*Submitted to the Annals of Statistics*

# OPTIMAL ESTIMATION OF HIGH-DIMENSIONAL GAUSSIAN LOCATION MIXTURES

BY NATALIE DOSS[1,a], YIHONG WU[1,b], PENGKUN YANG[2,d], AND HARRISON H. ZHOU[1,c]

[1]*Department of Statistics and Data Science, Yale University* , [a]*natalie.doss@aya.yale.edu*; [b]*yihong.wu@yale.edu*; [c]*huibin.zhou@yale.edu*

[2]*Center for Statistical Science, Department of Industrial Engineering, Tsinghua University* , [d]*yangpengkun@tsinghua.edu.cn*

This paper studies the optimal rate of estimation in a finite Gaussian location mixture model in high dimensions without separation conditions. We assume that the number of components $k$ is bounded and that the centers lie in a ball of bounded radius, while allowing the dimension $d$ to be as large as the sample size $n$. Extending the one-dimensional result of Heinrich and Kahn [38], we show that the minimax rate of estimating the mixing distribution in Wasserstein distance is $\Theta((d/n)^{1/4} + n^{-1/(4k-2)})$, achieved by an estimator computable in time $O(nd^2 + n^{5/4})$. Furthermore, we show that the mixture density can be estimated at the optimal parametric rate $\Theta(\sqrt{d/n})$ in Hellinger distance and provide a computationally efficient algorithm to achieve this rate in the special case of $k = 2$.

Both the theoretical and methodological development rely on a careful application of the method of moments. Central to our results is the observation that the information geometry of finite Gaussian mixtures is characterized by the moment tensors of the mixing distribution, whose low-rank structure can be exploited to obtain a sharp local entropy bound.

**1. Introduction.** Mixture models are useful tools for dealing with heterogeneous data. A mixture model posits that the data are generated from a collection of sub-populations, each governed by a different distribution. The Gaussian mixture model is one of the most widely studied mixture models because of its simplicity and wide applicability; however, optimal rates of both parameter and density estimation in this model are not well understood in high dimensions. Consider the $k$-component Gaussian location mixture model in $d$ dimensions:

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{k} w_j N(\mu_j, \sigma^2 I_d), \tag{1.1}$$

where $\mu_j \in \mathbb{R}^d$ and $w_j \geq 0$ are the center and the weight of the $j$th component, respectively, with $\sum_{j=1}^{k} w_j = 1$. Here the scale parameter $\sigma^2$ and $k$ as an upper bound on the number of components are assumed to be known; for simplicity, we assume that $\sigma^2 = 1$. Equivalently, we can view the Gaussian location mixture (1.1) as the convolution

$$P_\Gamma \triangleq \Gamma * N(0, I_d) \tag{1.2}$$

between the standard normal distribution and the *mixing distribution*

$$\Gamma = \sum_{j=1}^{k} w_j \delta_{\mu_j}, \tag{1.3}$$

*MSC2020 subject classifications:* Primary 62G05, 62G07; secondary 62C20.

*Keywords and phrases:* Gaussian mixture, finite mixture model, high-dimensional density estimation, method of moments, low-rank tensor, metric entropy, minimax optimality.

which is a $k$-atomic distribution on $\mathbb{R}^d$.

For the purpose of estimation, the most interesting regime is one in which the centers lie in a ball of bounded radius and are allowed to overlap arbitrarily. In this case, consistent clustering is impossible but the mixing distribution and the mixture density can nonetheless be accurately estimated. Indeed, [15, 38] provided optimal convergence rates for general one-dimensional mixtures, including Gaussian location mixtures, under weak conditions, while [81] proposed a practical algorithm that achieves the optimal rate specifically for one-dimensional Gaussian mixtures. This paper extends the procedure and results in [81] to high dimensions.

1.1. *Main results.* We start by defining the relevant parameter space. Let $\mathcal{G}_{k,d}$ denote the collection of $k$-atomic distributions supported on a ball of radius $R$ in $d$ dimensions,[1] i.e.,

$$\mathcal{G}_{k,d} \triangleq \left\{ \Gamma = \sum_{j=1}^{k} w_j \delta_{\mu_j} : \mu_j \in \mathbb{R}^d, \|\mu_j\|_2 \le R, w_j \ge 0, \sum_{j=1}^{k} w_j = 1 \right\}, \tag{1.4}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Throughout the paper, $R$ is assumed to be an absolute constant. The corresponding collection of $k$-Gaussian mixtures ($k$-GMs) is denoted by

$$\mathcal{P}_{k,d} = \{P_\Gamma : \Gamma \in \mathcal{G}_{k,d}\}, \qquad P_\Gamma = \Gamma * N(0, I_d). \tag{1.5}$$

Let $\phi_d(x) = (2\pi)^{-d/2} e^{-\|x\|_2^2/2}$ denote the standard normal density in $d$ dimensions. Then the density of $P_\Gamma$ is given by

$$p_\Gamma(x) = \sum_{j=1}^{k} w_j \phi_d(x - \mu_j). \tag{1.6}$$

We first discuss the problem of parameter estimation. The distribution (1.1) has $kd + k - 1$ parameters: $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and $w_1, \ldots, w_k$ that sum up to one. Without extra assumptions such as separation between centers or a lower bound on the weights, estimating individual parameters is clearly impossible; nevertheless, estimation of the mixing distribution $\Gamma = \sum w_i \delta_{\mu_i}$ is always well-defined. Reframing the parameter estimation problem in terms of estimating the mixing distribution allows for the development of a meaningful statistical theory in an assumption-free framework [38, 81] since the mixture model is uniquely identified through the mixing distribution.

For mixture models and deconvolution problems, the Wasserstein distance is a natural and commonly-used loss function ([15, 61, 39, 40, 38, 81]). For $q \ge 1$, the $q$-Wasserstein distance (with respect to the Euclidean distance) is defined as

$$W_q(\Gamma, \Gamma') \triangleq \left(\inf \mathbb{E} \left\|U - U'\right\|_2^q\right)^{\frac{1}{q}}, \tag{1.7}$$

where the infimum is taken over all couplings of $\Gamma$ and $\Gamma'$, i.e., joint distributions of random vectors $U$ and $U'$ with marginals $\Gamma$ and $\Gamma'$, respectively. We will mostly be concerned with the case of $q = 1$, although the $W_2$-distance will make a brief appearance in the proofs. In one dimension, the $W_1$-distance coincides with the $L_1$-distance between the cumulative distribution functions [79]. For multivariate distributions, there is no closed-form expression,

[1] If the mixing distributions have unbounded support, the minimax risk under the Wasserstein distance is infinite (see [81, Sec. 4.4]).

and the $W_1$-distance can be computed by linear programming. In the widely-studied case of the symmetric 2-GM in which

$$P_\mu = \frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d), \tag{1.8}$$

the mixing distribution is $\Gamma_\mu = \frac{1}{2}(\delta_{-\mu} + \delta_\mu)$, and the Wasserstein distance coincides with the commonly-used loss function $W_1(\Gamma_\mu, \Gamma_{\mu'}) = \min\{\|\mu - \mu'\|_2, \|\mu + \mu'\|_2\}$. In this paper we do not postulate any separation conditions or any lower bound on the mixing weights; nevertheless, given such assumptions, statistical guarantees in $W_1$-distance can be translated into those for the individual parameters cf. [81, Lemma 1].

For general $k$-GMs in one dimension where $k \geq 2$ is a constant, the minimax $W_1$-rate of estimating the mixing distribution is $n^{-1/(4k-2)}$, achieved by a minimum $W_1$-distance estimator [38] or the Denoised Method of Moments (DMM) approach [81]. This is the worst-case rate in the absence of any separation assumptions. In the case where the centers can be grouped into $k_0$ clusters each separated by a constant, the optimal rate improves to $n^{-1/(4(k-k_0)+2)}$, which reduces to the parametric rate $n^{-1/2}$ in the fully separated case.

