Polynomial Learning of Distribution Families

Mikhail Belkin
Dept. of Computer Science and Engineering
Ohio State University
Columbus, Ohio, USA
Email: mbelkin@cse.ohio-state.edu

Kaushik Sinha
Dept. of Computer Science and Engineering
Ohio State University
Columbus, Ohio, USA
Email: sinhak@cse.ohio-state.edu

***Abstract*—The question of polynomial learnability of probability distributions, particularly Gaussian mixture distributions, has recently received significant attention in theoretical computer science and machine learning. However, despite major progress, the general question of polynomial learnability of Gaussian mixture distributions still remained open. The current work resolves the question of polynomial learnability for Gaussian mixtures in high dimension with an arbitrary fixed number of components. Specifically, we show that parameters of a Gaussian mixture distribution with fixed number of components can be learned using a sample whose size is polynomial in dimension and all other parameters.**

The result on learning Gaussian mixtures relies on an analysis of distributions belonging to what we call "polynomial families" in low dimension. These families are characterized by their moments being polynomial in parameters and include almost all common probability distributions as well as their mixtures and products. Using tools from real algebraic geometry, we show that parameters of any distribution belonging to such a family can be learned in polynomial time and using a polynomial number of sample points. The result on learning polynomial families is quite general and is of independent interest.

To estimate parameters of a Gaussian mixture distribution in high dimensions, we provide a deterministic algorithm for dimensionality reduction. This allows us to reduce learning a high-dimensional mixture to a polynomial number of parameter estimations in low dimension. Combining this reduction with the results on polynomial families yields our result on learning arbitrary Gaussian mixtures in high dimensions.

***Index Terms*—Gaussian mixture learning, polynomial learnability**

I. Introduction

Estimating parameters of a model from sampled data is one of the oldest and most general problems of statistical inference. Given a number of samples, one needs to choose a distribution that best fits the observed data. While traditionally theoretical analysis in the statistical literature has concentrated on rates (e.g., minimax rates), in recent years other computational aspects of this problem, especially dependence on dimension of the space, have attracted attention. In particular, a recent line of work in the theoretical computer science and learning communities has been concerned with learning the distribution in time and using a number of samples, polynomial in parameters and the dimension of the space. This effort has been particularly directed at the family of Gaussian Mixture models due to their simple formulation and widespread use in applications spanning areas such as computer vision, speech recognition, and many others (see, e.g.,[1], [2], [3]). This line of research started with the work of Dasgupta [4], who was the first to show that learning the parameters of a Gaussian mixture distribution in time polynomial in the dimension of the space n was possible at all. This work has been refined and extended in a number of consequent papers. The results in [4] required separation between mixture components on the order of $\sqrt{n}$. That was later improved to of $\Omega(n^{\frac{1}{4}})$ in [5] for mixtures of spherical Gaussians and in [6] for general Gaussians. The separation requirement was further reduced and made independent of n to the order of $\Omega(k^{\frac{1}{4}})$ in [7] for a mixture of k spherical Gaussians and to the order of $\Omega(\frac{k^{\frac{3}{2}}}{\epsilon^2})$ in [8] for logconcave distributions. In [9] the separation requirement was further reduced to $\Omega(k+\sqrt{k\log n})$. An extension of PCA called isotropic PCA was introduced in [10] to learn mixtures of Gaussians when any pair of Gaussian components is separated by a hyperplane having very small overlap along the hyperplane direction (so-called "pancake layering problem"). A number of recent papers [11], [12], [13], [14], [15] addressed related problems, such as learning mixture of product distributions and heavy tailed distributions.

Author	Min. Separation	Description
Dasgupta [4], 1999	$\sqrt{n}$	Gaussian mixtures, mild assumptions
Dasgupta-Schulman [5], 2000	$n^{\frac{1}{4}}$	Spherical Gaussian mixtures
Arora-Kannan [6], 2001	$n^{\frac{1}{4}}$	Gaussian mixtures
Vempala-Wang [7], 2002	$k^{\frac{1}{4}}$	Spherical Gaussian mixtures
Kannan-Salmasian-Vempala [8], 2005	$k^{\frac{3}{2}}$	Gaussian mixtures, log-concave distributions
Achlioptas-McSherry [9], 2005	$k+\sqrt{k\log n}$	Gaussian mixtures
Feldman-Servedio-O'Donnell [16], 2006	> 0	Axis aligned Gaussians, no param. estimation
Belkin-Sinha [17], 2010	> 0	Identical spherical Gaussian mixtures
Kalai-Moitra-Valiant [18], 2010	≥ 0	Gaussian mixtures with two components
This paper, 2010	≥ 0	Gaussian mixtures

TABLE I
PARTIAL SUMMARY OF RESULTS ON GAUSSIAN MIXTURE MODEL LEARNING.

However all of these papers assumed a minimum separation between the components, which is an increasing function of the dimension n and/or the number of components k. The general question of learning parameters of a distribution without any separation conditions, remained open. The first result in that direction was obtained in Feldman, et al., [16], which showed that the density (but not the parameters) of mixtures of axis aligned Gaussians can be learned in polynomial time using the method of moments.

Very recently two papers [17], [18] independently addressed two special cases of Gaussian mixture learning without separation assumption. In Kalai, et al., [18] the authors showed that a mixture of two Gaussians with arbitrary covariance matrices can be learned in polynomial time. The technique relies on a

randomized algorithm to reduce the problem to one dimension. The key argument of the paper is based on deconvolving the one-dimensional mixture to increase the separation between the components and carefully analyzing the moments of the deconvolved mixture in order to apply the method of moments. In [17] it is shown that a mixture of k identical spherical Gaussians can be learned in time polynomial in dimension. The key result is based on analyzing the Fourier transform of the distribution in one dimension to give a lower bound on the norm. However, it is not clear whether the techniques of either [18] or [17] could be applied to the general case with an arbitrary number of components and covariance matrices.

In this paper we resolve the polynomial learnability problem by proving that there exists a polynomial algorithm to estimate parameters of a general high-dimensional mixture with arbitrary fixed number of Gaussians components without any additional assumptions. Table I briefly summarizes the progress in the area and our result.

Our main result for Gaussian mixtures relies on a quite general result of independent interest on learning what we call *polynomial families*. These families are characterized by their moments being polynomial in the parameters of a distribution. It turns out that almost all common distribution families, e.g., Gaussian, exponential, uniform, Laplace, binomial, Poisson and a number of others as well as their mixtures and (tensor) products have this property. A partial list of one dimensional distributions and description of their moments are presented in Table II in Appendix A. Our technique uses methods of real algebraic geometry and combines them with the classical method of moments (originally introduced by Pearson in [19] to analyze Gaussian mixtures).

We note that there have been applications of algebraic geometry in the field of statistics, particularly in conditional independence testing and likelihood estimation for discrete distributions and exponential families (see, e.g., [20]).

Below we give a brief summary of the main results and the structure of the paper.

Brief outline of the paper.

Section II. We start Section II by introducing the problem of parameter learning and defining the notion of a *polynomial family*. We proceed to prove the main result showing that parameters of a distribution from a polynomial family can be learned with confidence $1-\delta$ up to precision ϵ using the number of samples $\mathrm{poly}(\frac{1}{\delta}, \max(\frac{1}{\epsilon}, \frac{1}{\mathscr{R}}), B)$, where $\mathscr{R}$ is the *radius of identifiability*, a measure of intrinsic hardness of unique parameter identification for a distribution[1] and the set of parameters are contained within a ball of radius B. In fact, the result is more general, even if the radius of identifiability is zero, parameters can still be learned up to a certain equivalence relation defined in the paper.

The proof consists of the two main steps. The first step uses the Hilbert basis theorem for an appropriately defined ideal in the ring of polynomials to show that a fixed set of (possibly high-dimensional) moments uniquely identifies the distribution.

In the second step, we pose parameter estimation problem as a system of quantified algebraic equations and inequalities using the finite set of moments obtained in the first step. We use quantifier elimination for semi-algebraic sets (Tarski-Seidenberg theorem) to prove that there exists a polynomial algorithm for parameter learning.

Section III. In Section III we prove our main results on learning Gaussian mixture distributions in high dimensions. The main difficulty is that the general results of Section II cannot be applied directly since the number of parameters increases with the dimension of the space. To overcome this issue, we prove that the Gaussian family has the property that we call *polynomial reducibility*. That is the parameters of a distribution in n dimensions can be recovered from a $\mathrm{poly}(n)$ number of low-dimensional projections. Specifically, we show that a mixture of Gaussians with k components can be recovered using a polynomial number of projections to $(2k^2+2)$-dimensional space. This leads us to Theorem III.1, our main result for parameter learning on Gaussian mixtures. We show that parameters of a Gaussian mixture can be learned with precision ϵ and confidence $1-\delta$, using the number of samples polynomial in dimension n, $\frac{1}{\delta}$, $\max(\frac{1}{\epsilon}, \frac{1}{\mathscr{R}})$ and B. We also provide an explicit formula for the radius of identifiability of Gaussian mixtures, which, given an a priori bounds on the minimum mixing weight and the minimum separation between the mean/covariance pairs, leads to an upper bound on $\frac{1}{\mathscr{R}}$. For example, our results holds even in the extreme case where all components have the same mean, as long as the covariance matrices are different. In Theorem III.2 we also show that in the absence of an apriori lower bound, $\mathscr{R}$ can be estimated directly from the data.

