# PAC Learning Mixtures of Axis-Aligned Gaussians with No Separation Assumption

Jon Feldman⋆[1], Rocco A. Servedio⋆⋆[2], and Ryan O'Donnell⋆⋆⋆[3]

[1] Google jonfeld@google.com
[2] Columbia University rocco@cs.columbia.edu
[3] Microsoft Research odonnell@microsoft.com

**Abstract.** We propose and analyze a new vantage point for the learning of mixtures of Gaussians: namely, the PAC-style model of learning probability distributions introduced by Kearns et al. [12]. Here the task is to construct a hypothesis mixture of Gaussians that is statistically indistinguishable from the actual mixture generating the data; specifically, the KL divergence should be at most $\epsilon$.

In this scenario, we give a poly$(n/\epsilon)$ time algorithm that learns the class of mixtures of any constant number of axis-aligned Gaussians in $\mathbf{R}^n$. Our algorithm makes *no* assumptions about the separation between the means of the Gaussians, nor does it have any dependence on the minimum mixing weight. This is in contrast to learning results known in the "clustering" model, where such assumptions are unavoidable.

Our algorithm relies on the method of moments, and a subalgorithm developed in [8] for a discrete mixture-learning problem.

## 1 Introduction

In [12] Kearns et al. introduced an elegant and natural model of learning unknown probability distributions. In this framework we are given a class $\mathcal{C}$ of probability distributions over $\mathbf{R}^n$ and access to random data sampled from an unknown distribution $\mathbf{Z}$ that belongs to $\mathcal{C}$. The goal is to output a hypothesis distribution $\mathbf{Z}'$ which with high confidence is $\epsilon$-close to $\mathbf{Z}$ as measured by the the Kullback-Leibler (KL) divergence, a standard measure of the distance between probability distributions (see Section 2 for details on this distance measure). The learning algorithm should run in time poly$(n/\epsilon)$. This model is well-motivated by its close analogy to Valiant's classical Probably Approximately Correct (PAC) framework for learning Boolean functions [17].

Several notable results, both positive and negative, have been obtained for learning in the Kearns et al. framework of [12], see, e.g., [9, 14]. Here we briefly survey some of the positive results that have been obtained for learning various types of *mixture distributions*. (Recall that given distributions $\mathbf{X}^1, \ldots, \mathbf{X}^k$

⋆ Some of this work was done while supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship at Columbia University.

⋆⋆ Supported in part by NSF award CCF-0347282, by NSF award CCF-0523664, and by a Sloan Foundation Fellowship.

⋆⋆⋆ Some of this work was done while at the Institute for Advanced Study.

and mixing weights $\pi^1, \ldots, \pi^k$ that sum to 1, a draw from the corresponding mixture distribution is obtained by first selecting $i$ with probability $\pi^i$ and then making a draw from $\mathbf{X}^i$.) Kearns et al. gave an efficient algorithm for learning certain mixtures of *Hamming balls*; these are product distributions over $\{0,1\}^n$ in which each coordinate mean is either $p$ or $1-p$ for some $p$ fixed over all mixture components. Subsequently Freund and Mansour [10] and independently Cryan et al. [4] gave efficient algorithms for learning a mixture of two arbitrary product distributions over $\{0,1\}^n$. Recently, Feldman et al. [8] gave a poly($n$)-time algorithm that learns a mixture of any $k = O(1)$ many arbitrary product distributions over the discrete domain $\{0, 1, \ldots, b-1\}^n$ for any $b = O(1)$.

### 1.1 Results

As described above, research on learning mixture distributions in the PAC-style model of Kearns et al. has focused on distributions over discrete domains. In this paper we consider the natural problem of learning mixtures of Gaussians in the PAC-style framework of [12]. Our main result is the following theorem:

**Theorem 1. (Informal version)** *Fix any $k = O(1)$, and let $\mathbf{Z}$ be any unknown mixture of axis-aligned Gaussians over $\mathbf{R}^n$. There is an algorithm that, given samples from $\mathbf{Z}$ and any $\epsilon$, $\delta > 0$ as inputs, runs in time* $\mathrm{poly}(n/\epsilon) \cdot \log(1/\delta)$ *and with probability $1-\delta$ outputs a mixture $\mathbf{Z}'$ of $k$ axis-aligned Gaussians over $\mathbf{R}^n$ satisfying $KL(\mathbf{Z}||\mathbf{Z}') \leq \epsilon$.*

A signal feature of this result is that it requires no assumptions about the Gaussians being "separated" in space. It also has no dependence on the minimum mixing weight. We compare our result with other works on learning mixtures of Gaussians in the next section.

Our proof of Theorem 1 works by extending the basic approach for learning mixtures of product distributions over discrete domains from [8]. The main technical tool introduced in [8] is the WAM (Weights And Means) algorithm; the correctness proof of WAM is based on an intricate error analysis using ideas from the singular value theory of matrices. In this paper, we use this algorithm in a continuous domain to estimate the parameters of the Gaussian mixture. Dealing with this more complex class of distributions requires tackling a whole new set of issues around sampling error that did not exist in the discrete case.

Our results strongly suggest that the techniques introduced in [8] (and extended here) extend to PAC learning mixtures of other classes of product distributions, both discrete and continuous, such as exponential distributions or Poisson distributions. Though we have not explicitly worked out those extensions in this paper, we briefly discuss general conditions under which our techniques are applicable in Section 7.

### 1.2 Comparison with other frameworks for learning mixtures of Gaussians

There is a vast literature in statistics on modeling with mixture distributions, and on estimating the parameters of unknown such distributions from data. The case of mixtures of Gaussians is by far the most studied case; see, e.g., [13, 16] for surveys. Statistical work on mixtures of Gaussians has mainly focused on finding the distribution parameters (mixing weights, means, and variances) of *maximum likelihood*, given a set of data. Although one can write down equations whose solutions give these maximum likelihood values, solving the equations appears to be a computationally intractable problem. In particular, the most popular algorithm used for solving the equations, the *EM Algorithm* of Dempster et al. [7], has no efficiency guarantees and may run slowly or converge only to local optima on some instances.

A change in perspective led to the first provably efficient algorithm for learning: In 1999, Dasgupta [5] suggested learning in the *clustering* framework. In this scenario, the learner's goal is to group all the sample points according to which Gaussian in the mixture they came from. This is the strongest possible criterion for success one could demand; when the learner succeeds, it can easily recover accurate approximations of all parameters of the mixture distribution. However, a strong assumption is required to get such a strong outcome: it is clear that the learner cannot possibly succeed unless the Gaussians are guaranteed to be sufficiently "separated" in space. Informally, it must at least be the case that, with high probability, no sample point "looks like" it might have come from a different Gaussian in the mixture other than the one that actually generated it.

Dasgupta gave a polynomial time algorithm that could cluster a mixture of *spherical* Gaussians of *equal radius*. His algorithm required separation on the order of $n^{1/2}$ times the standard deviation. This was improved to $n^{1/4}$ by Dasgupta and Schulman [6], and this in turn was significantly generalized to the case of completely general (i.e., elliptical) Gaussians by Arora and Kannan [2]. Another breakthrough came from Vempala and Wang [18] who showed how the separation could be reduced, in the case of mixtures of $k$ spherical Gaussians (of different radii), to the order of $k^{1/4}$ times the standard deviation, times factors logarithmic in $n$. This result was extended to mixtures of general Gaussians (indeed, log-concave distributions) in works by Kannan et al. [11] and Achlioptas and McSherry [1], with some slightly worse separation requirements. It should also be mentioned that these results all have a running time dependence that is polynomial in $1/\pi_{\min}$, where $\pi_{\min}$ denotes the minimum mixing weight.

Our work gives another learning perspective that allows us to deal with mixtures of Gaussians that satisfy *no* separation assumption. In this case clustering is simply not possible; for any data set, there may be many different mixtures of Gaussians under which the data are plausible. This possibility also leads to the seeming intractability of finding the *maximum* likelihood mixture of Gaussians. Nevertheless, we feel that this case is both interesting and important, and that under these circumstances identifying *some* mixture of Gaussians which is statistically indistinguishable from the true mixture is a worthy task. This is precisely what the PAC-style learning scenario we work in requires, and what our main algorithm efficiently achieves.

Reminding the reader that they work in significantly different scenarios, we end this section with a comparison between other aspects of our algorithm and algorithms in the clustering model. Our algorithm works for mixtures of axis-aligned Gaussians. This is stronger than the case of spherical Gaussians considered in [5, 6, 18], but weaker than the case of general Gaussians handled in [2, 11, 1]. On the other hand, in Section 7 we discuss the fact that our methods should be readily adaptable to mixtures of a wide variety of discrete and continuous distributions — essentially, any distribution where the "method of moments" from statistics succeeds. The clustering algorithms discussed have polynomial running time dependence on $k$, the number of mixture components, whereas our algorithm's running time is polynomial in $n$ only if $k$ is a constant. We note that in [8], strong evidence was given that (for the PAC-style learning problem that we consider) such a dependence is unavoidable at least in the case of learning mixtures of product distributions on the Boolean cube. Finally, unlike the clustering algorithms mentioned, our algorithm has no running time dependence on $1/\pi_{\min}$.

### 1.3 Overview of the approach and the paper

An important ingredient of our approach is a slight extension of the WAM algorithm, the main technical tool introduced in [8]. The algorithm takes as input a parameter $\epsilon > 0$ and samples from an unknown mixture $\mathbf{Z}$ of $k$ product distributions $\mathbf{X}^1, \ldots, \mathbf{X}^k$ over $\mathbf{R}^n$. The output of the algorithm is a list of candidate descriptions of the $k$ mixing weights and $kn$ coordinate means of the distributions $\mathbf{X}^1, \ldots, \mathbf{X}^k$. Roughly speaking, the guarantee for the algorithm proved in [8] is that with high probability at least one of the candidate descriptions that the algorithm outputs is "good" in the following sense: it is an additive $\epsilon$-accurate approximation to each of the $k$ true mixing weights $\pi^1, \ldots, \pi^k$ and to each of the true coordinate means $\mu_j^i = \mathbf{E}[\mathbf{X}_j^i]$ for which the corresponding mixing weight $\pi^i$ is not too small. We give a precise specification in Section 3.

