter two cases, and in the first case the clustering scheme will choose the correct cluster with probability at least w_1, so the error probability of the clustering scheme is at most

$$D(F_1\|F_2) + (1 - D(F_1\|F_2))w_2 \leq D(F_1\|F_2) + w_2 \leq w_2 + \epsilon$$

And a similar argument implies that the optimal clustering scheme has error probability at least $w_2 - D(F_1\|F_2) \geq w_2 - \epsilon$. So this implies the claim. □

Claim 46. *Consider a clustering scheme $\hat{p} = \lceil \frac{\hat{w}_2\hat{F}_2}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2} \rceil + 1$. If there are indices $i, j \in \{1,2\}$ such that*

$$D(\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2\|w_1F_1 + w_2F_2) + D(\hat{w}_i\hat{F}_i\|w_jF_j) \leq \epsilon$$

then the error probability of $\hat{p}$ is at most $OPT + 2\epsilon$

Proof. Consider

$$E_{x\leftarrow F}[|\frac{w_1F_1}{w_1F_1 + w_2F_2} - \frac{\hat{w}_1\hat{F}_1}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2}|]$$

We can re-write this as

$$\begin{aligned}
& \int_{x\in\Re^n} |\frac{w_1F_1}{w_1F_1 + w_2F_2} - \frac{\hat{w}_1\hat{F}_1}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2}|(w_1F_1 + w_2F_2)dx \\
\leq & \int_{x\in\Re^n} |w_1F_1 - \hat{w}_1\hat{F}_1\frac{w_1F_1 + w_2F_2}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2}dx \\
\leq & \int_{x\in\Re^n} (|w_1F_1 - \hat{w}_1\hat{F}_1| + |\hat{w}_1\hat{F}_1 - \hat{w}_1\hat{F}_1(\frac{w_1F_1 + w_2F_2}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2})|dx \\
\leq & \int_{x\in\Re^n} (|w_1F_1 - \hat{w}_1\hat{F}_1| + |\frac{\hat{w}_1\hat{F}_1}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2}(w_1F_1 + w_2F_2 - \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2)|)dx \\
\leq & \int_{x\in\Re^n} (|w_1F_1 - \hat{w}_1\hat{F}_1| + |(w_1F_1 + w_2F_2 - \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2)|)dx
\end{aligned}$$

So

$$E_{x\leftarrow F}[|\frac{w_1F_1}{w_1F_1 + w_2F_2} - \frac{\hat{w}_1\hat{F}_1}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2}|] \leq D(\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2\|w_1F_1 + w_2F_2) + D(\hat{w}_i\hat{F}_i\|w_jF_j) \leq \epsilon$$

We will use this bound on the ℓ_1 difference between the conditional probabilities to give a bound on how much more the clustering scheme $\hat{p}$ can err, compared to the optimal. Consider any point $x \in \Re^n$. Suppose the optimal clustering scheme pays u at this point, i.e

$$\min(\frac{w_1F_1(x)}{w_1F_1(x) + w_2F_2(x)}, \frac{w_2F_2(x)}{w_1F_1(x) + w_2F_2(x)}) = u$$

Without loss of generality, assume that the optimal clustering scheme clusters this point at 1. Then, the clustering scheme $\hat{p}$ only pays more at this point if the clustering is different. Let

$$v = |\frac{w_1F_1}{w_1F_1 + w_2F_2} - \frac{\hat{w}_1\hat{F}_1}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2}|$$

If the clustering for $\hat{p}$ is different from p_{OPT} this implies

$$\frac{w_2F_2(x)}{w_1F_1(x) + w_2F_2(x)} + v = u + v \geq \frac{1}{2}$$

Algorithm 7. NEAR-OPTIMAL CLUSTERING
Input: Integer $n \geq 1$, sample oracle $\mathrm{SA}(F)$, $\epsilon, \delta > 0$
Output: For $\hat{p} : \mathbf{R}^n \to \{1, 2\}$.

1. Let ϵ_1 be as in Lemma 44, and let $\epsilon_2 << \epsilon_1$ Also let $m_1 = O(\frac{n^4 \ln \frac{1}{\delta}}{\epsilon_2^4}), m_2 = O(\frac{\ln \frac{1}{\delta}}{\epsilon_1})$
2. Run Algorithm 5$(n, \epsilon_2, \frac{\delta}{4}, \mathrm{SA}(F))$ to get approximation $\hat{F} = \hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2$
3. If $\min(\hat{w}_1, \hat{w}_2, D(\hat{F}_1 \| \hat{F}_2)) \leq \frac{\epsilon}{2}$, output $\hat{p} = 1$
4. Draw m_1 samples S_1 from $\mathrm{SA}(F)$, and set $\hat{S} = \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)$ where $\hat{\mu}_s$ and $\hat{\Sigma}_s$ are the mean and co-variance of the set S_1
5. Draw m_2 samples S_2 from $\mathrm{SA}(F)$. Let $U = \{x | x \in \mathbf{R}^n \text{ and } \hat{F}(x) > \hat{S}(x)\}$
6. Let p be the fraction of samples from S_2 in U
7. Construct a sample oracle $\mathrm{SA}(\hat{F})$ for $\hat{F}$. Draw m_2 samples from $\mathrm{SA}(\hat{F})$ and let q be the fraction of samples in U.
8. If $|p - q| \leq \frac{\epsilon_1}{4}$, Output $\hat{p} = \lceil \frac{\hat{w}_2\hat{F}_2}{\hat{w}_1\hat{F}_1 + \hat{w}_2 + \hat{F}_2} \rceil + 1$, and otherwise Output $\hat{p} = 1$

Figure 9: The near-optimal clustering algorithm.

