Tight Bounds for Learning a Mixture of Two Gaussians

Moritz Hardt
m@mrtz.org
IBM Research Almaden

Eric Price
ecprice@cs.utexas.edu
The University of Texas at Austin

May 16, 2015

Abstract

We consider the problem of identifying the parameters of an unknown mixture of two arbitrary d-dimensional gaussians from a sequence of independent random samples. Our main results are upper and lower bounds giving a computationally efficient moment-based estimator with an optimal convergence rate, thus resolving a problem introduced by Pearson (1894). Denoting by σ^2 the variance of the unknown mixture, we prove that $\Theta(\sigma^{12})$ samples are necessary and sufficient to estimate each parameter up to constant additive error when $d = 1$. Our upper bound extends to arbitrary dimension $d > 1$ up to a (provably necessary) logarithmic loss in d using a novel—yet simple—dimensionality reduction technique. We further identify several interesting special cases where the sample complexity is notably smaller than our optimal worst-case bound. For instance, if the means of the two components are separated by $\Omega(\sigma)$ the sample complexity reduces to $O(\sigma^2)$ and this is again optimal.

Our results also apply to learning each component of the mixture up to small error in total variation distance, where our algorithm gives strong improvements in sample complexity over previous work. We also extend our lower bound to mixtures of k Gaussians, showing that $\Omega(\sigma^{6k-2})$ samples are necessary to estimate each parameter up to constant additive error.

1 Introduction

Gaussian mixture models are among the most well-studied models in statistics, signal processing, and computer science with a venerable history spanning more than a century. Gaussian mixtures arise naturally as way of explaining data that arises from two or more homogeneous populations mixed in varying proportions. There have been numerous applications of gaussian mixtures in disciplines including astronomy, biology, economics, engineering and finance.

The most basic estimation problem when dealing with any mixture model is to approximately identify the parameters that specify the model given access to random samples. In the case of a gaussian mixture the model is determined by a collection of means, covariance matrices and mixture probabilities. A sample is drawn by first selecting a component according to the mixture probabilities and then sampling from the normal distribution specified by the corresponding mean and covariance. Already in 1894, Pearson [Pea94] proposed the problem of estimating the parameters of a mixture of two one-dimensional gaussians in the context of evolutionary biology. Pearson analyzed a population of crabs and found that a mixture of two gaussians faithfully explained the size of the crab "foreheads". He concluded that what he observed was a mixture of two species rather than a single species and further speculated that "*a family probably breaks up first into two species, rather than three or more, owing to the pressure at a given time of some particular form of natural selection.*"

Fitting a mixture of two gaussians to the observed crab data was a formidable task at the time that required Pearson to come up with a good approach. His approach is based on the *method of moments* which uses the empirical moments of a distribution to distinguish between competing models. Given n samples $x_1, \ldots, x_n$ the k-th empirical moment is defined as $\frac{1}{n}\sum_i x_i^k$, which for sufficiently large n will approximate the true moment $\mathbb{E}\, x_i^k$. A mixture of two one-dimensional gaussians has 5 parameters so one might hope that 5 moments are sufficient to identify the parameters. Pearson derived a ninth degree polynomial p_5 in the first 5 moments and located the roots of this polynomial. Each root gives a candidate mixture that matches the first 5 moments; there were two valid solutions, among which Pearson selected the one whose 6-th moment was closest to the observed empirical 6-th moment.

In this work, we extend the method proposed by Pearson and prove that the extended method reliably recovers the parameters of the unknown mixture. Moreover, we show that the sample complexity we achieve is essentially optimal. To illustrate the quantitative bound that we get, if the means and variances are separated by constants and the total variance of the mixture is σ^2, then we show that up to constant factors it is necessary and sufficient to use σ^{12} samples to recover the parameters up to small additive error. Our work can be interpreted as providing an extension of Pearson's 120-year old estimator that achieves an optimal convergence rate. We extend our result to arbitrary dimension d using an apparently novel but surprisingly simple dimensionality reduction technique. This allows us to obtain the same sample complexity in any dimension up to a logarithmic loss in d, which we can also show is necessary.

Closely related to our results is an important recent work of Kalai, Moitra and Valiant [KMV10] who gave the first proof of a computationally efficient estimator with an inverse-polynomial convergence rate for the problem we consider. In particular, they show that six moments suffice to identify a mixture of two one-dimensional gaussians. Moreover, the result is robust in the sense that if the parameters of a mixture with variance σ^2 are separated by constants, then one of the first 6 moments must differ by $1/\sigma^{66}$. In particular, the first six *empirical* moments suffice provided that they're within $1/\sigma^{66}$ of the true moments (which happens for $n \gg \sigma^{132}$). This then leads to an estimator up to some polynomial loss. They also show that a solution to the 1-dimensional problem extends to any dimension d up to some loss that's polynomial in d and σ using a suitable dimensionality reduction technique. In contrast to their result which is within a polynomial factor of optimal, our result is within a constant factor of optimal in one dimension and within a $\log_d \log \sigma$ factor of optimal in d dimensions.

1.1 Problem description

A mixture F of two d-dimensional gaussians is specified by mixing probabilities $p_1, p_2 \geq 0$ such that $p_1 + p_2 = 1$, two means $\mu_1, \mu_2 \in \mathbb{R}^d$ and two covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$. A sample from F is generated by first picking an integer $i \in \{1, 2\}$ from the distribution (p_1, p_2) and then sampling from the d-dimensional gaussian measure $\mathrm{N}(\mu_i, \Sigma_i)$.

The variance σ^2 of a 1-dimensional mixture of two gaussians is $p_1p_2(\mu_1 - \mu_2)^2 + p_1\sigma_1^2 + p_2\sigma_2^2$. For a d-dimensional mixture, it is useful to define its "variance" as the maximum variance of any coordinate,

$$\mathbb{V}(F) \stackrel{\text{def}}{=} p_1p_2\|\mu_1 - \mu_2\|_\infty^2 + p_1\|\Sigma_1\|_\infty + p_2\|\Sigma_2\|_\infty\,. \tag{1}$$

Given samples from F our goal is to recover the parameters that specify the mixture up to small additive error; this is known as *parameter distance*. It is easy to see that we can only hope to recover the components of the mixture up to permutation. For simplicity it is convenient to combine the error in estimating the parameters:

Definition 1.1. *We say that mixture $\widehat{F}$ is ϵ-close to mixture F if there is a permutation π for which*

$$\max\left(\|\mu^{(i)} - \widehat{\mu}^{(\pi(i))}\|_\infty^2, \|\Sigma^{(i)} - \widehat{\Sigma}^{(\pi(i))}\|_\infty\right) \leq \epsilon^2 \, \mathbb{V}(F)\,.$$

We say that an algorithm (ϵ, δ)-learns a mixture F of two gaussians from $f(\epsilon, \delta)$ samples, if given $f(\epsilon, \delta)$ i.i.d. samples from F, it outputs a mixture $\widehat{F}$ that is ϵ-close to F with probability $1 - \delta$.

Note that this definition does not require good recovery of the p_i. If the two components of the mixture are indistinguishable, one cannot hope to recover the p_i to additive error. On the other hand, if the components are well-separated, one can use that the overall mean is the p-weighted average of the component means—or an analogous statement for the variances—to estimate the p_i from estimates of the parameters. Our main theorem will give a more precise characterization of how well we estimate p, but for simplicity we ignore it in much of the paper.

We also consider learning a mixture of Gaussians component-wise in the total variation norm.

Definition 1.2. *We say that mixture $\widehat{F} = \widehat{p}_1\widehat{G}_1 + \widehat{p}_2\widehat{G}_2$ is component-wise ϵ-close to mixture $F = p_1G_1 + p_2G_2$ in total variation if there is a permutation π for which*

$$\max_{i\in\{1,2\}} \mathrm{TV}(G_i, \widehat{G}_{(\pi(i))}) \leq \epsilon\,.$$

We say that an algorithm (ϵ, δ)-learns a mixture F of two gaussians in total variation from $f(\epsilon, \delta)$ samples, if given $f(\epsilon, \delta)$ i.i.d. samples from F, it outputs a mixture $\widehat{F}$ that is component-wise ϵ-close to F in total variation with probability $1 - \delta$.

Why parameter distance? We believe that proper learning of each component of a Gaussian mixture in the parameter distance is the most natural objective. Consider the simple example of estimating the height distribution of adult men and women from unlabeled population data, which is well approximated by a mixture of Gaussians [SWW02]. By using parameter distance, our results give a tight characterization of how precisely you can estimate the average male and female heights from a given number of samples. A guarantee in total variation norm is less easily interpreted.

Focusing on parameter distance also has technical advantages in our context. First, it leads to a cleaner quantitative analysis. Second, if the covariance matrices are (close to) *sparse* we can recover only the dominant entries of the covariance matrices and ignore the rest, decreasing our sample complexity. An affine invariant measure such as total variation distance could not benefit from sparsity this way. Nevertheless, to facilitate comparison with previous work, we state our results for total variation norm as well.

Finally, it is important that we learn each component rather than the mixture distribution. Finding a distribution that closely approximates the mixture distribution is easier but less useful than approximating the individual components, and is the focus of a different line of work that we discuss in the related work section. In fact, the individual parameters are a strong reason for modeling with mixtures of gaussians in the first place.

1.2 Main results

One-dimensional algorithm. Our main theorem is a general result that achieves tight bounds in multiple parameter regimes. As a consequence it's a little cumbersome to state, so we start with two simpler corollaries. The first corollary is that the algorithm (ϵ, δ)-learns a mixture with $O(\epsilon^{-12}\log(1/\delta))$ samples.

Corollary 1.3. *Let F be any mixture of 1-dimensional gaussians where p_1 and p_2 are bounded away from zero. Then Algorithm 3.3 can (ϵ, δ)-learn F with $O(\epsilon^{-12} \log(1/\delta))$ samples.*

The 12th power dependence on ϵ arises because our algorithm uses the 6th moment. In fact, we will see that in general this result is tight: there exist distributions for which one cannot reliably estimate *either* the μ_i to $\pm\epsilon\sigma$ or the σ_i^2 to $\pm\epsilon^2\sigma^2$ with $o(1/\epsilon^{12})$ samples.

However, for many distributions one can estimate the parameters with fewer samples. One important special case is when the two gaussians have means that are separated by $\Omega(1)$ standard deviations. In this case, our algorithm requires only $O(1/\epsilon^2)$ samples.

Corollary 1.4. *Let F be a mixture of 1-dimensional gaussians where p_1 and p_2 are bounded away from zero and $|\mu_1 - \mu_2| = \Omega(\sigma)$. Then Algorithm 3.3 can (ϵ, δ)-learn F with $O(\epsilon^{-2} \log(1/\delta))$ samples.*

This result is also tight: even if the samples from the mixture were labeled, it still would take $\Omega(1/\epsilon^2)$ samples to estimate the mean and variance of each gaussian to the desired precision. Our main theorem gives a smooth tradeoff between these two corollaries.

Theorem 3.10. *Let F be any mixture of two gaussians with variance σ^2 and p_1, p_2 bounded away from 0. Then, given $O(\epsilon^{-2} n \log(1/\delta))$ samples Algorithm 3.3 with probability $1 - \delta$ outputs the parameters of a mixture $\widehat{F}$ so that for some permutation π and all $i \in \{1, 2\}$ we have the following guarantees:*

- *If $n \geq \left(\frac{\sigma^2}{|\mu_1 - \mu_2|^2}\right)^6$, then $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq \epsilon|\mu_1 - \mu_2|$, $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\mu_1 - \mu_2|^2$, and $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon$.*
- *If $n \geq \left(\frac{\sigma^2}{|\sigma_1^2 - \sigma_2^2|}\right)^6$, then $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\sigma_1^2 - \sigma_2^2| + |\mu_1 - \mu_2|^2$ and $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon + \frac{|\mu_1 - \mu_2|^2}{|\sigma_1^2 - \sigma_2^2|}$.*
- *For any $n \geq 1$, the algorithm performs as well as assuming the mixture is a single gaussian: $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq |\mu_1 - \mu_2| + \epsilon\sigma$ and $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq |\sigma_1^2 - \sigma_2^2| + |\mu_1 - \mu_2|^2 + \epsilon\sigma^2$.*

In essence, the theorem states that the algorithm can distinguish the two gaussians in the mixture if it has at least $(\frac{\sigma^2}{\max(|\mu_1 - \mu_2|^2, |\sigma_1^2 - \sigma_2^2|)})^6$ samples. Once this happens, the parameters can be estimated to $\pm\epsilon$ relative accuracy with only a $1/\epsilon^2$ factor more samples. If the means are reasonably separated, then the first clause of the theorem provides the strongest bounds. If there is no separation in the means, we cannot hope to learn the means to relative accuracy, but we can still learn the variances to relative accuracy provided that they're separated. This is the content of the second clause. If neither means nor variances are separated, our algorithm is no better or worse than treating the mixture as a single gaussian.

The only assumption present in our main theorem requires that $\min(p_1, p_2)$ be bounded away from zero. Making this assumption simplifies the proof on a syntactic level considerably. A polynomial dependence on the separation from 0 could be extracted from our techniques, but we don't know if this dependence would be optimal.

Lower bound. Our second main result is that the bound in Theorem 3.10 is essentially best possible among all estimators—even computationally inefficient ones. More concretely, we exhibit a pair of mixtures $F, \tilde{F}$ that satisfy the following strong bound on the squared Hellinger distance[1] between the two distributions.

[1]For probability measures P and Q with densities p and q, respectively, the squared Hellinger distance is defined as $\mathrm{H}^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}x$.

Lemma 1.5. *There are two one-dimensional gaussian mixtures $F, \tilde{F}$ with variances σ^2 and all of the μ_i, σ_i^2, and p_i separated by $\Theta(1)$ from each other such that the squared Hellinger distance satisfies*

$$\mathrm{H}^2(F, \tilde{F}) \leq O\left(\sigma^{-12}\right) .$$

Denoting by F^n the distribution obtained by taking n independent samples from F, the squared Hellinger distance satisfies the direct sum rule $\mathrm{H}^2(F^n, \tilde{F}^n) \leq n \cdot \mathrm{H}^2(F, \tilde{F})$. Moreover, if $\mathrm{H}^2(F^n, \tilde{F}^n) \leq o(1)$ then the total variation distance also satisfies $\mathrm{TV}(F^n, \tilde{F}^n) \leq o(1)$. In particular, in this case no statistical test can distinguish F and $\tilde{F}$ from n samples with high confidence and parameter estimation is therefore impossible. The following theorem follows, showing that Corollary 1.3 is optimal.

Theorem 2.5. *Consider any algorithm that, given n samples of any gaussian mixture with variance σ^2, with probability $1 - \delta$ learns* either *μ_i to $\pm\epsilon\sigma$ or σ_i^2 to $\pm\epsilon^2\sigma^2$. Then $n = \Omega(\epsilon^{-12} \log(1/\delta))$.*

Since (ϵ, δ)-learning the mixture requires learning *both* the μ_i and the σ_i^2 to this precision, we get that Corollary 1.3 is tight. This also justifies our definition of ϵ-approximation in parameter distance meaning approximating the means to $\pm\epsilon\sigma$ and the variances to $\pm\epsilon^2\sigma_i^2$.

Corollary 1.6. *Any algorithm that uses $f(\epsilon, \delta)$ samples to (ϵ, δ)-learn arbitrary mixtures of two 1-dimensional gaussians with p_1 and p_2 bounded away from zero requires $f(\epsilon, \delta) = \Omega(\epsilon^{-12} \log(1/\delta))$.*

We also note that our lower bound technique directly gives a lower bound of $\Omega(\epsilon^{-6k+2})$ for the problem of learning a mixture of k Gaussians for constant $k \geq 2$ (Theorem 2.11). This is incomparable to the lower bound of, roughly, $\exp(k/24)$ for $\epsilon < 1/k$ due to [MV10]. Our bound is useful when k is a small constant and ϵ is going to zero, while their bound is useful when both k and $1/\epsilon$ are large.

Upper bound in arbitrary dimensions. Our main result holds for the d-dimensional problem up to replacing $\log(1/\delta)$ by $\log(d \log(1/\epsilon)/\delta)$ in the sample complexity.

Theorem 4.11. *Let F be any mixture of d-dimensional gaussians where p_1 and p_2 are bounded away from zero. Then we can (ϵ, δ)-learn F with $O(\epsilon^{-12} \log(d \log(1/\epsilon)/\delta))$ samples.*

Notably, our bound is essentially dimension-free and incurs only a logarithmic dependence on d. The best previous bound for the problem is the bound due to [KMV10] that gives a polynomial dependence of $O((d/\epsilon)^c)$ for some large constant c. The proof of our theorem is based on a new dimension-reduction technique for the mixture problem that is quite different from the one in [KMV10]. Apart from the quantitative improvement that it yields, it is also notably simpler.

Lower bound in higher dimension. We can extend our lower bound (Theorem 2.5) to show that $\Omega(\epsilon^{-12} \log(d/\delta))$ samples are necessary to achieve the guarantee of Theorem 4.11; one can embed a different instance of the hard distribution in each of the d dimensions, and the guarantee requires that the algorithm solve all the copies. That this direct product is hard is shown in Theorem 2.7. Hence Theorem 4.11 is optimal up to the $\log\log(1/\epsilon)$ term, and optimal up to constant factors when $d \geq \log(1/\epsilon)$.

Learning in total variation norm. In Section 5 we derive various results for learning mixtures of gaussians in the total variation norm.

Theorem 5.1. *Let F be any mixture of d-dimensional gaussians where p_1 and p_2 are bounded away from zero. For any dimension $d \geq 1$, Algorithm 5.1 (ϵ, δ)-learns F in total variation with $O(\epsilon^{-36} d^{30} \log^6(d/\epsilon) \log(1/\delta))$ samples.*

While the d^{30} dependence here is probably not close to optimal, the exponent is nonetheless several orders of magnitude smaller than the exponent of the polynomial dependence that follows from previous work. Interestingly, this large sample complexity of the general case can be improved if the covariances of the two Gaussians have similar eigenvalues and eigenvectors (e.g., they are isotropic):

Theorem 5.2. *Let F be any mixture of d-dimensional gaussians with covariance matrices Σ_1 and Σ_2 where the mixing probabilities p_1 and p_2 are bounded away from zero. Further suppose that there exists a constant C such that*

$$\Sigma_1 \preceq C\Sigma_2 \preceq C^2\Sigma_1.$$

Then there is an algorithm that can (ϵ, δ)-learn F in total variation with $O(\epsilon^{-12} d^6 \log^6(d/\epsilon) \log(1/\delta))$ samples.

1.3 Related Work

The body of related work on gaussian mixture models is too broad to survey here. We refer the reader to [KMV10] for a helpful discussion of work prior to 2010. Since then a number of works have further contributed to the topic. Moitra and Valiant [MV10] gave polynomial bounds for estimating the parameters of a mixture of k gaussians based on the method of moments. Belkin and Sinha [BS10] achieved a similar result. It is an interesting question if our techniques extend to the case of k gaussians, but as our lower bounds show the sample size must be at least $\Omega(\epsilon^{-6k+2})$ which is prohibitive for small ϵ and even moderate k.

Work of Chan et al. [CDSS13, CDSS14] implies an *improper* learning algorithm for a mixture of two single-dimensional gaussians that learns the overall mixture (not the components) in total variation distance to error ϵ using $\tilde{O}(1/\epsilon^2)$ samples. An improper learning algorithm in general does not return a mixture of gaussians nor does it return an approximation to the individual components of the mixture.

Daskalakis and Kamath [DK14] strengthen this result by giving a *proper* learning algorithm for learning a one-dimensional mixture with the same sample complexity. However, unlike our algorithm, it does not learn the individal components of the mixture. Indeed, this is impossible in general given the stated sample complexity in light of our lower bound. Nonetheless, our bounds do imply a proper learning algorithm for the mixture itself (which is a strictly weaker task than learning both components). In the case where $d > 1$, our algorithm for learning under total variation norm implies the best known bounds also for this weaker task when no assumptions are placed on the mixture.

A number of recent works have considered gaussian mixture models under stronger assumptions on the components. See, for example, [HK13, AJOS14]. We are not aware of improvements over [KMV10] for the parameter estimation problem when no such assumptions are made.

1.4 Proof overview

We now give a high-level outline of our algorithmic approach (and the related approach of Pearson). The starting point for the method of moments is to set up a system of polynomial equations whose

coefficients are determined by the moments of the mixture and whose variables are the unknown parameters. Solving the system of polynomial equations recovers the unknown parameters. The main stumbling block is that the roots of polynomials are notoriously unstable with respect to small perturbations in the coefficients. A famous example is Wilkinson's polynomial. Perturbations arise inevitably in our context because we do not know the moments of the mixture model exactly but rather need to estimate them empirically from samples. Our main contribution is to exhibit a robust set of polynomial equations from which the parameters can be recovered. We hope that similar techniques may be useful in extending our results to other settings such as learning a mixture of more than two gaussians.

Reparametrization. We begin by reparametrizing the gaussian mixture in such a way to get parameters that are independent of adding gaussian noise to the mixture. Formally, adding or subtracting the same term from each of the variances leaves these parameters unchanged. Assuming the overall mean of the mixture is 0, this leaves us with 3 free parameters that we call α, β, γ. Since these parameters are independent of adding gaussian noise it is useful to also define the moments of the mixture in such a way that they are independent under adding gaussian noise. This is accomplished by considering what we call *excess moments*. The name is inspired by the term *excess kurtosis*, a well-known measure of "peakedness" of a distribution introduced in [Pea94] that corresponds to the fourth excess moment. At this point, the third through sixth excess moments give us four equations in the three variables α, β, γ.

Three different precision regimes. Our analysis distinguishes between three different parameter regimes. In the first parameter regime we know each excess moment X_i for $i \leq 6$ up to an additive error of $\epsilon|\mu_1 - \mu_2|^i$. This analysis is applicable when the means are separated and it leads to the first case in Theorem 3.10. The second regime is when the separation between the means is small, but we nevertheless know each excess moment up to error $\epsilon|\sigma_1^2 - \sigma_2^2|^{i/2}$. This analysis in this case applies when the variances are separated and leads to the second case in Theorem 3.10. Finally, when neither of the cases applies the two gaussians are indistinguishable and we simply fit a single gaussian. We show that we can figure out which parameter regime we're in and run the appropriate algorithm.

We focus here on a discussion of the first parameter regime, since it is the most interesting case. The full argument is in Section 3.3.

Robustifying Pearson's polynomial. Expressing the excess moments in terms of our new parameters α, β, γ, we can derive in a fairly natural manner a ninth degree polynomial $p_5(y)$ whose coefficients depend on X_3, X_4, and X_5 so that α has to satisfy $p_5(\alpha) = 0$. The polynomial p_5 was already used by Pearson. Unfortunately, p_5 can have multiple roots and this is to be expected since 5 moments are not sufficient to identify a mixture of two gaussians. Pearson computed the mixtures associated with each of the roots and threw out the invalid solutions (e.g. the ones that give imaginary variances), getting two valid mixtures that matched on the first five moments. He then chose the one whose sixth moment was closest to the observed sixth moment.