As described above, when WAM is run on a mixture distribution it generates candidate estimates of mixing weights and means. However, to describe a Gaussian we need not only its mean but also its variance. To achieve this we run WAM *twice*, once on $\mathbf{Z}$ and once on what might be called "$\mathbf{Z}^2$" — i.e., for the second run, each time a draw $(z_1, \ldots, z_n)$ is obtained from $\mathbf{Z}$ we convert it to $(z_1^2, \ldots, z_n^2)$ and use that instead. It is easy to see that $\mathbf{Z}^2$ corresponds to a mixture of the distributions $(\mathbf{X}^1)^2, \ldots, (\mathbf{X}^k)^2$, and thus this second run gives us estimates of the mixing weights (again) and also of the coordinate *second moments* $\mathbf{E}[(\mathbf{X}_j^i)^2]$. Having thus run WAM twice, we essentially take the "cross-product" of the two output lists to obtain a list of candidate descriptions, each of which specifies mixing weights, means, and second moments of the component Gaussians. In Section 4 we give a detailed description of this process and prove that with high probability at least one of the resulting candidates is a "good" description (in the sense of the preceding paragraph) of the mixing weights, coordinate means, and coordinate variances of the Gaussians $\mathbf{X}^1, \ldots, \mathbf{X}^k$.

To actually PAC learn the distribution $\mathbf{Z}$, we must find this good description among the candidates in the list. A natural idea is to apply some sort of maximum likelihood procedure. However, to make this work, we need to guarantee that the list contains a distribution that is close to the target in the sense of KL divergence. Thus, in Section 5, we show how to convert each "parametric" candidate description into a mixture of Gaussians such that any additively accurate description indeed becomes a mixture distribution with close KL divergence to the unknown target. (This procedure also guarantees that the candidate distributions satisfy some other technical conditions that are needed by the maximum likelihood procedure.) Finally, in Section 6 we put the pieces together and show how a maximum likelihood procedure can be used to identify a hypothesis mixture of Gaussians that has small KL divergence relative to the target mixture.

## 2 Preliminaries

**The PAC learning framework for probability distributions.** We work in the Probably Approximately Correct model of learning probability distributions which was proposed by Kearns et al. [12]. In this framework the learning algorithm is given access to samples drawn from the target distribution $\mathbf{Z}$ to be learned, and the learning algorithm must (with high probability) output an accurate approximation $\mathbf{Z}'$ of the target distribution $\mathbf{Z}$. Following [12], we use the *Kullback-Leibler (KL) divergence* (also known as the *relative entropy*) as our notion of distance. The KL divergence between distributions $\mathbf{Z}$ and $\mathbf{Z}'$ is

$$\mathrm{KL}(\mathbf{Z}||\mathbf{Z}') := \int \mathbf{Z}(x)\ln(\mathbf{Z}(x)/\mathbf{Z}'(x))\,dx$$

where here we have identified the distributions with their pdfs. The reader is reminded that KL divergence is not symmetric and is thus not a metric. KL divergence is a stringent measure of the distance between probability distances. In particular, it holds [3] that $0 \le \|\mathbf{Z}-\mathbf{Z}'\|_2 \le (2\ln 2)\sqrt{\mathrm{KL}(\mathbf{Z}||\mathbf{Z}')}$, where $\|\cdot\|_1$ denotes total variation distance; hence if the KL divergence is small then so is the total variation distance.

We make the following formal definition:

**Definition 1.** *Let $\mathcal{D}$ be a class of probability distributions over $\mathbf{R}^n$. An efficient (proper) learning algorithm for $\mathcal{D}$ is an algorithm which, given $\epsilon$, $\delta > 0$ and samples drawn from any distribution $\mathbf{Z} \in \mathcal{D}$, runs in $\mathrm{poly}(n, 1/\epsilon, 1/\delta)$ time and, with probability at least $1-\delta$, outputs a representation of a distribution $\mathbf{Z}' \in \mathcal{D}$ such that $\mathrm{KL}(\mathbf{Z}||\mathbf{Z}') \le \epsilon$.*

**Mixtures of axis-aligned Gaussians.** Here we recall some basic definitions and establish useful notational conventions for later.

A Gaussian distribution over $\mathbf{R}$ with mean $\mu$ and variance $\sigma$ has probability density function $f(x) = (1/\sqrt{2\pi}\sigma)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. An *axis-aligned* Gaussian over $\mathbf{R}^n$ is a product distribution over $n$ univariate Gaussians.

If we expect to learn a mixture of Gaussians, we need each Gaussian to have reasonable parameters in each of its coordinates. Indeed, consider just the problem of learning the parameters of a single one-dimensional Gaussian: If the variance is enormous, we could not expect to estimate the mean efficiently; or, if the variance was extremely close to 0, any slight error in the hypothesis would lead to a severe penalty in KL divergence. These issues motivate the following definition:

**Definition 2.** *We say that* $\mathbf{X}$ *is a* $d$-dimensional $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussian *if* $\mathbf{X}$ *is a* $d$*-dimensional axis-aligned Gaussian with the property that each of its one-dimensional coordinate Gaussians* $\mathbf{X}_j$ *has mean* $\mu_j \in [-\mu_{\max}, \mu_{\max}]$ *and variance* $(\sigma_j)^2 \in [\sigma^2_{\min}, \sigma^2_{\max}]$.

**Notational convention:** *Throughout the rest of the paper all Gaussians we consider are* $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$*-bounded, where for notational convenience we assume that the numbers* $\mu_{\max}$, $\sigma^2_{\max}$ *are at least 1 and that the number* $\sigma^2_{\min}$ *is at most* 1. *We will denote by* $L$ *the quantity* $\mu_{\max}\sigma_{\max}/\sigma_{\min}$*, which in some sense measures the bit-complexity of the problem. Given distributions* $\mathbf{X}^1, \dots, \mathbf{X}^k$ *over* $\mathbf{R}^n$*, we write* $\mu^i_j$ *to denote* $\mathbf{E}[\mathbf{X}^i_j]$*, the* $j$*-th coordinate mean of the* $i$*-th component distribution, and we write* $(\sigma^i_j)^2$ *to denote* $\mathrm{Var}[\mathbf{X}^i_j]$*, the variance in coordinate* $j$ *of the* $i$*-th distribution.*

A mixture of $k$ axis-aligned Gaussians $\mathbf{Z} = \pi_1\mathbf{X}^1 + \cdots + \pi_k\mathbf{X}^k$ is completely specified by the parameters $\pi^i$, $\mu^i_j$, and $(\sigma^i_j)^2$. Our learning algorithm for Gaussians will have a running time that depends polynomially on $L$; thus the algorithm is not strongly polynomial.

## 3 Listing candidate weights and means with WAM

We first recall the basic features of the WAM algorithm from [8] and then explain the extension we require. The algorithm described in [8] takes as input a parameter $\epsilon > 0$ and samples from an unknown mixture $\mathbf{Z}$ of $k$ distributions $\mathbf{X}^1, \dots, \mathbf{X}^k$ where each $\mathbf{X}^i = (\mathbf{X}^i_1, \dots, \mathbf{X}^i_n)$ is assumed to be a product distribution over the bounded domain $[-1, 1]^n$. The goal of WAM is to output accurate estimates for the mixing weights $\pi^i$ and coordinate means $\mu^i_j$; what the algorithm actually outputs is a list of candidate "parametric descriptions" of the means and mixing weights, where each candidate description is of the form $(\{\hat\pi^1, \dots, \hat\pi^k\}, \{\hat\mu^1_1, \hat\mu^1_2, \dots, \hat\mu^k_n\})$.

We now explain the notion of a "good" estimate of parameters from Section 1.3 in more detail. As motivation, note that if a mixing weight $\pi^i$ is very low then the WAM algorithm (or indeed any algorithm that only draws a limited number of samples from $\mathbf{Z}$) may not receive any samples from $\mathbf{X}^i$, and thus we would not expect WAM to construct an accurate estimate for the coordinate means $\mu^i_1, \dots, \mu^i_n$. We thus have the following definition from [8]:

**Definition 3.** *A candidate* $(\{\hat\pi^1, \dots, \hat\pi^k\}, \{\hat\mu^1_1, \hat\mu^1_2, \dots, \hat\mu^k_n\})$ *is said to be* parametrically $\epsilon$-accurate *if:*

1. $|\hat\pi^i - \pi^i| \le \epsilon$ *for all* $1 \le i \le k$*;*
2. $|\hat\mu^i_j - \mu^i_j| \le \epsilon$ *for all* $1 \le i \le k$ *and* $1 \le j \le n$ *such that* $\pi^i \ge \epsilon$.

Very roughly speaking, the WAM algorithm in [8] works by exhaustively "guessing" (to a certain prescribed granularity that depends on $\epsilon$) values for the mixing weights and for $k^2$ of the $kn$ coordinate means. Given a guess, the algorithm tries to approximately solve for the remaining $k(n-k)$ coordinate means using the guessed values and the sample data; in the course of doing this the algorithm uses estimates of the expectations $\mathbf{E}[\mathbf{Z}_j \mathbf{Z}_{j'}]$ that are obtained from the sample data. From each guess the algorithm thus obtains one of the candidates in the list that it ultimately outputs.

The assumption [8] that each distribution $\mathbf{X}^i$ in the mixture is over $[-1,1]^n$ has two nice consequences: each coordinate mean need only be guessed within a bounded domain $[-1,1]$, and estimating $\mathbf{E}[\mathbf{Z}_j \mathbf{Z}_{j'}]$ is easy for a mixture $\mathbf{Z}$ of such distributions. Inspection of the proof of correctness of the WAM algorithm shows that these two conditions are all that is really required. We thus introduce the following:

**Definition 4.** *Let $\mathbf{X}$ be a distribution over $\mathbf{R}$. We say that $\mathbf{X}$ is* $\lambda(\epsilon,\delta)$-samplable *if there is an algorithm $\mathcal{A}$ which, given access to draws from $\mathbf{X}$, runs for $\lambda(\epsilon,\delta)$ steps and outputs (with probability at least $1-\delta$ over the draws from $\mathbf{X}$) a quantity $\hat{\mu}$ satisfying $|\hat{\mu} - \mathbf{E}[\mathbf{X}]| \leq \epsilon$.*

With this definition in hand an obvious (slight) generalization of WAM, which we denote WAM$'$, suggests itself. The main result about WAM$'$ that we need is the following (the proof is essentially identical to the proof in [8] so we omit it):

**Theorem 2.** *Let $\mathbf{Z}$ be a mixture of product distributions $\mathbf{X}^1, \ldots, \mathbf{X}^k$ with mixing weights $\pi^1, \ldots, \pi^k$ where each $\mu^i_j = \mathbf{E}[\mathbf{X}^i_j]$ satisfies $|\mu^i_j| \leq U$ and $\mathbf{Z}_j \mathbf{Z}_{j'}$ is* $\mathrm{poly}(U/\epsilon) \cdot \log(1/\delta)$*-samplable for all $j \neq j'$. Given $U$ and any $\epsilon, \delta > 0$,* WAM$'$ *runs in time $(nU/\epsilon)^{O(k^3)} \cdot \log(1/\delta)$ and outputs a list of $(nU/\epsilon)^{O(k^3)}$ many candidates descriptions, at least one of which (with probability at least $1-\delta$) is parametrically $\epsilon$-accurate.*

## 4 Listing candidate weights, means, and variances

Through the rest of the paper we assume that $\mathbf{Z}$ is a $k$-wise mixture of independent $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussians $\mathbf{X}^1, \ldots, \mathbf{X}^k$, as discussed in Section 2. Recall also the notation $L$ from that section.