We briefly discuss other polynomially reducible families, where a similar approach would yield results on polynomial learnability.

In **Section IV** we conclude and discuss some limitations of our results and directions of future work.

II. Learning Polynomial Families

In this section we prove some general learnability results for a large class of probability distributions that we call *polynomial families*, which are characterized by the moments being polynomial functions of parameters. This class turns out to contain nearly all commonly used probability distributions, as well as their mixtures and (tensor) products. See Appendix A (Table II) for a partial list together with the description of their moments either explicitly or through a recurrence relation, as well as some examples of families, which are not polynomial.

The main result in this section is Theorem II.8, which shows that there exists an algorithm to learn the parameters of a polynomial distribution using a polynomial number of samples.

We start with the outline of the standard parameter learning problem. Let p_θ, $\theta = (\theta^1, \ldots, \theta^m), \theta \in \Theta \subset \mathbb{R}^m$ be a m-parametric family of probability distributions in $\mathbb{R}^l$. The

[1]For example, it is impossible to identify mixing coefficients of a mixture of two Gaussians with identical means and variances, thus in that case $\mathscr{R}=0$. See Section III for the detailed analysis of Gaussian mixtures.

problem of parameter learning is the following: given precision ϵ and confidence δ, and some number $n(\epsilon, \delta)$ of points sampled from p_θ, we need to provide an estimate $\hat{\theta}$, such that $\|\hat{\theta} - \theta\| < \epsilon$ with probability at least $1 - \delta$, where $\|\cdot\|$ is Euclidean distance in $\mathbb{R}^m$.

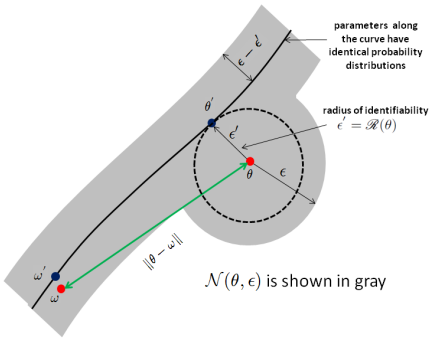

Fig. 1. If θ and ω are close to two values of parameters θ' and ω' with identical probability distribution, then it is be hard to distinguish between them from sampled data, even when $\|\theta - \omega\|$ is large.

However, for many families identifying the values of parameters uniquely is impossible, due to the fact that several different values of parameters may correspond to the same probability distribution. Moreover, if two values of parameters, say, θ and ω are close to two values of parameters, θ' and ω' respectively, which have identical probability distributions, then it may be hard to distinguish between them. This situation is illustrated in Fig. 1. These observations suggests that a more general formulation of learning distribution parameters needs to take these into account. A mathematical formalization of the more general of learnability will be given in Eq.1, which defines a notion of a neighborhood taking parameters with identical probability distribution into account. An ϵ-"neighborhood" of θ, $\mathcal{N}(\theta, \epsilon)$, is shown in gray in Fig. 1. We will also introduce the notion of the *radius of identifiability* $\mathscr{R}(\theta)$ (definition II.9) to give a quantification of how hard it may be to identify the parameters. For example, parameters θ for which $\mathscr{R}(\theta) = 0$ cannot be identified given any amount of data. In Fig. 1, the radius of identifiability $\mathscr{R}(\theta)$ is equal to ϵ'.

For mixtures of Gaussians any permutation of the mixture components has the same distribution, while a component with zero mixing weight may have arbitrary mean/covariance. If two components have the same mean/covariance pair, then the mixing coefficients are not defined uniquely. However, assuming that the mean/variance pairs for any two components are different and that the mixing coefficients are non-zero, the parameters are defined uniquely up to a permutation of components (see Section II).

Our main Theorem II.8 applies even when parameters of a probability distributions are not defined uniquly, including the standard definition of parameter learning as a special case (see Corollary II.10 and Corollary II.11).

In Subsection II-A we prove the basic properties of polynomial families, including the key result, Theorem II.3, which shows that a finite set of moments uniquely determines the distribution.

In Subsection II-B we define the extended notion of a neighborhood $\mathcal{N}(\theta, \epsilon)$ and discuss its basic properties. We proceed to obtain the main technical result, a lower bound in Theorem II.5. This, together with the upper bound in Proposition II.7 allows us to set up a grid search to prove the main Theorem II.8. We also define the radius of identifiability, and derive Corollary II.10 and Corollary II.11.

A. Polynomial Families and Finite Sets of Moments

We start by assuming that the parameter set Θ is a compact semi-algebraic subset of $\mathbb{R}^m$. Recall that a semi-algebraic set in $\mathbb{R}^m$ is a finite union of sets defined by a system of algebraic equations and inequalities. A sphere, a polytope, the sets of symmetric and orthogonal matrices are all examples of semi-algebraic sets.

The family of semi-algebraic sets is closed under finite union, intersection and taking complements. Importantly, the Tarski-Seidenberg theorem states that a linear projection of a semi-algebraic set is also semi-algebraic. This is equivalent to the *elimination of quantifiers* for semi-algebrac sets, which we will need shortly. See [21] for a review of results on real algebraic geometry.

Definition II.1 (Polynomial family)**.** *We call the family p_θ a polynomial family, if each (raw l-dimensional) moment $M_{i_1,\ldots,i_l}(\theta) = \int x_1^{i_1} \ldots x_l^{i_l}\, dp_\theta$ of the distribution exists and can be represented as a polynomial of the parameters $(\theta^1, \ldots, \theta^m)$. We also require that each p_θ should be defined uniquely by its moments*[2].

We will order the moments $M_{i_1,\ldots,i_l}$ lexicographically and denote them by $M_1(\theta), \ldots, M_n(\theta), \ldots$ In the one-dimensional case this corresponds to the standard ordering of the moments.

As it turns out, most of the common families of probability distributions are, in fact, polynomial (see Appendix A). Moreover, a mixture, a product or a linear transformation of polynomial families is also a polynomial family, as stated in the following

Lemma II.2. *Let p_θ, $\theta \in \Theta$ and q_ω, $\omega \in \Omega$ be polynomial families. Then the following families are also polynomial:*
(a) the family $w_1 p_\theta + w_2 q_\omega$, $w_1, w_2 \in \mathbb{R}$, $w_1 + w_2 = 1$.
(b) the family $p_{\theta,\omega}(x, y) = p_\theta(x) \times p_\omega(y)$, $(\theta, \omega) \in \Theta \times \Omega$.
(c) the family $p_{A\theta}$, where $A \in \mathbb{R}^{m \times m}$ is a fixed matrix and $A\theta \in \Theta \subset \mathbb{R}^m$.

The proof follows directly from the linearity of the integral, Fubini's theorem and the fact that polynomial functions stay polynomial under a linear change.

Note that a multivariate Gaussian distribution is a product of univariate Gaussians along its principal directions of the covariance matrix. Since the standard coordinates can be transformed to principal coordinates by a linear transformation, a multivariate Gaussian is a polynomial family. Hence a general mixture of k multivariate normal distributions in $\mathbb{R}^l$ is also a polynomial family with $lk + \frac{1}{2}l(l+1)k + k - 1$ parameters.

[2]This is true under some mild conditions, e.g., if the moment generating function converges in a neighborhood of zero [22].

Let us now recall that a family p_θ, is called *identifiable* if $p_{\theta_1} \neq p_{\theta_2}$ for any $\theta_1 \neq \theta_2$. We will now prove the following

Theorem II.3. *Let p_θ be a polynomial family of distributions. Then there exists a positive integer N, such that $p_{\theta_2} = p_{\theta_1}$ if and only if $M_i(\theta_1) = M_i(\theta_2)$ for all $i = 1, \ldots, N$. In the case when the family p_θ is identifiable, the first N moments are sufficient to uniquely identify the parameter θ.*

PROOF:
Since $p_\theta, \theta = (\theta^1, \ldots, \theta^m)$ is a polynomial family, each $M_i(\theta)$ is a polynomial of θ. Let $\theta_1 = (\theta_1^1, \ldots, \theta_1^m)$ and $\theta_2 = (\theta_2^1, \ldots, \theta_2^m)$. Let

$$P_i(\theta_1^1, \ldots, \theta_1^m, \theta_2^1, \ldots, \theta_2^m) \stackrel{\text{def}}{=} M_i(\theta_1) - M_i(\theta_2)$$

be a polynomial of $2m$ variables. Now let $\mathcal{I}_j$ be the ideal in the ring of polynomials of $2m$ variables generated by the polynomials $P_1, \ldots, P_j$. Thus we have an increasing sequence of ideals $\mathcal{I}_1 \subset \mathcal{I}_2 \subset \mathcal{I}_3 \ldots$ Let $\mathcal{I} = \cup_{j=1}^{\infty} \mathcal{I}_j$. By the Hilbert basis theorem, the ideal $\mathcal{I}$ is finitely generated, which implies that for some N large enough, $\mathcal{I}_N$ contains all of the generators. Therefore for any $M \geq N$ we can write

$$P_M(\theta_1, \theta_2) = \sum_{i=1}^{N} a_i(\theta_1, \theta_2) P_i(\theta_1, \theta_2)$$

for some polynomials a_i. Thus if $P_i(\theta_1, \theta_2) = 0$ for $i = 1, \ldots, N$ then $P_i(\theta_1, \theta_2) = 0$ for any i. Recalling the definition of P_M, we conclude that all moments of p_{θ_1} and p_{θ_2} coincide if and only if the first N moments of these distributions are the same. Since the sequence of moments defines the distribution uniquely, the statement of the theorem follows. □

B. Learning Polynomial Families

We will now introduce a notion of an ϵ-"neighbourhood" of a point, which takes into account that different parameters may have identical probability distribution. We proceed to prove the main Theorem II.8 and a few corollaries, showing that the standard parameter learning problem becomes a special case of the result.