We proceed somewhat differently from Pearson after computing p_5. First, we use the first 4 excess moments to compute an upper bound y_{max} on α. We show that the set of valid mixtures that match the first 5 moments correspond precisely to the roots y of $p_5(y)$ with $0 < y \leq y_{max}$. We then derive another ninth degree polynomial using X_3, X_4, X_5, and X_6 that we call $p_6(y)$. We prove that α is the only solution to the system of equations

$$\left\{p_5(y) = 0, \quad p_6(y) = 0, \quad 0 < y \leq y_{\max}\right\}.$$

This approach isn't yet robust to small perturbations in the moments; for example, if p_5 has a double root at α, it may have no nearby root after perturbation. We therefore consider the polynomial $r(y) = p_5^2(y) + p_6^2(y)$ which we know is zero at α. We argue that $r(y)$ is significantly nonzero for any y significantly far from α. This is the main technical claim we need.

For intuition of why this is the case, consider the normalization $|\mu_1 - \mu_2| = 1$ and the setting where $|\sigma_1^2 - \sigma_2^2| = O(1)$. Because the excess moments are polynomials in α, β, γ we can think of $r(y)$ as a polynomial in $(y, \alpha, \beta, \gamma)$. We are interested in some region $R \subset \mathbb{R}^4$ where every root of r corresponds to a mixture matching the first six moments. Because six moments suffice to identify the mixture by [KMV10], r has no roots in R outside $y = \alpha$. This lets us show that $r/(y-\alpha)^2$ has no roots over R, which for a compact R implies that $r/(y-\alpha)^2 = \Omega(1)$ over R. Thus $r = \Omega((y-\alpha)^2)$ over the region of interest.

Now, with $O(\sigma^{12}/\epsilon^2)$ samples we can estimate all the X_i to $\pm\epsilon$, which lets us estimate both $p_5(y)$ and $p_6(y)$ to $\pm O(\epsilon)$. This means $\sqrt{r(y)}$ is estimated to $\pm O(\epsilon)$. Since $\sqrt{r(y)} = \Omega(|y - \alpha|)$, this lets us find α to $\pm O(\epsilon)$. We then work back through our equations to get β and γ to $\pm O(\epsilon)$, which give the μ_i and σ_i^2 to $\pm O(\epsilon)$.

The analysis proceeds slightly differently in the setting where $|\sigma_1^2 - \sigma_2^2| \gg 1$. In this setting the region R of interest is not compact, because the parameter γ (which here equals $\sigma_1^2 - \sigma_2^2$) is unbounded. However, we can show directly that the highest (12th) degree coefficient of γ in $r/(y - \alpha)^2$ is bounded away from zero, getting that $r = \Omega(\gamma^{12}(y - \alpha)^2)$. Since the X_i are now not constant, while we can estimate each X_i to $\pm\epsilon$ with $O(\sigma^{12}/\epsilon^2)$ samples, we only estimate $p_5(y)$ and $p_6(y)$ to $\pm O(\epsilon\gamma^5)$. Since $\sqrt{r(y)} = \Omega(\gamma^6|y - \alpha|)$, this lets us estimate α to $\pm O(\epsilon/\gamma)$. This is sufficient to recover γ to $\pm O(\epsilon)$, which lets us recover β to $\pm O(\epsilon)$ and then the μ_i and σ_i^2 to $\pm O(\epsilon)$.

Dimension Reduction. In Section 4 we extend our theorem to arbitrary dimensional mixtures using two simple ideas. The first idea is used to reduce the d-dimensional case to the 4-dimensional case and is straightforward. The second argument reduces the 4-dimensional case to the 1-dimensional and is only slightly more involved. How can we use an algorithm for $d \leq 4$ to solve the problem in arbitrary dimension? Consider the case where $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ differ in some entry (i, j). We can find (i, j) by running our assumed algorithm for all pairs of variables. Each pair of variables leads to a two-dimensional mixture problem where the covariances are obtained by restricting Σ_1, Σ_2 to the corresponding entries. Once we have found an entry (i, j) where $|(\Sigma_1 - \Sigma_2)_{ij}| > \epsilon$, we are in good shape. We now iterate over all $k, l \in [d]$ and solve the 4-dimensional mixture problem on the variables (i, j, k, l) to within accuracy $\epsilon/10$. This not only reveals an additional entry k, l of the covariance matrix but it also tells us which of the two values for position (k, l) is associated with which of the values for position (i, j). This is because we solved the 4-dimensional problem to accuracy $\epsilon/10$ and we know that $|(\Sigma_1 - \Sigma_2)_{ij}| > \epsilon$. Hence, each newly recovered value for position (i, j) must be close to the value that we previously recovered. This ensures that we do not mix up any entries and so we recover the covariance matrices entry by entry. A similar but simpler argument works for the means.

Finally, the four-dimensional problem reduces to one dimension by brute forcing over an ϵ-net of all possible four-dimensional solutions (which is now doable in polynomial time) and using the algorithm for $d = 1$ to verify whether we picked a valid solution. The verification works by projecting the four-dimensional mixture in a random direction. Using anti-concentration results for quadratic forms in gaussian variables, we can show that any covariance matrix ϵ-far from the true covariance matrices will be ruled out with constant probability by each projection. Therefore $O(\log(1/\epsilon))$ projections will identify the covariance matrices among the poly$(1/\epsilon)$ possibilities. A union bound requires $\delta \approx 1/\log(1/\epsilon)$, giving $O(\log\log(1/\epsilon))$ overhead beyond the 1-dimensional

algorithm.

2 Lower bounds

2.1 Mixtures with matching moments are very close under Gaussian noise

Our main lemma shows that if we have two gaussian mixtures whose first k moments are matching and we add a gaussian random variable $N(0, \sigma^2)$ to each mixture, then the resulting distributions are $O(1/\sigma^{2k+2})$-close in squared Hellinger distance. The idea is illustrated in Figure 1.

Definition 2.1. *Let P, Q be probability distributions that are absolutely continuous with respect to the Lebesgue measure. Let p and q denote density functions of P and Q, respectively. Then, the* squared Hellinger distance *between P and Q is defined as*

$$\mathrm{H}^2(P, Q) = \frac{1}{2} \int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \mathrm{d}x \,.$$

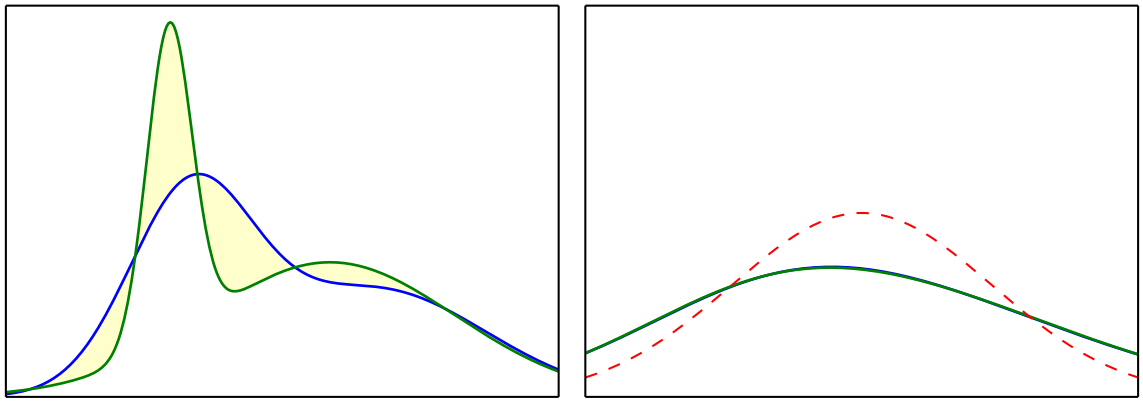

Figure 1: Mixtures with matching first 5 moments before and after adding $N(0, 2)$ (red).

Lemma 2.2. *Let F and G be distributions that are subgaussian with constant parameters and identical first k moments for $k = O(1)$. Let $P = F + N(0, \sigma^2)$ and $Q = G + N(0, \sigma^2)$ for $\sigma \gtrsim 1$. Then*

$$\mathrm{H}^2(P, Q) \lesssim 1/\sigma^{2k+2}.$$

Proof. We have that F and G are subgaussian with constant parameters, i.e., for any $d \geq 0$ we have

$$\left| \mathop{\mathbb{E}}_{t \sim F} t^d \right| \lesssim (O(\sqrt{d}))^d$$

and similarly for G. Denote by p, q, f, g density functions of P, Q, F, G, respectively. We would like to bound

$$\mathrm{H}^2(P, Q) = \frac{1}{2} \int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \mathrm{d}x \,. \tag{2}$$

We split the integral (2) into two regimes, $|x| \geq T$ and $|x| \leq T$ for $T \approx \sigma \sqrt{\log \sigma}$.

For the $|x| \geq T$ regime, we have

$$\begin{aligned}\frac{1}{2}\int_{|x|\geq T}\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 \mathrm{d}x &\leq \mathop{\mathbb{P}}_{x\sim P}\{|x|\geq T\} + Pr_{x\sim Q}\{|x|\geq T\} \\ &\lesssim e^{-T^2/(2(\sigma^2+O(1)))} \\ &\lesssim 1/\sigma^{2k+2}\,.\end{aligned}$$

The challenging part is the $|x| \leq T$ regime.

Claim 2.3. *For $|x| \leq T$, we have*

$$p(x) \gtrsim \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}. \tag{3}$$

Proof. Let x be such that $|x| \leq T$. Let t be such that $F([-t,t]) = 1/2$. Note that $t = O(1)$ since all the parameters of F are constant. In particular, denoting by $\nu(y)$ the density of $N(0,\sigma^2)$ we have for every $y \in [-t,t]$,

$$\nu(x-y) \geq \frac{e^{-(|x|+t)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \geq \frac{e^{-x^2/2\sigma^2}e^{-O(|x|/\sigma^2)}e^{-O(1/2\sigma^2)}}{\sqrt{2\pi}\sigma} \gtrsim \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}.$$

Hence,

$$p(x) \geq \int_{-t}^{t} f(x)\nu(x-y)\mathrm{d}y \geq \frac{1}{2}\cdot \min_{y\in[-t,t]}\nu(x-y) \gtrsim \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\,.$$

□

Now, we define

$$\Delta(x) = \frac{q(x)-p(x)}{p(x)}.$$

We have that

$$\begin{aligned}\Delta(x) &= \frac{1}{p(x)}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-t)^2}{2\sigma^2}}(g(t)-f(t))\mathrm{d}t \\ &= \frac{1}{p(x)}\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\int_{-\infty}^{\infty}e^{tx/\sigma^2}e^{-\frac{t^2}{2\sigma^2}}(g(t)-f(t))\mathrm{d}t \\ &\lesssim \int_{-\infty}^{\infty}e^{tx/\sigma^2}e^{-\frac{t^2}{2\sigma^2}}(g(t)-f(t))\mathrm{d}t. \end{aligned}\tag{4}$$

We take a power series expansion of the interior of the integral,

$$\begin{aligned}e^{tx/\sigma^2}e^{-\frac{t^2}{2\sigma^2}} &= \left(\sum_{d=0}^{\infty}\frac{(tx/\sigma^2)^d}{d!}\right)\left(\sum_{d=0}^{\infty}\frac{(-t^2/(2\sigma^2))^d}{d!}\right) \\ &= \sum_{d=0}^{\infty}\sum_{\substack{j\in\mathbb{Z}\\0\leq 2j\leq d}}\frac{(tx/\sigma^2)^{d-2j}}{(d-2j)!}\frac{(-t^2/(2\sigma^2))^j}{j!} \\ &= \sum_{d=0}^{\infty}(t/\sigma)^d\sum_{\substack{j\in\mathbb{Z}\\0\leq 2j\leq d}}\frac{(x/\sigma)^{d-2j}(-1/2)^j}{(d-2j)!j!}.\end{aligned}$$

Now, for all $j \in [0, d/2]$,

$$(d-2j)!j! \geq (d/2 - \sqrt{d})! = (\Omega(d))^{d/2-\sqrt{d}} = (\Omega(d))^{d/2}\,.$$

Therefore each term in the inner sum has magnitude bounded by $(x/\sigma)^{d-2j}(O(1/d))^{d/2}$, so the sum has magnitude bounded by $(1+(x/\sigma)^d)(O(1/d))^{d/2}$. Hence there exists a constant C and values $c_{x,d}$ with $|c_{x,d}| \leq 1$ such that

$$e^{tx/\sigma^2}e^{-\frac{t^2}{2\sigma^2}} = \sum_{d=0}^{\infty} c_{x,d}\left(\frac{C(1+x/\sigma)}{\sigma\sqrt{d}}\right)^d t^d.$$

Returning to (4), we have

$$\begin{aligned}
\Delta(x) &\lesssim \int_{-\infty}^{\infty}\sum_{d=0}^{\infty} c_{x,d}\left(\frac{C(1+x/\sigma)}{\sigma\sqrt{d}}\right)^d t^d(g(t)-f(t))\mathrm{d}t \\
&= \sum_{d=0}^{\infty} c_{x,d}\left(\frac{C(1+x/\sigma)}{\sigma\sqrt{d}}\right)^d \int_{-\infty}^{\infty} t^d(g(t)-f(t))\mathrm{d}t \\
&\leq \sum_{d=0}^{\infty} |c_{x,d}|\left(\frac{C(1+x/\sigma)}{\sigma\sqrt{d}}\right)^d \begin{cases} 0 & \text{if } d \leq k \\ (O(\sqrt{d}))^d & \text{otherwise.} \end{cases} \\
&\lesssim \sum_{d=k+1}^{\infty}\left(\frac{O(1)\cdot C(1+x/\sigma)}{\sigma}\right)^d \\
&\eqsim \left(\frac{1+x/\sigma}{\sigma}\right)^{k+1}.
\end{aligned}$$

for all $|x| \leq T \eqsim \sigma\log\sigma$. Note that this implies $|\Delta(x)| < 1$ which justifies the expansion

$$\sqrt{1+\Delta(x)} = 1 + \Delta(x)/2 \pm O(\Delta(x)^2).$$

Therefore, following the approach outlined in [Pol00], we can write

$$\begin{aligned}
\mathrm{H}^2(P,Q) &= 1 - \int_{-\infty}^{\infty}\sqrt{p(x)q(x)}\mathrm{d}x \\
&= 1 - O(1/\sigma^{2k+2}) - \int_{-T}^{T} p(x)\sqrt{1+\Delta(x)}\mathrm{d}x \\
&= 1 - O(1/\sigma^{2k+2}) - \int_{-T}^{T} p(x)(1+\Delta(x)/2 \pm O(\Delta(x)^2))\mathrm{d}x
\end{aligned}$$

Now, we have that

$$\begin{aligned}
\int_{-T}^{T} p(x)\left(1+\frac{1}{2}\Delta(x)\right)\mathrm{d}x &= \mathop{\mathbb{P}}_{x\sim P}[|x| \leq T] + \frac{1}{2}\left(\mathop{\mathbb{P}}_{x\sim Q}\{|x| \leq T\} - \mathop{\mathbb{P}}_{x\sim P}\{|x| \leq T]\}\right) \\
&= 1 - \frac{1}{2}\left(\mathop{\mathbb{P}}_{x\sim Q}\{|x| > T\} + \mathop{\mathbb{P}}_{x\sim P}\{|x| > T\}\right) \\
&= 1 - O(1/\sigma^{2k+2})
\end{aligned}$$

by our choice of T. We conclude,

$$\begin{aligned}
\mathrm{H}^2(P,Q) &\lesssim 1/\sigma^{2k+2} + \int_{-T}^{T} p(x)\left(\frac{1+x/\sigma}{\sigma}\right)^{2k+2}\mathrm{d}x \\
&\lesssim 1/\sigma^{2k+2}\left(1+\int_{-T}^{T} p(x)(x/\sigma)^{2k+2}\mathrm{d}x\right) \\
&\leq 1/\sigma^{2k+2}\left(1+\mathop{\mathbb{E}}_{x\sim P}(x/\sigma)^{2k+2}\right) \\
&\lesssim 1/\sigma^{2k+2}\,.
\end{aligned}$$

□

2.2 Lower bounds for mixtures of two Gaussians

Claim 2.4. *Let P and Q be distributions with $\mathrm{H}^2(P,Q) \leq \epsilon$. Then there exists a constant $c > 0$ such that $n = c\epsilon^{-1}\log(1/\delta)$ independent samples from P and Q have total variation distance less than $1-\delta$. In particular, we cannot distinguish the distributions from n samples with success probability greater than $1-\delta$.*

Proof. Let $x_1,\ldots,x_n \sim P$ and $y_1,\ldots,y_n \sim Q$ for $n = c\epsilon^{-1}\log(1/\delta)$. We will show that the total variation distance between $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$ is less than $1-\delta$.

We partition $[n]$ into k groups of size $1/(10\epsilon)$, for $k = 10c\log(1/\delta)$. Within each group, by sub-additivity of squared Hellinger distance we have that

$$\mathrm{H}^2\big((x_1,\ldots,x_{1/(10\epsilon)}),(y_1,\ldots,y_{1/(10\epsilon)})\big) \leq 1/10\,.$$

Appealing to the relation between total variation and Hellinger, this implies

$$\mathrm{TV}\big((x_1,\ldots,x_{1/(10\epsilon)}),(y_1,\ldots,y_{1/(10\epsilon)})\big) \leq 2\mathrm{H}\big((x_1,\ldots,x_{1/(10\epsilon)}),(y_1,\ldots,y_{1/(10\epsilon)})\big) \leq \frac{2}{3}\,.$$

Hence we may sample $(x_1,\ldots,x_{1/(10\epsilon)})$ and $(y_1,\ldots,y_{1/(10\epsilon)})$ in such a way that the two are identical with probability at least $1/3$. If we do this for all k groups, we have that $(x_1,\ldots,x_n) = (y_1,\ldots,y_n)$ with probability at least $1/3^k > 2\delta$ for sufficiently small constant c. □

Theorem 2.5. *Consider any algorithm that, given n samples of any gaussian mixture with variance σ^2, with probability $1-\delta$ learns* either *μ_i to $\pm\epsilon\sigma$ or σ_i^2 to $\pm\epsilon^2\sigma^2$. Then $n = \Omega(\epsilon^{-12}\log(1/\delta))$.*

Proof. Take any two gaussian mixtures F and G with constant parameters such that the four means and variances are all $\Omega(1)$ different from each other, but F and G match in the first five moments. One can find such mixtures by taking almost any mixture F with constant parameters and solving p_5 to find another root and the corresponding mixture (per Lemma B.1, this will cause the first five moments to match). We can find such an F and G in [Pea94], or alternatively take

$$\begin{aligned}
F &= \frac{1}{2}N(-1,1) + \frac{1}{2}N(1,2) \\
G &\approx 0.2968N(-1.2257,0.6100) + 0.7032N(0.5173,2.3960).
\end{aligned} \tag{5}$$

While G is expressed numerically, one can certainly prove that the p_5 derived from F has a second root that yields something close to this mixture. Plug the mixtures into Lemma 2.2. We get that

for any $\sigma > 0$, the mixtures

$$P = \frac{1}{2}N(-1, 1+\sigma^2) + \frac{1}{2}N(1, 2+\sigma^2)$$
$$Q \approx 0.2968N(-1.2257, 0.6100+\sigma^2) + 0.7032N(0.5173, 2.3960+\sigma^2).$$

have

$$\mathrm{H}^2(P, Q) \lesssim 1/\sigma^{12}. \tag{6}$$

Since by Claim 2.4 we cannot differentiate P and Q with $o(\sigma^{12}\log(1/\delta))$ samples, it requires $\Omega(\sigma^{12}\log(1/\delta))$ samples to learn either the μ_i or the σ_i^2 to $\pm 1/10$ with $1-\delta$ probability. Set $\sigma = 1/(10\epsilon)$ to get the result. □

Our argument extends to d dimensions. We gain a $\log(d)$ factor in our lower bound by randomly planting a hard mixture learning problem in each of the d coordinates.

Claim 2.6. *Let P and Q be distributions with $\mathrm{H}^2(P, Q) \le \epsilon$. Let $R_i \in \{P, Q\}$ uniformly at random for $i \in \{1, \ldots, d\}$. Then there exists a constant $c > 0$ such that given $n = c\epsilon^{-1}\log(d/\delta)$ no algorithm can identify all R_i with probability $1-\delta$.*

Proof. As in Claim 2.4, we have that the total variation distance between $B = 1/(10\epsilon)$ samples from P and B samples from Q is less than $2/3$.

Partition our samples into $k = 10c\log d$ groups $x^1, \ldots, x^k$, where for each group $j \in [k]$ and coordinate $i \in [d]$ we have $x_i^j \sim R_i^{\otimes B}$. By the total variation bound between $P^{\otimes B}$ and $Q^{\otimes B}$, we could instead draw x_i^j from a distribution independent of R_i with probability $1/3$ and a distribution dependent on R_i with probability $2/3$. Suppose we do this.

Then for any coordinate i, with probability $3^{-k} > \delta/d$ all of $x_i^1, \ldots, x_i^k$ are independent of R_i. Since the coordinates are independent, this means that with probability at least

$$1 - (1-\delta/d)^d \ge \delta/4$$

there will exist a coordinate i such that all of $x_i^1, \ldots, x_i^k$ are independent of R_i. The algorithm must then guess R_i incorrectly with probability at least $1/2$, for a $\delta/8$ probability of failure overall. Rescale δ to get the result. □

This immediately gives that Theorem 2.5 can be extended to d dimensions:

Theorem 2.7. *Consider any algorithm that, given n samples of any d-dimensional gaussian mixture F with $\mathbb{V}(F) = \sigma^2$, with probability $1-\delta$ for all $i \in [d]$ learns either μ_i to $\pm\epsilon\sigma$ or $\Sigma_{i,i}$ to $\pm\epsilon^2\sigma^2$. Then $n = \Omega(\epsilon^{-12}\log(d/\delta))$.*

Proof. Let P, Q be as in Theorem 2.5. We choose a mixture F to have independent coordinates, each of which is uniformly chosen from $\{P, Q\}$. Then $\mathbb{V}(F) \eqsim 1/\epsilon$, $H(P, Q) \le \epsilon^6$, and learning the parameters of the mixture in the ith coordinate to the specified precision would identify whether it is P or Q. Claim 2.6 gives the result. □

2.3 Lower bound for mixtures of k gaussians

We can extend these lower bounds to mixtures of k gaussians. The main issue is that for $k = 2$ we know that two explicit well-separated mixtures exist that match on 5 moments, by solving the method of moments on a random input and getting two solutions. For general k we would like to show the existence of two well-separated mixtures that match on nearly $3k$ moments.