As described in Section 1.3, we will run WAM$'$ twice, once on the original mixture of Gaussians $\mathbf{Z}$ and once on the squared mixture $\mathbf{Z}^2$. In order to do this, we must show that both $\mathbf{Z} = \pi_1 \mathbf{X}^1 + \cdots + \pi_k \mathbf{X}^k$ and $\mathbf{Z}^2 = \pi_1 (\mathbf{X}^1)^2 + \cdots + \pi_k (\mathbf{X}^k)^2$ satisfy the conditions of Theorem 2. The bound $|\mu^i_j| \leq \mu_{\max}$ on coordinate means is satisfied by assumption for $\mathbf{Z}$, and for $\mathbf{Z}^2$ we have that each $\mathbf{E}[(\mathbf{X}^i_j)^2]$ is at most $\sigma^2_{\max} + \mu^2_{\max}$. It remains to verify the required samplability condition on products of two coordinates for both $\mathbf{Z}$ and $\mathbf{Z}^2$; i.e. we must show that both the random variables $\mathbf{Z}_j \mathbf{Z}_{j'}$ are samplable and that the random variables $\mathbf{Z}^2_j \mathbf{Z}^2_{j'}$ are samplable. We do this in the following proposition, whose straightforward but technical proof appears in Appendix B:

**Proposition 1.** *Suppose* $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ *is the mixture of* $k$ *two-dimensional* $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$*-bounded Gaussians. Then both the random variable* $\mathbf{W} := \mathbf{Z}_1\mathbf{Z}_2$ *and the random variable* $\mathbf{W}^2$ *are* $\mathrm{poly}(L/\epsilon) \cdot \log(1/\delta)$*-samplable.*

The proof of the following theorem explains precisely how we can run WAM$'$ twice and how we can combine the two resulting lists (one containing candidate descriptions consisting of mixing weights and coordinate means, the other containing candidate descriptions consisting of mixing weights and coordinate second moments) to obtain a single list of candidate descriptions consisting of mixing weights, coordinate means, and coordinate variances.

**Theorem 3.** *Let* $\mathbf{Z}$ *be a mixture of* $k = O(1)$ *axis-aligned Gaussians* $\mathbf{X}^1, \ldots, \mathbf{X}^k$ *over* $\mathbf{R}^n$*, described by parameters* $(\{\pi^i\}, \{\mu^i_j\}, \{\sigma^i_j\})$*. There is an algorithm with the following property: For any* $\epsilon$*,* $\delta > 0$*, given samples from* $\mathbf{Z}$ *the algorithm runs in* $\mathrm{poly}(nL/\epsilon) \cdot \log(1/\delta)$ *time and with probability* $1-\delta$ *outputs a list of* $\mathrm{poly}(nL/\epsilon)$ *many candidates* $(\{\hat{\pi}^i\}, \{\hat{\mu}^i_j\}, \{\hat{\sigma}^i_j\})$ *such that for at least one candidate in the list, the following holds:*

1. $|\hat{\pi}^i - \pi^i| \le \epsilon$ *for all* $i \in [k]$*; and*
2. $|\hat{\mu}^i_j - \mu^i_j| \le \epsilon$ *and* $|(\hat{\sigma}^i_j)^2 - (\sigma^i_j)^2| \le \epsilon$ *for all* $i, j$ *such that* $\pi^i \ge \epsilon$*.*

*Proof.* First run the algorithm WAM$'$ with the random variable $\mathbf{Z}$, taking the parameter "$U$" in WAM$'$ to be $L$, taking "$\delta$" to be $\delta/2$, and taking "$\epsilon$" to be $\epsilon/(6\mu_{\max})$. By Proposition 1 and Theorem 2, this takes at most the claimed running time. WAM$'$ outputs a list List1 of candidate descriptions for the mixing weights and expectations, List1 $= [\ldots, (\hat{\pi}^i, \hat{\mu}^i_j), \ldots]$, which with probability at least $1 - \delta/2$ contains at least one candidate description which is parametrically $\epsilon/(6\mu_{\max})$-accurate.

Define $(s^i_j)^2 = \mathbf{E}[(\mathbf{X}^i_j)^2] = (\sigma^i_j)^2 + (\mu^i_j)^2$. Run the algorithm WAM$'$ again on the squared random variable $\mathbf{Z}^2$, with "$U$" $= \sigma^2_{\max} + \mu^2_{\max}$, "$\delta$" $= \delta/2$, and "$\epsilon$" $= \epsilon/2$. By Proposition 1, this again takes at most the claimed running time. This time WAM$'$ outputs a list List2 of candidates for the mixing weights (again) and second moments, List2 $= [\ldots, (\hat{\hat{\pi}}^i, (\hat{s}^i_j)^2) \ldots]$, which with probability at least $1 - \delta/2$ has a "good" entry which satisfies

1. $|\hat{\hat{\pi}}^i - \pi^i| \le \epsilon/2$ for all $i = 1 \ldots k$; and
2. $|(\hat{s}^i_j)^2 - (s^i_j)^2| \le \epsilon/2$ for all $i, j$ such that $\pi^i \ge \epsilon/2$.

We now form the "cross product" of the two lists. (Again, this can be done in the claimed running time.) Specifically, for each pair consisting of a candidate $(\hat{\pi}^i, \hat{\mu}^i_j)$ in List1 and a candidate $(\hat{\hat{\pi}}^i, (\hat{s}^i_j)^2)$ in List2, we form a new candidate consisting of mixing weights, means, and variances, namely $(\hat{\pi}^i, \hat{\mu}^i_j, (\hat{\sigma}^i_j)^2)$ where $(\hat{\sigma}^i_j)^2 = (\hat{s}^i_j)^2 - (\hat{\mu}^i_j)^2$. (Note that we simply discard $\hat{\hat{\pi}}^i$.)

When the "good" candidate from List1 is matched with the "good" candidate from List2, the resulting candidate's mixing weights and means satisfy the desired bounds. For the variances, we have that $|(\hat{\sigma}^i_j)^2 - (\sigma^i_j)^2|$ is at most

$$|(\hat{s}^i_j)^2 - (s^i_j)^2| + |(\hat{\mu}^i_j)^2 - (\mu^i_j)^2| \le \frac{\epsilon}{2} + |\hat{\mu}^i_j - \mu^i_j| \cdot |\hat{\mu}^i_j + \mu^i_j| \le \frac{\epsilon}{2} + \frac{\epsilon}{6\mu_{\max}} \cdot 3\mu_{\max} = \epsilon.$$

This proves the theorem.

## 5 From parametric estimates to bona fide distributions

At this point we have a list of candidate "parametric" descriptions $(\{\hat{\pi}^i\}, \{\hat{\mu}_j^i\}, \{(\hat{\sigma}_j^i)^2\})$ of mixtures of Gaussians, at least one of which is parametrically accurate in the sense of Theorem 3. In Section 5.1 we describe an efficient way to convert any parametric description into a true mixture of Gaussians such that:

(i) any parametrically accurate description becomes a distribution with close KL divergence to the target distribution; and
(ii) every mixture distribution that results from the conversion has a pdf that satisfies certain upper and lower bounds (that will be required for the maximum likelihood procedure).

The conversion procedure is conceptually straightforward — it essentially just truncates any extreme parameters to put them in a "reasonable" range — but the details establishing correctness are fairly technical. By applying this conversion to each of the parametric descriptions in our list from Section 4, we obtain a list of mixture distribution hypotheses all of which have bounded pdfs and at least one of which is close to the target $\mathbf{Z}$ in KL divergence (see Section 5.2). With such a list in hand, we will be able to use maximum likelihood (in Section 6) to identify a single hypothesis which is close in KL divergence.

### 5.1 The conversion procedure

In this section we prove:

**Theorem 4.** *There is a simple efficient procedure $\mathcal{A}$ which takes values $(\{\hat{\pi}^i\}, \{\hat{\mu}_j^i\}, \{(\hat{\sigma}_j^i)^2\})$ and a value $M > \mu_{\max}$ as inputs and outputs a true mixture $\dot{\mathbf{Z}}$ of $k$ many $n$-dimensional $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussians with mixing weights $\dot{\pi}^1, \ldots, \dot{\pi}^k$ satisfying*

*(a)* $\sum_{i=1}^k \dot{\pi}^i = 1$, *and*
*(b)* $\alpha_0 \le \dot{\mathbf{Z}}(x) \le \beta_0$ *for all* $x \in [-M, M]^n$,

*where* $\alpha_0 := \left[\frac{1}{\sqrt{2\pi}\sigma_{\max}} \cdot \exp\left(\frac{-2M^2}{\sigma^2_{\min}}\right)\right]^n$ *and* $\beta_0 := 1/(\sqrt{2\pi}\sigma_{\min})^n$.