Let $\mathcal{E}(\theta) = \{\omega | p_\omega = p_\theta\}$ be the set of parameters ω which have distributions same as p_θ. We note that the distributions corresponding to different values of parameters in the set $\mathcal{E}(\theta)$ are identical and hence cannot be distinguished from each other given any amount of sampled data. We now define

$$\begin{aligned} \mathcal{N}(\theta, \epsilon) \quad = \quad & \{\omega \in \Theta | \, \exists \, \omega', \theta' \in \Theta, \exists \, \epsilon' < \epsilon, \quad \text{such that} \\ & \|\omega - \omega'\| < \epsilon', \omega' \in \mathcal{E}(\theta'), \|\theta' - \theta\| < \epsilon - \epsilon'\} \end{aligned} \quad (1)$$

In other words, ω belongs to $\mathcal{N}(\theta, \epsilon)$ if it is within $\epsilon' < \epsilon$ distance of a parameter value which has the same probability distribution as a parameter value within $\epsilon - \epsilon'$ of θ. This definition is illustrated graphically in Fig. 1. We observe the following properties of $\mathcal{N}(\theta, \epsilon)$:

1) (Symmetry) If $\theta_1 \in \mathcal{N}(\theta_2, \epsilon)$ then $\theta_2 \in \mathcal{N}(\theta_1, \epsilon)$.
2) (ϵ-ball) An ϵ-ball $B(\theta, \epsilon)$ around θ is contained in $\mathcal{N}(\theta, \epsilon)$. If $B(\theta, \epsilon)$ is an identifiable family, then $B(\theta, \epsilon) = \mathcal{N}(\theta, \epsilon)$.
3) (Equivalence) If $p_{\theta_1} = p_{\theta_2}$, then $\theta_1 \in \mathcal{N}(\theta_2, \epsilon)$ for any $\epsilon > 0$.

Thus $\mathcal{N}(\theta, \epsilon)$ can be viewed as an "ϵ-ball" around θ taking probability distribution into account. For example, values of parameters with identical probability distributions cannot be distinguished by this metric, which is consistent with statistical identifiability.

Lemma II.4. *$\mathcal{N}(\theta, \epsilon)$ is an open semi-algebraic set.*

PROOF:$\mathcal{N}(\theta, \epsilon)$ is open since, a sufficiently small open ball around any point $\omega \in \mathcal{N}(\theta, \epsilon)$ is also contained in $\mathcal{N}(\theta, \epsilon)$. To see that it is algebraic we recall that by Theorem II.3 there exists an N, such that $\theta_1 \in \mathcal{E}(\theta_2)$ if and only if

$$Q(\theta_1, \theta_2) \stackrel{\text{def}}{=} \sum_{i=0}^{N} (M_i(\theta_1) - M_i(\theta_2))^2 = 0 \quad (2)$$

which is an algebraic condition. Hence, by applying the Tarski-Seidenberg theorem to eliminate the existential quantifiers in Eq. 1, we see that $\mathcal{N}(\theta, \epsilon)$ is semi-algebraic. □

Theorem II.5 (Lower bound)**.** *Let p_θ be a polynomial family. There exists $N \in \mathbb{N}$ and $t > 0$, such that for any sufficiently small $\epsilon > 0$ and any $\theta_1, \theta_2 \in \Theta$, if $|M_i(\theta_1) - M_i(\theta_2)| > \epsilon$ for at least one $i \leq N$, then $\theta_1 \notin \mathcal{N}(\theta_2, O(\epsilon^t))$.*

PROOF:
Choose N as in Theorem II.3. We start by observing we can replace the condition
$|M_i(\theta_1) - M_i(\theta_2)| > \epsilon$ by

$$Q(\theta_1, \theta_2) \stackrel{\text{def}}{=} \sum_{i=1}^{N} |M_i(\theta_1) - M_i(\theta_2)|^2 > N\epsilon^2$$

in the statement of the theorem. Since the existence of t is not affected by the substitution of $N\epsilon^2$, instead of ϵ, to simplify the matters we will assume that $Q(\theta_1, \theta_2) > \epsilon$.

From Theorem II.3 we recall that if for some $i \leq N$ $|Q(\theta_1, \theta_2)| \neq 0$ then $p_{\theta_1} \neq p_{\theta_2}$. Let δ be a positive real number. Consider the set $X = \{\theta_1, \theta_2 | \theta_1 \in \mathcal{N}(\theta_2, \delta)\}$. From Lemma II.4 and the fact that the relationship $\theta_1 \in \mathcal{N}(\theta_2, \delta)$ is symmetric, it follows that X is an open subset of $\Theta \times \Theta$. Hence the set $\Theta \times \Theta - X = \{\theta_1, \theta_2 \in \Theta, \theta_1 \notin \mathcal{N}(\theta_2, \delta))\}$ is compact and since $Q(\theta_1, \theta_2) > 0$ for any $(\theta_1, \theta_2) \in \Theta \times \Theta - X$ we have

$$\inf_{\theta_1, \theta_2 \in \Theta, \theta_1 \notin \mathcal{N}(\theta_2, \delta))} Q(\theta_1, \theta_2) > 0 \quad (3)$$

By an argument following that in Lemma II.4 we see that X and hence its complement are semi-algebraic sets.

Consider now the set S_δ, $\delta > 0$ given by the following expression

$$S_\delta = \{\epsilon > 0 \, | \, \forall_{\theta_1, \theta_2 \in \Theta} \quad (\theta_1 \notin \mathcal{N}(\theta_2, \delta)) \Rightarrow Q(\theta_1, \theta_2) > \epsilon\}. \quad (4)$$

Since these logical statements can be expressed as semi-algebraic conditions, by the Tarski-Seidenberg theorem S_δ is a semi-algebraic subset of $\mathbb{R}$. Let $\epsilon(\delta) = \inf S_\delta$. From Eq.3 we have that $\epsilon(\delta) > 0$ for any positive δ. Since the number

$\epsilon(\delta) > 0$ is easily written using quantifiers and algebraic conditions, the Tarski-Seidenberg theorem implies that it is a semi-algebraic set and hence satisfies some algebraic equation[3] whose coefficients are polynomial in δ.

We write this polynomial as $q(x) = q_M(\delta)x^M + \ldots + q_0(\delta)$, such that $q(\epsilon(\delta)) = 0$. We can assume that $q_0(\delta)$ is not identically zero (dividing by an appropriate power of x if necessary). From Lemma II.6 we see that if $q(\epsilon(\delta)) = 0$ then

$$\epsilon(\delta) > \frac{|q_0(\delta)|}{\sum_{i=1}^{M} |q_i(\delta)|}.$$

The last quantity is a ratio of two polynomials in δ and can thus be lower bounded by $C(\delta^{t'})$, so that $\epsilon(\delta) > C\delta^{t'}$ for some $t' > 0$, when δ is sufficiently small.

Putting $t = \frac{1}{t'}$ and recalling the definition of S_δ, we see that $Q(\theta_1, \theta_2) < \epsilon$, implies $\theta_1 \in \mathcal{N}(\theta_2, O(\epsilon^t))$, which completes the proof of the theorem. □

Lemma II.6. *Let δ be a positive root of the polynomial $q(x) = a_M x^M + \ldots + a_0$, $a_0 \neq 0$. Then $\delta > \min(\frac{\sum_{i=1}^{M} |a_i|}{|a_0|}, 1)$.*

PROOF:We have $\delta(\sum_{i=1}^{M} a_i \delta^{i-1}) = -a_0$. For $0 < \delta < 1$ we have $\sum_{i=1}^{M} a_i \delta^{i-1} < \sum_{i=1}^{M} |a_i|$, and the statement follows. □

Proposition II.7 (Upper bound)**.** *Let p_θ be a polynomial family. For any $N \in \mathbb{N}$ there exists a $C > 0$, such that*

$$\sum_{i=1}^{N} |M_i(\theta_1) - M_i(\theta_2)|^2 < C\|\theta_1 - \theta_2\|^2.$$

If Θ is contained in a ball of radius B, then C is bounded from above by a polynomial of B.