Given the one-dimensional result, it is reasonable to expect that the $d$-dimensional rate is given by $(d/n)^{1/(4k-2)}$. This conjecture turns out to be incorrect, as the following result shows.

THEOREM 1.1 (Estimating the mixing distribution). *Let $k \geq 2$ and $P_\Gamma$ be the $k$-GM defined in (1.2). Given $n$ i.i.d. observations from $P_\Gamma$, the minimax risk of estimating $\Gamma$ over the class $\mathcal{G}_{k,d}$ satisfies*

$$\inf_{\hat{\Gamma}} \sup_{\Gamma \in \mathcal{G}_{k,d}} \mathbb{E}_\Gamma W_1(\hat{\Gamma}, \Gamma) \asymp_k \left(\frac{d}{n}\right)^{1/4} \wedge 1 + \left(\frac{1}{n}\right)^{1/(4k-2)}, \tag{1.9}$$

*where the notation $\asymp_k$ means that both sides agree up to constant factors depending only on $k$. Furthermore, if $n \geq d$, there exists an estimator $\hat{\Gamma}$, computable in $O(nd^2) + O_k(n^{5/4})$ time, and an absolute constant $C$, such that for any $\Gamma \in \mathcal{G}_{k,d}$ and any $0 < \delta < \frac{1}{2}$, with probability at least $1 - \delta$,*

$$W_1(\hat{\Gamma}, \Gamma) \leq C\left(\sqrt{k}\left(\frac{d}{n}\right)^{1/4} + k^5\left(\frac{1}{n}\right)^{1/(4k-2)}\sqrt{\log\frac{1}{\delta}}\right). \tag{1.10}$$

The estimator in Theorem 1.1 achieves the minimax rate in the worst-case scenario where no lower bounds on the weights or separation or weights are imposed. While the main aim of this result is to demonstrate the achievability of the optimal rate in time that is polynomial in $n$ and $d$ for constant $k$, this estimator involves an exhaustive grid search and is far from being practical except for small $k$ as the hidden constant depends on $k$ as $k^{k^2}$. Finding a practical algorithm that provably achieves the optimal rate in Theorem 1.1 in the worst case remains an outstanding question.

We now explain the intuition behind the minimax rate (1.9). The atoms $\mu_1, \ldots, \mu_k$ of $\Gamma$ span a subspace $V$ in $\mathbb{R}^d$ of dimension at most $k$. We can identify $\Gamma$ with this subspace and its projection therein, which is a $k$-atomic mixing distribution in $k$ dimensions. This decomposition motivates a two-stage procedure which achieves the optimal rate (1.9):

- First, estimate the subspace $V$ by principal component analysis (PCA), then project the $d$-dimensional data onto the learned subspace. Since we do not impose any spectral gap assumptions, standard perturbation theory cannot be directly applied; instead, one needs to control the Wasserstein loss incurred by the subspace estimation error, which turns out to be $(d/n)^{1/4}$.

- Having reduced the problem to $k$ dimensions, a relevant notion is the *sliced Wasserstein distance* [64, 20, 63], which measures the distance of multivariate distributions by the maximal $W_1$-distance of their one-dimensional projections. We show that for $k$-atomic distributions in $\mathbb{R}^k$, the ordinary and the sliced Wasserstein distance are comparable up to constant factors depending only on $k$. This allows us to construct an estimator for a $k$-dimensional mixing distribution whose one-dimensional projections are simultaneously close to their estimates. We shall see that the resulting error is $n^{-1/(4k-2)}$, exactly as in the one-dimensional case.

Overall, optimal estimation in the general case is as hard as the special cases of $d$-dimensional symmetric 2-GM [82] and 1-dimensional $k$-GM [38, 81]. From (1.9), we see that there is a threshold $d^* = n^{(2k-3)/(2k-1)}$ (e.g., $d^* = n^{1/3}$ for $k = 2$). For $d > d^*$, the rate is governed by the subspace estimation error; otherwise, the rate is dominated by the error of estimating the low-dimensional mixing distribution. Note that Theorem 1.1 pertains to the optimal rate in the worst case. A faster rate is expected when the components are better separated, such as a parametric rate when the centers are separated by a constant, which, in one dimension, can be adaptively achieved by the estimators in [38, 81]. However, adaptation to the separation between components in $d$ dimensions remains an open problem; see the discussion in Section 6.

We note that the idea of using linear projections to reduce a multivariate Gaussian mixture to a univariate one has been previously explored in the context of parameter and density estimation (e.g., [60, 35, 2, 52, 81]); nevertheless, none of these results achieves the precision needed for attaining the optimal rate in Theorem 1.1. In particular, to avoid the unnecessary logarithmic factors, we use the denoised method of moments (DMM) algorithm introduced in [81] to simultaneously estimate many one-dimensional projections, which is amenable to sharp analysis via chaining techniques.

Next we discuss the optimal rate of density estimation for high-dimensional Gaussian mixtures, measured in the Hellinger distance. For distributions $P$ and $Q$, let $p$ and $q$ denote their respective densities with respect to some dominating measure $\mu$. The squared Hellinger distance between $P$ and $Q$ is $H^2(P,Q) \triangleq \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mu(dx)$. In this work, we focus on *proper learning*, in which the estimated density is required to be a $k$-GM. While there is no difference in the minimax rates for proper and improper density estimators, computationally the former is more challenging as it is not straightforward to find the best $k$-GM approximation to an improper estimate.

THEOREM 1.2 (Density estimation). *Let $P_\Gamma$ be as in (1.2). Then the minimax risk of estimating $P_\Gamma$ over the class $\mathcal{P}_{k,d}$ satisfies:*

$$\inf_{\hat{P}} \sup_{\Gamma \in \mathcal{G}_{k,d}} \mathbb{E}_\Gamma H(\hat{P}, P_\Gamma) \asymp_k \sqrt{\frac{d}{n}} \wedge 1. \tag{1.11}$$

*Furthermore, there exists a proper density estimate $P_{\hat{\Gamma}}$ and a positive constant $C$ not depending on $P_\Gamma$ such that for any $\Gamma \in \mathcal{G}_{k,d}$ and any $0 < \delta < \frac{1}{2}$, with $P_\Gamma$-probability at least $1 - \delta$,*

$$H(P_{\hat{\Gamma}}, P_\Gamma) \le C\sqrt{\frac{k^4 d + (2k)^{2k+2}}{n} \log \frac{1}{\delta}}. \tag{1.12}$$

Theorem 1.2, which follows a long line of research, is the first result we know of that establishes the sharp rate without logarithmic factors. The parametric rate $O_k(\sqrt{d/n})$ can be anticipated by noting that the model (1.1) is a smooth parametric family with $k(d+1)-1$

parameters. Justifying this heuristic, however, is not trivial, especially in high dimensions. To this end, we apply the Le Cam-Birgé construction of estimators from pairwise tests, which, as opposed to the analysis of the maximum likelihood estimator (MLE) based on bracketing entropy [77, 58, 32, 39], relies on bounding the local Hellinger entropy without brackets. By doing so, we also avoid the logarithmic slack in the existing result for the MLE; see Section 6 for more discussion.

The celebrated result of Le Cam-Birgé [51, 12, 13] shows that if the local covering number (the minimum number of Hellinger-balls of radius $\delta$ that cover any Hellinger-ball of radius $\epsilon$) is at most $(\frac{\epsilon}{\delta})^{O(D)}$, then there exists a density estimate that achieves a squared Hellinger risk $O\left(\frac{D}{n}\right)$. Here the crucial parameter $D$ is known as the *doubling dimension* (or the Le Cam dimension [76]), which serves as the effective number of parameters. In order to apply the theory of Le Cam-Birgé, we need to show that the doubling dimension of Gaussian mixtures is at most $O_k(d)$.