We will formalize the following intuition. With k gaussians there are $3k - 1$ free parameters–means, variances, and probabilities subject to the probabilities summing to one. Therefore if we only take $3k - 2$ moments, we are embedding a high dimensional space into a lower dimensional space, and expect lots of collisions. Therefore some pair that collide should be well separated.

First we show a few lemmas that will be useful. We use the following fact, shown in the appendix:

Lemma A.8. *Let $p(x) = p(x_1, \ldots, x_n)$ be a multivariate polynomial of degree d and smallest nonzero coefficient of magnitude a. Then*

$$\mathop{\mathbb{P}}_{x \sim N(0, I_n)}[|p(x)| < a\epsilon] < d\epsilon^{1/d}.$$

This lets us show the following:

Lemma 2.8. *Let $k \geq 1$ be a constant. There exists a set of k gaussians $N(\mu_i, \sigma_i^2)$ such that all the μ_i and σ_i are $O(1)$, all μ_i and σ_i are $\Omega(1)$ far from the others and from zero, and the matrix $A \in \mathbb{R}^{m \times k}$ of moments, given by*

$$A_{j,i} = \mathop{\mathbb{E}}_{x \sim N(\mu_i, \sigma_i^2)} x^j,$$

has minimum singular value $\Omega(1)$ for any $m \geq k$.

Proof. We will show this to be true with good probability for a randomly drawn set of k gaussians. We consider drawing the gaussians randomly, so $\mu_i \sim N(0, 1)$ and $\sigma_i \sim N(0, 1)^2$. This immediately gives the first two properties with probability arbitrarily close to 1, and we just need to bound the minimum singular value of A. We can assume without loss of generality that $m = k$, since this minimizes the singular values. Then, since A has $O(1)$ dimension and coefficients, it suffices to show that the determinant of A—which is the product of the eigenvalues—is $\Omega(1)$ with positive probability.

Now, consider the determinant of A as a formal polynomial in the μ_i and $\sqrt{\sigma_i}$. We know that this is nonzero, because (for example) the $\mu_1^1 \mu_2^2 \mu_3^3 \cdots \mu_k^k$ monomial appears only from the diagonal term. It is a fixed monomial, so its minimum coefficient is some constant. Hence Lemma A.8 shows for $\mu_i, \sqrt{\sigma_i} \sim N(0, 1)$ that the determinant will be $\Omega(1)$ with probability arbitrarily close to 1. In such cases, the minimum singular value is $\Omega(1)$ as well, giving the result. □

Lemma 2.9. *Let $k \geq 1$ be a constant. There exist two mixtures F and G of k gaussians each that match on the first $3k-2$ moments, for which all the parameters (p_i, μ_i, σ_i) are bounded by $O(1)$ for each mixture, and for which one of the mixtures has either a μ_i or a σ_i that is $\Omega(1)$ far from any μ_i or σ_i in the other mixture.*

Proof. There are $t = 3k - 1$ "free" parameters in a mixture of k gaussians: the means, variances, and relative probabilities subject to the sum of probabilities equalling 1. With t parameters, we expect there to be *lots* of mixtures that match on $t - 1$ moments.

Formally, consider the (μ_i, σ_i) given by Lemma 2.8. For the matrix of moments A as given by the lemma, and for any vector $p \in \mathbb{R}^k$ of probabilities for each gaussian, the first m moments of the mixture (p_i, μ_i, σ_i) are precisely Ap. Let $p_i = 1/k$ for all i.

Now, let $0 < \alpha < \beta \leq \frac{1}{k^2}$ be small constants to be determined later. For any set of free parameters $z \in \mathbb{R}^t$, consider a mixture of gaussians $f(z) = \{(p'_i, \mu'_i, \sigma'_i)\}$ given by $\mu'_i = \mu_i + \alpha z_i$, $\sigma'_i = \sigma_i + \alpha z_{k+i}$, and $p'_i = p_i + \beta z_{2k+i}$ for $i \in [k-1]$ and $p'_k = p_k - \sum_{i\in[k-1]} \beta z_{2k+i}$. For every z with $\|z\|_2 = 1$, this is a valid mixture of k gaussians (because β is small enough that $\sum_{i\in[k-1]} \beta z_{2k+i} < p_k$). For such a z, define $A(z) \in \mathbb{R}^{m\times k}$ to be the matrix of component moments, $A(z)_{j,i} = \mathbb{E}_{x\sim N(\mu'_i,\sigma'_i)} x^j$, and $p(z) \in \mathbb{R}^k$ to be the vector of probabilities p'. Then the first m moments of the mixture $f(z)$ are $M(z) = A(z)p(z)$.

Note that, for any $\alpha, \|z\| \leq 1$, we have $\|A(z) - A(0)\| \lesssim \alpha\|z\|$: every monomial in the matrix $A(z) - A(0)$ includes α times a coordinate of z in it, and there are only a constant number of monomials.

Because $M(z)$ is a polynomial in z, it is continuous. If $m = t-1 = 3k-2$, then the Borsuk-Ulam theorem states that there exist two antipodal points $z, -z$ with $\|z\|_2 = 1$ such that $M(z) = M(-z)$. This immediately gives two different mixtures that match on the first m moments. This is almost what we want, but we also need the mixtures to be different over σ or μ, not just p.

We have that $A(z)p(z) = A(-z)p(-z)$, or

$$A(0)(p(z) - p(-z)) = (A(-z) - A(0))p(-z) - (A(z) - A(0))p(z)$$

For $\sigma_{min} = \Omega(1)$ being the minimal singular value of $A(0)$, we have

$$\begin{aligned}
\sigma_{min}\|p(z) - p(-z)\|_2 &\leq \|A(0)(p(z) - p(-z))\| = \|(A(-z) - A(0))p(-z) - (A(z) - A(0))p(z)\| \\
\|p(z) - p(-z)\|_2 &\leq \frac{1}{\sigma_{min}}(\|A(-z) - A(0)\| + \|A(z) - A(0)\|)(\|p(-z)\| + \|p(z)\|) \\
&\leq \frac{1}{\Omega(1)} \cdot O(\alpha) \cdot 2 \lesssim \alpha.
\end{aligned}$$

On the other hand, $\|p(z) - p(-z)\|_2 \geq \beta\sqrt{\sum_{i>2k} z_i^2}$, so

$$\sqrt{\sum_{i>2k} z_i^2} \lesssim \frac{\alpha}{\beta}.$$

If we set β to some value (e.g. $1/k^2$), then if we choose α as a small enough constant we will have $\sqrt{\sum_{i>2k} z_i^2} \leq 1/2$. Since $\|z\|_2 = 1$, this means $\sqrt{\sum_{i\leq 2k} z_i^2} \geq \sqrt{3/4} \gtrsim 1$. Therefore the mixtures $f(z)$ and $f(-z)$ have at least one of their μ_i or σ_i perturbed by $\Theta(\alpha)$. For small enough constant α, the perturbations will be much less than the $\Omega(1)$ gap between all the different μ_i and σ_i in the mixture $f(0)$. Therefore the perturbed μ_i or σ_i in $f(z)$ is $\Omega(1)$ far from any corresponding μ_i or σ_i in $f(-z)$. So $f(z)$ and $f(-z)$ give the desired mixtures. □

The mixtures given by Lemma 2.9 let us extend the lower bounds to mixtures of k gaussians, but with a caveat: the mixtures differ in *at least one* of μ_i and σ_i, so the lower bound now only applies to algorithms that recover *both* μ_i and σ_i do the desired precision.

Theorem 2.10. *Consider any algorithm that, given n samples of any one dimensional gaussian mixture F of $k = O(1)$ components, with probability $1 - \delta$ learns both the μ_i to $\pm\epsilon\sigma$ and the σ_i^2 to $\pm\epsilon^2\sigma^2$. Then $n = \Omega_k(\epsilon^{2-6k}\log(1/\delta))$.*

Theorem 2.11. *Consider any algorithm that, given n samples of any d-dimensional gaussian mixture F of $k = O(1)$ components with $\mathbb{V}(F) = \sigma^2$, with probability $1 - \delta$ for all $i \in [d]$ learns both μ_i to $\pm\epsilon\sigma$ and $\Sigma_{i,i}$ to $\pm\epsilon^2\sigma^2$. Then $n = \Omega_k(\epsilon^{2-6k}\log(d/\delta))$.*

Proof. Identical to the proof of Theorems 2.5 and 2.7, but applying Lemma 2.2 to the F, G resulting from Lemma 2.9 rather than the explicit ones in Equation (5). □

3 Algorithm for one-dimensional mixtures

3.1 Preliminaries and Notation

Asymptotics. For any expressions f and g, we use $f \lesssim g$ to denote that there exists a constant $C > 0$ such that $f \leq Cg$. Similarly, $f \gtrsim g$ denotes that $g \lesssim f$, and $f \eqsim g$ denotes that $f \lesssim g \lesssim f$.

Parameters of the gaussian. The two gaussians have probabilities p_i, means μ_i, and standard deviations σ_i. The overall mean and variance of the mixture are $\mu = p_1\mu_1 + p_2\mu_2$ and

$$\sigma^2 = p_1((\mu_1 - \mu)^2 + \sigma_1^2) + p_2((\mu_2 - \mu)^2 + \sigma_2^2) = p_1p_2(\mu_1 - \mu_2)^2 + p_1\sigma_1^2 + p_2\sigma_2^2\,. \tag{7}$$

For almost all of the section, for simplicity of notation we will assume that the overall mean $\mu = 0$. We only need to consider $\mu \neq 0$ when showing that we can estimate the moments precisely enough.

We will also assume that $p_1, p_2 \in (0, 1)$ are both bounded away from zero. We define

$$\Delta_\mu = |\mu_2 - \mu_1| \qquad \Delta_{\sigma^2} = |\sigma_2^2 - \sigma_1^2|\,. \tag{8}$$

We also make use of a reparameterization of the gaussian distribution:

$$\alpha = -\mu_1\mu_2 \qquad \beta = \mu_1 + \mu_2 \qquad \gamma = \frac{\sigma_2^2 - \sigma_1^2}{\mu_2 - \mu_1} \tag{9}$$

Note that these are independent of adding gaussian noise, i.e. increasing both σ_1^2 and σ_2^2 by the same amount. Also we have $\alpha \geq 0$, since the mean is zero. With our assumption that p_1, p_2 are bounded away from zero we have that

$$\beta^2 \lesssim \alpha \eqsim \Delta_\mu^2.$$

Finally, we will make use of the parameter

$$\kappa = \max(1, \Delta_{\sigma^2}/\Delta_\mu^2)$$

which will relate to how well-conditioned our equations are. We have that $|\gamma| \lesssim \Delta_\mu\kappa$.

Excess moments. We define M_i to be the ith moment of our distribution, $\mathbb{E}[x^i]$, so $M_2 = \sigma^2$.

The *excess kurtosis* of a distribution is a standard statistical measure defined as $M_4/M_2^2 - 3$. It is designed to be independent of adding independent gaussian noise to the variable. Inspired by this, we define the *excess moments* X_i to be M_i plus a polynomial in $M_2, \ldots, M_{i-1}$ such that the result is independent of adding gaussian noise. We have that:

Lemma 3.1. *For $\alpha = -\mu_1\mu_2, \beta = \mu_1 + \mu_2, \gamma = \frac{\sigma_2^2 - \sigma_1^2}{\mu_2 - \mu_1}$ we have that*

$$\begin{aligned}
X_3 &:= M_3 && = \alpha\beta + 3\alpha\gamma \\
X_4 &:= M_4 - 3M_2^2 && = -2\alpha^2 + \alpha\beta^2 + 6\alpha\beta\gamma + 3\alpha\gamma^2 \\
X_5 &:= M_5 - 10M_3M_2 && = \alpha(\beta^3 - 8\alpha\beta + 10\beta^2\gamma + 15\gamma^2\beta - 20\alpha\gamma) \\
X_6 &:= M_6 - 15M_4M_2 + 30M_2^3 && = \alpha(16\alpha^2 - 12\alpha\beta^2 - 60\alpha\beta\gamma + \beta^4 + 15\beta^3\gamma + 45\beta^2\gamma^2 + 15\beta\gamma^3)
\end{aligned} \tag{10}$$

See Appendix B.2 for proof. Since α, β, γ are independent of adding gaussian noise, these definitions of the X_i are correct.

By inspection, we have for each $i \in \{3, 4, 5, 6\}$ that

$$|X_i| \lesssim \Delta_\mu^i \kappa^{i-2}. \tag{11}$$

For simplicity of notation, we also define $X_1 = \mu$ and $X_2 = \sigma^2$, despite them not technically being "excess," and refer to $\{X_1, X_2, X_3, X_4, X_5, X_6\}$ as the excess moments.

Estimation of moments. All our algorithms in this section proceed by first estimating the (excess) moments from the samples, then estimating the mixture from these moments. The relationship between sample complexity and estimation error is as follows:

Lemma 3.2. *Suppose p_1, p_2 are bounded away from zero and our mixture has variance σ^2. Given $O(\log(1/\delta)/\epsilon^2)$ samples, with probability $1-\delta$ we can compute estimates $\widehat{X}_i$ of the first $O(1)$ excess moments X_i satisfying $|\widehat{X}_i - X_i| \le \epsilon\sigma^i$.*

See Appendix A.2 for a proof. We will state our first two theorems in terms of the necessary error bound on $|\widehat{X}_i - X_i|$ rather than sample complexity. This is more general, since it supports other forms of perturbation of the inputs.

For a statistic f of the gaussian mixture, in general we use f to denote the true value of the statistic and $\widehat{f}$ to denote the estimate of f from estimates of the moments.

3.2 Algorithm overview

Our overall goal is to recover the μ_i to $\pm\epsilon\sigma$ and the σ_i^2 to $\pm\epsilon^2\sigma^2$ using roughly $O(1/\epsilon^{12})$ samples.

We have two different algorithms for different parameter regimes. The first algorithm, Algorithm 3.1, proceeds by first learning the μ_i, then using this to estimate the σ_i. However, it only works well if we have the X_i to within Δ_μ^i; without this, we cannot get a nontrivial estimate of the μ_i, which causes the algorithm to also not get a decent estimate of the σ_i.

Theorem 3.7 (Precision better than Δ_μ)**.** *Consider any mixture of gaussians where p_1 and p_2 are bounded away from zero, $c > 0$ a sufficiently small constant, and any $\epsilon < 1$. If $|\widehat{X}_i - X_i| \le c\epsilon\Delta_\mu^i$ for all $i \le 6$, Algorithm 3.1 recovers the p_i to $\pm\epsilon$, μ_i to $\pm\epsilon\Delta_\mu$, and σ_i^2 to $\pm\epsilon\Delta_\mu^2$.*

If $\Delta_{\sigma^2} \gg \Delta_\mu^2$, for example if $\Delta_\mu = 0$, one may hope to get a good estimate of the σ_i^2 despite having more than Δ_μ^i error in X_i. Algorithm 3.2 does this by solving for the σ_i under the assumption that $\mu_1 = \mu_2$. We can show that the solution is robust to Δ_μ being small but nonzero, so the algorithm does a good job when $\Delta_\mu^i \lesssim |\widehat{X}_i - X_i| \lesssim \Delta_{\sigma^2}^{i/2}$. (When $|\widehat{X}_i - X_i|$ goes below this bound, the performance doesn't degrade but does not improve as one would like.)

Theorem 3.9 (Precision between Δ_μ and $\sqrt{\Delta}_{\sigma^2}$)**.** *Consider any mixture of gaussians where p_1 and p_2 are bounded away from zero, and any $\epsilon < 1$. Suppose that $\epsilon\Delta_{\sigma^2} \gtrsim \Delta_\mu^2$. If $|\widehat{X}_i - X_i| \le \epsilon\Delta_{\sigma^2}^{i/2}$ for all $i \le 6$, Algorithm 3.2 recovers the p_i to $\pm O(\epsilon)$ and σ_i^2 to $\pm O(\epsilon)\Delta_{\sigma^2}$.*

In the remaining parameter regime, with $|\widehat{X}_i - X_i| > \Delta_\mu^i + \Delta_{\sigma^2}^{i/2}$, the two gaussians are in general indistinguishable and it suffices to just output the average mean and variance.

To get a general result, we just need to figure out which of the three parameter regimes we're in and apply the appropriate algorithm. Algorithm 3.3 does this by constructing sufficiently good estimates of σ, Δ_μ, and Δ_{σ^2}. We also invoke Lemma 3.2 to get bounds on sample complexity, showing:

Theorem 3.10. *Let F be any mixture of two gaussians with variance σ^2 and p_1, p_2 bounded away from 0. Then, given $O(\epsilon^{-2} n \log(1/\delta))$ samples Algorithm 3.3 with probability $1-\delta$ outputs the parameters of a mixture $\widehat{F}$ so that for some permutation π and all $i \in \{1,2\}$ we have the following guarantees:*

- *If $n \geq \left(\frac{\sigma^2}{|\mu_1-\mu_2|^2}\right)^6$, then $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq \epsilon|\mu_1-\mu_2|$, $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\mu_1-\mu_2|^2$, and $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon$.*
- *If $n \geq \left(\frac{\sigma^2}{|\sigma_1^2-\sigma_2^2|}\right)^6$, then $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\sigma_1^2-\sigma_2^2| + |\mu_1-\mu_2|^2$ and $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon + \frac{|\mu_1-\mu_2|^2}{|\sigma_1^2-\sigma_2^2|}$.*
- *For any $n \geq 1$, the algorithm performs as well as assuming the mixture is a single gaussian: $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq |\mu_1-\mu_2| + \epsilon\sigma$ and $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq |\sigma_1^2-\sigma_2^2| + |\mu_1-\mu_2|^2 + \epsilon\sigma^2$.*

The regimes can be unified to get the following, simpler but weaker, corollary:

Corollary 3.3. *Consider any mixture of gaussians where p_1 and p_2 are bounded away from zero, and any $\epsilon, \delta < 1$. With $n = O(\log(1/\delta)/\epsilon^{12})$ samples, Algorithm 3.3 returns $\{\mu_1, \mu_2\}$ to $\pm\epsilon\sigma$ additive error and the corresponding $\{\sigma_1^2, \sigma_2^2\}$ to $\pm\epsilon^2\sigma^2$ additive error with probability $1-\delta$.*

Proof. If $\epsilon\sigma = f\Delta_\mu$ for $f < 1$, then by the first clause of Theorem 3.10 with $\epsilon' = f^6$ we get the μ_i to within $f^6\Delta_\mu \leq f\Delta_\mu = \epsilon\sigma$ and the σ_i^2 to within $f^6\Delta_\mu^2 \leq f^2\Delta_\mu^2 = \epsilon^2\sigma^2$.

Otherwise, if $\epsilon\sigma > \Delta_\mu$ but $\epsilon^2\sigma^2 = f^2\Delta_{\sigma^2}$ for $f^2 < 1$, then by the second clause of Theorem 3.10 with $\epsilon' = f^6$ we get the σ_i to within $f^6\Delta_{\sigma^2} + \Delta_\mu^2 \lesssim \epsilon^2\sigma^2$.

And by the last clause, if $\epsilon\sigma \gtrsim \Delta_\mu$ we get the μ_i to within $\Delta_\mu + \epsilon^6\sigma \lesssim \epsilon\sigma$, and if $\epsilon^2\sigma^2 \gtrsim \max(\Delta_{\sigma^2}, \Delta_\mu^2)$ we get the σ_i^2 to within $\Delta_\mu^2 + \Delta_{\sigma^2} + \epsilon^6\sigma^2 \lesssim \epsilon^2\sigma^2$. □

3.3 Algorithm for better precision than Δ_μ

In this section we derive Pearson's polynomial and extend it into a robust algorithm.

Manipulation of X_3, X_4, X_5. Based on (10) we can remove β to get an equation in $\{\alpha, \gamma, X_3, X_4\}$:

$$6(\alpha\gamma)^2 = X_3^2 - 2\alpha^3 - X_4\alpha. \tag{12}$$

If we define $z := \alpha\gamma$, we can get another equation in $\{\alpha, z, X_3, X_4, X_5\}$.

$$\begin{aligned}\alpha^2 X_5 - X_3^3 + 8\alpha^3 X_3 &= \alpha^3\beta^2\gamma - 12\alpha^3\gamma^2\beta + 4\alpha^4\gamma - 27\alpha^3\gamma^3 \\ &= z(X_3 - 3z)^2 - 12z^2(X_3 - 3z) + 4\alpha^3 z - 27z^3 \\ &= 18z^3 - 18X_3 z^2 + (4\alpha^3 + X_3^2)z\end{aligned}$$

Substituting in (12) we make the equation linear in z:

$$\begin{aligned}\alpha^2 X_5 - X_3^3 + 8\alpha^3 X_3 &= 3(z - X_3)(X_3^2 - 2\alpha^3 - X_4\alpha) + (4\alpha^3 + X_3^2)z \\ &= (4X_3^2 - 2\alpha^3 - 3X_4\alpha)z + (3X_3X_4\alpha + 6\alpha^3 X_3 - 3X_3^3)\end{aligned}$$

or

$$z := \alpha\gamma = \frac{\alpha^2 X_5 + 2X_3^3 + 2\alpha^3 X_3 - 3X_3X_4\alpha}{4X_3^2 - 2\alpha^3 - 3X_4\alpha}. \tag{13}$$

▷ Recover gaussian mixture from (estimates of the) mean, variance, and excess moments

▷ Excess moments are function of moments, defined in (10)

procedure RecoverFromMoments($\epsilon, \mu, \sigma^2, X_3, X_4, X_5, X_6$)

$\alpha \leftarrow$ RecoverAlphaFromMoments($X_3, X_4, X_5, X_6, \epsilon$).

$\gamma \leftarrow \frac{1}{\alpha}\frac{\alpha^2 X_5 + 2X_3^3 + 2\alpha^3 X_3 - 3X_3 X_4 \alpha}{4X_3^2 - 2\alpha^3 - 3X_4\alpha}$.

$\beta \leftarrow \frac{1}{\alpha}(X_3 - 3\alpha\gamma)$.

$\mu_1, \mu_2 \leftarrow \frac{-\beta \mp \sqrt{\beta^2 + 4\alpha}}{2}$.

$p_1, p_2 \leftarrow \frac{\mu_2}{\mu_2 - \mu_1}, \frac{-\mu_1}{\mu_2 - \mu_1}$.

$\sigma_1^2 \leftarrow \sigma^2 - \big(p_1\mu_1^2 + p_2\mu_2^2 - \mu_1\gamma\big)$.

$\sigma_2^2 \leftarrow \sigma_1^2 + (\mu_2 - \mu_1)\gamma$.

return $(p_1, \mu_1 + \mu, \sigma_1), (p_2, \mu_2 + \mu, \sigma_2)$

end procedure

procedure RecoverAlphaFromMoments($X_3, X_4, X_5, X_6, \epsilon$)

Let y_{max} be the largest root of $2y^3 + X_4 y - X_3^2 = 0$.