PROOF:To prove the claim it is sufficient to show that each summand $|M_i(\theta_1) - M_i(\theta_2)|^2$ is bounded from above by $C'\|\theta_1 - \theta_2\|^2$, which is equivalent to proving that $\frac{|M_i(\theta_1) - M_i(\theta_2)|}{\|\theta_1 - \theta_2\|} < \sqrt{C'}$. We now observe that by the mean value theorem

$$\frac{|M_i(\theta_1) - M_i(\theta_2)|}{\|\theta_1 - \theta_2\|} \leq \sup_{\theta \in \Theta} \|\mathrm{grad}(M_i)(\theta)\|$$

where grad is the gradient of the function M_i. Since M_i is a polynomial, all elements of the vector $\mathrm{grad}(M_i)$ are polynomial in θ. Therefore

$$\sup_{\theta \in \Theta} \|\mathrm{grad}(M_i)(\theta)\| < C'' B^t$$

where t is the maximum degree of these polynomials and C'' is an appropriate constant. This implies the statement of the Proposition. □

Now we have the following:

Theorem II.8. *There exists an algorithm, which, given $\epsilon > 0$ and $1 > \delta > 0$, and $P(\frac{1}{\epsilon}, \frac{1}{\delta}, B)$ samples from $p_\theta, \theta \in \Theta$, where Θ is the set of parameters within a ball of radius B and P is a polynomial depending only on the distribution family, outputs $\hat{\theta}$, s.t. $\hat{\theta} \in \mathcal{N}(\theta, \epsilon)$ with probability at least $1 - \delta$. The algorithm also requires a polynomial number of operations.*

PROOF:From Theorem II.5 it follows that there exists an $N \in \mathbb{N}$ and $t > 0$, such that if $\forall_{i=1,\ldots,N} |M_i(\hat{\theta}) - M_i(\theta)| < \epsilon^t$, than $\hat{\theta} \in \mathcal{N}(\theta, \epsilon)$. Thus it is sufficient to estimate each moment within $O(\epsilon^t)$. From Lemma D.1 (moment estimation) this can be done with probability $1-\delta$ given a number of sample points $\mathrm{poly}(\frac{1}{\epsilon^t}, \frac{1}{\delta}, B) = \mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, B)$ by computing the empirical moments of the sample. Once we have precise estimates of the first N moments a simple grid search suffices to find the corresponding values of parameters. Indeed, suppose that Θ is contained in a ball of radius B in $\mathbb{R}^m$. Then the desired estimate can be obtained by conducting a grid search over a rectangular grid of size $O(\frac{\epsilon^t}{N\sqrt{m}})$ and invoking Proposition II.7. We see that the number of operations is polynomial in ϵ and the main theorem is proved. □

To simplify further discussion we will now define the *radius of identifiability*:

Definition II.9. *As before let p_θ, $\theta \in \Theta$ be a family of probability distributions. For each θ we define the radius of identifiability $\mathscr{R}(\theta)$ as the supremum of the following set*

$$\{r > 0 | \forall \theta_1 \neq \theta_2, (\|\theta_1 - \theta\| < r, \|\theta_2 - \theta\| < r) \Rightarrow (p_{\theta_1} \neq p_{\theta_2})\}$$

In other words, $\mathscr{R}(\theta)$ is the largest number, such that the open ball of radius $\mathscr{R}(\theta)$ around θ intersected with Θ is an identifiable (sub)family of probability distributions. If no such ball exists, $\mathscr{R}(\theta) = 0$.

From Theorem II.8 and the definition of the radius of identifiability we have the following

Corollary II.10. *There is an algorithm, which, given $\epsilon > 0$, for any identifiable $\theta \in \Theta$, where Θ is the set of parameters within a ball of radius B, outputs $\hat{\theta}$ within $\min(\epsilon, \mathscr{R}(\theta))$ of θ with probability $1 - \delta$, using a number of sample points from p_θ polynomial in $\max\left(\frac{1}{\epsilon}, \frac{1}{\mathscr{R}(\theta)}\right)$, $\frac{1}{\delta}$ and B.*

Corollary II.11. *More generally, if $\theta \in \Theta$, where Θ is the set of parameters within a ball of radius B, is not identifiable but, $\mathcal{E}(\theta) = \{\theta_1, \ldots, \theta_k\}$ is a finite set, there exists an algorithm, such that, given $\epsilon > 0$, it outputs $\hat{\theta}$ within $\min(\epsilon, \min_j \mathscr{R}(\theta_j))$ of θ_i for some $i \in \{1, \ldots, k\}$ with probability $1 - \delta$, using a number of sample points from p_θ polynomial in $\max\left(\frac{1}{\epsilon}, \frac{1}{\min_j \mathscr{R}(\theta_j)}\right)$, $\frac{1}{\delta}$ and B.*

This last result is what we need to analyze Gaussian mixture model in the next Section.

Remark: It is important to note that the radius of identifiability depends on the choice of family Θ. Specifically, the radius is a decreasing function on the family of the sets Θ ordered by inclusion.

III. Gaussian Distributions and Polynomially Reducible High Dimensional Families

The main result of this section is to show that there exists an algorithm for estimating parameters of high-dimensional

[3]Note that strict inequalities alone cannot define a set consisting of a single point.

Gaussian mixture distributions in time polynomial in the dimension n and other parameters. We note that the techniques from the previous section cannot be applied directly to high-dimensional distributions since the number of parameters generally increases with dimension. Instead our approach will be to show that parameters of high-dimensional Gaussians can be estimated using $\mathrm{poly}(n)$ linear projections to linear subspaces, whose dimension is independent of n. We will call this property *polynomial reducibility* and will also briefly discuss some other families satisfying this condition later in the section.

We will now specifically discuss the case of a mixture of Gaussian distributions. Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i)$ be a mixture of k Gaussian distributions in $\mathbb{R}^n$, with means μ_i and covariance matrices Σ_i. Let us consider the parameters of the distribution $\theta = (\mu_1, \Sigma_1, w_1, \ldots, \mu_k, \Sigma_k, w_k)$ as a single vector (thus flattening the covariance matrices). We take the usual Euclidean distance in this space (which, in fact, corresponds to the Frobenius distance for the covariance matrices).

We will assume that the number of components k is fixed. We note that any permutation of the mixture components leads to the same density function and hence cannot be identified from data. On the other hand, it is well known ([23]) that the density of the distribution determines the parameters uniquely up to a permutation, if and only if any two components with the same means have different covariance matrices and no mixing coefficient is equal to zero.

The main result of the section is given by the following

Theorem III.1. *Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i), \theta \in \Theta$, where Θ is the set of parameters within a ball of radius B, be a mixture of Gaussian distributions in $\mathbb{R}^n$ with radius of identifiability $\mathscr{R}(\theta)$. Then there exists an algorithm which given $\epsilon > 0$ and $1 > \delta > 0$, and poly $\left(n, \max\left(\frac{1}{\epsilon}, \frac{1}{\mathscr{R}(\theta)}\right), \frac{1}{\delta}, B\right)$ samples from p_θ, with probability greater than $(1-\delta)$, outputs a parameter vector $\hat{\theta} = \left((\hat{\mu}_1, \hat{\Sigma}_1, \hat{w}_1), \ldots, (\hat{\mu}_k, \hat{\Sigma}_k, \hat{w}_k)\right) \in \Theta$, such that there exists a permutation $\sigma : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ satisfying,*

$$\sum_{i=1}^k \left(\|\mu_i - \hat{\mu}_{\sigma(i)}\|^2 + \|\Sigma_i - \hat{\Sigma}_{\sigma(i)}\|^2 + |w_i - \hat{w}_{\sigma(i)}|^2 \right) \leq \epsilon^2$$

We note that the radius of identifiability $\mathscr{R}(\theta)$ can be calculated explicitly from the Proposition III.3:

$$(\mathscr{R}(\theta))^2 = \min\left(\frac{1}{4} \min_{i \neq j} \left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2 \right), \min_i w_i^2 \right)$$

Thus if the mean/variance pairs for any two components are different with difference bounded from below and the minimum mixing weight is is also bounded from below, then we have explicit lower bound for $\mathscr{R}(\theta)$.

In fact, even when $\mathscr{R}(\theta)$ is not known in advance, it can be estimated from data.

Theorem III.2. *Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i), \theta \in \Theta$, where Θ is the set of parameters within in a ball of radius B, be a mixture of Gaussian distributions in $\mathbb{R}^n$ with radius of identifiability $\mathscr{R}(\theta)$. Then there exists an algorithm, which, given $\epsilon > 0$ and $1 > \delta > 0$, and poly $\left(n, \frac{1}{\epsilon}, \frac{1}{\delta}, B\right)$ samples from p_θ, outputs whether $\mathscr{R}(\theta) < \epsilon$ with probability greater than $1 - \delta$.*

The rest of the section is structured as follows:

In subsection III-A we discuss various properties of Gaussian mixture distributions. In particular we derive the formula for the radius of identifiability (Proposition III.3) and show that there exists a low-dimensional projection such that the radius of identifiability changes by at most a linear factor (Theorem III.7).

In subsection III-B we give a sketch for the proof of the main theorem, showing how the parameters of a high-dimensional distribution can be estimated from a polynomial number of projections. The details of the proof as well as the proof of Theorem III.2 are given in the appendix C.

Finally, we note that our results apply to high-dimensional distributions which are not mixtures of Gaussians with a fixed number of components. For example, a product of n 1-dimensional Gaussian mixture distributions each with k components, which is a Gaussian mixture distribution in n dimensions with k^n components, can be easily learned using our methods. The same applies to other product distributions whose components are polynomial families.