Bounding the local entropy requires a sharp characterization of the information geometry of Gaussian mixtures, for which the *moment tensors* play a crucial role. To explain this, we begin with an abstract setting: Consider a parametric model $\{P_\theta : \theta \in \Theta\}$, where the parameter space $\Theta$ is a subset of the $D$-dimensional Euclidean space. We say a parameterization is *good* if the Hellinger distance satisfies the following *dimension-free* bound:

$$C_0\|\theta - \theta'\| \le H(P_\theta, P_{\theta'}) \le C_1\|\theta - \theta'\|, \tag{1.13}$$

for some norm $\|\cdot\|$ and constants $C_0, C_1$. The two-sided bound (1.13) leads to the desired result on the local entropy in the following way. First, given any $P_\theta$ in an $\epsilon$-Hellinger neighborhood of the true density $P_{\theta_*}$, the lower bound in (1.13) localizes the parameter $\theta$ in an $O(\epsilon)$-neighborhood (in $\|\cdot\|$-norm) of the true parameter $\theta_*$, which, thanks to the finite dimensionality, can be covered by at most $(\frac{\epsilon}{\delta})^{O(D)}$ $\delta$-balls. Then the upper bound in (1.13) shows that this covering constitutes an $O(\delta)$-covering for the Hellinger ball.

While satisfied by many parametric families, notably the Gaussian location model, (1.13) fails for their mixtures if we adopt the natural parametrization (in terms of the centers and weights), as shown by the simple counterexample of the symmetric 2-GM where $P_\theta = \frac{1}{2}N(-\theta, 1) + \frac{1}{2}N(\theta, 1)$, with $|\theta| \le 1$. Indeed, it is easy to show that [82]:

$$|\theta - \theta'|^2 \lesssim H(P_\theta, P_{\theta'}) \lesssim |\theta - \theta'|,$$

which is tight since the lower and upper bound are achieved when $\theta' \to \theta$ and for $\theta = 0$ and say $\theta = 0.1$, respectively. The behavior of the lower bound can be attributed to the zero Fisher information at $\theta = 0$. The importance of a two-sided comparison result like (1.13) and the difficulty in Gaussian mixtures were recognized by [27, 26] in their study of the local entropy of mixture models. See Section 4 for detailed discussion.

It turns out that for Gaussian mixture model (1.2), a good parametrization satisfying (1.13) is provided by the moment tensors. The degree-$\ell$ moment tensor of the mixing distribution $\Gamma$ is the symmetric tensor

$$M_\ell(\Gamma) \triangleq \mathbb{E}_{U\sim\Gamma}[U^{\otimes \ell}] = \sum_{j=1}^{k} w_j \mu_j^{\otimes \ell}. \tag{1.14}$$

It can be shown that any $k$-atomic distribution is uniquely determined by its first $2k-1$ moment tensors $\mathbf{M}_{2k-1}(\Gamma) = [M_1(\Gamma), \ldots, M_{2k-1}(\Gamma)]$. Consequently, moment tensors provides a valid parametrization of the $k$-GM in the sense that $\mathbf{M}_{2k-1}(\Gamma) = \mathbf{M}_{2k-1}(\Gamma')$ if and only if $P_\Gamma = P_{\Gamma'}$. At the heart of our proof of Theorem 1.2 is the following robust version of this identifiability result:

$$H^2(P_\Gamma, P_{\Gamma'}) \asymp_k \left\|\mathbf{M}_{2k-1}(\Gamma) - \mathbf{M}_{2k-1}(\Gamma')\right\|_{\mathrm{F}}^2 \tag{1.15}$$

which shows that the Hellinger distance between $k$-GMs are characterized by the Euclidean distance of their moment tensors up to dimension-free constant factors. Furthermore, the same result also holds for the Kullback-Leibler (KL) and the $\chi^2$ divergences. See Section 4.1 for details.

Note that moment tensors appear to be a gross overparameterization of $\mathcal{G}_{k,d}$ since the original number of parameters is only $kd + k - 1$ as compared to the size $d^{\Theta(k)}$ of moment tensors. The key observation is that the moment tensors (1.14) for $k$-atomic distributions are naturally low rank, so that the effective dimension remains $\Theta(kd)$. This observation underlies tensor decomposition methods for learning mixture models [4, 42]; here we use it for the information-theoretic purpose of bounding the local metric entropy of Gaussian mixtures.

Results similar to (1.15) were previously shown in [9] for the problem of multiple-reference alignment, a special case of Gaussian mixtures with mixing distribution being uniform over the cyclic shifts of a given vector. The crucial difference is that the characterization (1.15) involves moments tensors of degree at most $2k - 1$, while [9, Theorem 9] involves all moments.

The Le Cam-Birgé construction used to show Theorem 1.2 does not result in a computationally efficient estimator. In Section 4.3, we provide a variant of the algorithm in Section 3 that runs in $n^{O(k)}$ time which achieves the suboptimal rate of $O_k((d/n)^{1/4})$ for general $k$-GM (Theorem 4.7). Though not optimal, this result nonetheless improves the state of the art of [2] by logarithmic factors. Furthermore, in the special case of 2-GM, a slightly modified estimator is shown to achieve the optimal rate of $O(\sqrt{d/n})$ (Theorem 4.8). Finding a polynomial-time algorithm achieving the optimal rate in Theorem 1.2 for all $k$ is an open problem.

1.2. *Related work.* There is a vast literature on Gaussian mixtures; see [54, 38, 81] and the references therein for an overview. In one dimension, fast algorithms and optimal rates of convergence have already been achieved for both parameter and density estimation by, e.g., [81]. We therefore focus the following discussion on multivariate Gaussian mixtures, in both low and high dimensions.

*Parameter estimation.* For statistical rates, [39, Theorem 1.1] and [40, Theorem 4.3] obtained convergence rates for mixing distribution estimation in Wasserstein distances for low-dimensional location-scale Gaussian mixtures, both over- and exact-fitted. Their rates for over-fitted mixtures are determined by algebraic dependencies among a set of polynomial equations whose order depends on the level of overfitting and identifiability of the model; the rates are potentially much slower than $n^{-1/2}$. The estimator analyzed in [39, 40] is the MLE, which involves non-convex optimization and is typically approximated by the Expectation-Maximization (EM) algorithm.

In the computer science literature, a long line of research starting with [18] has developed fast algorithms for individual parameter estimation in multivariate Gaussian mixtures under fairly weak separation conditions, see, e.g., [78, 6, 11, 45, 60, 42, 35, 41]. Since these works focus on individual parameter estimation, some separation assumption on the mixing distribution is necessary.

*Density estimation.* Computational issues aside, there are several recent works addressing the minimax rate of density estimation for Gaussian mixtures. In low dimensions, an $O(\sqrt{\log n/n})$-Hellinger guarantee for the MLE is obtained for finite Gaussian mixtures [39, 40]. The near-optimal rate for high-dimensional location-scale mixtures was obtained recently in [7]. This work also provides a total variation guarantee of $\tilde{O}(\sqrt{kd/n})$ for location mixtures, where $\tilde{O}$ hides polylogarithmic factors, as compared to the sharp result in Theorem 1.2. The algorithm in [7] runs in time that is exponential in $d$.

To our knowledge, there is no polynomial-time algorithm that achieves the sharp density estimation guarantee in Theorem 1.2 (or the slightly suboptimal rate in [7]), even for constant $k$. The works of [45, 60] showed that their polynomial-time parameter estimation algorithms also provide density estimators without separation conditions, but the resulting rates of convergence are far from optimal. [23, 2, 52] provided polynomial-time algorithms for density estimation with improved statistical performance. In particular, [2] obtained an algorithm that runs in time $\tilde{O}_k(n^2d+d^2(n/d)^{3k^2/4})$ and achieves a total variation error of $\tilde{O}((d/n)^{1/4})$. The running time was further improved in [52], which achieves the rate $\tilde{O}((d/n)^{1/6})$ for 2-GM.