$\kappa \leftarrow 1 + \sqrt{|X_4|}/y_{max}$ ▷ $\kappa = \Theta(1 + |\gamma|/\sqrt{\alpha})$

Define the 18th degree polynomial $r(y) = p_5(y)^2 + p_6(y)^2$ for p_5 and p_6 given by (14) and (16).

▷ For the true moments, α is the only zero of r on $(0, y_{max}]$.

Compute the set of roots R of $r'(y)$.

Let α be maximal element of $R \cup \{(1 + \epsilon/\kappa)y_{max}\}$ satisfying

$$\alpha \le (1 + \epsilon/\kappa)y_{max}$$
$$r(\alpha) \le \epsilon^2 \alpha^{18} \kappa^{10}.$$

return α.

end procedure

Algorithm 3.1: Algorithm for recovery of mixture of two gaussians in one dimension when the means are separated.

We can substitute z back into (12) and clear the denominator to get a polynomial equation in the single variable α:[2]

$$p_5(\alpha) = 6(2X_3\alpha^3 + X_5\alpha^2 - 3X_3X_4\alpha + 2X_3^3)^2 + (2\alpha^3 + 3X_4\alpha - 4X_3^2)^2(2\alpha^3 + X_4\alpha - X_3^2) = 0. \tag{14}$$

Therefore, given the excess moments X_i, we can find a set of candidate α by solving for the positive roots of $p_5(y) = 0$. Unfortunately, there are in general multiple such roots. In fact, the first five moments do not suffice to uniquely identify a gaussian mixture, so we must incorporate the sixth moment.

Using the 6th moment. Analogously to the creation of (13), we take the expression for X_6 in (10), replace β with $X_4/\alpha - 3\gamma$, then remove γ^2 terms using (12) to get

$$z := \alpha\gamma = \frac{4X_3^4 - 4X_3^2X_4\alpha - 8X_3^2\alpha^3 - X_4^2\alpha^2 + 8X_4\alpha^4 + X_6\alpha^3 + 4\alpha^6}{10X_3^3 - 7X_3X_4\alpha - 2X_3\alpha^3} \tag{15}$$

(See Appendix B.3 for a more detailed explanation.) Combining with (13) and clearing the denominators gives that

$$\begin{aligned} p_6(\alpha) = (4X_3^2 - 3X_4\alpha - 2\alpha^3)(4X_3^4 - 4X_3^2X_4\alpha - 8X_3^2\alpha^3 - X_4^2\alpha^2 + 8X_4\alpha^4 + X_6\alpha^3 + 4\alpha^6) - \\ (10X_3^3 - 7X_3X_4\alpha - 2X_3\alpha^3)(2X_3^3 - 3X_3X_4\alpha + 2X_3\alpha^3 + X_5\alpha^2) = 0 \end{aligned} \tag{16}$$

which is another 9th degree equation in α in terms of the excess moments.

We would like to say that α is the only common positive root of p_5 and p_6, but this is not always true. Fortunately, we can exclude the other common roots if we enforce an upper bound on α.

Restricting the domain. Let y_{max} be the positive root of

$$2y^3 + X_4y - X_3^2 = 0. \tag{17}$$

There is at most one such root by Descartes' rule of signs. There exists such a root because if $X_3 = \alpha(\beta + 3\gamma)$ is zero, then $X_4 = \alpha(\beta + 3\gamma)^2 - 6\alpha\gamma^2 - 2\alpha^2$ is negative. And by (12), $\alpha \le y_{max}$.

Moreover,

$$y_{max} \lesssim \alpha. \tag{18}$$

(Since p_1 and p_2 are bounded away from zero, $\beta^2 \lesssim \alpha$. Then if $\gamma^2 \lesssim \alpha$, this follows from a cubic polynomial with bounded coefficients having bounded roots. Otherwise, $X_3^2 = \Theta(\alpha^2\gamma^2)$ and $X_4 = \Theta(\alpha\gamma^2)$ are positive so $y_{max} \le X_3^2/X_4 \lesssim \alpha$.)

Combining the equations. We will show that $y = \alpha$ is the only solution to the set of equations $0 < y \le y_{max}, p_5(y) = 0, p_6(y) = 0$. This statement would suffice to recover the mixture given the exact excess moments, but we also want the algorithm to have robustness to small perturbations in the X_i. We therefore define

$$r(y) := p_5(y)^2 + p_6(y)^2, \tag{19}$$

[2]This polynomial (14) is identical to (29) in Pearson's 1894 paper, up to rescaling variables by constant factors. Our (13) is similarly identical to Pearson's (27).

which we know is zero at α. We will show it is significantly non-zero for any candidate y that is far from α and still within $[cy_{max}, y_{max}]$ for any constant $c > 0$.

The robustness will depend on the parameter

$$\kappa := \max(1, \Delta_{\sigma^2}/\Delta_\mu^2) \eqsim 1 + |\gamma|/\sqrt{\alpha}. \tag{20}$$

This is intuitive because the excess moments X_i are bounded by $O(\Delta_\mu^i \kappa^{i-2})$, which implies by inspection of (14) and (16) that for all $|y| \lesssim y_{max}$, every monomial in p_5 and p_6 has magnitude bounded by

$$O(\Delta_\mu^{18} \kappa^6). \tag{21}$$

At this point, it is convenient to normalize so $\Delta_\mu \eqsim \alpha \eqsim 1$. While we state our lemmas in full generality, it is better to think about this normalization and we will use it in the proofs.

Lemma 3.4. *For any constant $c > 0$, and for all $\alpha \geq 0$ and $\beta, \gamma, y \in \mathbb{R}$ with $cy_{max} \leq y \leq y_{max}$ and $\beta^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim \kappa^{12}(y - \alpha)^2 \alpha^{16}$$

This is the key lemma of our proof, and shown in Section B.4. Note that the recovery algorithm will not know y_{max} exactly, so we need to extend the claim to slightly beyond y_{max}.

Lemma 3.5. *For any constant $c > 0$, there exists a constant $c' > 0$ such that for all $\alpha \geq 0$ and $\beta, \gamma, y \in \mathbb{R}$ with $cy_{max} \leq y \leq y_{max} + c'(y_{max} - \alpha)$ and $\beta^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim \kappa^{12}(y - \alpha)^2 \alpha^{16}$$

The proof is in Section B.4. This lemma lets us show that RecoverAlphaFromMoments returns a good approximation to α if it is given good approximations to the moments:

Lemma 3.6. *Suppose p_1, p_2 are bounded away from zero, let $c > 0$ be a sufficiently small constant, and let $\epsilon < 1$. Suppose further that $|\widehat{X}_i - X_i| \leq c\epsilon\Delta_\mu^i$ for all $i \in \{3, 4, 5, 6\}$. In this setting, the result $\widehat{\alpha} =$ RecoverAlphaFromMoments$(\widehat{X}_3, \widehat{X}_4, \widehat{X}_5, \widehat{X}_6, \epsilon)$ satisfies*

$$|\widehat{\alpha} - \alpha| \lesssim \epsilon\Delta_\mu^2/\kappa.$$

See Section B.6 for a proof. It is then easy to show that all the recovered parameters are good approximations to the true parameters, getting the theorem:

Theorem 3.7 (Precision better than Δ_μ)**.** *Consider any mixture of gaussians where p_1 and p_2 are bounded away from zero, $c > 0$ a sufficiently small constant, and any $\epsilon < 1$. If $|\widehat{X}_i - X_i| \leq c\epsilon\Delta_\mu^i$ for all $i \leq 6$, Algorithm 3.1 recovers the p_i to $\pm\epsilon$, μ_i to $\pm\epsilon\Delta_\mu$, and σ_i^2 to $\pm\epsilon\Delta_\mu^2$.*

Proof. We normalize so $\Delta_\mu = 1$. By Lemma 3.6,

$$|\widehat{\alpha} - \alpha| \lesssim \epsilon/\kappa.$$

Therefore α and the $\widehat{X}_i$ for $i \in \{3, 4, 5, 6\}$ all have error less than ϵ/κ times the corresponding upper bounds of 1 and κ^{i-2}. Then by Lemma A.1, the error in any monomial in α and the X_i is less than ϵ/κ times the upper bound on that monomial.

Let us consider the error in $\widehat{\gamma}$. The equation is

$$\widehat{\gamma} := \frac{1}{\widehat{\alpha}} \frac{\widehat{\alpha}^2\widehat{X}_5 + 2\widehat{X}_3^3 + 2\widehat{\alpha}^3\widehat{X}_3 - 3\widehat{X}_3\widehat{X}_4\widehat{\alpha}}{4\widehat{X}_3^2 - 2\widehat{\alpha}^3 - 3\widehat{X}_4\widehat{\alpha}}$$

For the true α, X_3, X_4, X_5, the numerator is $O(\kappa^3)$ and the denominator is $\Theta(\kappa^2)$, where to get the lower bound on the denominator we use from (12) that

$$4X_3^2 - 2\alpha^3 - 3X_4\alpha = X_3^2 + 4\alpha^3 + 3(X_3^2 - 2\alpha^3 - X_4\alpha) \geq 4\alpha^3 + 3(6\alpha\gamma)^2 \eqsim \kappa^2.$$

Hence for the estimates, we have

$$\widehat{\gamma} = \frac{O(\kappa^3) \pm O(\epsilon\kappa^2)}{\Theta(\kappa^2) \pm O(\epsilon\kappa)} = \gamma \pm O(\epsilon)$$

Then $\widehat{\beta}$ and the $\widehat{\mu}_i$ are trivial $\pm O(\epsilon)$ approximations. From this, the $\widehat{p}_i$ are $\pm O(\epsilon)$ approximations and the $\widehat{\sigma}_i^2$ are $\pm O(\epsilon)$-approximations. Rescaling ϵ gives the result. □

3.4 Algorithm for precision between Δ_μ and $\sqrt{\Delta}_{\sigma^2}$

▷ Recover gaussian mixture from (estimates of the) mean, variance, and excess moments
▷ Excess moments are function of moments, defined in (10)

procedure SameMeanRecoverFromMoments(μ, σ^2, X_4, X_6)
  $\Delta_{\sigma^2} \leftarrow \sqrt{\frac{4}{3}X_4 + \frac{X_6^2}{25X_4^2}}$ ▷ $\Delta_{\sigma^2} := \sigma_2^2 - \sigma_1^2$
  $p_i \leftarrow \frac{1}{2}(1 \mp \frac{X_6}{5X_4\Delta_{\sigma^2}})$ ▷ p_1 takes $-$ branch
  $\sigma_1^2 \leftarrow \sigma^2 - p_2\Delta_{\sigma^2}$.
  $\sigma_2^2 \leftarrow \sigma^2 + p_1\Delta_{\sigma^2}$.
  return $(p_1, \mu, \sigma_1), (p_2, \mu, \sigma_2)$
end procedure

Algorithm 3.2: Algorithm for recovery of mixture of two gaussians in one dimension, when $\mu_1 \approx \mu_2$.

Algorithm 3.2 solves for the gaussian mixture under the assumption that $\mu_1 = \mu_2$. First, we show that it is correct and robust to perturbations in the moments; we will then show that the moments are robust to perturbation of the means.

Lemma 3.8. *Suppose $\mu_1 = \mu_2$ and p_1, p_2 are bounded away from zero. Let $\Delta_{\sigma^2} := \sigma_1^2 - \sigma_2^2$. For any ϵ less than a sufficiently small constant, if $|\widehat{X}_i - X_i| \lesssim \epsilon\Delta_{\sigma^2}^{i/2}$ for all $i \in \{2, 4, 6\}$, then Algorithm 3.2 recovers σ_i^2 to $\pm O(\epsilon\Delta_{\sigma^2})$ additive error and p_i to $\pm O(\epsilon)$ additive error.*

Proof. First, we show that Algorithm 3.2 gives exact recovery when the moments are exact; then we show robustness.

We choose to disambiguate the mixtures by $\sigma_2^2 \geq \sigma_1^2$ so $\gamma = \Delta_{\sigma^2}/(\mu_2 - \mu_1)$. By examining Lemma 3.1 as $\mu_i \to 0$ and $\gamma \to \Delta_{\sigma^2}/\mu_i$, we observe that when $\Delta_\mu = 0$ we have

$$\begin{aligned} X_4 &\to 3\alpha\gamma^2 \\ X_6 &\to 15\alpha\beta\gamma^3 \end{aligned} \tag{22}$$

which, in terms of $p_1, p_2, \Delta_{\sigma^2}$, for $\Delta_\mu = 0$ implies that

$$\begin{aligned} X_4 &= 3p_1p_2\Delta_{\sigma^2}^2 \\ X_6 &= 15p_1p_2(p_2 - p_1)\Delta_{\sigma^2}^3 \end{aligned}$$

Therefore

$$\frac{4}{3}X_4 + \frac{X_6^2}{25X_4^2} = 4p_1p_2\Delta_{\sigma^2}^2 + (p_2-p_1)^2\Delta_{\sigma^2}^2$$
$$= (p_1+p_2)^2\Delta_{\sigma^2}^2 = \Delta_{\sigma^2}^2$$

and

$$p_2 - p_1 = \frac{X_6}{5X_4\Delta_{\sigma^2}}.$$

The algorithm is thus correct given the exact moments.

How robust is the algorithm? We have that $X_4 \eqsim \Delta_{\sigma^2}^2$ and $|X_6| \lesssim \Delta_{\sigma^2}^3$. Hence $\widehat{X_4} = (1 \pm O(\epsilon))X_4 \gtrsim \Delta_{\sigma^2}^2$ and $|\widehat{X}_6^2 - X_6^2| \lesssim \epsilon\Delta_{\sigma^2}^6$, and

$$|\frac{\widehat{X}_6^2}{\widehat{X}_4^2} - \frac{X_6^2}{X_4^2}| \le |\frac{\widehat{X}_6^2}{\widehat{X}_4^2} - \frac{\widehat{X}_6^2}{X_4^2}| + |\frac{\widehat{X}_6^2}{X_4^2} - \frac{X_6^2}{X_4^2}| \lesssim \epsilon\Delta_{\sigma^2}^2.$$

Therefore $\widehat{\Delta}_{\sigma^2} = (1 \pm O(\epsilon))\Delta_{\sigma^2}$. And since $|\frac{\widehat{X}_6}{\widehat{X}_4} - \frac{X_6}{X_4}| \lesssim \epsilon\Delta_{\sigma^2}$, this means

$$|\widehat{p}_i - p_i| \lesssim |\frac{\widehat{X}_6}{\widehat{X}_4\widehat{\Delta}_{\sigma^2}} - \frac{X_6}{X_4\Delta_{\sigma^2}}| \lesssim \epsilon$$

as desired. □

Theorem 3.9 (Precision between Δ_μ and $\sqrt{\Delta}_{\sigma^2}$)**.** *Consider any mixture of gaussians where p_1 and p_2 are bounded away from zero, and any $\epsilon < 1$. Suppose that $\epsilon\Delta_{\sigma^2} \gtrsim \Delta_\mu^2$. If $|\widehat{X}_i - X_i| \le \epsilon\Delta_{\sigma^2}^{i/2}$ for all $i \le 6$, Algorithm 3.2 recovers the p_i to $\pm O(\epsilon)$ and σ_i^2 to $\pm O(\epsilon)\Delta_{\sigma^2}$.*

Proof. Let G be the given gaussian mixture and G' be the mixture with probabilities p_i and variances σ_i^2 but both means moved to μ, which we may assume without loss of generality is 0. Then we can express $x \sim G$ as $y + z$ for $y \sim G'$ and $|z| \le \Delta_\mu$. Define X_i' to be the ith excess moment of G'.

Since the sign of y is independent of the mixture chosen, $\mathbb{E}[yz] = 0$. Therefore $|X_2' - X_2| = \mathbb{E}[z^2] = p_1\mu_1^2 + p_2\mu_2^2 \lesssim \Delta_\mu^2$. From (10) and (22), we have that

$$|X_4 - X_4'| = |-2\alpha^2 + \alpha\beta^2 + 6\alpha\beta\gamma| \lesssim \Delta_\mu^2\Delta_{\sigma^2}$$
$$|X_6 - X_6'| = |\alpha(16\alpha^2 - 12\alpha\beta^2 - 60\alpha\beta\gamma + \beta^4 + 15\beta^3\gamma + 45\beta^2\gamma^2)| \lesssim \Delta_\mu^2\Delta_{\sigma^2}^2$$

Therefore $|X_i' - X_i| \lesssim \Delta_\mu^2\Delta_{\sigma^2}^{i/2-1}$ for all $i \in \{2,4,6\}$, so $|X_i' - \widehat{X}_i| \lesssim \epsilon\Delta_{\sigma^2}^{i/2}$. Lemma 3.8 immediately implies the result. □

3.5 Combining the algorithms to get general precision

Theorem 3.10. *Let F be any mixture of two gaussians with variance σ^2 and p_1, p_2 bounded away from 0. Then, given $O(\epsilon^{-2}n\log(1/\delta))$ samples Algorithm 3.3 with probability $1-\delta$ outputs the parameters of a mixture $\widehat{F}$ so that for some permutation π and all $i \in \{1,2\}$ we have the following guarantees:*

- *If $n \ge \left(\frac{\sigma^2}{|\mu_1-\mu_2|^2}\right)^6$, then $|\mu_i - \widehat{\mu}_{\pi(i)}| \le \epsilon|\mu_1-\mu_2|$, $|\sigma_i^2 - \widehat{\sigma}_{\pi(i)}^2| \le \epsilon|\mu_1-\mu_2|^2$, and $|p_i - \widehat{p}_{\pi(i)}| \le \epsilon$.*

procedure RECOVER1DMIXTURE($x_1, \ldots, x_n, \delta$)
  Compute (excess) moments $\mu, \sigma^2, X_3, \ldots, X_6$
  $f \leftarrow (\frac{\log(1/\delta)}{n})^{1/12}$ ▷ Error $f^6\sigma^i$ in each X_i.
  $\overline{\Delta}_\mu \leftarrow \begin{cases} \min(|X_3|^{1/3} + |X_4|^{1/4}, \ X_3/\sqrt{X_4}) & \text{if } X_4 > 0 \\ |X_3|^{1/3} + |X_4|^{1/4} & \text{otherwise} \end{cases}$ ▷ $\Theta(\Delta_\mu) + O(f^{3/2}\sigma)$
  $\overline{\Delta}_{\sigma^2} \leftarrow \sqrt{|X_4|}$ ▷ $\Theta(\Delta_{\sigma^2}) \pm O(f^3\sigma^2) \pm O(\Delta_\mu^2)$
  if $f^2 \lesssim \overline{\Delta}_\mu^2/\sigma^2$ **then** ▷ Can get within $\pm\Delta_\mu^2$
    return result of Algorithm 3.1 with $\epsilon \eqsim \sqrt{\frac{(\sigma/\overline{\Delta}_\mu)^{12}\log(1/\delta)}{n}}$.
  else if $f^2 \lesssim \overline{\Delta}_{\sigma^2}/\sigma^2$ **then** ▷ Between $\pm\Delta_\mu^2$ and $\pm\Delta_{\sigma^2}$
    return result of Algorithm 3.2
  else ▷ Can't distinguish either μ_i or σ_i, so output a single gaussian.
    return $(1/2, \mu, \sigma^2), (1/2, \mu, \sigma^2)$
  end if
end procedure

Algorithm 3.3: Combined algorithm for recovery of mixture of two gaussians in one dimension.

- *If* $n \geq \left(\frac{\sigma^2}{|\sigma_1^2 - \sigma_2^2|}\right)^6$, *then* $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\sigma_1^2 - \sigma_2^2| + |\mu_1 - \mu_2|^2$ *and* $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon + \frac{|\mu_1 - \mu_2|^2}{|\sigma_1^2 - \sigma_2^2|}$.
- *For any* $n \geq 1$, *the algorithm performs as well as assuming the mixture is a single gaussian:* $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq |\mu_1 - \mu_2| + \epsilon\sigma$ *and* $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq |\sigma_1^2 - \sigma_2^2| + |\mu_1 - \mu_2|^2 + \epsilon\sigma^2$.

We compare the algorithm to the "ideal" algorithm which uses Δ_μ and Δ_{σ^2} instead of their estimates $\overline{\Delta}_\mu$ and $\overline{\Delta}_{\sigma^2}$ to decide which algorithm to use. We show that:

- If the first branch is taken in either the ideal or the actual setting, then $\overline{\Delta}_\mu \eqsim \Delta_\mu$.
- If the second branch is taken in either the ideal or the actual setting, then $\overline{\Delta}_{\sigma^2} \eqsim \Delta_{\sigma^2}$.

Therefore, up to constant factors in the sample complexity, the Algorithm 3.3 performs as well as the ideal algorithm, which performs as well as the best of Algorithm 3.1, Algorithm 3.2, and outputting a single gaussian. The proof is given in Appendix B.7.

4 Dimension Reduction

We first give a simple argument showing that the d-dimensional problem reduces to the 4-dimensional problem. We then give a separate result showing that the 4-dimensional problem reduces to the 1-dimensional problem. Since we previously saw a solution to the 1-dimensional problem, our reductions show how to solve the general d-dimensional problem.

To describe our reduction we need to know $\mathbb{V}(F)$ up to a constant factor. This can be accomplished with few samples as shown next.

Lemma 4.1. *Given* $n = O(\log(1/\delta))$ *samples from a mixture* F *we can output a parameter* σ^2 *such that* $\mathbb{P}\left\{\sigma^2 \in [\mathbb{V}(F), 2\,\mathbb{V}(F)]\right\} \geq 1 - \delta$.

Proof. This follows from estimating the second moment of the distribution up to constant multiplicative error and is shown in the proof of Lemma 3.2. □

Theorem 4.2 ($(d$ to 4)-reduction)**.** *Assume there is a polynomial time algorithm that (ϵ, δ)-learns mixtures of two gaussians in $\mathbb{R}^4$ from $f(\epsilon, \delta)$ samples. Then, for every $d \geq 4$, there is a polynomial time algorithm that (ϵ, δ)-learns mixtures of two gaussians using $f(\epsilon/20, \delta/10d^2) + O(\log(1/\delta))$ samples.*

Proof. Let $\mathcal{A}$ denote the assumed algorithm for $\mathbb{R}^4$. We give an algorithm $\mathcal{B}$ for the d-dimensional problem. The algorithm is given sample access to a mixture F of variance $\sigma^2 = \mathbb{V}(F)$. The algorithm $\mathcal{B}$ always invokes $\mathcal{A}$ with error parameter $\epsilon/20$ and failure probability $\delta/10d^2$.

Algorithm $\mathcal{B}$:

1. Use Lemma 4.1 to obtain a parameter $\widehat{\sigma}^2$ such that $\widehat{\sigma}^2 \in [\sigma^2, 2\sigma^2]$ with probability $1 - \delta/2$.