*Furthermore, suppose $\mathbf{Z}$ is a mixture of Gaussians $\mathbf{X}^1, \ldots, \mathbf{X}^k$ with mixing weights $\pi^i$, means $\mu_j^i$, and variances $(\sigma_j^i)^2$ and that the following are satisfied:*

*(c) for $i = 1 \ldots k$ we have $|\pi^i - \hat{\pi}^i| \le \epsilon_{\mathrm{wts}}$ where $\epsilon_{\mathrm{wts}} \le 1/(12k)^3$; and*
*(d) for all $i, j$ such that $\pi^i \ge \epsilon_{\mathrm{minwt}}$ we have $|\mu_j^i - \hat{\mu}_j^i| \le \epsilon_{\mathrm{means}}$ and $|(\sigma_j^i)^2 - (\hat{\sigma}_j^i)^2| \le \epsilon_{\mathrm{vars}}$.*

*Then $\dot{\mathbf{Z}}$ will satisfy* $\mathrm{KL}(\mathbf{Z}||\dot{\mathbf{Z}}) \le \eta(\epsilon_{\mathrm{means}}, \epsilon_{\mathrm{vars}}, \epsilon_{\mathrm{wts}}, \epsilon_{\mathrm{minwt}})$, *where*

$$\eta(\epsilon_{\mathrm{means}}, \epsilon_{\mathrm{vars}}, \epsilon_{\mathrm{wts}}, \epsilon_{\mathrm{minwt}}) := n \cdot \left(\frac{\epsilon_{\mathrm{vars}}}{2\sigma^2_{\min}} + \frac{\epsilon^2_{\mathrm{means}} + \epsilon_{\mathrm{vars}}}{2(\sigma^2_{\min} - \epsilon_{\mathrm{vars}})}\right) + k\epsilon_{\mathrm{minwt}} \cdot n \cdot \left(\frac{\sigma^2_{\max} + 2\mu^2_{\max}}{\sigma^2_{\min}}\right) + 13k\epsilon_{\mathrm{wts}}^{1/3}.$$

*Proof.* We construct a mixture $\dot{\mathbf{Z}}$ of product distributions $\dot{\mathbf{X}}^1, \ldots, \dot{\mathbf{X}}^k$ by defining new mixing weights $\dot{\pi}^i$, expectations $\dot{\mu}^i_j$, and variances $(\dot{\sigma}^i_j)^2$. The procedure $\mathcal{A}$ is defined as follows:

1. For all $i, j$, set

$$\dot{\mu}^i_j = \begin{cases} -\mu_{\max} & \text{if } \hat{\mu}^i_j < -\mu_{\max} \\ \mu_{\max} & \text{if } \hat{\mu}^i_j > \mu_{\max} \\ \hat{\mu}^i_j & \text{o.w.} \end{cases} \qquad \text{and} \qquad \dot{\sigma}^i_j = \begin{cases} \sigma_{\min} & \text{if } \hat{\sigma}^i_j < \sigma_{\min} \\ \sigma_{\max} & \text{if } \hat{\sigma}^i_j > \sigma_{\max} \\ \hat{\sigma}^i_j & \text{o.w.} \end{cases}$$

2. For all $i = 1, \ldots, k$ let $\ddot{\pi}^i = \begin{cases} \hat{\pi}^i & \text{if } \hat{\pi}^i \geq \epsilon_{\text{wts}} \\ \epsilon_{\text{wts}} & \text{if } \hat{\pi}^i < \epsilon_{\text{wts}}. \end{cases}$

   Let $s$ be such that $s \sum_{i=1}^k \ddot{\pi}^i = 1$. Take $\dot{\pi}^i = s\ddot{\pi}^i$. (This is just a normalization so the mixing weights sum to precisely 1.)

It is clear from this construction that condition (a) is satisfied. For (b), the bounds on $\dot{\sigma}^i_j$ are easily seen to imply that $\dot{\mathbf{X}}^i(x) \leq 1/(\sqrt{2\pi}\sigma_{\min})^n =: \beta_0$ for all $x \in \mathbf{R}^n$, and hence the same upper bound holds for the mixture $\dot{\mathbf{Z}}(x)$, being a convex combination of the values $\dot{\mathbf{X}}^i(x)$. Similarly, using the fact that $M \geq \mu_{\max}$ together with the bounds on $\dot{\mu}^i_j$ and $\dot{\sigma}^i_j$, we have that $\dot{\mathbf{X}}^i(x) \geq \left[\frac{1}{\sqrt{2\pi}\sigma_{\max}} \cdot \exp\left(\frac{-2M^2}{\sigma^2_{\min}}\right)\right]^n =: \alpha_0$, for all $x \in [-M, M]^n$, and this lower bound holds for $\dot{\mathbf{Z}}(x)$ as well.

We now prove the second half of the theorem; so suppose that conditions (c) and (d) hold. Our goal is to apply the following proposition (proved in [8]) to bound $\mathrm{KL}(\mathbf{Z}||\dot{\mathbf{Z}})$:

**Proposition 2.** *Let $\pi^1, \ldots, \pi^k$, $\gamma^1, \ldots, \gamma^k \geq 0$ be mixing weights satisfying $\sum \pi^i = \sum \gamma^i = 1$. Let $\mathcal{I} = \{i : \pi^i \geq \epsilon_3\}$. Let $\mathbf{P}^1, \ldots, \mathbf{P}^k$ and $\mathbf{Q}^1, \ldots, \mathbf{Q}^k$ be distributions. Suppose that*

1. $|\pi^i - \gamma^i| \leq \epsilon_1$ *for all* $i \in [k]$;
2. $\gamma^i \geq \epsilon_2$ *for all* $i \in [k]$;
3. $\mathrm{KL}(\mathbf{P}^i||\mathbf{Q}^i) \leq \epsilon_{\mathcal{I}}$ *for all* $i \in \mathcal{I}$;
4. $\mathrm{KL}(\mathbf{P}^i||\mathbf{Q}^i) \leq \epsilon_{\text{all}}$ *for all* $i \in [k]$.

*Then, letting $\mathbf{P}$ denote the $\pi$-mixture of the $\mathbf{P}^i$'s and $\mathbf{Q}$ the $\gamma$-mixture of the $\mathbf{Q}^i$'s, for any $\epsilon_4 > \epsilon_1$ we have* $\mathrm{KL}(\mathbf{P}||\mathbf{Q}) \leq \epsilon_{\mathcal{I}} + k\epsilon_3\epsilon_{\text{all}} + k\epsilon_4 \ln\frac{\epsilon_4}{\epsilon_2} + \frac{\epsilon_1}{\epsilon_4 - \epsilon_1}$.

More precisely, our goal is to apply this proposition with parameters

$$\epsilon_1 = 3k\epsilon_{\text{wts}}; \quad \epsilon_2 = \epsilon_{\text{wts}}/2; \quad \epsilon_3 = \epsilon_{\text{minwt}}; \quad \epsilon_{\mathcal{I}} = n \cdot \left(\frac{\epsilon_{\text{vars}}}{2\sigma^2_{\min}} + \frac{\epsilon^2_{\text{means}} + \epsilon_{\text{vars}}}{2(\sigma^2_{\min} - \epsilon_{\text{vars}})}\right);$$
$$\epsilon_{\text{all}} = n \cdot \left(\frac{\sigma^2_{\max} + 2\mu^2_{\max}}{\sigma^2_{\min}}\right); \quad \epsilon_4 = \epsilon^{2/3}_{\text{wts}}/2.$$

To satisfy the conditions of the proposition, we must (1) upper bound $|\pi^i - \dot{\pi}^i|$ for all $i$; (2) lower bound $\dot{\pi}^i$ for all $i$; (3) upper bound $\mathrm{KL}(\mathbf{X}^i||\dot{\mathbf{X}}^i)$ for all $i$ such that $\pi^i \geq \epsilon_{\text{minwt}}$; and (4) upper bound $\mathrm{KL}(\mathbf{X}^i||\dot{\mathbf{X}}^i)$ for all $i$. We now do this.

**(1) Upper bounding $|\pi^i - \dot{\pi}^i|$.** A straightforward argument given in [8] shows that assuming $\epsilon_{\text{wts}} \le 1/(2k)$, we get $|\pi^i - \dot{\pi}^i| \le 3k\epsilon_{\text{wts}}$.

**(2) Lower bounding $\dot{\pi}^i$.** In [8] it is also shown that $\dot{\pi}^i \ge \frac{\epsilon_{\text{wts}}}{2}$ assuming that $\epsilon_{\text{wts}} \le 1/k$.

**(3) Upper bounding $\mathrm{KL}(\mathbf{X}^i||\dot{\mathbf{X}}^i)$ for all $i$ such that $\pi^i \ge \epsilon_{\text{minwt}}$.** Fix an $i$ such that $\pi^i \ge \epsilon_{\text{minwt}}$ and fix any $j \in [n]$. Consider some particular $\mu^i_j$ and $\dot{\mu}^i_j$ and $\sigma^i_j$ and $\dot{\sigma}^i_j$, so we have $|\mu^i_j - \hat{\mu}^i_j| \le \epsilon_{\text{means}}$ and $|(\sigma^i_j)^2 - (\hat{\sigma}^i_j)^2| \le \epsilon_{\text{vars}}$. Since $|\mu^i_j| \le \mu_{\max}$, by the definition of $\dot{\mu}^i_j$ we have that $|\mu^i_j - \dot{\mu}^i_j| \le \epsilon_{\text{means}}$, and likewise we have $|(\sigma^i_j)^2 - (\dot{\sigma}^i_j)^2| \le \epsilon_{\text{vars}}$. Let $\mathbf{P}$ and $\mathbf{Q}$ be the one-dimensional Gaussians with means $\mu^i_j$ and $\dot{\mu}^i_j$ and variances $\sigma^i_j$ and $\dot{\sigma}^i_j$ respectively. By Corollary 1, we have

$$\mathrm{KL}(\mathbf{P}||\mathbf{Q}) \le \frac{\epsilon_{\text{vars}}}{2\sigma^2_{\min}} + \frac{\epsilon^2_{\text{means}} + \epsilon_{\text{vars}}}{2(\sigma^2_{\min} - \epsilon_{\text{vars}})}.$$

Each $\dot{\mathbf{X}}^i$ is the product of $n$ such Gaussians. Since KL divergence is additive for product distributions (see Proposition 4) we have the following bound for each $i$ such that $\pi^i \ge \epsilon_{\text{minwt}}$:

$$\mathrm{KL}(\mathbf{X}^i||\dot{\mathbf{X}}^i) \le n \cdot \left( \frac{\epsilon_{\text{vars}}}{2\sigma^2_{\min}} + \frac{\epsilon^2_{\text{means}} + \epsilon_{\text{vars}}}{2(\sigma^2_{\min} - \epsilon_{\text{vars}})} \right).$$

**(4) Upper bounding $\mathrm{KL}(\mathbf{X}^i||\dot{\mathbf{X}}^i)$ for all $i \in [k]$.** Using the fact that both $\mathbf{X}^i$ and $\dot{\mathbf{X}}^i$ are $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded, it follows from Fact 8 and Proposition 4 that we have

$$\mathrm{KL}(\mathbf{X}^i||\dot{\mathbf{X}}^i) \le n \left( \frac{\sigma^2_{\max} + 2\mu^2_{\max}}{\sigma^2_{\min}} \right).$$

Proposition 2 now gives us

$$\mathrm{KL}(\mathbf{Z}||\dot{\mathbf{Z}}) \le n \cdot \left( \frac{\epsilon_{\text{vars}}}{2\sigma^2_{\min}} + \frac{\epsilon^2_{\text{means}} + \epsilon_{\text{vars}}}{2(\sigma^2_{\min} - \epsilon_{\text{vars}})} \right) + k\epsilon_{\text{minwt}} \cdot n \cdot \left( \frac{\sigma^2_{\max} + 2\mu^2_{\max}}{\sigma^2_{\min}} \right) + R,$$

where $R = k\epsilon_4 \ln \frac{\epsilon_4}{\epsilon_2} + \frac{\epsilon_1}{\epsilon_4 - \epsilon_1} = \frac{k}{2}\epsilon^{2/3}_{\text{wts}} \ln(\epsilon^{-1/3}_{\text{wts}}) + \frac{3k\epsilon_{\text{wts}}}{\epsilon^{2/3}_{\text{wts}}/2 - 3k\epsilon_{\text{wts}}}$. Using the fact that $\ln x \le x^{1/2}$ for $x > 1$, the first of these two terms is at most $\frac{k}{2}\epsilon^{1/2}_{\text{wts}}$. Using the fact that $\epsilon_{\text{wts}} < 1/(12k)^3$, the second of these terms is at most $12k\epsilon^{1/3}_{\text{wts}}$. So $R$ is at most $13k\epsilon^{1/3}_{\text{wts}}$ and the theorem is proved.