*Nonparametric mixtures.* The above-mentioned works all focus on finite mixtures, which is also the scenario considered in this paper. A related strain of research (e.g., [28, 32, 83, 70]) studies the so-called *nonparametric mixture model*, in which the mixing distribution $\Gamma$ may be an arbitrary probability measure.

In this case, the nonparametric maximum likelihood estimator (known as the NPMLE) entails solving a convex (but infinite-dimensional) optimization problem, which, in principle, can be solved by discretization [48]. For statistical rates, it is known that in one dimension, the optimal $L_2$-rate for density estimation is $\Theta((\log n)^{1/4}/\sqrt{n})$ and the Hellinger rate is at least $\Omega(\sqrt{\log n/n})$ [43, 47], which shows that the parametric rate (1.11) is only achievable for finite mixture models. For the NPMLE, [83] proved the Hellinger rate of $O(\log n/\sqrt{n})$ in one dimension; this was extended to the multivariate case by [70]. In particular, [70, Theorem 2.3] obtained a Hellinger rate of $C_d\sqrt{k(\log n)^{d+1}/n}$ for the NPMLE when the true model is a $k$-GM. In high dimensions, this is highly suboptimal compared to the parametric rate in (1.11), although the dependency on $k$ is optimal.

*Bayesian methods.* There is a rich literature on the asymptotic behavior of the posterior distribution of mixture models in Bayesian settings. Analysis of the posterior in Wasserstein distance, for both finite and infinite Dirichlet-process mixture models, was investigated [62], which provides posterior contraction rates under various conditions. Asymptotic posterior contraction rates, both under Dirichlet process priors and in more general Bayesian nonparametric settings, were considered for instance in [29, 30, 72, 31, 71]. [32], mentioned previously for its maximum likelihood results, also considered rates of contraction for infinite Gaussian mixtures in a Bayesian nonparametric setting.

For overfitted mixtures in multiple dimensions, where an upper bound on the number of components is known, [68, Theorem 1] showed that with a Dirichlet prior with sufficiently small hyperparameters, and under certain regularity conditions, the redundant weights vanish at a near $n^{-1/2}$ rate under the posterior. Classical posterior contraction results under Dirichlet process priors do not in fact show that the posteriors converge to distributions with the correct number of components, and a standard practice for inferring the number of components was demonstrated to be inconsistent in [59]. This, however, can be corrected by a post-processing procedure in [33], which moreover provides an alternative prior that yields both posterior contraction at the correct rate and convergence to the correct number of components.

1.3. *Organization.* The rest of the paper is organized as follows. Section 3 presents a polynomial-time algorithm for estimating the mixing distribution and provides the theoretical justification for Theorem 1.1. Section 4 introduces the necessary background on moment tensors and proves the optimal rate of density estimation in Theorem 1.2. Section 5 provides simulations that support the theoretical results. Section 6 provides further discussion on the connections between this work and the Gaussian mixture literature.

**2. Notation.** Let $[n] \triangleq \{1,\dots,n\}$. Let $S^{d-1}$ and $\Delta^{d-1}$ denote the unit sphere and the probability simplex in $\mathbb{R}^d$, respectively. Let $e_j$ be the vector with a 1 in the $j$th coordinate

and zeros elsewhere. For a matrix $A$, let $\|A\|_2 = \sup_{x:\|x\|_2=1} \|Ax\|_2$ and $\|A\|_F^2 = \mathrm{tr}(A^\top A)$. For two positive sequences $\{a_n\}, \{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there exists a constant $C$ such that $a_n \le C b_n$ and we write $a_n \lesssim_k b_n$ and $a_n = O_k(b_n)$ to emphasize that $C$ may depend on a parameter $k$.

For $\epsilon > 0$, an $\epsilon$-covering of a set $A$ with respect to a metric $\rho$ is a set $\mathcal{N}$ such that for all $a \in A$, there exists $b \in \mathcal{N}$ such that $\rho(a,b) \le \epsilon$; denote by $N(\epsilon, A, \rho)$ the minimum cardinality of $\epsilon$-covering sets of $A$. An $\epsilon$-packing in $A$ with respect to the metric $\rho$ is a set $\mathcal{M} \subset A$ such that $\rho(a,b) > \epsilon$ for any distinct $a, b$ in $\mathcal{M}$; denote by $M(\epsilon, A, \rho)$ the largest cardinality of $\epsilon$-packing sets in $A$.

For distributions $P$ and $Q$, let $p$ and $q$ denote their relative densities with respect to some dominating measure $\mu$, respectively. The total variation distance is defined as $\mathsf{TV}(P,Q) \triangleq \frac{1}{2}\int |p(x) - q(x)| \mu(dx)$. If $P \ll Q$, the KL divergence and the $\chi^2$-divergence are defined as $\mathsf{KL}(P||Q) \triangleq \int p(x) \log \frac{p(x)}{q(x)} \mu(dx)$ and $\chi^2(P\|Q) \triangleq \int \frac{(p(x)-q(x))^2}{q(x)} \mu(dx)$, respectively. Let $\mathrm{supp}(P)$ denote the support set of a distribution $P$. Let $\mathcal{L}(U)$ denote the distribution of a random variable $U$. For a one-dimensional distribution $\nu$, denote the $r$th moment of $\nu$ by $m_r(\nu) \triangleq \mathbb{E}_{U\sim\nu}[U^r]$ and its moment vector $\mathbf{m}_r(\nu) \triangleq (m_1(\nu), \ldots, m_r(\nu))$. Given a $d$-dimensional distribution $\Gamma$, for each $\theta \in \mathbb{R}^d$, we denote

$$\Gamma_\theta \triangleq \mathcal{L}(\theta^\top U), \quad U \sim \Gamma; \tag{2.1}$$

in other words, $\Gamma_\theta$ is the pushforward of $\Gamma$ by the projection $u \mapsto \theta^\top u$; in particular, the $i$th marginal of $\Gamma$ is denoted by $\Gamma_i \triangleq \Gamma_{e_i}$, with $e_i$ being the $i$th coordinate vector. Similarly, for $V \in \mathbb{R}^{d\times k}$, denote

$$\Gamma_V \triangleq \mathcal{L}(V^\top U), \quad U \sim \Gamma. \tag{2.2}$$

**3. Mixing distribution estimation.** In this section we present the algorithm that achieves the optimal rate for estimating the mixing distribution in Theorem 1.1. The procedure is described in Sections 3.1 and 3.2. The proof of correctness is given in Sections 3.3, with supporting lemmas proved in the supplemental material [22]. Throughout this section we assume that $n \ge d$.

3.1. *Dimension reduction via PCA.* In this section we assume $d > k$ and reduces the dimension from $d$ to $k$. For $d \le k$, we will directly apply the procedure in Section 3.2. Recall that the atoms $\Gamma$ are $\mu_1, \ldots, \mu_k$; they span a subspace of $\mathbb{R}^d$ of dimension at most $k$. Therefore, there exists $V = [v_1, \ldots, v_k]$ consisting of orthonormal columns, such that for each $j = 1, \ldots, k$, we have $\mu_j = V\psi_j$, where $\psi_j = V^\top \mu_j \in \mathbb{R}^k$ encodes the coefficients of $\mu_j$ in the basis vectors in $V$. Therefore, we can identify a $k$-atomic distribution $\Gamma$ on $\mathbb{R}^d$ with a pair $(V, \gamma)$, where $\gamma = \sum_{j\in[k]} w_j \delta_{\psi_j}$ is a $k$-atomic distribution on $\mathbb{R}^k$. This perspective motivates the following two-step procedure. First, we estimate the subspace $V$ using PCA, relying on the fact that the covariance matrix satisfies $\mathbb{E}[XX^\top] = I_d + \sum_{j=1}^k w_j \mu_j \mu_j^\top$. We then project the data onto the estimated subspace, reducing the dimension from $d$ to $k$, and apply an estimator of $k$-GM in $k$ dimensions. The precise execution of this idea is described below.