A. Gaussian Distributions

Proposition III.3. *Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i)$, $\theta \in \Theta$ be a family of mixtures of Gaussian distributions in $\mathbb{R}^n$ with non-zero mixing weights. Then the following inequality is satisfied:*

$$(\mathscr{R}(\theta))^2 \geq \min\left(\frac{1}{4} \min_{i \neq j} \left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2 \right), \min_i w_i^2 \right) \quad (5)$$

Moreover, suppose Θ is a convex set[4] such that it contains all possible mixing coefficients $(w_1, \ldots, w_k)$ for any fixed set of means and variances[5]

In this case the inequality becomes an equality:

$$(\mathscr{R}(\theta))^2 = \min\left(\frac{1}{4} \min_{i \neq j} \left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2 \right), \min_i w_i^2 \right) \quad (6)$$

In particular, the radius of identifiability is invariant under the permutation of components.

PROOF:We will start by proving the inequality 5. Suppose that the distributions $p_{\theta'}$ and $p_{\theta''}$ have the same density. To prove the inequality, we need to show that at least one of θ', θ'' is no closer to θ then the right hand side of the inequality 5.

Let us first consider the case when there is no pair $i \neq j$, s.t. $\mu_i' = \mu_j^{''}$ and $\Sigma_i' = \Sigma_j^{''}$. In that case that case at least one of the mixing coefficients for one of the mixtures must be equal to zero. That implies that either $\|\theta - \theta'\| \geq \min_i w_i$ or $\|\theta - \theta''\| \geq \min_i w_i$, which is consistent with the 5.

[4]Note that requiring convexity is natural, since the set of positive definite matrices is a convex cone.

[5]This requirement is unnecessarily strong, however, the precise condition, evident from the proof, is awkward to state.

Alternatively, suppose that for some $i \neq j$ we have $(\mu_i', \Sigma_i') = (\mu_j'', \Sigma_j'')$. Put $v' = (\mu_i', \Sigma_i') = (\mu_j'', \Sigma_j'')$, $v_1 = (\mu_i, \Sigma_i)$, $v_2 = (\mu_j, \Sigma_j)$. We see that

$$\|\theta''-\theta\|^2+\|\theta'-\theta\|^2 \geq \|v'-v_1\|^2+\|v'-v_2\|^2 \geq \frac{1}{2}\|v_1-v_2\|^2$$

$$= \frac{1}{2}\|\mu_i - \mu_j\|^2 + \frac{1}{2}\|\Sigma_i - \Sigma_j\|^2$$

Therefore, $\max\{\|\theta'-\theta\|^2, \|\theta''-\theta\|^2\} \geq \frac{1}{4}(\|\mu_i-\mu_j\|^2+\|\Sigma_i-\Sigma_j\|^2)$ which is again consistent with Inequality 5 and together with the first case implies the inequality.

To show Eq. 6 we need to observe that the bound is tight. Again we consider two possible cases. If the minimum in the right hand side of Eq. 6 is equal to the square of one of the mixing weights, say, w_i, construct θ' by putting $w_i' = 0$ and keeping the rest of the parameters of θ. We see that $\|\theta'-\theta\| = w_i$. By slightly perturbing μ', we see that there exists a θ'' arbitrarily close (but not equal)to θ' with the same probability density. Thus the radius of identifiability cannot exceed w_i.

Alternatively the minimum in the right hand side of Eq. 6 could be equal to $\frac{1}{4}\left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2\right)$ for some $i \neq j$. Construct θ' by putting $\mu_i' = \mu_j' = \frac{1}{2}(\mu_i - \mu_j)$ and $\Sigma_i' = \Sigma_j' = \frac{1}{2}(\Sigma_i-\Sigma_j)$ and keeping the rest of the parameters of θ. It is easy to see that $\|\theta'-\theta\|^2 = \frac{1}{4}\left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2\right)$. Note that $\theta' \in \Theta$ by the convexity condition. By perturbing w_i and w_j slightly, and keeping the rest of parameters fixed, we can obtain θ'' arbitrarily close to θ' with the same probability density. Hence the radius of identifiability does not exceed $\frac{1}{4}\left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2\right)$, which completes the proof. □

From the discussion above we have the following

Corollary III.4. *Let Θ be a convex set, such that for any $\theta \in \Theta$ all mixing coefficients w_i are nonzero. Then*

$$(\mathscr{R}(\theta))^2 = \frac{1}{4}\min_{i\neq j}\left(\|\mu_i - \mu_j\|^2 + \|\Sigma_i - \Sigma_j\|^2\right) \qquad (7)$$

It is also easy to see that the radius of identifiability satisfies a type of triangle inequality and that under any permutation of (mean, covariance matrix, mixing weight) triples the radius of identifiability does not change. This is expressed in the following two lemmas (the straightforward proofs are omitted):

Lemma III.5. *Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i)$, $\theta \in \Theta$ be a family of mixtures of Gaussian distributions in $\mathbb{R}^n$. For any $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 \in \mathcal{N}(\theta_2, \epsilon)$ for some $\epsilon > 0$, $|\mathscr{R}(\theta_1) - \mathscr{R}(\theta_2)| \leq \epsilon$.*

Lemma III.6. *Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i), \theta \in \Theta$ be a mixture of Gaussian distributions in $\mathbb{R}^n$. Suppose θ is represented as $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, where $\theta_i = (\mu_i, \Sigma_i, w_i)$ is the mean, covariance matrix, mixing weight triple. Let $\theta' = (\theta_{\sigma(1)}, \theta_{\sigma(2)}, \ldots, \theta_{\sigma(2)})$, where $\sigma : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ is a permutation. Then $\mathscr{R}(\theta) = \mathscr{R}(\theta')$.*

From now on, we will assume that Θ is a sufficiently large ball or cube (with the necessary conditions to make p_θ a valid probability distribution), so that we do not have to worry about convexity and other technical properties.

We now recall that a projection of a Gaussian mixture distribution onto a subspace is a lower-dimensional Gaussian mixture distribution. Specifically, if $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i)$, the Gaussian mixture distributions in $\mathbb{R}^n$, is projected onto a subspace S then the projection is a lower-dimensional Gaussian mixture distribution family $\pi_S(p_\theta)$, parameterized by $P_S(\theta)$. In particular, if S is a coordinate plane then P_S is a projection operator, which is an identity mapping for the mixing weights, an orthogonal projection onto S for the means and the restriction operator for the covariance matrices of the components, where each covariance matrix is projected to its minor corresponding to the coordinates in S.

We will now state the following Theorem whose proof is omitted due to space limitation.

Theorem III.7. *Let $p_\theta = \sum_{i=1}^k w_i N(\mu_i, \Sigma_i)$, $\theta \in \Theta$ be a Gaussian mixture distribution in $\mathbb{R}^n$ with radius of identifiability $\mathscr{R}(\theta)$. Then there exists a $2k^2$-dimensional coordinate plane S, such that $\mathscr{R}(P_S(\theta)) \geq \frac{1}{n}\mathscr{R}(\theta)$.*

B. Sketch of the Proof of Theorem III.1

We present a brief overview of the proof. The technical details can be found in Appendix C. The main idea is to show that parameters of high-dimensional Gaussian mixture can be estimated arbitrarily well using $\mathrm{poly}(n)$ projections to coordinate subspaces, whose dimension only depends on k. Since the dimension of these lower dimensional subspaces is independent of n, results from Section II can be used to estimate the parameters.

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, where $\theta_i = (\mu_i, \Sigma_i, w_i)$, be the parameter vector after flattening the covariance matrices. Recall that projection of p_θ onto a $2k^2$-coordinate plane T, will result in a mixture $\pi_T(p_\theta)$, parameterized (with a slight abuse of notation) by $P_T(\theta) = (P_T(\theta_1), P_T(\theta_2), \ldots, P_T(\theta_k))$.
Step 1: Let $\mathscr{R}(\theta)$ be the radius of identifiability. Theorem III.7 guarantees the existence of a $2k^2$-dimensional coordinate subspace S, such that $\mathscr{R}(P_S(\theta)) \geq \frac{1}{n}\mathscr{R}(\theta)$.

To identify such a subspace, we take all $\binom{n}{2k^2}$ coordinate projections. For each projection to a subspace T we estimate the parameters using the Theorem II.8. It is important to note that given ϵ' as input, Theorem II.8 is guaranteed to produce a value of parameter $\widehat{P_T(\theta)}$, such that $|\mathscr{R}(\widehat{P_T(\theta)}) - \mathscr{R}(P_T(\theta))| < \epsilon'$ (Lemma III.5) using a number of samples polynomial in k and $\frac{1}{\epsilon'}$. Applying the union bound for all $\binom{n}{2k^2}$ projections provides an estimate for the radius of identifiability for each projection within ϵ'. Choosing ϵ' appropriately (say, $\frac{\mathscr{R}(\theta)}{2n}$), and choosing the projection with the largest estimated radius of identifiability, yields a coordinate subspace S with a lower bounded $\mathscr{R}(P_S(\theta))$. We use S as a starting point for Step 2.
Step 2: By applying Corollary II.11 to the projection $P_S(\theta)$, we can estimate the mixing weights, projections of the original means and $2k^2 \times 2k^2$ minors of the covariance matrices corresponding to the coordinates within S. We now need to estimate the rest of the parameters using a sample size polynomial on n. We do this by estimating each additional coordinate

separately. That is for each coordinate i not in S we take $S_i = span(S, e_i)$, where e_i is the corresponding coordinate vector. It can be seen that the radius of identifiability does not decrease going from S to S_i. We show that the i'th coordinate of each component mean can be estimated by applying Corollary II.11 to the projection to S_i. We repeat this procedure for each of the $n - 2k^2$ coordinates not in S.