Determine $\widehat{\mu}^{(i)}$:

2. For every $i \in [d]$ use $\mathcal{A}$ to obtain numbers ξ_i^p for every $p \in \{1, 2\}$ there is $q \in \{1, 2\}$ such that $|\xi_i^p - \mu_i^{(q)}| \leq \epsilon\widehat{\sigma}/20$. For each i this can be done by invoking $\mathcal{A}$ to solve the 1-dimensional mixture problem obtained by restricting the samples to coordinate i.
3. If for all i we have $|\xi_i^1 - \xi_i^2| \leq \epsilon\widehat{\sigma}/4$, then put $\widehat{\mu}^{(1)} = \widehat{\mu}^{(2)} = (\xi_i^1)_{i \in [d]}$
4. Otherwise, let i be the first index such that $|\xi_i^1 - \xi_i^2| > \epsilon\widehat{\sigma}/4$ and do for each $j \in [d]$:
   (a) Use $\mathcal{A}$ to solve the 2-dimensional mixture problem obtained by restricting to the coordinates i, j to accuracy $\epsilon\widehat{\sigma}/20$ in order to obtain numbers (ν_i^p, ν_j^p) for $p = 1, 2$ as the estimate for the two-dimensional means.
   (b) Determine $p \in \{1, 2\}$ such that $|\xi_i^1 - \nu_i^p| \leq \epsilon\widehat{\sigma}/10$. If no such p exists, terminate and output "failure".
   (c) Put $\widehat{\mu}_j^{(1)} = \nu_j^p$ and put $\widehat{\mu}_j^{(2)} = \nu_j^{3-p}$.

Determine $\widehat{\Sigma}^{(i)}$:

5. For every $i, j \in [d]$ use $\mathcal{A}$ to obtain numbers ξ_{ij}^p for every $p \in \{1, 2\}$ there is $q \in \{1, 2\}$ such that $|\xi_{ij}^p - \Sigma_{ij}^{(q)}| \leq \epsilon^2\widehat{\sigma}^2/20$. For each i, j this can be done by using $\mathcal{A}$ to solve the 2-dimensional mixture problem obtained by restricting the samples to coordinates i, j.
6. If for all i, j we have $|\xi_{ij}^1 - \xi_{ij}^2| \leq \epsilon^2\widehat{\sigma}^2/4$, then put $\widehat{\Sigma}^{(1)} = \widehat{\Sigma}^{(2)} = [\xi_{ij}^1]_{i,j \in [d]}$
7. Otherwise, let i, j be the first indices in lexicographic order such that $|\xi_{ij}^1 - \xi_{ij}^2| > \epsilon^2\widehat{\sigma}^2/4$ and do for each $k, l \in [d]$:
   (a) Invoke $\mathcal{A}$ to solve the 4-dimensional mixture problem obtained by restricting to the coordinates i, j, k, l to accuracy $\epsilon^2\widehat{\sigma}^2/20$ in order to obtain numbers $\sigma_{ij}^p, \sigma_{kl}^p$ for $p = 1, 2$.
   (b) Determine $p \in \{1, 2\}$ such that $|\xi_{ij}^1 - \sigma_{ij}^p| \leq \epsilon^2\widehat{\sigma}^2/10$. If no such p exists, terminate and output "failure".
   (c) Put $\widehat{\Sigma}_{kl}^{(1)} = \sigma_{kl}^p$ and put $\widehat{\Sigma}_{kl}^{(2)} = \sigma_{kl}^{3-p}$.

Matching up $\widehat{\Sigma}$ and $\widehat{\mu}$.

8. If there exist an i, j, k with $|\xi_{ij}^1 - \xi_{ij}^2| \geq \epsilon\sigma/2$ and $|\xi_k^1 - \xi_k^2| \geq \epsilon\sigma/2$, then run $\mathcal{A}$ on $\{i, j, k\}$ to get estimates (σ_{ij}^p, ν_k^p) of (Σ_{ij}^p, μ_k^p) for $p = 1, 2$. If there exists a permutation $\pi : \{1, 2\} \to \{1, 2\}$ with $|\widehat{\Sigma}_{ij}^{(1)} - \sigma_{ij}^{(\pi(1))}| < |\widehat{\Sigma}_{ij}^{(1)} - \sigma_{ij}^{(\pi(2))}|$ and $|\widehat{\mu}_k^{(1)} - \nu_k^{(\pi(1))}| > |\widehat{\mu}_k^{(1)} - \nu_k^{(\pi(2))}|$, then switch $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$.

Correctness of $\widehat{\sigma}^2$ and invocations of $\mathcal{A}$. Appealing Lemma 4.1 we have with probability $1-\delta/2$,

$$\sigma^2 \le \widehat{\sigma}^2 \le 2\sigma^2\,.$$

Moreover, we know that each invocation of $\mathcal{A}$ is on a mixture problem of variance at most σ^2 and we run the algorithm with accuracy parameter $\epsilon/20$ and error probability $\delta/10d^2$. The total number of invocations is at most $5d^2$ and therefore every invocation is successful with probability $1-\delta/2$. In this case, we have that all the "mean parameters" returned by $\mathcal{A}$ are $\epsilon\sigma/20$-accurate and all the "variance parameters" are $\epsilon^2\sigma^2/400$-accurate. Both events described above occur with probability $1-\delta$ and we will show that $\mathcal{B}$ succeeds in outputting a mixture that's ϵ-close to F assuming that these events occur.

Correctness of means. On the one hand, suppose that the case described in Step 3 occurs. In this case, each pair of parameters is within distance $\epsilon\widehat{\sigma}/4 \le \epsilon\sigma/2$ and the estimates are $\epsilon\sigma/20$ accurate. Hence, the output is $\epsilon\sigma$-close for both means.

On the other hand, consider the case described in Step 4 and let i denote the coordinate found by the algorithm. Since $|\xi_i^1 - \xi_i^2| > \epsilon\widehat{\sigma}/4 \ge \epsilon\sigma/4$ it must be the case that

$$\left|\mu_i^{(1)} - \mu_i^{(2)}\right| \ge \frac{\epsilon\sigma}{4} - \frac{\epsilon\sigma}{10} = \frac{\epsilon\sigma}{8}\,.$$

Further since all estimates are $(\epsilon\sigma/20)$-accurate, there always must exist a $p \in \{1,2\}$ such that $|\xi_i^1 - \nu_i^p| \le \epsilon\sigma/10 \le \epsilon\widehat{\sigma}/10$. There is at most one such p since $|\xi_i^1 - \xi_i^2| > \epsilon\widehat{\sigma}/4$. For this p we have ν_i^p and ξ_i^1 are either both $(\epsilon\sigma/20)$-close to $\mu_i^{(1)}$ or they are both $(\epsilon\sigma/20)$-close to $\mu_i^{(2)}$ but not both. It follows that for every $j \in [d]$ our estimates ν_j^p all belong to the same d-dimensional mean. This shows that we correctly identify $\mu^{(1)}, \mu^{(2)} \in \mathbb{R}^d$ up to additive error $\epsilon\sigma/20$ in each coordinate.

Correctness of covariances. The argument for $\widehat{\Sigma}^{(1)}, \widehat{\Sigma}^{(2)}$ is analogous. Suppose that the case described in Step 6 occurs. In this case, each pair of parameters is within distance $\epsilon^2\sigma^2/2$ and the estimates are $\epsilon^2\sigma^2/20$ accurate. Hence, the output is $\epsilon^2\sigma^2$ accurate for both covariance matrices.

Now consider the case described in Step 7 and let (i,j) denote the pair of coordinates found by the algorithm. Since $|\xi_{ij}^1 - \xi_{ij}^2| > \epsilon^2\widehat{\sigma}^2/4 \ge \epsilon^2\sigma^2/4$ it must be the case that

$$\left|\Sigma_{ij}^{(1)} - \Sigma_{ij}^{(2)}\right| > \epsilon^2\sigma^2/4 - \epsilon^2\sigma^2/10 = \epsilon^2\sigma^2/8.$$

Further since all estimates are $(\epsilon^2\sigma^2/20)$-accurate, there always must exist a $p \in \{1,2\}$ such that $|\xi_{ij}^1 - \sigma_{ij}^p| \le \epsilon^2\sigma^2/10$. For this p we know that σ_{ij}^p and ξ_{ij}^1 are either both $(\epsilon^2\sigma^2/10)$-close to $\Sigma_{ij}^{(1)}$ or they are both $(\epsilon^2\sigma^2/10)$-close to $\Sigma_{ij}^{(2)}$. In particular, for every $k,l \in [d]$ our estimates σ_{kl}^p all belong to the same n-dimensional covariance matrix. This shows that we correctly identify $\Sigma^{(1)}, \Sigma^{(2)} \in \mathbb{R}^{d\times d}$ up to additive error $\epsilon^2\sigma^2/20$ in each coordinate.

Correctness of matching Σ to μ. If there does not exist an i,j,k with $|\xi_{ij}^1 - \xi_{ij}^2| \ge \epsilon\sigma/2$ and $|\xi_k^1 - \xi_k^2| \ge \epsilon\sigma/2$, then either the means or the variances are indistinguishable and the order of matching doesn't matter. Otherwise, since $\mathcal{A}$ gives accuracy $(\epsilon\sigma/20, \epsilon^2\sigma^2/20)$ and the true parameters are separated by at least $(\epsilon\sigma/8, \epsilon^2\sigma^2/8)$, the correct pairing will only have $|\widehat{\mu}_k^{(1)} - \nu_k^{(\pi(1))}| < |\widehat{\mu}_k^{(1)} - \nu_k^{(\pi(2))}|$ or $|\widehat{\Sigma}_{ij}^{(1)} - \sigma_{ij}^{(\pi(1))}| < |\widehat{\Sigma}_{ij}^{(1)} - \sigma_{ij}^{(\pi(2))}|$ when π is the correct permutation for $\widehat{\mu}$ and $\widehat{\Sigma}$, respectively.

□

4.1 From 4 to 1 dimension

For our reduction from $\mathbb{R}^4$ to $\mathbb{R}$ we invoke a powerful anti-concentration result for polynomials in gaussian variables due to Carbery and Wright.

Theorem 4.3 ([CW01])**.** *Let $p(x_1, \ldots, x_d)$ be a degree r polynomial, normalized such that $\mathrm{Var}(p) = 1$ under the normal distribution. Then, for any $t \in \mathbb{R}$ and $\delta > 0$, we have*

$$\mathop{\mathbb{P}}_{x \sim N(0,1)^d} \{|p(x) - t| \leq \tau\} \leq O(r) \cdot \tau^{1/r} .$$

Lemma 4.4. *Let N_9^d be the d-dimensional normal distribution $N(0,1)^d$ conditioned on vectors of norm at most 9. There is a constant $c > 0$ such that for every $B \in \mathbb{R}^{d \times d}$,*

$$\mathop{\mathbb{P}}_{a \sim \mathrm{N}_9^d} \left\{ \left| a^\top B a \right| \leq c \cdot \|B\|_\infty \right\} \leq \frac{1}{3} .$$

Proof. Observe that $p(x) = x^\top B x$ is a degree 2 polynomial in n gaussian variables. It is easy to see that the variance of p under the normal distribution is at least the square of the largest entry of B. That is, $\|B\|_\infty^2$. Hence, we can apply Theorem 4.3 to $p(x)/V$ for some number $V \geq \|B\|_\infty$ to conclude that

$$\mathop{\mathbb{P}}_{x \sim N(0,1)^d} \left\{ \left| x^\top B x \right| \leq c \cdot \|B\|_\infty \right\} \leq \frac{1}{6} .$$

On the other hand, $\|x\|^2 > 9$ with probability less than $1/6$. Hence, the claim follows. □

The next lemma is a direct consequence.

Lemma 4.5. *There is a constant $c > 0$ such that for every $\epsilon > 0, m \in \mathbb{N}$ and $\widehat{\Sigma}, \Sigma^{(1)}, \Sigma^{(2)} \in \mathbb{R}^{d \times d}$ such that $\left\| \widehat{\Sigma} - \Sigma^{(1)} \right\|_\infty > \epsilon$ and $\left\| \widehat{\Sigma} - \Sigma^{(2)} \right\|_\infty > \epsilon$ we have*

$$\mathop{\mathbb{P}}_{a_1, \ldots, a_m \sim \mathrm{N}_9^d} \left\{ \exists i \in [m] \colon \left| a_i^\top (\widehat{\Sigma} - \Sigma^{(1)}) a_i \right| > c\epsilon \quad \text{and} \quad \left| a_i^\top (\widehat{\Sigma} - \Sigma^{(2)}) a_i \right| > c\epsilon \right\} \geq 1 - \left(\frac{1}{3} \right)^m .$$

Proof. For every fixed $i \in [m]$, by Lemma 4.4 and the union bound we have,

$$\mathop{\mathbb{P}}_{a_i \sim \mathrm{N}_9^d} \left\{ \left| a_i^\top (\widehat{\Sigma} - \Sigma^{(1)}) a_i \right| \leq c\epsilon \quad \text{or} \quad \left| a_i^\top (\widehat{\Sigma} - \Sigma^{(2)}) a_i \right| \leq c\epsilon \right\} \leq 2/3 .$$

The claim therefore follows since the samples are independent. □

We have the analogous statement for vectors instead of matrices.

Lemma 4.6. *There is a constant $c > 0$ such that for every $\epsilon > 0, m \in \mathbb{N}$ and $\widehat{\mu}, \mu^{(1)}, \mu^{(2)} \in \mathbb{R}^d$ such that $\left\| \widehat{\mu} - \mu^{(1)} \right\|_\infty > \epsilon$ and $\left\| \widehat{\mu} - \mu^{(2)} \right\|_\infty > \epsilon$ we have*

$$\mathop{\mathbb{P}}_{a_1, \ldots, a_m \sim \mathrm{N}_9^d} \left\{ \exists i \in [m] \colon \left| \langle \widehat{\mu} - \mu^{(1)}, a_i \rangle \right| > c\epsilon \quad \text{and} \quad \left| \langle \widehat{\mu} - \mu^{(2)}, a_i \rangle \right| > c\epsilon \right\} \geq 1 - \left(\frac{1}{3} \right)^m .$$

Proof. The proof is analogous to that of Lemma 4.5, but instead of Lemma 4.4 we directly appeal to the anti-concentration properties of the one-dimensional Normal distribution. □

We also note two obvious bounds.

Lemma 4.7. *Let $B \in \mathbb{R}^{4\times 4}$ and $\mu \in \mathbb{R}^4$. Then,*

1. $\mathbb{P}_{a\sim \mathrm{N}_9^4}\left\{|a^\top B a| \le O(\|B\|_\infty)\right\} = 1$, *and*
2. $\mathbb{P}_{a\sim \mathrm{N}_9^4}\left\{|\langle a, \mu\rangle| \le O(\|\mu\|_\infty)\right\} = 1$.

Proof. This is immediate because the dimension is constant and the norm of a is at most 9 with probability 1. □

We now have all the ingredients for our reduction from four to one dimension.

Theorem 4.8 ((4 to 1)-reduction)**.** *Assume there is a polynomial time algorithm that (ϵ, δ)-learns a mixture of two gaussians in $\mathbb{R}$ from $f(\epsilon, \delta)$ samples. Then for some constant $c > 0$ there is a polynomial time algorithm that (ϵ, δ)-learns mixtures of two gaussians in $\mathbb{R}^4$ from $f(c\epsilon, c\delta/\log(\epsilon/\delta)) + O(\log(1/\delta))$ samples.*

Proof. Let $\mathcal{A}$ denote the assumed algorithm for one-dimensional mixtures. We give an algorithm $\mathcal{C}$ for the 4-dimensional problem. We prove that the algorithm $(O(\epsilon), O(\delta))$-learns mixtures of two gaussians in $\mathbb{R}^4$ given the stated sample bounds. We get the statement of the theorem by rescaling ϵ, δ.

We use Lemma 4.1 to obtain a parameter $\widehat{\sigma}^2$ such that $\widehat{\sigma}^2 \in [\sigma^2, 2\sigma^2]$ with probability $1 - \delta$.

We will locate the unknown mixture parameters by doing a grid search and checking each solution using the previous lemmas that we saw. To find a suitable grid for the means, we first find an estimate $\widehat{\mu} \in \mathbb{R}^4$ so that $\mu^{(1)}$ and $\mu^{(2)}$ are both within $2\widehat{\sigma}$ of $\widehat{\mu}$ in each coordinate. This can be done by invoking $\mathcal{A}$ on each of the 4 coordinates with error parameter $1/2$ and success probability $1 - \delta$. For $i = 1, \ldots, 4$ we take $\widehat{\mu}_i$ to be either of the two estimates for the means in the i-th invocation of $\mathcal{A}$. Since $\|\mu^{(1)} - \mu^{(2)}\|_\infty \le \sigma$ by assumption we know that $\mu^{(1)}, \mu^{(2)}$ are both within distance 2σ of $\widehat{\mu}_i$.

Let N_μ be a $(c\epsilon\widehat{\sigma})$-net in ℓ_∞-distance around the point $\widehat{\mu}$ of width $2\widehat{\sigma}$ in every coordinate. For small enough $c > 0$, the true parameters must be $(\epsilon\widehat{\sigma}/20)$ close to a point contained in N_μ. This is because $\sigma^2 \ge \|\mu^{(1)} - \mu^{(2)}\|_\infty^2$ by definition. Similarly, we let $N_\sigma = ([-\widehat{\sigma}^2, \widehat{\sigma}^2] \cap (c\epsilon\widehat{\sigma}^2)\mathbb{Z})^{4\times 4}$. Since $\|\Sigma^{(i)}\|_\infty \le \sigma^2 \le \widehat{\sigma}^2$ this net must contain an $(\epsilon\sigma/20)$-close point to each true covariance matrix. Note that $|N_\mu| \times |N_\sigma| = \mathrm{poly}(1/\epsilon)$.

Algorithm $\mathcal{C}$:

1. Let $m = 10\log((|N_\mu| \times |N_\sigma|)/\delta)$. Sample $a_1, \ldots, a_m \sim \mathrm{N}_9^4$ and sample $x_1, \ldots, x_{m'} \sim F$ where $m' = f(\epsilon', \delta')$ with $\epsilon' = c\epsilon$ and $\delta' = \delta/m$ for sufficiently small constant $c > 0$. For each a_i run $\mathcal{A}$ on $\{\langle a_i, x_j\rangle \colon j \in [m']\}$ with error parameter ϵ' and confidence δ'.

   Denote the outputs of $\mathcal{A}$ by $\left\{(\widehat{\mu}^{1,i}, \widehat{\Sigma}^{1,i}, \widehat{\mu}^{2,i}, \widehat{\Sigma}^{2,i}) \colon i \in [m]\right\}$.

2. For every vector $\widehat{\mu} \in N_\mu$, do the following:

   (a) If there exists an $i \in [m]$ such that $|\langle a_i, \widehat{\mu}\rangle - \widehat{\mu}^{1,i}| > \epsilon\widehat{\sigma}/2$ and $|\langle a_i, \widehat{\mu}\rangle - \widehat{\mu}^{2,i}| > \epsilon\widehat{\sigma}/2$, then label $\widehat{\mu}$ as "rejected". Otherwise if there is no such $i \in [m]$, label $\widehat{\mu}$ as "accepted".

3. Let M be the set of accepted vectors. If $M = \emptyset$ output "failure" and terminate. Otherwise, choose $\widehat{\mu}^{(1)}, \widehat{\mu}^{(2)} \in M$ to maximize $\|\widehat{\mu}_1 - \widehat{\mu}_2\|_\infty$.

4. For every symmetric $\widehat{\Sigma} \in N_\sigma$, do the following:

(a) If there exists an $i \in [m]$ such that $|\langle a_i, \widehat{\Sigma} a_i\rangle - \widehat{\Sigma}^{1,i}| > \epsilon^2\widehat{\sigma}^2/2$ and $|\langle a_i, \widehat{\Sigma} a_i\rangle - \widehat{\Sigma}^{2,i}| > \epsilon^2\widehat{\sigma}^2/2$, then label $\widehat{\Sigma}$ as "rejected". Otherwise if there is no such $i \in [m]$, label $\widehat{\Sigma}$ as "accepted".

5. Let S be the set of accepted matrices. If $S = \emptyset$ output "failure" and terminate. Otherwise, choose $\widehat{\Sigma}^{(1)}, \widehat{\Sigma}^{(2)} \in S$ to maximize $\|\widehat{\Sigma}^{(1)} - \widehat{\Sigma}^{(2)}\|_\infty$.

6. If there exists an $i \in [m]$ where there does not exist a permutation $\pi : \{1,2\} \to \{1,2\}$ such that for all $p \in \{1,2\}$ we have $|\langle a_i, \widehat{\Sigma}^{(p)} a_i\rangle - \widehat{\Sigma}^{\pi(p),i}| \le \epsilon^2\widehat{\sigma}^2/2$ and $|\langle a_i, \widehat{\mu}^{(p)}\rangle - \widehat{\mu}^{\pi(p),i}| \le \epsilon\widehat{\sigma}/2$, then switch $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$.

Claim 4.9. *Let σ_i^2 denote the variance of the mixture problem induced by a_i. Then we have that $\max_{i\in[m]} \sigma_i^2 \le O(\sigma^2)$ with probability 1.*

Proof. This follows directly from the concentration bounds in Lemma 4.7. □

We need the following claim which shows that with high probability the estimates obtained in Step 1 are $(\epsilon/10)$-accurate.

Claim 4.10. *With probability $1-\delta$, for all $i \in [m]$, there is a permutation π_i so that:*

1. $|\widehat{\mu}^{1,i} - \langle a_i, \mu^{(\pi_i(1))}\rangle| \le \epsilon\sigma/10$ *and* $|\widehat{\mu}^{2,i} - \langle a_i, \mu^{(\pi_i(2))}\rangle| \le \epsilon\sigma/10$
2. $|\widehat{\Sigma}^{1,i} - \langle a_i, \Sigma^{(\pi_i(1))} a_i\rangle| \le \epsilon^2\sigma^2/100$ *and* $|\widehat{\Sigma}^{2,i} - \langle a_i, \Sigma^{(\pi_i(2))} a_i\rangle| \le \epsilon^2\sigma^2/100$

Proof. If $x \sim F$ then $\langle a, x\rangle$ is sampled from a 1-dimensional mixture model with means $\langle a, \mu^{(1)}\rangle, \langle a, \mu^{(2)}\rangle$ and variances $\langle a, \Sigma^{(1)} a\rangle, \langle a, \Sigma^{(2)} a\rangle$. Note that we chose the error probability of $\mathcal{A}$ small enough so that we can take a union bound over all m invocations of the algorithm. Moreover, by Claim 4.9, all of these mixtures have variance at most $O(\sigma^2)$. □

We suppose in what follows that the result of Claim 4.10 holds.