### 5.2 Getting a list of distributions one of which is KL-close to the target

In this section we show that combining the conversion procedure from the previous subsection with the results of Section 4 lets us obtain the following:

**Theorem 5.** *Let $\mathbf{Z}$ be any unknown mixture of $k = O(1)$ axis-aligned Gaussians over $\mathbf{R}^n$. There is an algorithm with the following property: for any $\epsilon, \delta > 0$, given samples from $\mathbf{Z}$ the algorithm runs in $\mathrm{poly}(nL/\epsilon) \cdot \log(1/\delta)$ time and with probability $1 - \delta$ outputs a list of $\mathrm{poly}(nL/\epsilon)$ many mixtures of Gaussians with the following properties:*

1. *For any $M > \mu_{\max}$ such that $M = \mathrm{poly}(nL/\epsilon)$, every distribution $\mathbf{Z}'$ in the list satisfies $\exp(-\mathrm{poly}(nL/\epsilon)) \le \mathbf{Z}'(x) \le \mathrm{poly}(L)^n$ for all $x \in [-M, M]^n$.*
2. *Some distribution $\mathbf{Z}^\star$ in the list satisfies $\mathrm{KL}(\mathbf{Z}||\mathbf{Z}^\star) \le \epsilon$.*

Note that Theorem 5 guarantees that $\mathbf{Z}'(x)$ has bounded mass only on the range $[-M, M]^n$, whereas the support of $\mathbf{Z}$ goes beyond this range. This issue is addressed in the proof of Theorem 7, where we put together Theorem 5 and the maximum likelihood procedure.

**Proof of Theorem 5:** We will use a specialization of Theorem 3 in which we have different parameters for the different roles that $\epsilon$ plays:

**Theorem 3′** *Let $\mathbf{Z}$ be a mixture of $k = O(1)$ axis-aligned Gaussians $\mathbf{X}^1, \dots, \mathbf{X}^k$ over $\mathbf{R}^n$, described by parameters $(\{\pi^i\}, \{\mu^i_j\}, \{\sigma^i_j\})$. There is an algorithm with the following property: for any $\epsilon_{\mathrm{means}}, \epsilon_{\mathrm{vars}}, \epsilon_{\mathrm{wts}}, \epsilon_{\mathrm{minwt}}, \delta > 0$, given samples from $\mathbf{Z}$, with probability $1 - \delta$ it outputs a list of candidates $(\{\hat\pi^i\}, \{\hat\mu^i_j\}, \{\hat\sigma^i_j\})$ such that for at least one candidate in the list, the following holds:*

1. *$|\hat\pi^i - \pi^i| \le \epsilon_{\mathrm{wts}}$ for all $i \in [k]$; and*
2. *$|\hat\mu^i_j - \mu^i_j| \le \epsilon_{\mathrm{means}}$ and $|(\hat\sigma^i_j)^2 - (\sigma^i_j)^2| \le \epsilon_{\mathrm{vars}}$ for all $i, j$ such that $\pi^i \ge \epsilon_{\mathrm{minwt}}$.*

*The algorithm runs in time $\mathrm{poly}(nL/\epsilon')\cdot\log(1/\delta)$ where $\epsilon' = \min\{\epsilon_{\mathrm{wts}}, \epsilon_{\mathrm{means}}, \epsilon_{\mathrm{vars}}, \epsilon_{\mathrm{minwt}}\}$.*

Let $\epsilon, \delta > 0$ be given. We run the algorithm of Theorem 3′ with parameters $\epsilon_{\mathrm{means}} = \frac{\epsilon\sigma^2_{\min}}{12n}$, $\epsilon_{\mathrm{vars}} = 2\epsilon_{\mathrm{means}}$, $\epsilon_{\mathrm{minwt}} = \frac{\epsilon\sigma^2_{\min}}{3kn(\sigma^2_{\max}+2\mu^2_{\max})}$ and $\epsilon_{\mathrm{wts}} = \frac{\epsilon^3}{(39k)^3}$. With these parameters the algorithm runs in time $\mathrm{poly}(nL/\epsilon)\cdot\log(1/\delta)$. By Theorem 3′, we get as output a list of $\mathrm{poly}(nL/\epsilon)$ many candidate parameter settings $(\{\hat\pi^i\}, \{\hat\mu^i_j\}, \{\hat\sigma^i_j\})$ with the guarantee that with probability $1-\delta$ at least one of the settings satisfies

- $|\pi^i - \hat\pi^i| \le \epsilon_{\mathrm{wts}}$ for all $i \in [k]$, and
- $|\hat\mu^i_j - \mu^i_j| \le \epsilon_{\mathrm{means}}$ and $|(\hat\sigma^i_j)^2 - (\sigma^i_j)^2| \le \epsilon_{\mathrm{vars}}$ for all $i, j$ such that $\pi^i \ge \epsilon_{\mathrm{minwt}}$.

We now pass each of these candidate parameter settings through Theorem 4. (Note that $\epsilon_{\mathrm{wts}} < 1/(12k^3)$ as required by Theorem 4.) By Theorem 4, for any $M = \mathrm{poly}(nL/\epsilon)$ all the resulting distributions will satisfy $\exp(-\mathrm{poly}(nL/\epsilon)) \le \mathbf{Z}'(x) \le \mathrm{poly}(L)^n$ for all $x \in [-M, M]^n$. It is easy to check that under our parameter settings, each of the three component terms of $\eta$ (namely $n\cdot\left(\frac{\epsilon_{\mathrm{vars}}}{2\sigma^2_{\min}} + \frac{\epsilon^2_{\mathrm{means}}+\epsilon_{\mathrm{vars}}}{2(\sigma^2_{\min}-\epsilon_{\mathrm{vars}})}\right)$, $k\epsilon_{\mathrm{minwt}}\cdot n\left(\frac{\sigma^2_{\max}+2\mu^2_{\max}}{\sigma^2_{\min}}\right)$, and $13k\epsilon^{1/3}_{\mathrm{wts}}$) is at most $\epsilon/3$. Thus $\eta(\epsilon_{\mathrm{means}}, \epsilon_{\mathrm{vars}}, \epsilon_{\mathrm{wts}}, \epsilon_{\mathrm{minwt}}) \le \epsilon$, so at least one of the resulting distributions $\mathbf{Z}^\star$ satisfies $\mathrm{KL}(\mathbf{Z}||\mathbf{Z}^\star) \le \epsilon$.

## 6 Putting it all together

### 6.1 Identifying a good distribution using maximum likelihood

Theorem 5 gives us a list of distributions at least one of which is close to the target distribution we are trying to learn. Now we must *identify* some distribution in the list which is close to the target. We use a natural maximum likelihood algorithm described in [8] to help us accomplish this:

**Theorem 6.** *[8] Let $\beta$, $\alpha$, $\epsilon > 0$ be such that $\alpha < \beta$. Let $\mathcal{Q}$ be a set of hypothesis distributions for some distribution $\mathbf{P}$ over the space $X$ such that at least one $\mathbf{Q}^* \in \mathcal{Q}$ has $\mathrm{KL}(\mathbf{P}||\mathbf{Q}^*) \leq \epsilon$. Suppose also that $\alpha \leq \mathbf{Q}(x) \leq \beta$ for all $\mathbf{Q} \in \mathcal{Q}$ and all $x$ such that $\mathbf{P}(x) > 0$.*

*Run the ML algorithm on $\mathcal{Q}$ using a set $\mathcal{S}$ of independent samples from $\mathbf{P}$, where $\mathcal{S} = m$. Then, with probability $1-\delta$, where $\delta \leq (|\mathcal{Q}|+1)\cdot\exp\left(-2m\frac{\epsilon^2}{\log^2(\beta/\alpha)}\right)$, the algorithm outputs some distribution $\mathbf{Q}^{\mathrm{ML}} \in \mathcal{Q}$ which has $\mathrm{KL}(\mathbf{P}||\mathbf{Q}^{\mathrm{ML}}) \leq 4\epsilon$.*

### 6.2 The main result

Here we put the pieces together and give our main learning result for mixtures of Gaussians.

**Theorem 7.** *Let $\mathbf{Z}$ be any unknown mixture of $k$ $n$-dimensional Gaussians. There is a $(nL/\epsilon)^{O(k^3)} \cdot \log(1/\delta)$ time algorithm which, given samples from $\mathbf{Z}$ and any $\epsilon, \delta > 0$ as inputs, outputs a mixture $\mathbf{Z}'$ of $k$ Gaussians which with probability at least $1-\delta$ satisfies $\mathrm{KL}(\mathbf{Z}||\mathbf{Z}') \leq \epsilon$.*

*Proof.* Run the algorithm given by Theorem 5. With probability $1-\delta$ this produces a list of $T = (nL/\epsilon)^{O(k^3)} \cdot \log(1/\delta)$ hypothesis distributions, one of which, $\mathbf{Z}^\star$, has KL divergence at most $\epsilon$ from $\mathbf{Z}$ and all of which have their pdfs bounded between $\exp(-\mathrm{poly}(nL/\epsilon))$ and $\mathrm{poly}(L)^n$ for all $x \in [-M, M]^n$, where $M > \mu_{\max}$ is any $\mathrm{poly}(nL/\epsilon)$.