To estimate the covariance matrices we proceed similarly, except that we need to estimate entries corresponding to pairs of coordinates (i, j). Now we have two possibilities, since either one of i, j or both of them may not be in S. If exactly one of them, say i, is not in S, projection to S_i defined above can be used to estimate the corresponding entry of each covariance matrix. If both i, j are not in S, we take the projection onto $S_{ij} = span(S, e_i, e_j)$. By applying Corollary II.11, we show that the ij'th entry of covariance matrices can also be estimated.

Thus, after obtaining the initial space S, the complete set of parameters can be estimated using at most $n - 2k^2 + \binom{n-2k^2}{2}$ parameter estimations for $2k^2 + 1$ or $2k^2 + 2$-dimensional subspaces.

IV. Conclusion and Discussion

The results of this paper resolve the general problem of polynomial learning of Gaussian mixture distributions in high dimension. Our results do not require any separation assumptions and apply as long as the mixture is identifiable. For example, they apply even if all components of the mixture have the same mean distribution, as long as the covariance matrices are different and the mixing coefficients are non-zero.

We also provide quite general results applicable to learning various fixed-dimensional families, including mixtures and products of a number of standard distributions.

In high dimensions, our results can be applied beyond Gaussian mixture distributions with a fixed number of components. For example, one can also learn a product of n number of d-dimensional Gaussians mixtures with k components each (which is a nd-dimensional Gaussian mixture distribution with k^n components).

The proof brings the techniques of algebraic geometry to the classical method of moments, an approach that, as far as we know, is new to this domain.

We are planning to investigate other applications in learning of the framework presented in this paper. We also note that the methods proposed in the paper can be turned into implementable (and potentially practical) algorithms through the use of tools from computational algebraic geometry ([21]). This is also a direction of future investigation.

References

[1] B. G. Lindsay, *Mixture Models: Theory Geometry and Applications*. Institute of Mathematical Statistics, 1995.

[2] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[3] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, 1985.

[4] S. Dasgupta, "Learning Mixture of Gaussians," in *40th Annual Symposium on Foundations of Computer Science*, 1999.

[5] S. Dasgupta and L. Schulman, "A Two Round Variant of EM for Gaussian Mixtures," in *16th Conference on Uncertainty in Artificial Intelligence*, 2000.

[6] S. Arora and R. Kannan, "Learning Mixtures of Arbitrary Gaussians," in *33rd ACM Symposium on Theory of Computing*, 2001.

[7] S. Vempala and G. Wang, "A Spectral Algorithm for Learning Mixtures of Distributions," in *43rd Annual Symposium on Foundations of Computer Science*, 2002.

[8] R. Kannan, H. Salmasian, and S. Vempala, "The Spectral Method for General Mixture Models," in *18th Annual Conference on Learning Theory*, 2005.

[9] D. Achlioptas and F. McSherry, "On Spectral Learning of Mixture of Distributions," in *18th Annual Conference on Learning Theory*, 2005.

[10] S. C. Brubaker and S. Vempala, "Isotropic PCA and Affine-Invariant Clustering," in *49th Annual Symposium on Foundations of Computer Science*, 2008.

[11] K. Chaudhuri, S. Dasgupta, and A. Vattani, "Learning Mixtures of Gaussians Using the k-means Algorithm," *CoRR*, vol. abs/0912.0086, `http://arxiv.org/abs/0912.0086`, 2009.

[12] K. Chaudhuri and S. Rao, "Beyond Gaussians: Spectral Methods for Learning Mixtures of Heavy Tailed Distributions," in *The 21st Annual Conference on Learning Theory*, 2008.

[13] ——, "Learning Mixtures of Product Distributions Using Correlations and Independence," in *21st Annual Conference on Learning Theory*, 2008.

[14] A. Dasgupta, J. E. Hopcroft, J. Kleinberg, and M. Sandler, "On Learning Mixture of Heavy Tailed Distributions," in *46th Annual Symposium on Foundations of Computer Science*, 2005.

[15] J. Feldman, R. O'Donnell, and R. A. Servedio, "Learning Mixture of Product Distributions over Discrete Domains," *SIAM Journal on Computing*, vol. 37(5), pp. 1536–1564, 2008.

[16] J. Feldman, R. A. Servedio, and R. O'Donnell, "PAC Learning Axis Aligned Mixtures of Gaussians with No Separation Assumption," in *The 19th Annual Conference on Learning Theory*, 2006.

[17] M. Belkin and K. Sinha, "Toward Learning Gaussian Mixtures with Arbitrary Separation," in *23 Annual Conference on Learning Theory*, 2010.

[18] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently Learning Mixtures of Two Gaussians," in *41st ACM Symposium on Theory of Computing*, 2010.

[19] K. Pearson, "Contributions to the Mathematical Theory of Evolution," *Phil. Trans. Royal Soc.*, pp. 71–110, 1894.

[20] M. Drton, B. Strumfels, and S. Sullivant, *Lectures on Algebraic Statistics*. Birkhauser, 2009.

[21] S. Basu, R. Pollack, and M.-F. Roy, *Algorithms in Real Algebraic Geometry*. Springer., 2003.

[22] W. Feller, *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons Inc., 1971.

[23] H. Teicher, "Identifiability of Finite Mixtures," *Annals of Mathematical Statistics*, vol. 34(4), pp. 1265–1269, 1963.

Appendix A
Some Polynomial Families of Distributions

In Table II, we present a partial list of probability distributions which form polynomial families. Standard probability distributions that are not polynomial are more rare to find and include Weibull distribution and Cauchy distribution.

Appendix B
Separation Preserving Coordinate Planes

Let $p_\theta = \sum_{i=1}^{k} w_i N(\mu_i, \Sigma_i)$ be a mixture of k Gaussian distributions in $\mathbb{R}^l$, with means μ_i and covariance matrices Σ_i. When this distribution is projected onto any lower dimensional coordinate plane S, the corresponding Gaussian mixture $\pi_S(p_\theta)$, parameterized by $P_S(\theta)$, has means and covariance matrices represented by $P_S(\mu_i)$ and $P_S(\Sigma_i)$ respectively. We first show that if any pair of means or pair of covariance matrices of the original component Gaussian distributions are

Distribution	θ	Pdf/Pmf $f(x;\theta)$	Mgf $M(t)$	Moments Expression
				$\mathbb{E}(X^i) = \mu\mathbb{E}(X^{i-1}) + (i-1)\sigma^2\mathbb{E}(X^{i-2})$
Gaussian	μ, σ	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$e^{\mu t + a\frac{\sigma^2 t^2}{2}}$	$\mathbb{E}(X) = \mu$ $\mathbb{E}(X^2) = \mu^2 + \sigma^2$ $\mathbb{E}(X^3) = \mu^3 + 3\mu\sigma^2$
				$\mathbb{E}(X^i) = \frac{1}{i+1}\sum_{j=1}^{i} a^j b^{i-j}$
Uniform	a, b	$\frac{1}{b-a}$, $a \le x \le b$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\mathbb{E}(X) = \frac{a+b}{2}$ $\mathbb{E}(X^2) = \frac{a^2+ab+b^2}{3}$ $\mathbb{E}(X^3) = \frac{a^3+a^b+ab^2+b^3}{4}$
				$\mathbb{E}(X^i) = \prod_{j=0}^{i-1}(m+j)\beta^i$
Gamma	β, m	$\frac{x^{m-1}e^{-x/\beta}}{\beta^m \Gamma(m)}$, $x > 0$	$(1-\beta t)^{-m}$	$\mathbb{E}(X) = m\beta$ $\mathbb{E}(X^2) = m(m+1)\beta^2$ $\mathbb{E}(X^3) = m(m+1)(m+2)\beta^3$
				$\mathbb{E}(X^i) = \sum_{j=0}^{i} \frac{i! b^j \mu^{i-j}}{(i-j)!} 1\{j \text{ is even}\}$
Laplace	μ, b	$\frac{1}{2b} e^{-\frac{\lvert x-\mu \rvert}{b}}$	$\frac{e^{\mu t}}{1 - b^2 t^2}$	$\mathbb{E}(X) = \mu$ $\mathbb{E}(X^2) = \mu^2 + 2b^2$ $\mathbb{E}(X^3) = \mu^3 + 6\mu b^2$
				$\mathbb{E}(X^i) = i!\lambda^i$
Exponential	λ	$\frac{1}{\lambda} e^{-\frac{x}{\lambda}}$, $x > 0$	$(1-\lambda t)^{-1}$	$\mathbb{E}(X) = \lambda$ $\mathbb{E}(X^2) = 2\lambda^2$ $\mathbb{E}(X^3) = 6\lambda^3$
				$\mathbb{E}(X^i) = k(k+2)\cdots(k+2i-2)$
Chi-Square	k	$\frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{k/2}\Gamma(k/2)}$, $x > 0$	$(1-2t)^{-\frac{k}{2}}$	$\mathbb{E}(X) = k$ $\mathbb{E}(X^2) = k(k+2)$ $\mathbb{E}(X^3) = k(k+2)(k+4)$
				$\mathbb{E}(X^i) = (2i-3)\lambda\mu^2\mathbb{E}(X^{i-1}) + \mu^2\mathbb{E}(X^{i-2})$
Inverse Gaussian	μ, λ	$\sqrt{\frac{1}{2\pi\lambda x^3}} e^{-\frac{(x-\mu)^2}{2\lambda\mu^2 x}}$	$e^{\frac{(1-\sqrt{1-2\lambda\mu^2 t})}{\lambda\mu}}$	$\mathbb{E}(X) = \mu$ $\mathbb{E}(X^2) = \lambda\mu^3$ $\mathbb{E}(X^3) = 3\lambda^2\mu^5$
				$\mathbb{E}(X^i) = \lambda\mathbb{E}(X^{i-1}) + \lambda\frac{d(\mathbb{E}(X^{i-1}))}{d\lambda}$
Poisson	λ	$\frac{\lambda^x e^{-\lambda}}{x!}$	$e^{\lambda(e^t - 1)}$	$\mathbb{E}(X) = \lambda$ $\mathbb{E}(X^2) = \lambda^2 + \lambda$ $\mathbb{E}(X^3) = \lambda^3 + 3\lambda^2 + \lambda$
				$\mathbb{E}(X^i) = np\mathbb{E}(X^{i-1}) + p(1-p)\frac{d(\mathbb{E}(X^{i-1}))}{dp}$
Binomial	n, p	$\binom{n}{x} p^x (1-p)^{n-x}$	$(1-p+pe^t)^n$	$\mathbb{E}(X) = np$ $\mathbb{E}(X^2) = n(n-1)p^2 + np$ $\mathbb{E}(X^3) = (n^3 - 3n^2 + 2n)p^3 + 3n(n-1)p^2 + np$
				$\mathbb{E}(X^i) = \sum_{j=0}^{\infty} \frac{1}{p}\left(1 - \frac{1}{p}\right)^j j^i$
Geometric	p	$(1-\frac{1}{p})^x (\frac{1}{p})$	$\frac{1}{p-(p-1)e^t}$	$\mathbb{E}(X) = (p-1)$ $\mathbb{E}(X^2) = (p-1)(2p-1)$ $\mathbb{E}(X^3) = (p-1)(6p^2 - 6p + 1)$
				$\mathbb{E}(X^i) = rm\mathbb{E}(X^{i-1}) + m(m+1)\frac{d(\mathbb{E}(X^{i-1}))}{dm}$
Negative Binomial	r, m	$\binom{x+r-1}{r-1}\frac{m^x}{(m+1)^{r+x}}$	$\left(\frac{1}{m+1-me^t}\right)^r$	$\mathbb{E}(X) = rm$ $\mathbb{E}(X^2) = r(r+1)m^2 + rm$ $\mathbb{E}(X^3) = (r^3 + 3r^2 + 2r)m^3 + 3r(r+1)m^2 + rm$