Correctness of means. Let

$$A = \left\{\mu\colon \left\|\mu^{(1)} - \mu\right\|_\infty \le C\epsilon\sigma\right\} \cup \left\{\mu\colon \left\|\mu^{(2)} - \mu\right\|_\infty \le C\epsilon\sigma\right\}.$$

for a sufficiently large constant C. We first claim that with probability $1-\delta$, every element that gets accepted is in A. To establish the claim we need to show that with probability $1-\delta$ every element in $A^c \cap N_\mu$ gets rejected. For any $\mu \in A^c$, we have from Lemma 4.6 that for some constant $c' > 0$, with probability $1 - 1/3^m$ we have for some $i \in [m]$ that

$$\min(|\langle \mu - \mu^{(1)}, a_i\rangle|, |\langle \mu - \mu^{(2)}, a_i\rangle|) > c' C\epsilon\sigma \ge \epsilon\widehat{\sigma}$$

if C is sufficiently large, and hence μ is rejected. By our choice of m and a union bound, with $1-\delta$ probability all $\mu \in A^c \cap N_\mu$ are rejected.

We also need to show there exists $\bar{\mu}^{(1)}, \bar{\mu}^{(2)}$ that get accepted such that $\left\|\bar{\mu}^{(1)} - \mu^{(1)}\right\| \le \epsilon\sigma$ and $\left\|\bar{\mu}^{(2)} - \mu^{(2)}\right\| \le \epsilon\sigma$. To see this take $\bar{\mu}^{(1)}$ to be the nearest neighbor of $\mu^{(1)}$, which has $|\bar{\mu}^{(1)} - \mu^{(1)}| \le c\epsilon\sigma$. By Lemma 4.7, it follows that

$$\max_{i\in[m]} |\langle \bar{\mu}^{(1)} - \mu^{(1)}, a_i\rangle| \le O(c\epsilon\sigma)$$

and hence $\bar{\mu}^{(1)}$ is accepted if c is sufficiently small. The symmetric argument holds for the nearest neighbor of $\mu^{(2)}$.

Now we can finish the argument by distinguishing two cases. Consider the case where $\left\|\mu^{(1)} - \mu^{(2)}\right\|_\infty \leq 5C\epsilon\sigma$. In this case any accepted element must be $6C\epsilon\sigma$-close to both means. The other case is when $\left\|\mu^{(1)} - \mu^{(2)}\right\|_\infty > 5C\epsilon\sigma$. In this case, A contains two distinct clusters of elements centered around each mean. Each pair within a single cluster has distance at most $2C\epsilon\sigma$, while any pair spanning the two clusters has distance at least $3C\epsilon\sigma$. Hence the pair of largest distance are in different clusters and within $C\epsilon\sigma$ of the corresponding means.

Correctness of covariances. The argument is very similar to the previous one. Let

$$A = \left\{\Sigma\colon \left\|\Sigma^{(1)} - \Sigma\right\|_\infty \leq C\epsilon^2\sigma^2\right\} \cup \left\{\Sigma\colon \left\|\Sigma^{(2)} - \Sigma\right\|_\infty \leq C\epsilon^2\sigma^2\right\}.$$

for a sufficiently large constant C. We first claim that with probability $1-\delta$, every element that gets accepted is in A. To establish the claim we need to show that with probability $1-\delta$ every element in $A^c \cap N_\sigma$ gets rejected. For any $\Sigma \in A^c$, we have from Lemma 4.5 that for some constant $c' > 0$, with probability $1 - 1/3^m$ we have for some $i \in [m]$ that

$$\min(|a_i^\top(\Sigma - \Sigma^{(1)})a_i|, |a_i^\top(\Sigma - \Sigma^{(2)})a_i|) > c'C\epsilon^2\sigma^2 \geq \epsilon^2\widehat{\sigma}^2$$

if C is sufficiently large, and hence Σ is rejected. By our choice of m and a union bound, with $1-\delta$ probability all $\Sigma \in A^c \cap N_\sigma$ are rejected.

We also need to show there exists $\bar{\Sigma}^{(1)}, \bar{\Sigma}^{(2)}$ that get accepted such that $\left\|\bar{\Sigma}^{(1)} - \Sigma^{(1)}\right\| \leq \epsilon^2\sigma^2$ and $\left\|\bar{\Sigma}^{(2)} - \Sigma^{(2)}\right\| \leq \epsilon^2\sigma^2$. To see this take $\bar{\Sigma}^{(1)}$ to be the nearest neighbor of $\Sigma^{(1)}$, which has $|\bar{\Sigma}^{(1)} - \Sigma^{(1)}| \leq c\epsilon^2\sigma^2$. By Lemma 4.7, it follows that

$$\max_{i\in[m]} |\langle \bar{\Sigma}^{(1)} - \Sigma^{(1)}, a_i\rangle| \leq O(c\epsilon^2\sigma^2)$$

and hence $\bar{\Sigma}^{(1)}$ is accepted if c is sufficiently small. The symmetric argument holds for the nearest neighbor of $\Sigma^{(2)}$. Now we can finish the argument by distinguishing two cases. Consider the case where $\left\|\Sigma^{(1)} - \Sigma^{(2)}\right\|_\infty \leq 5C\epsilon^2\sigma^2$. In this case any accepted element must be $6C\epsilon^2\sigma^2$-close to both $\Sigma^{(i)}$. The other case is when $\left\|\Sigma^{(1)} - \Sigma^{(2)}\right\|_\infty > 5C\epsilon^2\sigma^2$. In this case, A contains two distinct clusters of elements centered around each $\Sigma^{(i)}$. Each pair within a single cluster has distance at most $2C\epsilon^2\sigma^2$, while any pair spanning the two clusters has distance at least $3C\epsilon^2\sigma^2$. Hence the pair of largest distance are in different clusters and within $C\epsilon^2\sigma^2$ of the corresponding $\Sigma^{(i)}$.

Correctness of matching Σ to μ. If either $\left\|\Sigma^{(1)} - \Sigma^{(2)}\right\|_\infty \leq 5C\epsilon^2\sigma^2$ or $\left\|\mu^{(1)} - \mu^{(2)}\right\|_\infty \leq 5C\epsilon\sigma$, then matching the $\widehat{\Sigma}^{(p)}$ to the $\widehat{\mu}^{(p)}$ is unnecessary. Otherwise, we have for each $i \in [m]$ that the probability that either $\widehat{\Sigma}^{(1)}$ or $\mu^{(1)}$ matches the wrong mean under a_i is at most $2/3$. Hence with $1-\delta$ probability, one of the a_i will disambiguate the two. (One must be careful because $\widehat{\mu}$ and $\widehat{\Sigma}$ depend on the randomness in a_i, but m is large enough that we can union bound over all $|N_\mu| \times |N_\sigma|$ possibilities.) □

Combining our two reductions we immediately have the following result.

Theorem 4.11. *Let F be any mixture of d-dimensional gaussians where p_1 and p_2 are bounded away from zero. Then we can (ϵ, δ)-learn F with $O(\epsilon^{-12}\log(d\log(1/\epsilon)/\delta))$ samples.*

Proof. By Corollary 1.3, there is an algorithm that (ϵ, δ)-learns mixtures of two gaussians in R from $O(\epsilon^{-12}\log(1/\delta))$ samples. Hence, by Theorem 4.8, there is an algorithm that (ϵ, δ)-learns mixtures of two gaussians in $\mathbb{R}^4$ using sample size

$$O(\epsilon^{-12}\log(\log(1/\epsilon\delta)/\delta) + \log(1/\delta)) = O(\epsilon^{-12}\log(\log(1/\epsilon\delta)/\delta)).$$

Finally, by Theorem 4.2, there is an algorithm that (ϵ, δ)-learns mixtures of two gaussians in $\mathbb{R}^d$ from using a number of samples that is bounded by

$$O(\epsilon^{-12} \log(2d \log(d/\epsilon\delta)/\delta) + \log(1/\delta)) = O(\epsilon^{-12} \log(2d \log(d/\epsilon\delta)/\delta)) .$$

□

5 Algorithm in the TV norm

procedure RECOVERTVAPPROX(ϵ, δ)
  Take $O(d^2 \log(1/\delta))$ samples and compute the empirical overall covariance $\widehat{\Sigma}$.
  Factor $\widehat{\Sigma}$ as AA^T.
  Take $O((\epsilon')^{-12} \log(1/\delta))$ samples $x_1, \ldots, x_n$ for $\epsilon' = \Theta(\frac{\epsilon^3}{d^{2.5}\sqrt{\log(d/\epsilon)}})$
  $(\widehat{\mu}_1, \widehat{\Sigma}_1), (\widehat{\mu}_2, \widehat{\Sigma}_2) \leftarrow$ RecoverParamApprox$(A^{-1}x_1, \ldots, A^{-1}x_n)$
  $\widehat{\lambda}^2, v \leftarrow$ minimum eigenvalue/eigenvector in either $\widehat{\Sigma}_1$ or $\widehat{\Sigma}_2$ (WLOG of $\widehat{\Sigma}_1$)
  if $\widehat{\lambda}^2 > 2(\epsilon')^2 d/\epsilon^2$ **then**
    return $(\widehat{\mu}_1, A\widehat{\Sigma}_1 A^T), (\widehat{\mu}_2, A\widehat{\Sigma}_2 A^T)$.
  else
    for $t \in [T] = [O(\log(1/\delta))]$ **do**
      Take $O(d^2/\epsilon^2)$ samples $x_1, \ldots, x_m$.
      Partition into $S = \{i : |v^T(A^{-1}x_i - \widehat{\mu}_1)| \lesssim \epsilon' \sqrt{d \log \frac{d}{\epsilon}}/\epsilon\}$, $\overline{S} = [n] \setminus S$.
      $G_1^{(t)}, G_2^{(t)} \leftarrow$ empirical means/covariances of x_S and $x_{\overline{S}}$.
    end for
    For $i \in \{1, 2\}$,

$$\widehat{G}_i \leftarrow \arg\min_{G_i^{(t)} : t \in [T]} \operatorname{median}_{j \in [T]} \mathrm{TV}(G_i^{(t)}, G_i^{(j)})$$

    return $\widehat{G}_1, \widehat{G}_2$
  end if
end procedure

Algorithm 5.1: Algorithm for TV approximation

In order to get a good approximation in total variation distance, we use the following lemma (proven in the appendix):

Lemma A.4. *Let G, G' be d-dimensional gaussians with means and covariance matrices (μ, Σ) and (μ', Σ'), respectively. Suppose that the minimum eigenvalue of Σ is λ^2. Then*

$$\mathrm{TV}(G, G') \lesssim \frac{1}{\lambda}(\|\mu - \mu'\| + \|\Sigma - \Sigma'\|_F)$$

The intuition for the algorithm is that either the minimum eigenvalue is large (in which case parameter estimation works well) or it is small (in which case we can cluster based on that direction).

Theorem 5.1. *Let F be any mixture of d-dimensional gaussians where p_1 and p_2 are bounded away from zero. For any dimension $d \geq 1$, Algorithm 5.1 (ϵ, δ)-learns F in total variation with $O(\epsilon^{-36} d^{30} \log^6(d/\epsilon) \log(1/\delta))$ samples.*

Proof. Because the overall mixture is subgaussian, our estimate $\widehat{\Sigma}$ of the overall covariance Σ will be good: for all v we have $v^T\widehat{\Sigma}v = (1 \pm 0.25)v^T\Sigma v$. Hence the normalized samples $y_1, \ldots, y_n = A^{-1}x_1, \ldots, A^{-1}x_n$ have covariance Σ' with all eigenvalues between $3/4$ and $5/4$. Because our algorithm is linear in normalization factor A, we may assume that A is the identity and Σ has all eigenvalues between $3/4$ and $5/4$.

By Corollary 1.3, therefore, the parameter estimation algorithm will learn (up to permutation) the μ_i to $\pm\epsilon'$ in each coordinate and Σ_i to $\pm(\epsilon')^2$ in each coordinate.

Now, let λ^2 be the minimum eigenvalue of either Σ_1 or Σ_2. We have by Lemma A.4 that there is a permutation with

$$\begin{aligned}\mathrm{TV}(\widehat{G_1}, G_{\pi_1}) + \mathrm{TV}(\widehat{G_2}, G_{\pi_2}) &\lesssim \frac{1}{\lambda}(\epsilon'\sqrt{d} + (\epsilon')^2 d) \\ &\lesssim \epsilon'\sqrt{d}/\lambda\end{aligned}$$

where the second step follows from $\epsilon'\sqrt{d} < 1$ in order for the result to be nontrivial (since $\mathrm{TV} \le 1$ always). Hence the results from parameter approximation are sufficient if $\lambda \gtrsim \epsilon'\sqrt{d}/\epsilon$. Since our approximations $\widehat{\Sigma_i}$ of Σ_i are within $d(\epsilon')^2$ in Frobenius norm, $|\widehat{\lambda}^2 - \lambda^2| \le d(\epsilon')^2$. Hence the first branch of the conditional will be taken if and only if $\lambda \gtrsim \epsilon'\sqrt{d}/\epsilon$, in which case it gives an ϵ approximation in the TV norm.

If the other branch is taken, it finds a v for which $v^T(\Sigma_1)v - \mu_1\mu_1^T \le 3(\epsilon'\sqrt{d}/\epsilon)^2$ (where we choose Σ_1 without loss of generality). Hence

$$\mathop{\mathbb{E}}_{x\sim G_1}(v^T(x - \mu_1))^2 \le 3(\epsilon'\sqrt{d}/\epsilon)^2.$$

Since $v^T G_1$ is Gaussian, this means for $x \sim G_1$ that $|v^T(x - \mu_1)| \lesssim \epsilon'\sqrt{d\log(\frac{d}{\epsilon})}/\epsilon$ with probability $1 - O(\frac{\epsilon^2}{d^2})$. By a union bound, in any inner loop, with $9/10$ probability all the samples in $x_1, \ldots, x_m$ that are drawn from G_1 will then lie in S.

On the other hand, $\mathbb{E}_{x\sim F}(v^T(x - \mu_1))^2 \ge 3/4$ by our assumption on the overall covariance, so

$$\mathop{\mathbb{E}}_{x\sim G_2}(v^T(x - \mu_1))^2 \ge 3/4.$$

Therefore for $x \sim G_2$, $\mathbb{P}[|v^T(x - \mu_1)| \lesssim \epsilon'\sqrt{d\log(\frac{d}{\epsilon})}/\epsilon] \lesssim \epsilon'\sqrt{d\log(\frac{d}{\epsilon})}/\epsilon$. For our choice of ϵ', this is less than ϵ^2/d^2, so by a union bound, in any inner loop, with $9/10$ probability none of the samples drawn from G_2 will lie in S.

Overall, this means that each inner loop will correctly classify S as the samples from G_1 and $\overline{S}$ as the samples from G_2 with $8/10$ probability. Furthermore, given the correct classifications, Lemma A.5 shows that $G_1^{(t)}$ and $G_2^{(t)}$ will be $\epsilon/3$ approximations to G_1 and G_2 in the TV norm with at least $19/20$ probability. By Lemma A.6, this means that the result $\widehat{G}_1$ and $\widehat{G}_2$ of the procedure will be a ϵ approximation to G_1 and G_2 with $1 - \delta/2$ probability.

Thus regardless of which branch is taken, the result of `RecoverTVApprox` will be an ϵ approximation in the TV norm with $1 - \delta$ probability. The measurement complexity is dominated by the parameter estimation call. □

This sample complexity can be improved if the covariances of the two Gaussians have similar eigenvalues and eigenvectors (e.g., they are isotropic):

Theorem 5.2. *Let F be any mixture of d-dimensional gaussians with covariance matrices Σ_1 and Σ_2 where the mixing probabilities p_1 and p_2 are bounded away from zero. Further suppose that there exists a constant C such that*

$$\Sigma_1 \preceq C\Sigma_2 \preceq C^2\Sigma_1.$$

Then there is an algorithm that can (ϵ, δ)-learn F in total variation with $O(\epsilon^{-12}d^6 \log^6(d/\epsilon)\log(1/\delta))$ samples.

Proof. The algorithm is just Algorithm 5.1, changed to have $\epsilon' = \Theta(\frac{\epsilon}{\sqrt{d\log(d/\epsilon)}})$. The only difference from the proof of Theorem 5.1 is that now, when we consider

$$\mathop{\mathbb{E}}_{x\sim G_2}(v^T(x-\mu_1))^2 \geq 3/4$$

we also have

$$\mathop{\mathbb{E}}_{x\sim G_2}(v^T(x-\mu_2))^2 \leq C \mathop{\mathbb{E}}_{x\sim G_1}(v^T(x-\mu_1))^2 \leq 3(\epsilon'\sqrt{d}/\epsilon)^2.$$

Thus $v^T\mu_1$ is $\Theta(\frac{\epsilon}{\epsilon'\sqrt{d}})$ standard deviations away from the mean of v^TG_2. So

$$\mathop{\mathbb{P}}_{x\sim G_2}[|v^T(x-\mu_1)| \lesssim \epsilon'\sqrt{d\log(d/\epsilon)}] \lesssim \exp(-\Omega(\frac{\epsilon}{\epsilon'\sqrt{d}})^2)$$

which is $O(\epsilon^2/d^2)$ for the given ϵ'. □

By Theorem **??**, $\Omega(d^6/\epsilon^{12})$ samples are *necessary* for a TV approximation even in the approximately isotropic case. Hence Theorem 5.2 is tight up to logarithmic factors.

References

[AJOS14] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Near-optimal-sample estimators for spherical gaussian mixtures. *CoRR*, abs/1402.4746, 2014.

[BS10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112. IEEE Computer Society, 2010.

[CDSS13] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proc. 24th ACM-SIAM SODA*, pages 1380–1394, 2013.

[CDSS14] Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proc. 46th Symposium on Theory of Computing (STOC)*. ACM, 2014.

[CW01] A Carbery and J Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

[DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proc. 27th COLT*, pages 1183–1213, 2014.

[HK13] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In Robert D. Kleinberg, editor, *ITCS*, pages 11–20. ACM, 2013.

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *STOC*, pages 553–562. ACM, 2010.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102. IEEE Computer Society, 2010.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 185, 1894.

[Pol00] D. Pollard. *Asymptopia*. in progress, 2000.

[SWW02] Mark F Schilling, Ann E Watkins, and William Watkins. Is human height bimodal? *The American Statistician*, 56(3):223–229, 2002.

[Sym14] SymPy Development Team. *SymPy: Python library for symbolic mathematics*, 2014.

A Utility Lemmas

A.1 Approximating the coefficients of a polynomial will approximate its value

The following lemma shows that a constant size monomial is robust to perturbations of its inputs.

Lemma A.1. *Let $a, b, c \in \mathbb{R}^k$ for constant k and $0 \le \epsilon \le 1$. If $|a_i| \le b_i$ and $|c_i| \le \epsilon b_i$, we have that*

$$\left| \prod_{i=1}^{k} (a_i + c_i) - \prod_{i=1}^{k} a_i \right| \lesssim \epsilon \prod_{i=1}^{k} b_i.$$

Proof. All $2^k - 1 \lesssim 1$ terms on the left are bounded by the value on the right. □

Lemma A.2. *Let p be any constant-degree polynomial in a constant number of variables $a_1, \ldots, a_t$ with constant coefficients. Let $\overline{q}$ equal p except with all the coefficients having their absolute value taken. Suppose $|a_i| \le b_i$ and $|c_i| \le \epsilon b_i$ for some $b, c \in \mathbb{R}^t$ and $0 \le \epsilon \le 1$. Then*

$$|p(a_1, \ldots, a_t) - p(a_1 + c_1, \ldots, a_t + c_t)| \lesssim \epsilon \overline{p}(b_1, \ldots, b_t).$$

Proof. Apply Lemma A.1 to each monomial. □

A.2 Estimating moments of a distribution from samples

The following lemma shows that we can estimate moments well.

Lemma 3.2. *Suppose p_1, p_2 are bounded away from zero and our mixture has variance σ^2. Given $O(\log(1/\delta)/\epsilon^2)$ samples, with probability $1 - \delta$ we can compute estimates $\widehat{X}_i$ of the first $O(1)$ excess moments X_i satisfying $|\widehat{X}_i - X_i| \le \epsilon \sigma^i$.*

Proof. We partition the samples x_i into $O(\log(1/\delta))$ groups of size $k = O(1/\epsilon^2)$, then compute the median (over groups) of the empirical excess moment of the group. We will show that this gives the desired result.

Suppose we want to compute $t = O(1)$ moments. Because our samples x_i are the sum of a gaussian and a bounded variable and hence subgaussian, $\mathbb{E}[x_i^p] \lesssim \sigma^p$ for any $p \le t$. Therefore $\mathrm{Var}(x_i^p) \le \mathbb{E}[x_i^{2p}] \lesssim \sigma^{2p}$.

For a group of k samples x_i, consider how well the empirical pth moment $\widehat{M}_p = \frac{1}{k}\sum x_i^p$ approximates the true moment M_p. We have that $\mathrm{Var}(\widehat{M}_p) \lesssim \sigma^{2p}/k$. By Chebyshev's inequality, then, for any $c > 0$ we have

$$\mathbb{P}[|\widehat{M}_p - M_p| > O(\sigma^p/\sqrt{ck})] \le c.$$

Setting the constant $c = 1/(4t)$ and then choosing $k = O(1/(c\epsilon^2))$, we have with $3/4$ probability that

$$|\widehat{M}_p - M_p| \le \epsilon\sigma^p$$

for all $p \le t$. Then since the X_p are polynomials in the M_p with total degree p, by Lemma A.2 we have for all $p \le t$ that

$$|\widehat{X}_p - X_p| \lesssim \epsilon\sigma^p. \tag{23}$$

Call a block where this happens "good." Since each block is good with $3/4$ probability and there are $O(\log(1/\delta))$ blocks, with $1-\delta$ probability more than half the blocks are "good." If this is the case, then for each p the median $\widehat{X}_p$ will also satisfy (23). Rescaling ϵ gives the result. □

A.3 Relationship between parameter distance and TV distance

The next few lemmas are elementary results relating parameter distance and total variation distance of Gaussians.