The clustering scheme $\hat{p}$ pays $1 - u$ which is $1 - 2u$ larger than what the optimal clustering scheme pays at this point. But using the above inequality, the difference in payment is at most $2v$, and from the above argument we know that under the best permutation $\pi : \{1, 2\} \to \{1, 2\}$, the expected difference in conditional probabilities is at most:

$$D(\hat{w}_1\hat{F}_1 + \hat{w}_2 + \hat{F}_2 \| w_1F_1 + w_2F_2) + D(\hat{w}_i\hat{F}_i \| w_jF_j) \leq \epsilon$$

for any indices $i, j \in \{1, 2\}$. So this implies that the expected cost difference between $\hat{p}$ and p_{OPT} is at most 2ϵ

□

Theorem 47. *For any $n \geq 1, \epsilon, \delta > 0$ and any n-dimensional GMM $F = w_1F_1 + w_2F_2$, using m independent samples from F, Algorithm 7 outputs a clustering scheme $\hat{p}$ such that with probability $\geq 1 - \delta$ (over the samples and randomization of the algorithm) the error probability of $\hat{p}$ is at most ϵ larger than the error probability of the optimal clustering scheme p_{OPT}. And the runtime (in the Real RAM model) and number of samples drawn from the oracle is at most poly$(n, \frac{1}{\epsilon}, \frac{1}{\delta})$.*

Proof. Consider the actual mixture of Gaussians $F = w_1F_1 + w_2F_2$. Either $\min(w_1, w_2, D(F_1 \| F_2)) \geq \epsilon_2$, or $\min(w_1, w_2, D(F_1 \| F_2)) \leq \epsilon_2$ and in either case one of the two clustering schemes will have error probability at most ϵ larger than the optimal clustering scheme.

In order to show this, consider each case. Suppose $\min(w_1, w_2, D(F_1 \| F_2)) \geq \epsilon_2$. In this case, the conditions of Algorithm 5 are met. This implies

$$D(\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2 \| w_1F_1 + w_2F_2) + D(\hat{w}_i\hat{F}_i \| w_jF_j) \leq \frac{\epsilon}{2}$$

And using the Claim 46 this implies that $\hat{p} = \lceil \frac{\hat{w}_2\hat{F}_2}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2} \rceil + 1$ has error probability at most ϵ larger than the optimal clustering scheme.

Suppose that $\min(w_1, w_2, D(F_1 \| F_2)) \leq \epsilon_2$. Then we can immediately invoke Claim 45 and this implies that $\hat{p} = 1$ has error probability at most ϵ larger than the optimal clustering scheme.

All that we need to prove is that Algorithm 7 outputs a good clustering scheme. It is easy to see that if $\min(\hat{w}_1, \hat{w}_2, D(\hat{F}_1 \| \hat{F}_2)) \leq \frac{\epsilon}{2}$, then the algorithm might as well output the clustering scheme $\hat{p} = 1$ because even if $\min(w_1, w_2, D(F_1 \| F_2)) \geq \frac{\epsilon_1}{4}$ and $\hat{p} = \lceil \frac{\hat{w}_2\hat{F}_2}{\hat{w}_1\hat{F}_1 + \hat{w}_2\hat{F}_2} \rceil + 1$ has error probability at most ϵ larger than the optimal clustering scheme, based on the guarantees of Algorithm 5 $\min(w_1, w_2, D(F_1 \| F_2)) \leq \epsilon$ and so the clustering scheme $\hat{p} = 1$ will have error probability at most ϵ larger than optimal, again invoking Claim 45.

And in the remaining case, (after applying the transformation T) we can invoke Lemma 44 and with probability at least $1 - \frac{\delta}{4}$, when projected on the random direction r the statistical distance between the two distributions $\hat{F}$ and $\hat{S}$ when projected onto r is at least ϵ_1. And projecting cannot increase statistical distance, so the statistical distance between $\hat{F}$ and $\hat{S}$ is at least ϵ_1. Notice that the statistical distance between F and $\hat{F}$ is at most $\frac{\epsilon_2}{4}$ in the case in which $\min(w_1, w_2, D(F_1 \| F_2)) \geq \frac{\epsilon_2}{4}$. So with sufficiently many samples p will be within $O(\epsilon_2) << \epsilon_1$ of q, and we will output the clustering scheme based on $\hat{F}$ in this case. The analysis is identical to the analysis in Algorithm 6. And a similar argument follows for the remaining case.

□

O Moments for Gaussians

For completeness, we give the first six (raw) moments of a univariate normal random variable, $\mathcal{N}(\mu, \sigma^2)$:

$$
\begin{aligned}
&\mathrm{E}[x^0] = 1 \\
&\mathrm{E}[x^1] = \mu \\
&\mathrm{E}[x^2] = \mu^2 + \sigma^2 \\
&\mathrm{E}[x^3] = \mu^3 + 3\mu\sigma^2 \\
&\mathrm{E}[x^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \\
&\mathrm{E}[x^5] = \mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4 \\
&\mathrm{E}[x^6] = \mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6
\end{aligned}
$$

Additionally, for the univariate normal distribution $\mathcal{N}(0, \sigma^2)$, the ith raw moment is,

$$
\mathrm{E}_{x \sim \mathcal{N}(0,\sigma^2)}[x^i] = \begin{cases} 0 & \text{if } i \text{ is odd} \\ (i-1)!!\sigma^i = \frac{i!}{2^i (i/2)!}\sigma^i & \text{if } i \text{ is even.} \end{cases} \tag{17}
$$