TABLE II
COMMON POLYNOMIAL FAMILIES AND THEIR MOMENTS

separated, then they remain so after projecting the mixture distribution onto some suitable lower dimensional coordinate plane.

Lemma B.1. *For any $\mu_1, \mu_2, ..., \mu_k \in \mathbb{R}^l$, there exists a k^2-coordinate plane S such that,*

$$\forall_{i,j}, \|P_S(\mu_i) - P_S(\mu_j)\| \ge \|\mu_i - \mu_j\| \frac{1}{\sqrt{l}}$$

Lemma B.2. *For any $\Sigma_1, \Sigma_2, ..., \Sigma_k \in \mathbb{R}^{l \times l}$, there exists a k^2-coordinate plane S such that,*

$$\forall_{i,j}, \|P_S(\Sigma_i) - P_S(\Sigma_j)\| \ge \|\Sigma_i - \Sigma_j\| \frac{1}{l}$$

.

APPENDIX C
PROOF OF THEOREM III.1 AND THEOREM III.2

In this appendix we give the detailed proof of Theorem III.1 and Theorem III.2. We start with some preliminary Lemmas.

Lemma C.1. *Let $p_\theta = \sum_{i=1}^{k} w_i N(\mu_i, \Sigma_i)$, where θ is the set of parameters within a ball of radius B, be a mixture of Gaussian distributions in $\mathbb{R}^n$ with the radius of identifiability $\mathscr{R}(\theta)$. If θ is represented as $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, where $\theta_i = (\mu_i, \Sigma_i, w_i)$ is the mean, covariance matrix, mixing weight triple, (after flattening the covariance matrices) then for any $i \ne j$, $\|\theta_i - \theta_j\| \ge 2\mathscr{R}(\theta)$.*

Lemma C.2. *Let $p_\theta = \sum_{i=1}^{k} w_i N(\mu_i, \Sigma_i)$, where θ is the set of parameters within a ball of radius B, be a mixture of Gaussian distributions in $\mathbb{R}^n$ with radius of identifiability $\mathscr{R}(\theta)$. Let S and T be two lower-dimensional subspaces such that $S \subset T$. Then $\mathscr{R}(P_T(\theta)) \ge \mathscr{R}(P_S(\theta))$.*

Proof of Theorem III.1 :

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, where $\theta_i = (\mu_i, \Sigma_i, w_i)$, be parameter vector after flattening the covariance matrices. Recall that projection of p_θ onto any $2k^2$-coordinate plane T, will result in a mixture $\pi_T(p_\theta)$, which is parameterized (with a little abuse of notation) by $P_T(\theta) = (P_T(\theta_1), P_T(\theta_2), \ldots, P_T(\theta_k))$.

Details of Step 1: Let $\gamma = \min\left(\frac{\mathscr{R}(\theta)}{n}, \frac{\epsilon}{n}\right)$ where $\mathscr{R}(\theta)$ is the radius of identifiability. Theorem III.7 guarantees the existence of a $2k^2$-dimensional coordinate subspace S, such that $\mathscr{R}(P_S(\theta)) \ge \frac{1}{n}\mathscr{R}(\theta) \ge \gamma$. To identify such a subspace, consider all $\binom{n}{2k^2}$ coordinate projections. For any fixed projection to a $2k^2$-dimensional subspace T, invoking Theorem II.8 using a sample of size $\text{poly}(\frac{1}{\gamma}, \frac{1}{\delta}, B)$, (setting the precision parameter to $\frac{\gamma}{3}$) produces a value of parameters $\widehat{P_T(\theta)}$ such that $|\mathscr{R}(\widehat{P_T(\theta)}) - \mathscr{R}(P_T(\theta))| < \frac{\gamma}{3}$ (Lemma III.5). Applying the union bound for all $\binom{n}{2k^2}$ coordinate projections provides an estimate for the radius of identifiability for each projection within $\frac{\gamma}{3}$. Thus invoking Theorem II.8 $\binom{n}{2k^2}$ times, each time using a sample of size $\text{poly}\left(\frac{1}{\gamma}, \frac{1}{(\delta/4n^2)}, B\right)$, (setting the precision parameters to $\frac{\gamma}{3}$ and $\frac{\delta}{4n^2}$ respectively) and choosing the projection with the largest estimated radius of identifiability, yields a coordinate subspace S such that with probability at least $(1 - \frac{\delta}{4})$, $\mathscr{R}(P_S(\theta)) \ge \frac{\gamma}{3}$.

Details of Step 2: By applying Corollary II.11 to the mixture $\pi_S(p_\theta)$, where S is obtained in Step 1, using a sample of size $\text{poly}(\frac{1}{\gamma}, \frac{1}{\delta}, B)$, (setting the precision parameters to $\frac{\gamma}{9}$ and $\frac{\delta}{4}$ respectively) with probability greater than $(1 - \frac{\delta}{4})$ we can get an estimate of $\widehat{P_S(\theta)}$ satisfying $\|\widehat{P_S(\theta)} - P_S(\theta)\| \le \frac{\gamma}{9}$. Note that these estimates encompass the mixing weights, projections of the original means and $2k^2 \times 2k^2$ minors of the covariance matrices corresponding to the coordinates within S. If we let θ' to be $P_S(\theta)$ then the estimate $\hat{\theta}' = \widehat{P_S(\theta)}$ is, up to a permutation, within $\frac{\gamma}{9}$ of θ' with probability greater than $(1 - \frac{\delta}{4})$. Note that the dimension of θ' is $(k-1) + k\left(2k^2 + \frac{2k^2(2k^2+1)}{2}\right)$.

We now need to estimate the rest of the parameters using a sample size polynomial on n. This procedure explained in the following two sub-steps.