We now consider $\mathbf{Z}_M$, the $M$-truncated version of $\mathbf{Z}$; this is simply the distribution obtained by restricting the support of $\mathbf{Z}$ to be $[-M, M]^n$ and scaling so that $\mathbf{Z}_M$ is a distribution (see Appendix D for a precise definition of $\mathbf{Z}_M$). We prove the following proposition in Appendix D:

**Proposition 3.** *Let $\mathbf{P}$ and $\mathbf{Q}$ be any mixtures of $n$-dimensional Gaussians. Let $\mathbf{P}_M$ denote the $M$-truncated version of $\mathbf{P}$. For some $M = \mathrm{poly}(nL/\epsilon)$ we have $|\mathrm{KL}(\mathbf{P}_M||\mathbf{Q}) - \mathrm{KL}(\mathbf{P}||\mathbf{Q})| \leq 4\epsilon + 2\epsilon \cdot \mathrm{KL}(\mathbf{P}||\mathbf{Q})$.*

This proposition implies that $\mathrm{KL}(\mathbf{Z}_M||\mathbf{Z}^\star) \leq 7\epsilon$.

Now run the ML algorithm with $m = \mathrm{poly}(nL/\epsilon)\log(M/\delta)$ on this list of hypothesis distributions *using $\mathbf{Z}_M$ as the target distribution.* (We can obtain draws from $\mathbf{Z}_M$ using rejection sampling from $\mathbf{Z}$; with probability $1-\delta$ this incurs only a negligible increase in the time required to obtain $m$ draws.) Note that running the algorithm with $\mathbf{Z}_M$ as the target distribution lets us assert that all hypothesis distributions have pdfs bounded above and below on the support of the target distribution, as is required by Theorem 6. (In contrast, since the support of $\mathbf{Z}$ is all of $\mathbf{R}^n$, we cannot guarantee that our hypothesis distributions have pdf bounds on the support of $\mathbf{Z}$.) By Theorem 6, with probability at least $1-\delta$ the ML algorithm outputs a hypothesis $\mathbf{Z}^{\mathrm{ML}}$ such that $\mathrm{KL}(\mathbf{Z}_M||\mathbf{Z}^{\mathrm{ML}}) \leq 28\epsilon$.

It remains only to bound $\mathrm{KL}(\mathbf{Z}||\mathbf{Z}^{\mathrm{ML}})$. By Proposition 3 we have

$$\mathrm{KL}(\mathbf{Z}||\mathbf{Z}^{\mathrm{ML}}) \leq 28\epsilon + 4\epsilon + 2\epsilon \cdot \mathrm{KL}(\mathbf{Z}||\mathbf{Z}^{\mathrm{ML}})$$

which implies that $\mathrm{KL}(\mathbf{Z}||\mathbf{Z}^{\mathrm{ML}}) \leq 33\epsilon$. The running time of the overall algorithm is $(nL/\epsilon)^{O(k^3)} \cdot \log(1/\delta)$ and the theorem is proved.

## 7 Extensions to other distributions

In this paper we have shown how to PAC learn mixtures of any constant number of distributions, each of which is an $n$-dimensional Gaussian product distribution. This expands upon the work by Feldman et al. [8] which worked for discrete distributions in place of Gaussians. It should be clear from our work that in fact many "nice" univariate distributions can be handled similarly. Also, it should be noted that the $n$ coordinates need not come from the same family of distributions; for example, our methods would handle mixtures where some attributes had discrete distributions and the remainder had Gaussian distributions.

What level of "niceness" do our methods require for a parameterized family of univariate distributions on $\mathbf{R}$? First and foremost, it should be amenable to the "method of moments" from statistics. By this it is meant that it should be possible to solve for the parameters of the distribution given a constant number of the moments. Distributions in this category include gamma distributions, chi-square distributions, beta distributions, exponential — more generally, Weibull — distributions, and more. As a trivial example, the unknown parameter of an exponential distribution is simply its mean. As a slightly more involved example, given a beta distribution with unknown parameters $\alpha$ and $\beta$ (the pdf for which is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on $[0,1]$), these parameters can be determined from mean and variance estimates via

$$\alpha = \mathbf{E}[\mathbf{X}]\left(\frac{\mathbf{E}[\mathbf{X}](1-\mathbf{E}[\mathbf{X}])}{\mathrm{Var}[\mathbf{X}]} - 1\right), \qquad \beta = (1-\mathbf{E}[\mathbf{X}])\left(\frac{\mathbf{E}[\mathbf{X}](1-\mathbf{E}[\mathbf{X}])}{\mathrm{Var}[\mathbf{X}]} - 1\right).$$

So long as the univariate distribution family can be determined by a constant number of moments, our basic strategy of running WAM multiple times to determine moment estimates and then taking the cross-products of these lists can be employed.

There are only two more concerns that need to be addressed for a given parameterized family of distributions. First, one needs an analogue of Proposition 1, showing that products of independent random variables from the distribution family are efficiently samplable. (In fact, this should hold for *mixtures* of such, but this is very likely to be implied in any reasonable case.) This immediately holds for any distribution with bounded support; it will also typically hold for "reasonable" probability distributions that have pdfs with rapidly decaying tails.

Second, one needs an analogue of Theorem 4. This requires that it should be possible to convert accurate candidate parameter values into a KL-close actual distribution. It seems that this will typically be possible so long as the distributions in the family are not highly concentrated at any particular point. The conversion procedure should also have the property that the distributions it output have pdfs that are bounded below/above by at most exponentially small/large values, at least on polynomially-sized domains. This again seems to be a mild constraint, satisfiable for reasonable distributions with rapidly decaying tails.

In summary, we believe that for most parameterized distribution families "$D$" of interest, performing a small amount of technical work should be sufficient to show that our methods can learn "mixtures of products of $D$'s". We leave the problem of checking these conditions for distribution families of interest as an avenue for future research.

## References


[1] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual COLT*, pages 458–469, 2005.

[2] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[5] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

[6] S. Dasgupta and L. Schulman. A Two-round Variant of EM for Gaussian Mixtures. In *Proceedings of the 16th Conf. on UAI*, pages 143–151, 2000.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Ser. B*, 39:1–38, 1977.

[8] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th IEEE FOCS*, pages 501–510, 2005.

[9] Y. Freund, M. Kearns, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. Efficient learning of typical finite automata from random walks. *Information and Computation*, 138(1):23–48, 1997.

[10] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual COLT*, pages 183–192, 1999.

[11] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Annual COLT*, pages 444–457, 2005.

[12] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th STOC*, pages 273–282, 1994.

[13] B. Lindsay. *Mixture models: theory, geometry and applications*. Institute for Mathematical Statistics, 1995.

[14] M. Naor. Evaluation may be easier than generation. In *Proceedings of the 28th Symposium on Theory of Computing (STOC)*, pages 74–83, 1996.

[15] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, Univ. Edinburgh, 2003.

[16] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions*. Wiley & Sons, 1985.

[17] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[18] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd IEEE FOCS*, pages 113–122, 2002.


## A Notational convention on Gaussians

Recall that all Gaussians we consider are $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded. In dealing with Gaussians it will be very useful to define a function $M(\theta)$ which satisfies

$$\int_{|x|\geq M} \mathbf{X}(x)dx < \theta, \int_{|x|\geq M} |x|\mathbf{X}(x)dx < \theta, \quad \text{and} \quad \int_{|x|\geq M} x^2\mathbf{X}(x)dx < \theta$$

for any one-dimensional $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussians $\mathbf{X}$. Straightforward arguments show that this can be achieved with $M(\theta) = \mathrm{poly}(L/\theta)$.

**Notational convention:** *Throughout the appendices $M(\theta) = \mathrm{poly}(L/\theta)$ denotes a function satisfying the conditions above.*

## B Proof of Proposition 1

*Proof.* We shall prove the proposition for $\mathbf{W}^2$; the proof for $\mathbf{W}$ is similar but slightly simpler.

Let the mixing weights be $\pi^1, \dots, \pi^k$ and suppose that $\mathbf{Z}_j$ is a mixture of $\mathbf{X}^1_j, \dots, \mathbf{X}^k_j$ for $j = 1, 2$. Let $s = \mathbf{E}[\mathbf{W}^2]$.

Recall the quantity $M = M(\theta)$ and take $C = M^4 = \mathrm{poly}(L/\theta)$. Let $\mathbf{W}^2_C$ denote the random variable $\mathbf{W}^2$ conditioned on the event $|\mathbf{W}^2| \le C$. Observe that

$$\Pr[\mathbf{W}^2 > C] = \Pr[\mathbf{W}^2 > M^4] \le \Pr[|\mathbf{Z}_1| > M] + \Pr[|\mathbf{Z}_2| > M] \le 2\theta, \quad (1)$$

using the fact that $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussians and the definition of $M$.

We shall show that $|\mathbf{E}[\mathbf{W}^2_C] - s| \le \epsilon/2$. Our sampling algorithm for $\mathbf{W}^2$ will be to sample from $\mathbf{W}^2_C$ using rejection sampling and to compute and output the empirical mean of $\mathbf{W}^2_C$. Since the random variable $\mathbf{W}^2_C$ is bounded in the range $[-C, C]$, by the Hoeffding bound if we take $\mathrm{poly}(C/\epsilon) \cdot \log(1/\delta) = \mathrm{poly}(L/\epsilon\theta) \cdot \log(1/\delta)$ samples from $\mathbf{W}^2_C$ then with probability $1 - \delta$ the empirical mean of $\mathbf{W}^2_C$ will be within $\epsilon/2$ of the true mean $\mathbf{E}[\mathbf{W}^2_C]$. (Technically, we must also note that since $\theta$ is much smaller than 1 we can do rejection sampling with very little slowdown.) Thus it remains to show that indeed $|\mathbf{E}[(\mathbf{W}_C)^2] - s| \le \epsilon/2$.

Observe that $\mathbf{E}[(\mathbf{W}_C)^2] = \sum_{i=1}^k \pi^i \mathbf{E}[(\mathbf{W}_C)^2 \mid i \text{ is chosen}]$ and $s = \sum_{i=1}^k \pi^i \mathbf{E}[\mathbf{W}^2 \mid i \text{ is chosen}]$. Thus by convexity it is sufficient to prove $|\mathbf{E}[(\mathbf{X}^i_1)^2(\mathbf{X}^i_2)^2 \mid (\mathbf{X}^i_1)^2(\mathbf{X}^i_1)^2 \le C] - \mathbf{E}[(\mathbf{X}^i_1)^2(\mathbf{X}^i_2)^2]| \le \epsilon/2$ for all $i = 1 \dots k$. For simplicity we now write $\mathbf{X}_j = \mathbf{X}^i_j$ for $j = 1, 2$. Recall that $\mathbf{X}_1$ and $\mathbf{X}_2$ are one-dimensional $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussians.