For simplicity, consider a sample of $2n$ observations $X_1, \ldots, X_{2n} \overset{\text{i.i.d.}}{\sim} P_\Gamma$. We construct an estimator $\hat{\Gamma}$ of $\Gamma$ in the following way:

(a) Estimate the subspace $V$ using the first half of the sample. Let $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_k] \in \mathbb{R}^{d\times k}$ be the matrix whose columns are the top $k$ orthonormal left singular vector of $[X_1, \ldots, X_n]$.

(b) Project the second half of the sample onto the learned subspace $\hat{V}$:

$$x_i \triangleq \hat{V}^\top X_{i+n}, \quad i = 1, \dots, n. \tag{3.1}$$

Thanks to independence, conditioned on $\hat{V}$, $x_1, \dots, x_n$ are i.i.d. observations from a $k$-GM in $k$ dimensions, with mixing distribution

$$\gamma \triangleq \Gamma_{\hat{V}} = \sum_{j=1}^{k} w_j \delta_{\hat{V}^\top \mu_j} \tag{3.2}$$

obtained by projecting the original $d$-dimensional mixing distribution $\Gamma$ onto $\hat{V}$.

(c) To estimate $\hat{\gamma}$, we apply a multivariate version of the denoised method of moments to $x_1, \dots, x_n$ to obtain a $k$-atomic distribution on $\mathbb{R}^k$:

$$\hat{\gamma} = \sum_{j=1}^{k} \hat{w}_j \delta_{\hat{\psi}_j}. \tag{3.3}$$

This procedure is explained next and detailed in Algorithm 1.

(d) Lastly, we report

$$\hat{\Gamma} = \hat{\gamma}_{\hat{V}^\top} = \sum_{j=1}^{k} \hat{w}_j \delta_{\hat{V}\hat{\psi}_j} \tag{3.4}$$

as the final estimate of $\Gamma$.

A slightly better dimension reduction can be achieved by first centering the data by subtracting the sample mean, then projecting to a subspace of dimension $k-1$ rather than $k$, and finally adding back the sample mean after obtaining the final estimator. As this only affect constant factors, we forgo the centering step in this section. Later in Section 4.3, it turns out that centering is important for achieving the optimal density estimation for 2-GM (see Theorem 4.8).

The usefulness of dimension reduction has long been recognized in the literature of mixture models [18, 78, 46, 3, 36, 56], where the mixture data is projected to a good low-dimensional subspace (by either random projection or spectral methods) and parameter estimation or clustering are carried out subsequently in the low-dimensional mixture model. For such methods, the error bound typically depends on those of these two steps, analogously to the analysis of mixing distribution estimation in Theorem 1.1.

3.2. *Estimating the mixing distribution in low dimensions.* We now explain how we estimate a $k$-GM in $k$ dimensions from i.i.d. observations. As mentioned in Section 1, the idea is to use many projections to reduce the problem to one dimension. We first present a conceptually simple estimator $\hat{\gamma}^\circ$ with an optimal statistical performance but unfavorable run time $n^{O(k)}$. We then describe an improved estimator $\hat{\gamma}$ that retains the statistical optimality and can be executed in time $n^{O(1)}$. These procedures are also applicable to estimating a $k$-GM in $d < k$ dimensions using fewer projections.

To make precise the reduction to one dimension, a relevant metric is the *sliced Wasserstein distance* [64, 20, 63], which measures the distance of two $d$-dimensional distributions by the maximal $W_1$-distance of their one-dimensional projections:

$$W_1^{\mathsf{sliced}}(\Gamma, \Gamma') \triangleq \sup_{\theta \in S^{d-1}} W_1(\Gamma_\theta, \Gamma'_\theta). \tag{3.5}$$

Here we recall that $\Gamma_\theta$ defined in (2.1) denotes the projection, or pushforward, of $\Gamma$ onto the direction $\theta$. A related definition was introduced earlier by [67], where the supremum

over $\theta$ in (3.5) is replaced by the average. Computing the sliced Wasserstein distance can be difficult and in practice is handled by gradient descent heuristics [20]; we will, however, only rely on its theoretical properties. The following result, which is proved in Section 2 of the supplement [22], shows that for low-dimensional distributions with few atoms, the full Wasserstein distance and the sliced one are comparable up to constant factors. Related results are obtained in [64, 10]. For instance, [10, Theorem 2.1(ii)] showed that $W_1(\Gamma,\Gamma') \le C_d \cdot W_1^{\mathsf{sliced}}(\Gamma,\Gamma')$ holds for all distributions $\Gamma,\Gamma'$ for some non-explicit constant $C_d$.

LEMMA 3.1 (Sliced Wasserstein distance). *For any $k$-atomic distributions $\Gamma,\Gamma'$ on $\mathbb{R}^d$,*

$$W_1^{\mathsf{sliced}}(\Gamma,\Gamma') \le W_1(\Gamma,\Gamma') \le k^2\sqrt{d}\cdot W_1^{\mathsf{sliced}}(\Gamma,\Gamma').$$

Having obtained via PCA the reduced sample $x_1,\dots,x_n \sim \gamma * N(0,I_k)$ in (3.1), Lemma 3.1 suggests the following "meta-procedure": Suppose we have an algorithm (call it a 1-D algorithm) that estimates the mixing distribution based on $n$ i.i.d. observations drawn from a $k$-GM in one dimension. Then

1. For each $\theta \in S^{k-1}$, since $\langle \theta, x_i\rangle \overset{\text{i.i.d.}}{\sim} \gamma_\theta * N(0,1)$, we can apply the 1-D algorithm to obtain an estimate $\widehat{\gamma}_\theta \in \mathcal{G}_{k,1}$;
2. We obtain an estimate of the multivariate distribution by minimizing a proxy of the sliced Wasserstein distance:

$$\hat{\gamma}^\circ = \operatorname*{argmin}_{\gamma' \in \mathcal{G}_{k,k}} \sup_{\theta\in S^{k-1}} W_1(\gamma'_\theta, \widehat{\gamma}_\theta). \tag{3.6}$$

Then by Lemma 3.1 (with $d=k$) and the optimality of $\hat{\gamma}^\circ$, we have

$$\begin{aligned} W_1(\hat{\gamma}^\circ,\gamma) \lesssim_k W_1^{\mathsf{sliced}}(\hat{\gamma}^\circ,\gamma) &= \sup_{\theta\in S^{k-1}} W_1(\hat{\gamma}^\circ_\theta,\gamma_\theta) \\ &\le \sup_{\theta\in S^{k-1}} W_1(\widehat{\gamma}_\theta,\gamma_\theta) + \sup_{\theta\in S^{k-1}} W_1(\widehat{\gamma}_\theta,\hat{\gamma}^\circ_\theta) \\ &\le 2\sup_{\theta\in S^{k-1}} W_1(\widehat{\gamma}_\theta,\gamma_\theta). \end{aligned} \tag{3.7}$$

Recall that the optimal $W_1$-rate for $k$-atomic one-dimensional mixing distribution is $O(n^{-\frac{1}{4k-2}})$. Suppose there is a 1-D algorithm that achieves the optimal rate *simultaneously* for all projections, in the sense that

$$\mathbb{E}\left[\sup_{\theta\in S^{k-1}} W_1(\widehat{\gamma}_\theta,\gamma_\theta)\right] \lesssim_k n^{-\frac{1}{4k-2}}. \tag{3.8}$$

This immediately implies the desired

$$\mathbb{E}[W_1(\hat{\gamma}^\circ,\gamma)] \lesssim_k n^{-\frac{1}{4k-2}}. \tag{3.9}$$

However, it is unclear how to solve the min-max problem in (3.6) where the feasible sets for $\gamma$ and $\theta$ are both non-convex. The remaining tasks are two-fold: (a) provide a 1-D algorithm that achieves (3.8); (b) replace $\hat{\gamma}^\circ$ by a computationally feasible version.