2a: Estimating means and part of covariance matrices

In this sub-step we estimate each additional coordinate separately. That is for each coordinate i not in S we take $S_i = span(S, e_i)$, where e_i is the corresponding coordinate vector. For each such S_i (note that there are $(n - 2k^2)$ such $(2k^2 + 1)$-coordinates planes corresponding to each i not in S), we project p_θ onto S_i and invoke the algorithm of Corollary II.11 (setting the precision parameters to $\frac{\gamma}{9}$ and $\frac{\delta}{4n}$ respectively) using a sample of size $\text{poly}\left(\frac{1}{\gamma}, \frac{1}{(\delta/4n)}, B\right)$. Noting that radius of identifiability does not decrease going from S to S_i (Lemma C.2) and applying union bound ensures that, with probability at least $(1 - \frac{\delta}{4})$, each time we get an estimate $\widehat{P_{S_i}(\theta)}$ such that $\|\widehat{P_{S_i}(\theta)} - P_{S_i}(\theta)\| \le \frac{\gamma}{9}$. Since $P_S(\theta) \subset P_{S_i}(\theta)$ letting $\phi_i = P_{S_i}(\theta) \setminus P_S(\theta)$ to be the extra

parameters, we have for each i,

$$\|\hat{\phi}_i - \phi_i\| = \|\widehat{P_{S_i}(\theta)} - P_{S_i}(\theta)\| \leq \frac{\gamma}{9}$$

with probability greater than $(1-\frac{\delta}{4})$, where $\hat{\phi}$ is the estimate of ϕ. Since for each S_i, $\forall_{m\neq n}, \|P_{S_i}(\theta_m) - P_{S_i}(\theta_n)\| \geq \frac{2\gamma}{3}$, (using the fact $\mathscr{R}(P_S(\theta)) \geq \frac{\gamma}{3}$, Lemma C.1 and Lemma C.2), estimates of the extra parameters can be uniquely associated to the parameters of the component Gaussian distributions estimated in Step 2.

Letting $\theta^{''}$ to be $\cup_{i=1}^{n-2k^2}\phi_i$, we have

$$\|\hat{\theta}^{''} - \theta^{''}\| = \left(\sqrt{\sum_{i=1}^{n-2k^2}\|\hat{\phi}_i - \phi_i\|^2}\right) \leq \sqrt{\left(\frac{\gamma}{9}\right)^2 (n-2k^2)} < \left(\frac{\gamma}{9}\right) n$$

with probability greater than $(1-\frac{\delta}{4})$, where $\hat{\theta}^{''}$ is the estimate of $\theta^{''}$.

Note that the dimension of $\theta^{''}$ is $k(n-2k^2)(2+2k^2)$, where each $P_{S_i}(\theta) \setminus P_S(\theta)$ encompasses i'th coordinate for each component mean, i'th diagonal entry for each component covariance matrix and $2k^2$ extra off diagonal entries for each component covariance matrix.

2b: Estimating the remaining entries of covariance matrices

To estimate the the remaining parameters of the covariance matrices we need to estimate entries corresponding to pairs of coordinates (i,j) when both i and j are not in S. We consider projection onto $S_{ij} = span(S, e_i, e_j)$. For each such projection S_{ij} (note that there are $\binom{n-2k^2}{2}$ such $(2k^2+2)$-coordinate planes corresponding to i,j not in S), we project p_θ onto S_{ij} and invoke the algorithm of Corollary II.11, (setting the precision parameters to $\frac{\gamma}{9}$ and $\frac{\delta}{4n^2}$ respectively) using a sample of size $\text{poly}\left(\frac{1}{\gamma}, \frac{1}{(\delta/4n^2)}, B\right)$. Noting that radius of identifiability does not decrease going from S to S_{ij} (Lemma C.2) and applying union bound ensures that, with probability at least $(1-\frac{\delta}{4})$, each time we get an estimate $\widehat{P_{S_{ij}}(\theta)}$ such that $\|\widehat{P_{S_{ij}}(\theta)} - P_{S_{ij}}(\theta)\| \leq \frac{\gamma}{9}$. Since $P_S(\theta) \subset P_{S_{ij}}(\theta)$ in each case and there are $\binom{n-2k^2}{2}$ such cases, letting $\psi_t, t = 1,\ldots,\binom{n-2k^2}{2}$ to be the extra parameters in each case, we have for each t, $\|\hat{\psi}_t - \psi_t\| \leq \frac{\gamma}{9}$ with probability greater than $(1-\frac{\delta}{4})$, where $\hat{\psi}_t$ is the estimate of ψ_t. As before estimates of these extra parameters can be uniquely associated to the parameters of the component Gaussian distributions estimated in Step 2.

Letting $\theta^{'''}$ to be the $k\binom{n-2k^2}{2}$ covariance parameters that have not been estimates in the previous steps, we have $\theta^{'''} \subset \cup_{t=1}^{\binom{n-2k^2}{2}}\psi_t$, and in particular,

$$\|\hat{\theta}^{'''} - \theta^{'''}\| \leq \left(\sqrt{\sum_{t=1}^{\binom{n-2k}{2}}\|\hat{\psi}_t - \psi_t\|^2}\right) \leq \sqrt{\left(\frac{\gamma}{9}\right)^2 \binom{n-2k^2}{2}} < \left(\frac{\gamma}{9}\right) n$$

with probability greater than $(1-\frac{\delta}{4})$, where $\hat{\theta}^{'''}$ is the estimate of $\theta^{'''}$.

In Step 1 we need to invoke Theorem II.8 $\binom{n}{2k^2}$ times. In step 2 we need to invoke Corollary II.11 $\left(1+(n-2k^2)+\binom{n-2k^2}{2}\right)$ times. Thus total invocation of Theorem II.8 and Corollary II.11 combined is $\text{poly}(n)$. It is easy to see that $\gamma \leq \frac{\epsilon}{n}$.

Since $\theta^{'} \cup \theta^{''} \cup \theta^{'''} = \theta$, the corresponding estimate (with a little abuse of notation) $\hat{\theta} = \hat{\theta}^{'} \cup \hat{\theta}^{''} \cup \hat{\theta}^{'''}$, with probability greater than $(1-\delta)$, is within ϵ of θ only up to a permutation using a sample of size $\text{poly}\left(n, \max\left(\frac{1}{\epsilon}, \frac{1}{\mathscr{R}(\theta)}\right), \frac{1}{\delta}, B\right)$. This completes the proof of the Theorem III.1. □

Proof of Theorem III.2 :

Theorem III.7, guarantees the existence of a $2k^2$-coordinate plane S such that when p_θ is projected onto S, the corresponding mixture $\pi_S(p_\theta)$, parameterized by $P_S(\theta)$, satisfies $\mathscr{R}(P_S(\theta)) \geq \mathscr{R}(\theta)\frac{1}{n}$. Since S is not known in advance, projecting p_θ on to all $\binom{n}{2k^2}$, $2k^2$-coordinate planes, each time invoking the algorithm of Theorem II.8 (setting the precision parameters to $\frac{\epsilon}{3n}$ and $\frac{\delta}{n^2}$ respectively) with a sample of size $\text{poly}\left(\frac{1}{(\epsilon/3n)}, \frac{1}{(\delta/n^2)}, B\right)$ and using union bound ensures that for each $2k^2$-coordinate plane T, Theorem II.8 produces a value of parameters $\widehat{P_T(\theta)}$ such that $\widehat{P_T(\theta)} \in \mathcal{N}\left(P_T(\theta), \frac{\epsilon}{3n}\right)$ with probability greater than $(1-\delta)$. Now for each such $2k^2$-coordinate plane T, Lemma III.5 guarantees that $|\widehat{\mathscr{R}P_T(\theta)} - \mathscr{R}P_T(\theta)| \leq \frac{\epsilon}{3n}$. Thus there must exist at least one $2k^2$-coordinate plane (say T_*) such that, $\mathscr{R}(\widehat{P_{T_*}(\theta)}) \geq \mathscr{R}(\theta)\frac{1}{n} - \frac{\epsilon}{3n}$. Thus,

$$(\mathscr{R}(\theta) \geq \epsilon) \Rightarrow \left(\mathscr{R}(\widehat{P_{T_*}(\theta)}) \geq \frac{2\epsilon}{3n}\right)$$

The desired algorithm now works as follows. For each of the $\binom{n}{2k^2}$ values of parameters $\widehat{P_T(\theta)}$ outputted by Theorem II.8, we compute $\mathscr{R}(\widehat{P_T(\theta)})$ using Equation 6. Now set $\mathscr{R}_* = \max_T \mathscr{R}(\widehat{P_T(\theta)})$. If $\mathscr{R}_* < \frac{2\epsilon}{3n}$ then output $\mathscr{R}(\theta) < \epsilon$ otherwise output $\mathscr{R}(\theta) \geq \epsilon$. □

APPENDIX D
MOMENT CONCENTRATION

Lemma D.1. *Let $p_\theta, \theta \in \Theta \subset \mathbb{R}^m$ be a m-parametric family of probability distributions in $\mathbb{R}^l$ where Θ is contained in a ball of radius B in $\mathbb{R}^m$ and let $X_1, X_2, \ldots, X_M$ be iid random vectors drawn from p_θ. Suppose the moments $M_{i_1\ldots i_l}(\theta) = \int x_1^{i_1}\ldots x_l^{i_l} dp_\theta$ and the corresponding empirical moments $\hat{M}_{i_1\ldots i_l}(\theta) = \frac{\sum_{i=1}^M X_{i,1}^{i_1}\ldots X_{i,l}^{i_l}}{M}$ are lexicographically ordered as $M_1(\theta), M_2(\theta), \ldots$ and $\hat{M}_1(\theta), \hat{M}_2(\theta), \ldots$ respectively. Then given any positive integer N, and sample size $M > \frac{CNB^{2\lceil \frac{N}{l} \rceil}}{\epsilon^2\delta}$ where C is a constant, for any $\epsilon > 0$ and $0 < \delta < 1$, $|\hat{M}_i(\theta) - M_i(\theta)| \leq \epsilon$ for all $i \leq N$ with probability greater than $1-\delta$.*