Lemma A.3. *Let G, G' be d-dimensional gaussians with means and covariance matrices (μ, Σ) and (μ', Σ'), respectively. Suppose that the maximum eigenvalue of Σ is λ^2. Then*

$$\mathrm{TV}(G, G') \gtrsim \min(1, \|\mu - \mu'\|/\lambda).$$

Proof. Project onto $(\mu - \mu')$ and solve the $d = 1$ case. □

Lemma A.4. *Let G, G' be d-dimensional gaussians with means and covariance matrices (μ, Σ) and (μ', Σ'), respectively. Suppose that the minimum eigenvalue of Σ is λ^2. Then*

$$\mathrm{TV}(G, G') \lesssim \frac{1}{\lambda}(\|\mu - \mu'\| + \|\Sigma - \Sigma'\|_F)$$

Proof. First, note that by symmetry $\mathrm{TV}(N(\mu, \Sigma), N(\mu', \Sigma')) \ge \mathrm{TV}(N(\mu, \Sigma), N(\mu, \Sigma')) = \mathrm{TV}(N(0, \Sigma), N(0, \Sigma'))$. Combined with the triangle inequality we get

$$\mathrm{TV}(N(\mu, \Sigma), N(\mu', \Sigma')) \le \mathrm{TV}(N(0, \Sigma), N(0, \Sigma')) + \mathrm{TV}(N(\mu, \Sigma), N(\mu', \Sigma))$$

so it suffices to consider the variation in μ and Σ separately. From Pollard [Pol00] we have that $\mathrm{TV}(N(\mu, I_d), N(\mu', I_d)) \lesssim \|\mu - \mu'\|$, so

$$\mathrm{TV}(N(\mu, \Sigma), N(\mu', \Sigma)) \lesssim \|\mu - \mu'\|/\lambda.$$

For the covariance term, in one dimension it is an easy calculation that $\mathrm{TV}(N(0, \lambda^2), N(0, \lambda^2 + \epsilon^2)) \lesssim \epsilon/\lambda$. In general, by rotation invariance and scaling we may assume that $\Sigma = I$, and by decomposing Σ' we may assume that $\Sigma' - \Sigma$ is very sparse: either a single diagonal term or a single pair of symmetric off-diagonal terms. If $\Sigma' - \Sigma$ is a single diagonal term it reduces directly to the one dimensional case, and the remaining option reduces to the two dimensional case

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma' = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$$

which is just two one-dimensional cases on the eigenvectors $(1, 1)$ and $(1, -1)$, giving the result. □

Lemma A.5. *Given n samples from a single Gaussian $G = (\mu, \Sigma)$ in d dimensions, the empirical $\widehat{G} = (\widehat{\mu}, \widehat{\Sigma})$ satisfies $\mathrm{TV}(G, \widehat{G}) \lesssim d\sqrt{\frac{\log(1/\delta)}{n}}$ with probability $1 - \delta$.*

Proof. The TV distance is independent of linear transformations, so assume $\mu = 0$ and $\Sigma = I_d$ without loss of generality. Then $\widehat{\mu} \sim N(0, I_d/n)$, so $\|\widehat{\mu}\| \le \sqrt{\frac{d+\log(1/\delta)}{n}}$ with probability $1 - \delta$. For the covariance, for $i \neq j$ we have that $\widehat{\Sigma}_{i,j} \sim N(0,1)^2/n$ which is subexponential with constant parameters, so and $\|\Sigma - \widehat{\Sigma}\|_F \le d\sqrt{\frac{\log(1/\delta)}{n}}$ with probability $1 - \delta$. Apply Lemma A.4 to get the result. □

A.4 Getting a TV approximation with arbitrarily high probability

The next lemma shows that we can amplify our probability of getting a TV approximation using $O(\log(1/\delta))$ samples.

Lemma A.6. *Let X be an arbitrary space with norm $\|\cdot\|$. Let $x_1, \ldots, x_m$ be drawn independently from X so that $\mathbb{P}[\|x_i - x^*\| \le \epsilon] \ge 3/4$ for some x^*. Then*

$$\widehat{x} := \arg\min_{x_i : i \in [m]} \operatorname*{median}_{j \in [m]} \|x_i - x_j\|$$

satisfies $\|x^* - \widehat{x}\| \le 3\epsilon$ with $1 - 2^{-\Omega(m)}$ probability.

Proof. With the given probability, more than $m/2$ of the x_i will have $\|x_j - x^*\| \le \epsilon$. Each of these will have $\mathrm{median}_{j \in [m]} \|x_i - x_j\| \le 2\epsilon$. Therefore the resulting $\widehat{x}$ is within 2ϵ of at least one of these $m/2$ elements, or within 3ϵ of x^*. □

A.5 Gaussian anticoncentration

This section gives some lemmas showing for a polynomial p that $p(x)$ is not too concentrated around zero if x is drawn from a univariate or multivariate Gaussian.

Lemma A.7. *Let $p(x)$ be a univariate polynomial of degree d and with leading coefficient ax^d. Then for any $\epsilon > 0$,*

$$\mathbb{P}_{x \sim N(0,1)}[|p(x)| < |a|\epsilon] < d\epsilon^{1/d}.$$

Proof. By scaling we can assume $a = 1$. Then we can factor $p(x)$ over $\mathbb{C}$, getting $p(x) = \prod_{i=1}^{d}(x - z_i)$ for some complex numbers z_i. For any $t > 0$, we have

$$\mathbb{P}[|x - z_i| < t] \le \frac{2t}{\sqrt{2\pi}} < t$$

Hence

$$\mathbb{P}[|p(x)| < \epsilon] \le \mathbb{P}[\exists i \text{ such that } |x - z_i| < \epsilon^{1/d}] < d\epsilon^{1/d}.$$

□

Lemma A.8. *Let $p(x) = p(x_1, \ldots, x_n)$ be a multivariate polynomial of degree d and smallest nonzero coefficient of magnitude a. Then*

$$\mathbb{P}_{x \sim N(0, I_n)}[|p(x)| < a\epsilon] < d\epsilon^{1/d}.$$

Proof. We proceed by induction on n. The base case is Lemma A.7. If the degree of x_n in p is d', then we write

$$p(x_1, \ldots, x_n) = \sum_{i=0}^{d'} q_i(x_1, \ldots, x_{n-1}) x_n^i$$

where $q_{d'}$ is a nonzero formal polynomial of degree $d - d'$ with smallest nonzero coefficient a. By induction, we have for any t that

$$|q_i(x_1, \ldots, x_{n-1})| \geq at$$

with probability at least $1 - (d - d')t^{1/(d-d')}$. Consider any setting of $x_1, \ldots, x_{n-1}$ for which this happens. Conditioned on this, $p(x)$ is a univariate polynomial in x_n, so Lemma A.7 shows for any ϵ that

$$\mathbb{P}[|p(x)| < a\epsilon] < d'(\epsilon/t)^{1/d'}.$$

Hence, unconditionally and using a union bound, we have that

$$\mathbb{P}[|p(x)| < a\epsilon] < d'(\epsilon/t)^{1/d'} + (d - d')t^{1/(d-d')}$$

for an arbitrary t. Optimizing with $t = \epsilon^{(d-d')/d}$, we get

$$\mathbb{P}[|p(x)| < a\epsilon] < d\epsilon^{1/d}.$$

This shows the inductive step, getting the result. □

B Algorithm for $d = 1$

B.1 Sympy

The proofs in this section involve a fair amount of algebraic manipulation. To make these computations more reliable and easier to verify, in some cases we provide code for a computer to do them. We use Sympy [Sym14], a standard Python package for symbolic manipulation. We only use Sympy for simple tasks – multiplying and adding polynomials, substituting expressions for variables – that can be verified by hand.

B.2 Excess Moments of a gaussian Mixture

Lemma 3.1. *For $\alpha = -\mu_1\mu_2, \beta = \mu_1 + \mu_2, \gamma = \frac{\sigma_2^2 - \sigma_1^2}{\mu_2 - \mu_1}$ we have that*

$$\begin{aligned}
X_3 &:= M_3 && = \alpha\beta + 3\alpha\gamma \\
X_4 &:= M_4 - 3M_2^2 && = -2\alpha^2 + \alpha\beta^2 + 6\alpha\beta\gamma + 3\alpha\gamma^2 \\
X_5 &:= M_5 - 10M_3M_2 && = \alpha(\beta^3 - 8\alpha\beta + 10\beta^2\gamma + 15\gamma^2\beta - 20\alpha\gamma) \\
X_6 &:= M_6 - 15M_4M_2 + 30M_2^3 && = \alpha(16\alpha^2 - 12\alpha\beta^2 - 60\alpha\beta\gamma + \beta^4 + 15\beta^3\gamma + 45\beta^2\gamma^2 + 15\beta\gamma^3)
\end{aligned}$$

Proof. For a standard $N(\mu, \sigma^2)$ gaussian we have moments

$$\begin{aligned}
M_2 &= \mu^2 + \sigma^2 \\
M_3 &= \mu^3 + 3\mu\sigma^2 \\
M_4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \\
M_5 &= \mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4 \\
M_6 &= \mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6.
\end{aligned}$$

and the mixture has probability $p_1 = \mu_2/(\mu_2 - \mu_1)$ and $p_2 = -\mu_1/(\mu_2 - \mu_1)$.

Therefore the following Sympy code (see Section B.1 for an explanation of Sympy) can be used to formally verify the result:

```
#! /usr/bin/python
from sympy import *

# Define variables
mu1 = Symbol(r'\mu_1')
sigma1 = Symbol(r'\sigma_1')
mu2 = Symbol(r'\mu_2')
sigma2 = Symbol(r'\sigma_2')
alpha = Symbol(r'\alpha')
beta = Symbol(r'\beta')
gamma = Symbol(r'\gamma')

p1 = mu2/(mu2 - mu1)

# Moments of single (mu1, sigma1) gaussian
M2 = mu1**2 +               sigma1**2
M3 = mu1**3 +  3 * mu1    * sigma1**2
M4 = mu1**4 +  6 * mu1**2 * sigma1**2 +  3 *          sigma1**4
M5 = mu1**5 + 10 * mu1**3 * sigma1**2 + 15 * mu1    * sigma1**4
M6 = mu1**6 + 15 * mu1**4 * sigma1**2 + 45 * mu1**2 * sigma1**4 + 15*sigma1**6

# Convert to moments of mixture
M2 = p1 * M2 + (1-p1)*M2.subs({mu1:mu2, sigma1:sigma2})
M3 = p1 * M3 + (1-p1)*M3.subs({mu1:mu2, sigma1:sigma2})
M4 = p1 * M4 + (1-p1)*M4.subs({mu1:mu2, sigma1:sigma2})
M5 = p1 * M5 + (1-p1)*M5.subs({mu1:mu2, sigma1:sigma2})
M6 = p1 * M6 + (1-p1)*M6.subs({mu1:mu2, sigma1:sigma2})

# Claimed excess moments
x3 = alpha*beta + 3*alpha*gamma
x4 = -2*alpha**2 + alpha*beta**2 + 6*alpha*beta*gamma  + 3*alpha*gamma**2
x5 = alpha * (beta**3 - 8*alpha*beta + 10*beta**2*gamma + 15*gamma**2*beta
              - 20*alpha*gamma)
x6 = alpha*(16*alpha**2 - 12*alpha*beta**2 - 60*alpha*beta*gamma + beta**4 +
            15*beta**3*gamma + 45*beta**2*gamma**2 + 15*beta*gamma**3)

# Check that they match

alphadefs = {alpha: -mu1*mu2, beta: mu1+mu2,
             gamma: (sigma2**2-sigma1**2)/(mu2-mu1)}

print (M3 - x3.subs(alphadefs)).factor()
print (M4-3*M2**2 - x4.subs(alphadefs)).factor()
print (M5-10*M3*M2 - x5.subs(alphadefs)).factor()
```

```
print (M6-15*M4*M2 + 30*M2**3 - x6.subs(alphadefs)).factor()
```

All the results are zero, so the claimed X_i are correct. □

B.3 Expressing $\alpha\gamma$ using X_3, X_4, X_6, α

We prove (15), which is analogous to (13). We demonstrate using the following Sympy code (see Section B.1 for an explanation of Sympy):

```
#! /usr/bin/python
from sympy import *

# define variables
alpha = Symbol(r'\alpha')
beta = Symbol(r'\beta')
gamma = Symbol(r'\gamma')
X3 = Symbol('X3')
X4 = Symbol('X4')
X5 = Symbol('X5')
X6 = Symbol('X6')
z = Symbol('z')

# define expressions for X_i in terms of alpha, beta, gamma
x3 = alpha*beta + 3*alpha*gamma
x4 = -2*alpha**2 + alpha*beta**2 + 6*alpha*beta*gamma  + 3*alpha*gamma**2
x5 = alpha * (beta**3 - 8*alpha*beta + 10*beta**2*gamma + 15*gamma**2*beta
              - 20*alpha*gamma)
x6 = alpha*(16*alpha**2 - 12*alpha*beta**2 - 60*alpha*beta*gamma + beta**4 +
            15*beta**3*gamma + 45*beta**2*gamma**2 + 15*beta*gamma**3)

# we know that this should be zero.
eqn = alpha**3 * (x6 - X6)
print eqn, '= 0'
```

$$-\alpha^3(X_6 - 16\alpha^3 + 12\alpha^2\beta^2 + 60\alpha^2\beta\gamma - \alpha\beta^4 - 15\alpha\beta^3\gamma - 45\alpha\beta^2\gamma^2 - 15\alpha\beta\gamma^3) = 0$$

```
eqn = eqn.expand().subs(alpha*beta, X_3-3*alpha*gamma) # remove beta
print eqn, '= 0'
```

$$X_3^4 + 3X_3^3\alpha\gamma - 12X_3^2\alpha^3 - 36X_3^2\alpha^2\gamma^2 + 12X_3\alpha^4\gamma + 42X_3\alpha^3\gamma^3 - X_6\alpha^3 + 16\alpha^6 + 72\alpha^5\gamma^2 + 36\alpha^4\gamma^4 = 0$$

```
# Use (7) to remove gamma**2 terms
eqn = eqn.expand().subs(alpha**2*gamma**2, (X3**2 - 2*alpha**3 - X4*alpha)/6)
print eqn, '= 0'
```

$$-4X_3^4 + 10X_3^3\alpha\gamma + 4X_3^2X_4\alpha + 8X_3^2\alpha^3 - 7X_3X_4\alpha^2\gamma - 2X_3\alpha^4\gamma + X_4^2\alpha^2 - 8X_4\alpha^4 - X_6\alpha^3 - 4\alpha^6 = 0$$

```
eqn = eqn.subs(alpha*gamma, z).expand().collect(z)  # this of the form f*z + g
print eqn, '= 0'
```

$$z(10X_3^3 - 7X_3X_4\alpha - 2X_3\alpha^3) - 4X_3^4 + 4X_3^2X_4\alpha + 8X_3^2\alpha^3 + X_4^2\alpha^2 - 8X_4\alpha^4 - X_6\alpha^3 - 4\alpha^6 = 0$$

```
answer = -eqn.subs(z, 0) / eqn.coeff(z)
print 'z = ', answer
```

$$z = \frac{4X_3^4 - 4X_3^2X_4\alpha - 8X_3^2\alpha^3 - X_4^2\alpha^2 + 8X_4\alpha^4 + X_6\alpha^3 + 4\alpha^6}{X_3(10X_3^2 - 7X_4\alpha - 2\alpha^3)}$$

which is (15).

B.4 Bounding r away from zero

This section proves the following lemma:

Lemma 3.4. *For any constant $c > 0$, and for all $\alpha \geq 0$ and $\beta, \gamma, y \in \mathbb{R}$ with $cy_{max} \leq y \leq y_{max}$ and $\beta^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim \kappa^{12}(y - \alpha)^2\alpha^{16}$$

We start by showing that $r(y) = 0$ has a unique solution on $(0, y_{max}]$. The following lemma shows that for any such solution y there exists a gaussian mixture with $\alpha = y$ and matching excess moments; since the first six moments uniquely identify a gaussian mixture, this gives uniqueness.

Lemma B.1. *For any solution y to the system of equations*

$$\begin{aligned} p_5(y) &= 0 \\ p_6(y) &= 0 \\ y &> 0 \\ X_3^2 - 2y^3 + X_4 y &\geq 0 \end{aligned} \tag{24}$$

there exists a mixture of gaussians with $\alpha = y$ and excess moments $X_3, \ldots, X_6$.

Proof. We set the recovered $\widehat{\alpha} = y$, recover γ via (13):

$$\widehat{\gamma} = \frac{1}{\widehat{\alpha}} \frac{\widehat{\alpha}^2 X_5 + 2X_3^3 + 2\widehat{\alpha}^3 X_3 - 3X_3X_4\widehat{\alpha}}{4X_3^2 - 3X_4\widehat{\alpha} - 2\widehat{\alpha}^3}.$$

which is well defined, using (12) and that the denominator is

$$4X_3^2 - 3X_4\widehat{\alpha} - 2\widehat{\alpha}^3 = 4\widehat{\alpha}^3 + X_3^2 + 3(X_3^2 - 2\widehat{\alpha}^3 - X_4\widehat{\alpha}) > 0.$$

We then recover

$$\widehat{\beta} = \frac{1}{\widehat{\alpha}}(X_3 - 3\widehat{\alpha}\widehat{\gamma}).$$

Now, our $\widehat{\alpha}$ and $\widehat{\gamma}$ satisfy (13) and (14), which implies that (12) is satisfied as well.

Now, consider the excess moments X_i' of the gaussian mixture with parameters $(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma})$. By choice of $\widehat{\beta}$, $X_3' = X_3$. Then since (12) is satisfied by both (X_3, X_4) and (X_3, X_4') and the coefficient of X_4' is nonzero, $X_4' = X_4$. Similarly with (13), $X_5' = X_5$.

What remains is to show $p_6(\widehat{\alpha}) = 0$ implies $X_6' = X_6$. The coefficient of X_6 in p_6 is

$$-2\alpha^2(4X_3^2 - 3X_4\widehat{\alpha} - 2\widehat{\alpha}^3) < 0.$$

Thus since $p_6 = 0$ is satisfied by both $(\widehat{\alpha}, X_3, X_4, X_5, X_6)$ and $(\widehat{\alpha}, X_3, X_4, X_5, X_6')$, $X_6' = X_6$. □

Corollary B.2. *The set of equations* (24) *has exactly one solution, $y = \alpha$.*

Proof. By construction, $y = \alpha$ is a solution to (24). Suppose there existed another solution $y' \neq y$. Then by Lemma B.1, there exist two mixtures of gaussians with different α and matching $X_3, \ldots, X_6$.

The excess moments are constructed to be indifferent to adding gaussian noise, i.e. increasing σ_1^2 and σ_2^2 by the same amount. Hence, by "topping off" the second moment, we can construct two different mixtures of gaussians with identical second moment as well as identical $X_3, \ldots, X_6$; these mixtures also both have mean zero. Such mixtures would have identical first six moments. But by [KMV10], any two different mixtures of gaussians differ in their first six moments. So this is a contradiction. □

Lemma B.3. *For no α, β, γ with $\alpha > 0$ is it the case that both p_5 and p_6 have a double root at $y = \alpha$.*

Proof. Define $q_5(y) = p_5(y)/(y - \alpha)$, $q_6(y) = p_6(y)/(y - \alpha)$. We need to show that it is not true that both $q_5(\alpha) = 0$ and $q_6(\alpha) = 0$. We show this by repeatedly adding multiples of one to the other, essentially taking the GCD. The below is a transcript of a Sympy session (see Section B.1 for an explanation of Sympy) proving this claim.

```
#! /usr/bin/python
# Setup variables and expressions

from sympy import *

alpha = Symbol(r'\alpha')
beta = Symbol(r'\beta')
gamma = Symbol(r'\gamma')
y = Symbol('y')

X3 = alpha * beta + 3 * alpha * gamma
X4 = -2*alpha**2 + alpha*beta**2 + 6*alpha*beta*gamma  + 3*alpha*gamma**2
X5 = alpha * (beta**3 - 8*alpha*beta + 10*beta**2*gamma + 15*gamma**2*beta
              - 20*alpha*gamma)
X6 = alpha*(16*alpha**2 - 12*alpha*beta**2 + beta**4 + 45*beta**2*gamma**2 +
            15*beta*gamma**3 + 15*gamma*(-4*alpha*beta + beta**3))
p5 = (6*(2*X3*y**3 + X5*y**2 - 3*X3*X4*y + 2*X3**3)**2 +
      (2*y**3+3*X4*y - 4*X3**2)**2*(2*y**3 + X4*y - X3**2))
p6 = ((4*X3**2 - 3*X4*y - 2*y**3) *
      (4*X3**4 - 4*X3**2*X4*y - 8*X3**2*y**3 - X4**2*y**2 +
       8*X4*y**4 + X6*y**3 + 4*y**6) -
      (10*X3**3 - 7*X3*X4*y - 2*X3*y**3) *
      (2*X3**3 - 3*X3*X4*y + 2*X3*y**3 + X5*y**2))

# Start actual code

q5 = (p5 / (y - alpha)).factor().subs(y, alpha).factor()
q6 = (p6 / (y - alpha)).factor().subs(y, alpha).factor()

# Our goal is to show that (q5, q6) != (0, 0) for any alpha, beta, gamma.
```

```
print q5
```

$$\alpha^5(4\alpha+\beta^2+6\beta\gamma+27\gamma^2)(16\alpha^2+8\alpha\beta^2+72\alpha\gamma^2+\beta^4+18\beta^2\gamma^2-135\gamma^4)$$

```
# and since α^5(4α+β^2+6βγ+27γ^2) > 0, we can reduce by
q5a = q5 / (alpha**5*(4*alpha + beta**2 + 6*beta*gamma + 27*gamma**2))
print q5a.subs(gamma, 0).factor()
```

$$(4\alpha+\beta^2)^2$$

```
# nonzero, so must have γ ≠ 0.
print q6
```

$$\alpha^5\gamma(16\alpha^2\beta-48\alpha^2\gamma+8\alpha\beta^3-24\alpha\beta^2\gamma-72\alpha\beta\gamma^2-72\alpha\gamma^3+\beta^5-3\beta^4\gamma-18\beta^3\gamma^2-18\beta^2\gamma^3+405\beta\gamma^4+2025\gamma^5)$$

```
q6a = (q6 / (alpha**5*gamma) - beta * q5a).factor()
print q6a
```

$$-3\gamma(16\alpha^2+8\alpha\beta^2+48\alpha\beta\gamma+24\alpha\gamma^2+\beta^4+12\beta^3\gamma+6\beta^2\gamma^2-180\beta\gamma^3-675\gamma^4)$$

```
q6b = (q6a / (-3*gamma) - q5a).factor()
print q6b
```

$$12\gamma(4\alpha\beta-4\alpha\gamma+\beta^3-\beta^2\gamma-15\beta\gamma^2-45\gamma^3)$$

```
q6c = q6b / (12*gamma)

z = 4*alpha + beta**2
q5b = (z + 9*gamma**2)**2 - 216*gamma**4
q6d = (z - 15*gamma**2)*(beta - gamma) - 60*gamma**3
print (q5a - q5b).factor(), (q6d - q6c).factor()
```

$$(0,0)$$

So the solution must have $(z+9\gamma^2)^2-216\gamma^4=0$ and $(z-15\gamma^2)(\beta-\gamma)-60\gamma^3=0$ for $z=4\alpha+\beta^2$. From the first equation,

$$z=-9\gamma^2\pm 6\sqrt{6}\gamma^2$$

and since $z>0$, this means

$$z=(6\sqrt{6}-9)\gamma^2.$$

Plugging into the second equation and dividing by γ^2,

$$(6\sqrt{6}-24)\beta-(36+6\sqrt{6})\gamma=0$$

and so

$$\begin{aligned}\gamma&=-\frac{4-\sqrt{6}}{6+\sqrt{6}}\beta\\ z&=\frac{(6\sqrt{6}-9)(4-\sqrt{6})^2}{(6+\sqrt{6})^2}\beta^2\approx 0.19\beta^2\\ &\le\beta^2<4\alpha+\beta^2=z\end{aligned}$$

a contradiction. □

Lemma B.4 (r is large when γ is unbounded)**.** *There exists a constant C such that for all $\beta, \gamma \in \mathbb{R}$ and $\alpha, y > 0$ with $\gamma^2 \geq C\alpha$, $y \lesssim y_{max}$, and $\beta^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim (y-\alpha)^2\gamma^{12}\alpha^{10}$$

Proof. As in Lemma B.3, define $q_5(y) = p_5(y)/(y-\alpha)$ and $q_6(y) = p_6(y)/(y-\alpha)$ Per (17), we have that $y_{max} \lesssim \alpha + \beta^2 \lesssim \alpha$.