Let $p(w)$ be the pdf for the random variable $(\mathbf{X}_1)^2(\mathbf{X}_2)^2$. Note that

$$\begin{aligned}
\left|\int_{|w|>C} wp(w)dw\right| &= \int_{x_1}\int_{x_2} \mathbf{1}_{\{x_1^2x_2^2\geq C\}} x_1^2x_2^2\mathbf{X}_1(x_1)\mathbf{X}_2(x_2)dx_1dx_2 \\
&\leq \int_{x_1}\int_{x_2} (\mathbf{1}_{\{|x_1|\geq C^{1/4}\}} + \mathbf{1}_{\{|x_2|\geq C^{1/4}\}}) x_1^2x_2^2\mathbf{X}_1(x_1)\mathbf{X}_2(x_2)dx_1dx_2 \\
&= \int_{x_2} x_2^2\mathbf{X}_2(x_2)dx_2 \int_{|x_1|\geq M} x_1^2\mathbf{X}_1(x_1)dx_1 \\
&\qquad + \int_{x_1} x_1^2\mathbf{X}_1(x_1)dx_1 \int_{|x_2|\geq M} x_2^2\mathbf{X}_2(x_2)dx_2 \\
&= \mathbf{E}[(\mathbf{X}_2)^2] \int_{|x_1|\geq M} x_1^2\mathbf{X}_1(x_1)dx_1 \\
&\qquad + \mathbf{E}[(\mathbf{X}_1)^2] \int_{|x_2|\geq M} x_2^2\mathbf{X}_2(x_2)dx_2 \\
&\leq 2L^2\left(\int_{|x_1|\geq M} x_1^2\mathbf{X}_1(x_1)dx_1 + \int_{|x_2|\geq M} x_2^2\mathbf{X}_2(x_2)dx_2\right) \\
&\leq 4\theta L^2, \qquad (2)
\end{aligned}$$

using the definitions of $M$ and $L$.

Let $\eta = 1/(1-\Pr[(\mathbf{X}_1)^2(\mathbf{X}_2)^2 > C])-1$, so $\eta \leq 3\theta$ using the same argument as in (1). Note that the pdf $p_C(w)$ for the random variable $(\mathbf{X}_1)^2(\mathbf{X}_2)^2$ conditioned on $|(\mathbf{X}_1)^2(\mathbf{X}_2)^2| \leq C$ is given by

$$p_C(w) = \begin{cases} (1+\eta)p(w) & \text{if } |w| \leq C, \\ 0 & \text{if } |w| > C. \end{cases}$$

Let $t = \mathbf{E}[(\mathbf{X}_1)^2(\mathbf{X}_2)^2]$; finally, we can show that $|\mathbf{E}[(\mathbf{X}_1)^2(\mathbf{X}_2)^2 \mid (\mathbf{X}_1)^2(\mathbf{X}_2)^2 \leq C] - t| \leq \epsilon/2$, as desired:

$$\begin{aligned}
|\mathbf{E}[(\mathbf{X}_1)^2(\mathbf{X}_2)^2 \mid (\mathbf{X}_1)^2(\mathbf{X}_2)^2 \leq C] - t| &= \left|\int_{\mathbf{R}} wp_C(w) - \int_{\mathbf{R}} wp(w)\right| \\
&= \left|(1+\eta)\int_{|w|\leq C} wp(w) - \int_{|w|\leq C} wp(w) - \int_{|w|>C} wp(w)\right| \\
&= \left|\eta\int_{|w|<C} wp(w) - \int_{|w|\geq C} wp(w)\right| \\
&\leq \eta t + \theta \leq (3\theta)\mathrm{poly}(L) + \theta,
\end{aligned}$$

once more using the definition of $M$ (note: $C \geq M$). Choosing $\theta = \mathrm{poly}(\epsilon/L)$, we get that this is bounded by $\epsilon/2$; consequently $M = \mathrm{poly}(L/\epsilon)$ and the sampling time is as claimed.

## C Auxiliary facts about KL divergence

The following fact gives the KL divergence between two univariate Gaussians; it can be found in, e.g., [15].

**Fact 8** *Let* $\mathbf{P}, \mathbf{Q}$ *each be a one-dimensional normal distribution with means and variances* $\mu_{\mathbf{P}}, \sigma_{\mathbf{P}}$ *and* $\mu_{\mathbf{Q}}, \sigma_{\mathbf{Q}}$ *respectively. Then we have*

$$\mathrm{KL}(\mathbf{P}||\mathbf{Q}) = \frac{1}{2}\ln\left(\frac{\sigma_{\mathbf{Q}}^2}{\sigma_{\mathbf{P}}^2}\right) + \frac{(\mu_{\mathbf{P}} - \mu_{\mathbf{Q}})^2 + \sigma_{\mathbf{P}}^2 - \sigma_{\mathbf{Q}}^2}{2\sigma_{\mathbf{Q}}^2}.$$

An easy consequence is the following bound on the KL divergence between two Gaussians:

**Corollary 1.** *Let* $\mathbf{P}, \mathbf{Q}$ *be one-dimensional Gaussians as above and suppose that* $|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}| \leq \epsilon_{\text{means}}$, $|\sigma_{\mathbf{P}}^2 - \sigma_{\mathbf{Q}}^2| < \epsilon_{\text{vars}}$, *and* $\sigma_{\mathbf{P}}^2 \geq \sigma_{\min}^2$. *Then*

$$\mathrm{KL}(\mathbf{P}||\mathbf{Q}) \leq \frac{\epsilon_{\text{vars}}}{2\sigma_{\min}^2} + \frac{\epsilon_{\text{means}}^2 + \epsilon_{\text{vars}}}{2(\sigma_{\min}^2 - \epsilon_{\text{vars}})}.$$

*Proof.* We have

$$\frac{\sigma_{\mathbf{Q}}^2}{\sigma_{\mathbf{P}}^2} \leq \frac{\sigma_{\min}^2 + \epsilon_{\text{vars}}}{\sigma_{\min}^2} = 1 + \frac{\epsilon_{\text{vars}}}{\sigma_{\min}^2}$$

which implies

$$\frac{1}{2}\ln\left(\frac{\sigma_{\mathbf{Q}}^2}{\sigma_{\mathbf{P}}^2}\right) \leq \frac{\epsilon_{\text{vars}}}{2\sigma_{\min}^2}.$$

The bound easily follows observing that $\sigma_{\mathbf{Q}}^2 \geq \sigma_{\min}^2 - \epsilon_{\text{vars}}$.

**Proposition 4.** *Suppose* $\mathbf{P}_1, \ldots, \mathbf{P}_n$ *and* $\mathbf{Q}_1, \ldots, \mathbf{Q}_n$ *are distributions satisfying* $\mathrm{KL}(\mathbf{P}_i||\mathbf{Q}_i) \leq \epsilon_i$ *for all* $i$. *Then* $\mathrm{KL}(\mathbf{P}_1 \times \cdots \times \mathbf{P}_n||\mathbf{Q}_1 \times \cdots \times \mathbf{Q}_n) \leq \sum_{i=1}^n \epsilon_i$.

*Proof.* We prove the case $n = 2$:

$$\begin{aligned}
\mathrm{KL}(\mathbf{P}_1 \times \mathbf{P}_2||\mathbf{Q}_1 \times \mathbf{Q}_2) &= \iint \mathbf{P}_1(x)\mathbf{P}_2(y)\ln\frac{\mathbf{P}_1(x)\mathbf{P}_2(y)}{\mathbf{Q}_1(x)\mathbf{Q}_2(y)}dxdy \\
&= \iint \mathbf{P}_1(x)\mathbf{P}_2(y)\ln\frac{\mathbf{P}_1(x)}{\mathbf{Q}_1(x)}dxdy + \iint \mathbf{P}_1(x)\mathbf{P}_2(y)\ln\frac{\mathbf{P}_2(y)}{\mathbf{Q}_2(y)}dxdy \\
&= \int \mathbf{P}_2(y)\mathrm{KL}(\mathbf{P}_1||\mathbf{Q}_1)dy + \int \mathbf{P}_1(x)\mathrm{KL}(\mathbf{P}_2||\mathbf{Q}_2)dx \\
&\leq \epsilon_1 + \epsilon_2.
\end{aligned}$$

The general case follows by induction.

## D Truncated versus untruncated mixtures of Gaussians

**Definition 5.** *Let $\mathbf{X}$ be a distribution over $\mathbf{R}^n$. The $M$-*truncated version *of $\mathbf{X}$ is the distribution $\mathbf{X}_M$ obtained by restricting the support of $\mathbf{X}$ to be $[-M, M]^n$ and scaling so that $\mathbf{X}_M$ is a distribution. More precisely, for $x \in \mathbf{R}^n$ we have*

$$\mathbf{X}_M(x) = \begin{cases} 0 & \text{if } \|x\|_\infty > M, \\ c\mathbf{X}(x) & \text{if } \|x\|_\infty \leq M \end{cases}$$

*where $c = 1/\left(\int_{x \in [-M,M]^n} \mathbf{X}(x)\right)$ is chosen so that $\int \mathbf{X}_M(x) = 1$.*

In this section we prove Proposition 3:

**Proposition 3** *Let $\mathbf{P}$ and $\mathbf{Q}$ be any mixtures of $n$-dimensional Gaussians. Let $\mathbf{P}_M$ denote the $M$-truncated version of $\mathbf{P}$. For some $M = \text{poly}(nL/\epsilon)$ we have $|\text{KL}(\mathbf{P}_M||\mathbf{Q}) - \text{KL}(\mathbf{P}||\mathbf{Q})| \leq 4\epsilon + 2\epsilon \cdot \text{KL}(\mathbf{P}||\mathbf{Q})$.*

*Proof.* We will take $M = M(\theta)$ (recall Appendix A). As we go through the proof various conditions will be set on $\theta$. At the end of the proof we will see that we can take $\theta = \text{poly}(\epsilon/nL)$ and obtain the desired bound on $|\text{KL}(\mathbf{P}_M||\mathbf{Q}) - \text{KL}(\mathbf{P}||\mathbf{Q})|$ and satisfy all the conditions on $\theta$. This proves the theorem.

We have that $\mathbf{P}_M(x)$ satisfies

$$\mathbf{P}_M(x) = \begin{cases} (1+\delta)\mathbf{P}(x) & \text{if } x \in [-M, M]^n, \\ 0 & \text{if } x \notin [-M, M]^n, \end{cases}$$

where $\delta > 0$ is chosen so that $\frac{1}{1+\delta} = \int_{x \in [-M,M]^n} \mathbf{P}(x)$. Using the definition of $M$ we have

$$\int_{x \notin [-M,M]^n} \mathbf{P}(x) = \Pr_{\mathbf{P}}[x \notin [-M, M]^n] \leq \sum_{j=1}^{n} \Pr_{\mathbf{P}}[|x_j| \geq M] \leq n\theta \leq \epsilon$$

where we have used the fact that $\theta \leq \epsilon/n$ (this is our first condition on $\theta$). Consequently we have $\frac{1}{1+\delta} \geq 1 - \epsilon$, so $\delta \leq 2\epsilon$.