*Achieving (3.8) by denoised method of moments.* In principle, any estimator for a one-dimensional mixing distribution with exponential concentration can be used as a black box; this achieves (3.8) up to logarithmic factors by a standard covering and union bound argument. In order to attain the sharp rate in (3.8), we consider the Denoised Method of Moments (DMM) algorithm introduced in [81], which allows us to use the chaining technique to obtain a tight control of the fluctuation over the sphere (see Lemma 3.2).

DMM is an optimization-based approach that introduces a denoising step before solving the method of moments equations. For location mixtures, it provides an exact solver to the non-convex optimization problem arising in generalized method of moments [34]. For Gaussian location mixtures with unit variance, the DMM algorithm proceeds as follows:

(a) Given $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \nu * N(0,1)$ for some $k$-atomic distribution $\nu$ supported on $[-R, R]$, we first estimate the moment vector $\mathbf{m}_{2k-1}(\nu) = (m_1(\nu), \dots, m_{2k-1}(\nu))$ by their unique unbiased estimator $\tilde{m} = (\tilde{m}_1, \dots, \tilde{m}_{2k-1})$, where $\tilde{m}_r = \frac{1}{n}\sum_{i=1}^n H_r(Y_i)$, and $H_r$ is the degree-$r$ Hermite polynomial defined via

$$H_r(x) \triangleq r! \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^i}{i!(r-2i)!} x^{r-2i}. \tag{3.10}$$

Then $\mathbb{E}[\tilde{m}_r] = m_r(\nu)$ for all $r$. This step is common to all approaches based on the method of moments.

(b) In general the unbiased estimate $\tilde{m}$ is not a valid moment vector, in which case the method-of-moment-equation lacks a meaningful solution. The key idea of the DMM method is to denoise $\tilde{m}$ by its projection onto the space of moments:

$$\hat{m} \triangleq \operatorname{argmin}\{\|\tilde{m} - m\| : m \in \mathcal{M}_r\}, \tag{3.11}$$

where the moment space

$$\mathcal{M}_r \triangleq \{\mathbf{m}_r(\pi) : \pi \text{ supported on } [-R, R]\} \tag{3.12}$$

consists of the first $r$ moments of all probability measures on $[-R, R]$. The moment space is a convex set and characterized by positive semidefinite constraints (of the associated Hankel matrix); we refer the reader to the monograph [73] or [81, Sec. 2.1] for details. This means that the optimization problem (3.11) can be solved efficiently as a semidefinite program (SDP); see [81, Algorithm 1].

(c) Use Gauss quadrature to find the unique $k$-atomic distribution $\hat{\nu}$ such that $\mathbf{m}_{2k-1}(\hat{\nu}) = \hat{m}$. We denote the final output $\hat{\nu}$ by $\mathsf{DMM}(Y_1, \dots, Y_n)$.

The following result shows the DMM estimator achieves the optimal rate in (3.8) simultaneously for all one-dimensional projections. (For a single $\theta$, this is shown in [81, Theorem 1].)

LEMMA 3.2. *For each $\theta \in S^{k-1}$, let $\widehat{\gamma}_\theta = \mathsf{DMM}(\langle\theta, x_1\rangle, \dots, \langle\theta, x_n\rangle)$ where $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \gamma * N(0, I_k)$ as in (3.1). There is a positive constant $C$ such that, for any $\delta \in (0, \frac{1}{2})$, with probability at least $1 - \delta$,*

$$\max_{\theta \in S^{k-1}} W_1(\widehat{\gamma}_\theta, \gamma_\theta) \le C k^{7/2} n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}.$$

*Solving (3.6) efficiently using marginal estimates.* We first note that in order to achieve the optimal rate in (3.9), it is sufficient to consider any approximate minimizer of (3.6) up to an additive error of $\epsilon$, as long as $\epsilon = O(n^{-\frac{1}{4k-2}})$. Therefore, to find an $\epsilon$-optimizer, it suffices to maximize over $\theta$ in an $\epsilon$-net (in $\ell_2$) of the sphere, which has cardinality $(\frac{1}{\epsilon})^k = n^{O(1)}$, and, likewise, minimize $\gamma$ over an $\epsilon$-net (in $W_1$) of $\mathcal{G}_{k,k}$. The $W_1$-net can be constructed by combining an $\epsilon$-net (in $\ell_2$) for each of the $k$ centers and an $\epsilon$-net (in $\ell_1$) for the weights, resulting in a set of cardinality $(\frac{1}{\epsilon})^{O(k^2)} = n^{O(k)}$. This naïve discretization scheme leads to an estimator of $\gamma$ with optimal rate but time complexity $n^{O(k)}$. We next improve it to $n^{O(1)}$.

The key idea is to first estimate the marginals of $\gamma$, which narrows down its support set. It is clear that a $k$-atomic joint distribution is not determined by its marginal distributions, as shown by the example of $\frac{1}{2}\delta_{(-1,-1)} + \frac{1}{2}\delta_{(1,1)}$ and $\frac{1}{2}\delta_{(-1,1)} + \frac{1}{2}\delta_{(1,-1)}$, which have identical marginal distributions. Nevertheless, the support of the joint distribution must be a $k$-subset of the Cartesian product of the marginal support sets. This suggests that we can select the atoms from this Cartesian product and weights by fitting all one-dimensional projections, as in (3.6).

Specifically, for each $j \in [k]$, we estimate the $j$th marginal distribution of $\gamma$ by $\widehat{\gamma}_j$, obtained by applying the DMM algorithm on the coordinate projections $\langle e_j, x_1\rangle, \ldots, \langle e_j, x_n\rangle$. Consider the Cartesian product of the support of each estimated marginal as the candidate set of atoms:

$$\mathcal{A} \triangleq \operatorname{supp}(\widehat{\gamma}_1) \times \cdots \times \operatorname{supp}(\widehat{\gamma}_k).$$

Throughout this section, let

$$\epsilon_{n,k} \triangleq n^{-\frac{1}{4k-2}}, \tag{3.13}$$

and fix an $(\epsilon_{n,k}, \|\cdot\|_2)$-covering $\mathcal{N}$ for the unit sphere $S^{k-1}$ and an $(\epsilon_{n,k}, \|\cdot\|_1)$-covering $\mathcal{W}$ for the probability simplex $\Delta^{k-1}$, such that[2]

$$\max\{|\mathcal{N}|, |\mathcal{W}|\} \lesssim \left(\frac{C}{\epsilon_{n,k}}\right)^{k-1}. \tag{3.14}$$

Define the following set of candidate $k$-atomic distributions on $\mathbb{R}^k$:

$$\mathcal{S} \triangleq \left\{ \sum_{j\in[k]} w_j \delta_{\psi_j} : (w_1, \ldots, w_k) \in \mathcal{W}, \psi_j \in \mathcal{A} \right\}. \tag{3.15}$$

Note that $\mathcal{S}$ is a random set which depends on the sample; furthermore, each $\psi_j \in \mathcal{A}$ has coordinates lying in $[-R, R]$ by virtue of the DMM algorithm.

The next lemma shows that with high probability there exists a good approximation of $\gamma$ in the set $\mathcal{S}$.