Then q_5 is a homogeneous polynomial in $\sqrt{y}, \sqrt{\alpha}, \beta, \gamma$, all of which are $O(\sqrt{\alpha})$ except γ. The leading γ^6 term of q_5 is

```
>>> q5 = (p5 / (y - alpha)).factor()
>>> q5.expand().coeff(gamma**6).factor()
```

$$243\alpha^3((y+4\alpha)^2 - 40\alpha^2)\gamma^6$$

Hence for $y \notin (2\alpha, 3\alpha)$, $|q_5(y)| \gtrsim \gamma^6\alpha^5$.

Doing the same for p_6, for $p_6/(y-\alpha)$ the leading γ^6 term is

```
>>> q6 = (p6 / (y - alpha)).factor()
>>> q6.expand().coeff(gamma**6).factor()
```

$$162\alpha^3(y-6\alpha)^2\gamma^6$$

so for $y \notin (5\alpha, 7\alpha)$, $|q_5(y)| \gtrsim \gamma^6\alpha^5$.

Thus for all y with γ^2 sufficiently much greater than α,

$$r(y) = (y-\alpha)^2(q_5(y)^2 + q_6(y)^2) \gtrsim (y-\alpha)^2\gamma^{12}\alpha^{10}.$$

□

Lemma B.5 (r is large when γ is bounded)**.** *For any constant $c > 0$, and for all $\alpha \geq 0$ and $\beta, \gamma, y \in \mathbb{R}$ with $cy_{max} \leq y \leq y_{max}$ and $\beta^2, \gamma^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim (y-\alpha)^2\alpha^{16}$$

Proof. Define the polynomial $q(y, \alpha, \beta, \gamma) = r(y)/(y-\alpha)^2 = q_5^2 + q_6^2$. By homogeneity of the constraints and result, we may normalize so that $\alpha = 1$ and consider $q(y, \beta, \gamma) := q(y, 1, \beta, \gamma)$. Our goal is to show that $q(y, \beta, \gamma) \gtrsim 1$.

Define R to be the region of (y, β, γ) allowed by the lemma constraints. By our normalization, $y \eqsim y_{max} \eqsim \alpha = 1$ and $\gamma^2, \beta^2 \lesssim 1$ over R. This means R is closed and bounded and hence compact.

By Corollary B.2, q can only be zero over R when $y = 1$. By Lemma B.3, $q(1, \beta, \gamma) \neq 0$ for all β, γ. Hence q has no roots over R. Because R is compact, this means $q \gtrsim 1$ over R, giving the result. □

Lemma 3.4. *For any constant $c > 0$, and for all $\alpha \geq 0$ and $\beta, \gamma, y \in \mathbb{R}$ with $cy_{max} \leq y \leq y_{max}$ and $\beta^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim \kappa^{12}(y-\alpha)^2\alpha^{16}$$

Proof. This is simply the union of Lemma B.4 and Lemma B.5. □

Lemma 3.5. *For any constant $c > 0$, there exists a constant $c' > 0$ such that for all $\alpha \geq 0$ and $\beta, \gamma, y \in \mathbb{R}$ with $cy_{max} \leq y \leq y_{max} + c'(y_{max} - \alpha)$ and $\beta^2 \lesssim \alpha$ we have*

$$r(y) \gtrsim \kappa^{12}(y-\alpha)^2\alpha^{16}$$

Proof. By homogeneity of the equations, we may assume $y_{max} = 1$ and $\alpha \eqsim 1$.

By Lemma 3.4, $r(1) \gtrsim (1-\alpha)^2\kappa^{12}$. That is, one of $|p_5(1)|$ and $|p_5(1)|$ is $\Omega((1-\alpha)\kappa^6)$. Since $p_5(y)$ and $p_6(y)$ are constant degree polynomials with coefficients of magnitude $O(\kappa^6)$, their derivatives over $[1,2]$ are bounded in magnitude by $O(\kappa^6)$. Hence for all $y \in [1, 1 + c'(1-\alpha)]$,

$$|p_5(y)| \geq |p_5(1)| - O(\kappa^6) \cdot |y-1| \geq |p_5(1)| - O(c'(1-\alpha)\kappa^6)$$

and similarly for p_6. For sufficiently small c, if $|p_5(1)|$ is the $\Omega((1-\alpha)^2\kappa^6)$ term, this is $\Omega((1-\alpha)\kappa^6)$. If instead $|p_6(1)|$ is the $\Omega((1-\alpha)\kappa^6)$ term, then $|p_6(y)|$ is $\Omega((1-\alpha)\kappa^6)$; regardless, the conclusion holds. □

B.5 Accuracy of estimating r

Lemma B.6. *Suppose that $|\widehat{X}_i - X_i| \lesssim \epsilon\kappa^{i-2}\Delta_\mu^i$ for all $i \in \{3,4,5,6\}$ and some $\epsilon < 1$. Then for any $y \lesssim \Delta_\mu^2$,*

$$|\widehat{p_5}(y) - p_5(y)| \lesssim \epsilon\kappa^6\Delta_\mu^{18}$$
$$|\widehat{p_6}(y) - p_6(y)| \lesssim \epsilon\kappa^6\Delta_\mu^{18}$$

Hence

$$\sqrt{\widehat{r}(y)} - \sqrt{r(y)} \lesssim \epsilon\kappa^6\Delta_\mu^{18}.$$

Proof. Recall from (11) that $|X_i| \lesssim \kappa^{i-2}\Delta_\mu^i$ and from (21) that this means each monomial of $p_5(y)$ and $p_6(y)$ is bounded by $O(\kappa^6\Delta_\mu^{18})$. Since p_5 and p_6 are constant size polynomials, the first result follows from by Lemma A.2.

For the second claim, we use that $\sqrt{r(y)} = \sqrt{p_5(y)^2 + p_6(y)^2}$ is Lipschitz in $p_5(y)$ and $p_6(y)$. □

B.6 Recovering α

Lemma B.7. *In* RecoverAlphaFromMoments*, for any ϵ if $|\widehat{X}_i - X_i| \leq \epsilon\Delta_\mu^i$ for $i \in \{3,4\}$, then the estimation $\widehat{y}_{max}$ of y_{max} satisfies*

$$|\widehat{y}_{max} - y_{max}| \lesssim \epsilon\Delta_\mu^2/\kappa^2.$$

Proof. We would like to know the stability of the largest root y_{max} of the polynomial $s(y) := 2y^3 + X_4 y - X_3^2$ to perturbations in X_4 and X_3. Without loss of generality we normalize so that $y_{max} \eqsim \Delta_\mu^2 \eqsim 1$.

Since y_{max} is a root, $2y_{max}^3 + X_4 y_{max} = X_3^2 \geq 0$ so $2y_{max}^2 + X_4 \geq 0$. Hence for all $y \geq (2/3)y_{max}$,

$$s'(y) = 6y^2 + X_4 \geq (8/3)y_{max}^2 + X_4 \geq (2/3)y_{max}^2 \gtrsim 1.$$

But we also have that

$$s'(y) = 6y^2 + X_4 \geq X_4 \gtrsim \gamma^2 - O(1) \eqsim (\kappa^2 - O(1)).$$

Combining gives that for all $y \geq (2/3)y_{max}$,

$$s'(y) \gtrsim \kappa^2.$$

This implies for any parameter $t > 0$ that

$$s(y_{max} - t\epsilon) \lesssim -t\epsilon\kappa^2$$

as long as $t\epsilon \le y_{max}/3$, and

$$s(y_{max} + t\epsilon) \gtrsim t\epsilon\kappa^2$$

for all $t > 0$. On the other hand, for all $y \eqsim 1$ we have

$$|\widehat{s}(y) - s(y)| \le |\widehat{X}_4 - X_4|y + |\widehat{X}_3^2 - X_3^2| \lesssim \epsilon.$$

Let $t = C/\kappa^2$ for sufficiently large constant C.

If $t\epsilon \le y_{max}/3$, then combining gives that $\widehat{s}$ must have a root within $y_{max} \pm t\epsilon$ and no root above this range. This is the desired result.

On the other hand, if $t\epsilon \ge y_{max}/3$, then it is still true that $\widehat{s}$ has no root above $y_{max} + t\epsilon$. Since $\widehat{s}(0) = -\widehat{X}_3^2 \le 0$, we also have that $\widehat{y}_{max} \ge 0$. Hence the result lies in $[0, y_{max} + t\epsilon]$, which is still $y_{max} \pm O(\epsilon/\kappa^2)$ in this parameter regime. □

Lemma 3.6. *Suppose p_1, p_2 are bounded away from zero, let $c > 0$ be a sufficiently small constant, and let $\epsilon < 1$. Suppose further that $|\widehat{X}_i - X_i| \le c\epsilon\Delta_\mu^i$ for all $i \in \{3, 4, 5, 6\}$. In this setting, the result $\widehat{\alpha} =$ RecoverAlphaFromMoments$(\widehat{X}_3, \widehat{X}_4, \widehat{X}_5, \widehat{X}_6, \epsilon)$ satisfies*

$$|\widehat{\alpha} - \alpha| \lesssim \epsilon\Delta_\mu^2/\kappa.$$

Proof. We will suppose $|\widehat{X}_i - X_i| \lesssim \epsilon\Delta_\mu^i$, and show for a sufficiently large constant C that $\widehat{\alpha} =$ RecoverAlphaFromMoments$(\widehat{X}_3, \widehat{X}_4, \widehat{X}_5, \widehat{X}_6, C\epsilon)$ satisfies $|\widehat{\alpha} - \alpha| \lesssim \epsilon\alpha/\kappa$. Rescaling ϵ gives the result.

We normalize so $\alpha \eqsim \Delta_\mu^2 \eqsim 1$.

By Lemma B.7, $|\widehat{y}_{max} - y_{max}| \lesssim \epsilon/\kappa^2 \le \epsilon/\kappa$. Since $\widehat{y}_{max} \eqsim y_{max} \eqsim 1$, $\widehat{X}_4 \lesssim \kappa^2$, and $\widehat{X}_4 \eqsim \kappa^2$ if $\kappa \gg 1$, the estimation $\widehat{\kappa}$ of κ is always

$$\widehat{\kappa} := 1 + \sqrt{|\widehat{X}_4|}/\widehat{y}_{max} \eqsim \kappa.$$

Hence $\alpha \le y_{max} < (1 + O(\epsilon/\kappa))\widehat{y}_{max}$.

We have by Lemma B.6 with $\epsilon' = \epsilon/\kappa$ that $\sqrt{\widehat{r}(y)} = \sqrt{r(y)} \pm O(\epsilon\kappa^5)$ for all $y \lesssim 1$. In particular, this means that

$$\sqrt{\widehat{r}(\alpha)} \lesssim \epsilon\kappa^5. \tag{25}$$

Moreover, by Lemma 3.4 for all $y \in [\alpha/2, y_{max}]$ we have $r(y) \gtrsim (y - \alpha)^2\kappa^{12}$. Therefore for some sufficiently large constant c, for all $y \in [\alpha/2, y_{max}]$ with $|y - \alpha| > c\epsilon/\kappa$ we have

$$\sqrt{\widehat{r}(y)} \gtrsim |y - \alpha|\kappa^6 - O(\epsilon\kappa^5) \ge \frac{1}{2}|y - \alpha|\kappa^6 \gtrsim c\epsilon\kappa^5 > \sqrt{\widehat{r}(\alpha)}. \tag{26}$$

And by Lemma 3.5, if $\alpha < (1 - O(c\epsilon/\kappa))y_{max}$ then $r(y') \gtrsim c^2\epsilon^2\kappa^{10}$ for all $y' \in [y_{max}, (1 + O(c\epsilon/\kappa))y_{max}$.

This implies (A) that a local minimum of $\widehat{r}$ over $[0, (1 + O(c\epsilon/\kappa))\widehat{y}_{max}]$ is $\alpha \pm O(c\epsilon/\kappa)$, and (B) that any larger local minimum y' has $\widehat{r}(y') \gtrsim c^2\epsilon^2\kappa^{10}$.

By definition, RecoverAlphaFromMoments$(\widehat{X}_3, \widehat{X}_4, \widehat{X}_5, \widehat{X}_6, C\epsilon)$ finds the largest local minimum $\widehat{\alpha}$ of $\widehat{r}$ with $\widehat{\alpha} \le (1 + C\epsilon/\kappa)\widehat{y}_{max} \le (1 + (C + O(1))\epsilon/\kappa)y_{max}$ and $\widehat{r}(\widehat{\alpha}) \le C^2\epsilon^2\kappa^{10}$. For sufficiently large C and c, (A) and (B) imply that $\widehat{\alpha} = \alpha \pm O(\epsilon/\kappa)$. □

B.7 Proof of Theorem 3.10

Theorem 3.10. *Let F be any mixture of two gaussians with variance σ^2 and p_1, p_2 bounded away from 0. Then, given $O(\epsilon^{-2} n \log(1/\delta))$ samples Algorithm 3.3 with probability $1-\delta$ outputs the parameters of a mixture $\widehat{F}$ so that for some permutation π and all $i \in \{1,2\}$ we have the following guarantees:*

- *If $n \geq \left(\frac{\sigma^2}{|\mu_1-\mu_2|^2}\right)^6$, then $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq \epsilon|\mu_1-\mu_2|$, $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\mu_1-\mu_2|^2$, and $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon$.*
- *If $n \geq \left(\frac{\sigma^2}{|\sigma_1^2-\sigma_2^2|}\right)^6$, then $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon|\sigma_1^2-\sigma_2^2| + |\mu_1-\mu_2|^2$ and $|p_i - \widehat{p}_{\pi(i)}| \leq \epsilon + \frac{|\mu_1-\mu_2|^2}{|\sigma_1^2-\sigma_2^2|}$.*
- *For any $n \geq 1$, the algorithm performs as well as assuming the mixture is a single gaussian: $|\mu_i - \widehat{\mu}_{\pi(i)}| \leq |\mu_1-\mu_2| + \epsilon\sigma$ and $|\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq |\sigma_1^2-\sigma_2^2| + |\mu_1-\mu_2|^2 + \epsilon\sigma^2$.*

Proof. Suppose the number of samples is $f^{-12}\log(1/\delta)$ so $f^{-12} \approx \epsilon^{-2} n$. By Lemma 3.2 we have with probability $1-\delta$ that all the $\widehat{X}_i$ are within $\pm O(f^6\sigma^i)$ of the true moments X_i. Suppose this happens.

First, we show that $\overline{\Delta}_{\sigma^2}$ and $\overline{\Delta}_\mu$ are good approximations to Δ_{σ^2} and Δ_μ, and therefore the conditionals are followed roughly in the same cases as they would if they were not approximations.

The first conditional. We have that, if $\Delta_{\sigma^2} \gg \Delta_\mu^2$, then $X_4 \eqsim \Delta_{\sigma^2}^2$ and $X_3 \eqsim \Delta_\mu \Delta_{\sigma^2}$. Then $X_4 \eqsim \Delta_{\sigma^2}^2 > 0$ and

$$\min(|X_3|^{1/3} + |X_4|^{1/4}, \quad X_3/\sqrt{X_4}) \eqsim \min(\Delta_{\sigma^2}^{1/2}, \Delta_\mu) = \Delta_\mu$$

Therefore, as long as $f^6\sigma^4 \ll \Delta_{\sigma^2}^2$, in the $\Delta_{\sigma^2} \gg \Delta_\mu^2$ setting we have $\overline{\Delta}_\mu \eqsim \Delta_\mu$.

Otherwise, i.e. when $\Delta_{\sigma^2} \lesssim \Delta_\mu^2$, we have that $|X_4| \lesssim \Delta_\mu^4$ and $|X_3| \lesssim \Delta_\mu^3$. Moreover, in this case, since $X_4 = X_3^2/\alpha - 6\alpha\gamma^2 - 2\alpha^2 = \Theta(X_3^2/\Delta_\mu^2) - \Theta(\Delta_\mu^4)$, either $|X_3| \eqsim \Delta_\mu^3$ or $|X_4| \eqsim \Delta_\mu^4$. If $X_4 > 0$, then $X_3^2/X_4 \gtrsim \Delta_\mu^2$. This is because the single positive root y_{max} of (17) is $\Theta(\Delta_\mu^2)$ and the polynomial is positive at X_3^2/X_4.

Thus, when $\Delta_{\sigma^2} \lesssim \Delta_\mu^2$,

$$|X_3|^{1/3} + |X_4|^{1/4} \eqsim \Delta_\mu$$
$$X_3/\sqrt{X_4} \gtrsim \Delta_\mu \text{ if } X_4 > 0$$

As long as $f^6\sigma^4 \lesssim \Delta_\mu^4$, we will have $|\widehat{X}_3|^{1/3} + |\widehat{X}_4|^{1/4} \eqsim |X_3|^{1/3} + |X_4|^{1/4} \eqsim \Delta_\mu$. For $\widehat{X}_3/\sqrt{\widehat{X}_4}$, if $X_3 \ll \Delta_\mu^3$ then $X_4 \eqsim -\Delta_\mu^4$ so $\widehat{X}_4 < 0$. Otherwise, $\widehat{X}_3/\sqrt{\widehat{X}_4} \gtrsim (\Delta_\mu^3 - f^6\sigma^3)/\sqrt{\Delta_\mu^4 + f^6\sigma^4} \gtrsim \Delta_\mu$. Therefore, as long as $f^6\sigma^4 \ll \Delta_\mu^4$, in the $\Delta_{\sigma^2} \lesssim \Delta_\mu^2$ setting we have $\overline{\Delta}_\mu \eqsim \Delta_\mu$.

Thus, regardless of the relationship between Δ_μ^2 and Δ_{σ^2}, we have that $\overline{\Delta}_\mu \eqsim \Delta_\mu$ as long as $f^6\sigma^4 \ll \Delta_\mu^4 + \Delta_{\sigma^2}^2$. In particular, $\overline{\Delta}_\mu \eqsim \Delta_\mu$ whenever $f^2 \lesssim \Delta_\mu^2/\sigma^2$.

Therefore the first conditional in Algorithm 3.3 is followed if and only if $f^2 \lesssim \Delta_\mu^2/\sigma^2$; that is to say, if the first conditional is that $f^2 \leq C_1\overline{\Delta}_\mu^2/\sigma^2$, then it is followed whenever $f^2 \leq (C_1 - O(1))\Delta_\mu^2/\sigma^2$ and not followed whenever $f^2 \geq (C_1 + O(1))\Delta_\mu^2/\sigma^2$.

The second conditional. We have that $X_4 = \Theta(\Delta_{\sigma^2}^2) \pm O(\Delta_\mu^4)$. Therefore $\overline{\Delta}_{\sigma^2} \lesssim \Delta_{\sigma^2} + \Delta_\mu^2 + f^3\sigma^2$ in general, and if $f^3\sigma^2 \leq c\Delta_{\sigma^2}$ and $\Delta_\mu^2 \leq c\Delta_{\sigma^2}$ for some sufficiently small constant $c > 0$, then $\overline{\Delta}_{\sigma^2} \eqsim \Delta_{\sigma^2}$.

Therefore the second conditional in Algorithm 3.3 is taken if and only if $\Delta_\mu^2/\sigma^2 \lesssim f^2 \lesssim \Delta_{\sigma^2}/\sigma^2$; that is to say, if the second conditional is that $f^2 \leq C_2\overline{\Delta}_\mu^2/\sigma^2$, then that branch is taken whenever

$$(C_1 + O(1))\Delta_\mu^2/\sigma^2 \leq f^2 \leq (C_2 - O(1))\Delta_{\sigma^2}/\sigma^2$$

and not taken whenever

$$(C_1 - O(1))\Delta_\mu^2/\sigma^2 \geq f^2 \text{ or } f^2 \geq (C_2 + O(1))\Delta_{\sigma^2}/\sigma^2$$

The third branch. The remainder is the third branch, which is taken when $f^2 \geq (C_2 + O(1))\Delta_{\sigma^2}/\sigma^2$ and $f^2 \geq (C_1 + O(1))\Delta_\mu^2/\sigma^2$, and not taken if either condition is false after replacing $+O(1)$ by $-O(1)$.

Completing the theorem. We have shown that the three branches are taken in the same settings as they would be taken with the true Δ_μ and Δ_{σ^2}. We now show that the clauses of the theorem correspond to the branches.

In the first clause of the theorem, we have that $f^{-12} = C\frac{1}{\epsilon^2}n$ for sufficiently large C and $n > (\sigma^2/\Delta_\mu^2)^6$. Then $f^2 < C^{-1/6}\Delta_\mu^2/\sigma^2$, so the first branch will be taken and Algorithm 3.1 is run. By Theorem 3.7, running Algorithm 3.1 with $O(\frac{1}{\epsilon^2}(\frac{\sigma}{\Delta_\mu})^{12}\log(1/\delta))$ samples will recover the p_i to additive ϵ error, the μ_i to additive $\epsilon\Delta_\mu$ error, and the σ_i^2 to additive $\epsilon\Delta_\mu^2$ error.

In the second clause of the theorem, we have that $f^2 < C^{-1/6}\Delta_{\sigma^2}/\sigma^2$ for sufficiently large C. Hence, if the first branch is not taken, then the second branch is taken. The first branch is only taken if it can recover the σ_i^2 to better than $\pm|\mu_1 - \mu_2|^2$, which satisfies the second clause of the theorem. The second branch invokes Algorithm 3.2, which by Theorem 3.9 uses $O(\frac{1}{\epsilon^2}(\frac{\sigma}{\sqrt{\Delta_{\sigma^2}}})^{12}\log(1/\delta))$ samples to estimate the p_i to additive ϵ error and the σ_i^2 to additive $\epsilon\Delta_{\sigma^2}$ error, again satisfying the theorem.

Finally, the last clause of the theorem is satisfied by both algorithms and by outputting the single gaussian $N(\mu, \sigma^2)$. (If the second branch is taken, the mean outputted is the sample mean, which suffices for this purpose.) □