We have

$$\begin{aligned}
&|\text{KL}(\mathbf{P}_M||\mathbf{Q}) - \text{KL}(\mathbf{P}||\mathbf{Q})| \\
&\quad = \left| \int_{x \in [-M,M]^n} (1+\delta)\mathbf{P}(x) \ln \frac{(1+\delta)\mathbf{P}(x)}{\mathbf{Q}(x)} - \int_{x \in \mathbf{R}^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)} \right| \\
&\quad = \left| (1+\delta)\ln(1+\delta) \int_{x \in [-M,M]^n} \mathbf{P}(x) + \delta \int_{x \in [-M,M]^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)} - \int_{x \notin [-M,M]^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)} \right| \\
&\quad \leq (1+\delta)\ln(1+\delta) + \delta \left| \int_{x \in [-M,M]^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)} \right| + \left| \int_{x \notin [-M,M]^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)} \right| \\
&\quad = \delta(1+\delta) + \delta|R| + |S|,
\end{aligned}$$

where $R := \int_{x \in [-M,M]^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}$ and $S := \int_{x \notin [-M,M]^n} \mathbf{P}(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}$. For succinctness let $\kappa$ denote $\mathrm{KL}(\mathbf{P}||\mathbf{Q})$. Note that we have $\kappa = R + S$.

Suppose we show that $|S| \leq \epsilon$. Then since $\kappa = R+S$, we must have $|R| \leq \kappa + \epsilon$, and hence $|\mathrm{KL}(\mathbf{P}_M||\mathbf{Q}) - \kappa| \leq \delta(1+\delta) + \delta(\kappa+\epsilon) + \epsilon \leq 4\epsilon + 2\epsilon\kappa$ (using $\delta \leq 2\epsilon$), as desired. Thus we can complete the proof by showing $|S| \leq \epsilon$.

Let us analyze the integrand of $S$. Decompose $\mathbf{P}$ into its mixture components, i.e. $\mathbf{P}(x) = \sum_{i=1}^k \pi^i \mathbf{P}^i(x)$, where $\mathbf{P}^1, \ldots, \mathbf{P}^k$ are $n$-dimensional Gaussians. Hence

$$S = \sum_{i=1}^k \pi^i \int_{x \notin [-M,M]^k} \mathbf{P}^i(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}.$$

We will show that for each $i$ we have $|\int_{x \notin [-M,M]^k} \mathbf{P}^i(x) \ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}| \leq \epsilon$. It then follows that $|S| \leq \epsilon$ since $|S|$ is upper bounded by a convex combination of these quantities.

Let us now analyze the quantity $\ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}$. We will show that for any $x \notin [-M,M]^k$, neither $\mathbf{P}(x)$ nor $\mathbf{Q}(x)$ can be either "too small" or "too large" as a function of $||x||_2^2$; hence $|\ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}|$ will be of moderate size. We will prove this for $\mathbf{P}(x)$ using the fact that it is a mixture of $n$-dimensional $(\mu_{\max}, \sigma^2_{\min}, \sigma^2_{\max})$-bounded Gaussians; since this is also true of $\mathbf{Q}(x)$, the same bound will hold for it.

We will show that for all $i = 1, \ldots, k$ and all $x \in \mathbf{R}^n$ we have $\mathbf{P}^i(x) \in [t(x), T]$ where $T$ is a quantity and $t(x)$ is a function that will both be defined below. Since $\mathbf{P}(x) = \sum_{i=1}^k \pi^i \mathbf{P}^i(x)$ is a convex combination of the $\mathbf{P}^i(x)$'s, the same bound will hold for $\mathbf{P}(x)$. Fix any $i$ and consider the Gaussian $\mathbf{P}^i$. Since this Gaussian is axis-aligned, we have $\mathbf{P}^i(x) = \prod_{j=1}^n \phi_{\mu_j, \sigma_j^2}(x_j)$ for some pairs $(\mu_1, \sigma_1^2), \ldots, (\mu_n, \sigma_n^2)$ satisfying $|\mu_j| \leq \mu_{\max}$, $\sigma_j^2 \in [\sigma^2_{\min}, \sigma^2_{\max}]$. (Here $\phi_{\mu,\sigma^2}(x)$ is the usual pdf $\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ for a one-dimensional Gaussian.) It is easy to see that for any $x_j$,

$$\frac{1}{\sqrt{2\pi}\sigma_{\max}} \exp\left(-\frac{x_j^2}{\sigma^2_{\min}} - \frac{\mu^2_{\max}}{\sigma^2_{\min}}\right) \leq \phi_{\mu_j, \sigma_j^2}(x_j) \leq \frac{1}{\sqrt{2\pi}\sigma_{\min}}.$$

Hence for all $x \in \mathbf{R}^n$ we have

$$t(x) := \left(\frac{\exp(-\mu^2_{\max}/\sigma^2_{\min})}{\sqrt{2\pi}\sigma_{\max}}\right)^n \exp\left(-\frac{||x||_2^2}{\sigma^2_{\min}}\right) \leq \mathbf{P}^i(x) \leq \left(\frac{1}{\sqrt{2\pi}\sigma_{\min}}\right)^n =: T \tag{3}$$

for all $i$, and so (3) holds true for $\mathbf{P}(x)$ as well. As stated earlier, the same argument also shows that (3) holds for $\mathbf{Q}(x)$. We conclude that for any $x$,

$$\begin{aligned}\left|\ln \frac{\mathbf{P}(x)}{\mathbf{Q}(x)}\right| &\leq |\ln t(x)| + |\ln T| \\ &= \left|-n\frac{\mu_{\max}^2}{\sigma_{\min}^2} - n\ln(\sqrt{2\pi}\sigma_{\max}) - \frac{||x||_2^2}{\sigma_{\min}^2}\right| + n\ln(1/\sqrt{2\pi}\sigma_{\min}) \\ &\leq O\left(n\frac{\mu_{\max}^2}{\sigma_{\min}^2}\ln\frac{\sigma_{\max}}{\sigma_{\min}}||x||_2^2\right).\end{aligned}$$

Recall that we want to show $|\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)\ln\frac{\mathbf{P}(x)}{\mathbf{Q}(x)}| \leq \epsilon$. It clearly suffices to show that $\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)|\ln\frac{\mathbf{P}(x)}{\mathbf{Q}(x)}| \leq \epsilon$. By the above it suffices to show

$$O\left(n\frac{\mu_{\max}^2}{\sigma_{\min}^2}\ln\frac{\sigma_{\max}}{\sigma_{\min}}\right)\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)||x||_2^2 \leq \epsilon.$$

We have

$$\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)||x||_2^2 = \sum_{j=1}^n\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)x_j^2 \tag{4}$$

Fix $j$; we now bound $\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)x_j^2$. Recall that $\mathbf{P}^i(x) = \mathbf{P}_1^i(x_1)\cdots\mathbf{P}_n^i(x_n)$. We have

$$\begin{aligned}\int_{x\notin[-M,M]^n}\mathbf{P}^i(x)x_j^2 &\leq \sum_{\ell=1}^n\int_{x\in\mathbf{R}^n:|x_\ell|>M}\mathbf{P}^i(x)x_j^2 \\ &= \int_{x\in\mathbf{R}^n:|x_j|>M}\mathbf{P}^i(x)x_j^2 + \sum_{\ell\neq j}\int_{x\in\mathbf{R}^n:|x_\ell|>M}\mathbf{P}^i(x)x_j^2\end{aligned}$$

For the first integral of (5) above we have

$$\begin{aligned}\int_{x\in\mathbf{R}^n:|x_j|>M}\mathbf{P}^i(x)x_j^2 = \left(\prod_{\ell\neq j}\left[\int_{x_\ell\in\mathbf{R}}\mathbf{P}_\ell^i(x_\ell)dx_\ell\right]\right)\cdot\int_{|x_j|>M}\mathbf{P}_j^i(x_j)x_j^2dx_j &= \int_{|x_j|>M}\mathbf{P}_j^i(x_j)x_j^2dx_j \\ &\leq \theta \end{aligned}\tag{5}$$

where the inequality is by the definition of $M$. For the second term of (5) above we have

$$\sum_{\ell\neq j}\int_{x\in\mathbf{R}^n:|x_\ell|>M}\mathbf{P}^i(x)x_j^2 = \sum_{\ell\neq j}\left[\left(\int_{|x_\ell|>M}\mathbf{P}_\ell^i(x_\ell)dx_\ell\right)\left(\int_{x_j\in\mathbf{R}}\mathbf{P}_j^i(x_j)x_j^2dx_j\right)\right] \tag{6}$$

where we have used the fact that for any $\ell'$ which is neither $\ell$ nor $j$ we have

$$\int_{x_{\ell'}\in\mathbf{R}}\mathbf{P}_{\ell'}^i(x_{\ell'})dx_{\ell'} = 1.$$

Again using the definition of $M$ to bound the integral over variable $x_\ell$ in (6) above by $\theta$, we have that (6) is at most

$$\begin{aligned}(n-1)\theta\int_{x_j\in\mathbf{R}}\mathbf{P}_j^i(x_j)x_j^2dx_j = (n-1)\theta\mathbf{E}_{\mathbf{P}_j^i}[x^2] &= (n-1)\theta\left(\mathrm{Var}_{\mathbf{P}_j^i}[x]+\mathbf{E}_{\mathbf{P}_j^i}[x]^2\right)\\ &= (n-1)\theta((\sigma_j^i)^2+(\mu_j^i)^2)\\ &\leq (n-1)\theta(\sigma_{\max}^2+\mu_{\max}^2) \qquad (7)\end{aligned}$$

where the inequality holds since $\mathbf{P}_j^i$ is a one-dimensional $(\mu_{\max},\sigma_{\min}^2,\sigma_{\max}^2)$-bounded Gaussian.

Putting all the pieces together, we find that (4) is at most

$$n[\theta+(n-1)\theta(\sigma_{\max}^2+\mu_{\max}^2)]\leq n^2\theta(\sigma_{\max}^2+\mu_{\max}^2)$$

It follows that $|S|\leq n^2\theta(\sigma_{\max}^2+\mu_{\max}^2)\cdot O(n^2\frac{\mu_{\max}^2}{\sigma_{\min}^2}\ln\frac{\sigma_{\max}}{\sigma_{\min}})$. We can take $\theta=\mathrm{poly}(\epsilon/nL)$ and have this quantity be at most $\epsilon$.