LEMMA 3.3. *Let $\mathcal{S}$ be given in (3.15). There is a positive constant $C$ such that, for any $\delta \in (0, \frac{1}{2})$, with probability $1-\delta$,*

$$\min_{\gamma' \in \mathcal{S}} W_1\left(\gamma', \gamma\right) \le C k^5 n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}. \tag{3.16}$$

We conclude this subsection with Algorithm 1, which provides a full description of an estimator for $k$-atomic mixing distributions in $k$ dimensions. The following result shows its optimality under the $W_1$ loss:

LEMMA 3.4. *There is a positive constant $C$ such that the following holds. Let $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \gamma * N(0, I_k)$ for some $\gamma \in G_{k,k}$. Then Algorithm 1 produces an estimator $\hat{\gamma} \in \mathcal{G}_{k,k}$ such that, for any $\delta \in (0, \frac{1}{2})$, with probability $1-\delta$,*

$$W_1(\gamma, \hat{\gamma}) \le C k^5 n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}. \tag{3.17}$$

[2]This is possible by, e.g., [69, Prop. 2.1] and [32, Lemma A.4] for the sphere and probability simplex, respectively.

**Algorithm 1:** Parameter estimation for $k$-GM in $k$ dimensions

**Input:** Dataset $\{x_i\}_{i\in[n]}$ with each point in $\mathbb{R}^k$, order $k$, radius $R$.
**Output:** Estimate $\hat{\gamma}$ of $k$-atomic distribution in $k$ dimensions.
**For** $j = 1, \dots, k$**:**
Compute the marginal estimate $\widehat{\gamma_j} = \mathsf{DMM}(\{e_j^\top x_i\}_{i\in[n]})$ ;
Form the set $\mathcal{S}$ of $k$-atomic candidate distributions on $\mathbb{R}^k$ as in (3.15) ;
**For each** $\theta \in \mathcal{N}$**:**
Estimate the projection by $\widehat{\gamma_\theta} = \mathsf{DMM}(\{\theta^\top x_i\}_{i\in[n]})$ ;
**For each candidate distribution** $\gamma' \in \mathcal{S}$ **and each direction** $\theta \in \mathcal{N}$**:**
Compute $W_1(\gamma'_\theta, \widehat{\gamma_\theta})$ ;
Report

$$\hat{\gamma} = \arg\min_{\gamma'\in\mathcal{S}} \max_{\theta\in\mathcal{N}} W_1(\gamma'_\theta, \widehat{\gamma_\theta}). \tag{3.18}$$

REMARK 1. The total time complexity to compute the estimator (3.4) is $O(nd^2) + O_k(n^{5/4})$. Indeed, the time complexity of computing the sample covariance matrix is $O(nd^2)$, and the time complexity of performing the eigendecomposition is $O(d^3)$, which is dominated by $O(nd^2)$ since $d \le n$. By (3.14), both $\mathcal{W}$ and $\mathcal{N}$ have cardinality at most $(C/\epsilon_{n,k})^{k-1} = O_k(n^{1/4})$. Each one-dimensional DMM estimate takes $O_k(n)$ time to compute [81, Theorem 1]. Thus computing the one-dimensional estimator $\widehat{\gamma_\theta}$ for all $\theta = e_i$ and $\theta \in \mathcal{N}$ takes time $O_k(n^{5/4})$. Since both $\gamma'_\theta$ and $\widehat{\gamma_\theta}$ are $k$-atomic distributions by definition, their $W_1$ distance can be computed in $O_k(1)$ time. Finally, $|\mathcal{A}| = k^k$, and to form $S$ we select all sets of $k$ atoms from $\mathcal{A}$, so $|\mathcal{S}| \le |\mathcal{W}|\binom{k^k}{k} = O_k\left(n^{1/4}\right)$. Thus searching over $\mathcal{S}\times\mathcal{N}$ takes time at most $O_k(n^{1/4}) * O_k(n^{1/4}) = O_k\left(n^{1/2}\right)$. Therefore, the overall time complexity of Algorithm 1 is $O_k(n^{5/4})$.

3.3. *Proof of Theorem 1.1.* The proof is outlined as follows. Recall that the estimate $\hat{\Gamma}$ in (3.4) is supported on the subspace spanned by the columns of $\hat{V}$, whose projection is $\hat{\gamma}$ in (3.3). Similarly, the projection of the ground truth $\Gamma$ on the space $\hat{V}$ is denoted by $\gamma = \Gamma_{\hat{V}}$ in (3.2). Note that both $\gamma$ and $\hat{\gamma}$ are $k$-atomic distributions in $k$ dimensions. Let $\hat{H} = \hat{V}\hat{V}^\top$ be the projection matrix onto the space spanned by the columns of $\hat{V}$. By the triangle inequality,

$$\begin{aligned} W_1(\Gamma, \hat{\Gamma}) &\le W_1(\Gamma, \Gamma_{\hat{H}}) + W_1(\Gamma_{\hat{H}}, \hat{\Gamma}) \\ &\le W_1(\Gamma, \Gamma_{\hat{H}}) + W_1(\gamma, \hat{\gamma}). \end{aligned} \tag{3.19}$$

We will upper bound the first term by $(d/n)^{1/4}$ (using Lemmas 3.5 and 3.6 below) and the second term by $n^{-1/(4k-2)}$ (using the previous Lemma 3.4).

We first control the difference between $\Gamma$ and its projection onto the estimated subspace $\hat{V}$. Since we do not impose any lower bound on $\|\mu_j\|_2$, we cannot directly show the accuracy of $\hat{V}$ by means of perturbation bounds such as the Davis-Kahan theorem [19]. Instead, the following general lemma bounds the error by the difference of the covariance matrices. For a related result, see [78, Corollary 3].

LEMMA 3.5. *Let* $\Gamma = \sum_{j=1}^k w_j \delta_{\mu_j}$ *be a* $k$*-atomic distribution. Let* $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top] = \sum_{j=1}^k w_j \mu_j \mu_j^\top$ *with eigenvalues* $\lambda_1 \ge \cdots \ge \lambda_d$*. Let* $\Sigma'$ *be a symmetric matrix and* $\Pi'_r$ *be the projection matrix onto the subspace spanned by the top* $r$ *eigenvectors of* $\Sigma'$*. Then,*

$$W_2^2(\Gamma, \Gamma_{\Pi'_r}) \le k\left(\lambda_{r+1} + 2\|\Sigma - \Sigma'\|_2\right).$$

We will apply Lemma 3.5 with $\Sigma'$ being the sample covariance matrix $\hat{\Sigma}$. The following lemma provides the concentration of $\hat{\Sigma}$ we need to prove the upper bound on the high-dimensional component of the error in Theorem 1.1.

LEMMA 3.6. *Let $\Gamma \in \mathcal{G}_{k,d}$ and $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top]$. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - I_d$, where $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P_\Gamma$. Then there exists a positive constant $C$ such that, with probability at least $1-\delta$,*

$$\|\hat{\Sigma} - \Sigma\|_2 \le C\left(\sqrt{\frac{d}{n}} + k\sqrt{\frac{\log(k/\delta)}{n}} + \frac{\log(1/\delta)}{n}\right).$$

**Proof of Theorem 1.1.** We first show that the estimator (3.4) achieves the tail bound stated in (1.10), which, after integration, implies the average risk bound in (1.9). To bound the first term in (3.19), note that the rank of $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top]$ is at most $k$. Furthermore, the top $k$ left singular vectors of $[X_1, \dots, X_n]$ coincide with the top $k$ eigenvectors of $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - I_d$. Applying Lemmas 3.5 and 3.6 yields that, with probability $1-\delta$,

$$W_1(\Gamma, \Gamma_{\hat{H}}) \le \sqrt{2Ck}\left(\left(\frac{d}{n}\right)^{1/4} + \left(\frac{k^2\log(k/\delta)}{n}\right)^{1/4} + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{3.20}$$

where we used the fact that $W_1(\Gamma, \Gamma') \le W_2(\Gamma, \Gamma')$ by the Cauchy-Schwarz inequality. To upper bound the second term in (3.19), recall that $\hat{V}$ was obtained from $\{X_1, \dots, X_n\}$ and hence is independent of $\{X_{n+1}, \dots, X_{2n}\}$. Thus conditioned on $\hat{V}$,

$$x_i = \hat{V}^\top X_{i+n} \overset{\text{i.i.d.}}{\sim} \gamma * N(0, I_k), \quad i = 1, \dots, n.$$

Let $\hat{\gamma}$ be obtained from Algorithm 1 with input $x_1, \dots, x_n$. By Lemma 3.4, with probability $1-\delta$,

$$W_1(\gamma, \hat{\gamma}) \le Ck^5 n^{-1/(4k-2)}\sqrt{\log\frac{1}{\delta}}. \tag{3.21}$$

Note that $(k^2\log(k/\delta)/n)^{1/4} + (\log(1/\delta)/n)^{1/2}$ in (3.20) is dominated by the right-hand side of (3.21). The desired (1.10) follows from combining (3.19), (3.20), and (3.21).

Finally, we prove the lower bound in (1.9). For any subset $\mathcal{G} \subseteq \mathcal{G}_{k,d}$, we have

$$\inf_{\hat{\Gamma}} \sup_{\Gamma\in\mathcal{G}_{k,d}} \mathbb{E}W_1(\hat{\Gamma}, \Gamma) \ge \inf_{\hat{\Gamma}} \sup_{\Gamma\in\mathcal{G}} \mathbb{E}W_1(\hat{\Gamma}, \Gamma) \ge \frac{1}{2}\inf_{\hat{\Gamma}\in\mathcal{G}} \sup_{\Gamma\in\mathcal{G}} \mathbb{E}W_1(\hat{\Gamma}, \Gamma), \tag{3.22}$$

where the second inequality follows from replacing an arbitrary estimator $\hat{\Gamma}$ by its $W_1$-projection $\arg\min_{\tilde{\Gamma}\in\mathcal{G}} W_1(\tilde{\Gamma}, \hat{\Gamma})$ and applying the triangle inequality for $W_1$. To obtain the lower bound in (1.9), we apply the $\Omega((d/n)^{1/4} \wedge 1)$ lower bound in [82, Theorem 10] for $d$-dimensional symmetric 2-GM (by taking $\mathcal{G}$ to the mixtures of the form (1.8)) and the $\Omega(n^{-1/(4k-2)})$ lower bound in [81, Proposition 7] for 1-dimensional $k$-GM (by taking $\mathcal{G}$ to be the set of mixing distributions whose atoms are zero except for their coordinates.) □

**4. Density estimation.** In this section we prove the density estimation guarantee of Theorem 1.2 for finite Gaussian mixtures. The lower bound simply follows from the minimax quadratic risk of the Gaussian location model (corresponding to $k=1$), since $H^2(N(\theta, I_d), N(\theta', I_d)) = 2 - 2e^{\|\theta-\theta'\|_2^2/8} \asymp \|\theta-\theta'\|^2$ when $\theta, \theta' \in B(0, R)$. Thus, we focus on the attainability of the parametric rate of density estimation. Departing from the prevailing approach of maximum likelihood, we aim to apply the estimator of Le Cam and Birgé which requires bounding the local entropy of Hellinger balls for $k$-GMs. This is given by the following lemma.

LEMMA 4.1 (Local entropy of $k$-GM). *For any $\Gamma_0 \in \mathcal{G}_{k,d}$, let $\mathcal{P}_\epsilon(\Gamma_0) = \{P_\Gamma : \Gamma \in \mathcal{G}_{k,d}, H(P_\Gamma, P_{\Gamma_0}) \le \epsilon\}$. Then, for any $\delta \le \epsilon/2$,*

$$N(\delta, \mathcal{P}_\epsilon(\Gamma_0), H) \le \left(\frac{\epsilon}{\delta}\right)^{c(dk^4+(2k)^{2k+2})}, \tag{4.1}$$

*where the constant $c$ only depends only on $R$.*

Lemma 4.1 shows that any $\epsilon$-Hellinger ball $\mathcal{P}_\epsilon(\Gamma_0)$ in the space of $k$-GMs can be covered by at most $(\frac{C\epsilon}{\delta})^{Cd}$ $\delta$-Hellinger balls for some $C = C(k, R)$. This result is uniform in $\Gamma_0$ and depends optimally on $d$ but the dependency on the number of components $k$ is highly suboptimal. Lemma 4.1 should be compared with the local entropy bound in [27] obtained using a different approach than ours based on moment tensor. Specifically, [27, Example 3.4] shows that for Gaussian location mixtures, the local bracketing entropy centered at $P_{\Gamma_0}$ is bounded by $N_{[]}(\delta, \mathcal{P}_\epsilon(\Gamma_0), H) \le (\frac{C'\epsilon}{\delta})^{20kd}$, for some constant $C'$ depending on $P_{\Gamma_0}$ and $R$. This result yields optimal dependency on both $d$ and $k$ but lacks uniformity in the center of the Hellinger neighborhood (which is needed for applying the theory of Le Cam and Birgé).

Given the local entropy bound in Lemma 4.1, the upper bound $O_k(\frac{d}{n})$ in the squared Hellinger loss in Theorem 1.2 immediately follows by invoking the Le Cam-Birgé construction [13, Theorem 3.1]; see also [80, Lec. 18] for a self-contained exposition. For a high-probability bound that leads to (1.12), see, e.g., [80, Theorem 18.3].

Before proceeding to the proof of Lemma 4.1, we note that the Le Cam-Birgé construction, based on (exponentially many) pairwise tests, does not lead to a computationally efficient scheme for density estimation. This problem is much more challenging than estimating the mixing distribution, for which we have already obtained a polynomial-time optimal estimator in Section 3. (In fact, we show in Section 4.4, estimation of the mixing distribution can be reduced to proper density estimation both statistically and computationally.) Finding a computationally efficient proper density estimate that attains the parametric rate in Theorem 1.2 for arbitrary $k$, or even within logarithmic factors thereof, is open. Section 4.3 presents some partial progress on this front: We show that the estimator in Section 3 with slight modifications achieves the optimal rate of $O(\sqrt{d/n})$ for 2-GMs and the rate of $O((d/n)^{1/4})$ for general $k$-GMs; the latter result slightly improves (by logarithmic factors only) the state of the art in [2], but is still suboptimal.

Both the construction of the Hellinger covering for Lemma 4.1 and the analysis of density estimation in Section 4.3 rely on the notion of moment tensors, which we now introduce.

4.1. *Moment tensors and information geometry of Gaussian mixtures.* We recall some basics of tensors; for a comprehensive review, see [49]. The *rank* of an order-$\ell$ tensor $T \in (\mathbb{R}^d)^{\otimes \ell}$ is defined as the minimum $r$ such that $T$ can be written the sum of $r$ rank-one tensors, namely [50]:

$$\operatorname{rank}(T) \triangleq \min\left\{ r : T = \sum_{i=1}^r \alpha_i u_i^{(1)} \otimes \cdots \otimes u_i^{(\ell)}, \quad u_i^{(j)} \in \mathbb{R}^d, \alpha_i \in \mathbb{R} \right\}, \tag{4.2}$$

We will also use the *symmetric rank* [16]:

$$\operatorname{rank}_s(T) \triangleq \min\left\{ r : T = \sum_{i=1}^r \alpha_i u_i^{\otimes \ell}, \quad u_i \in \mathbb{R}^d, \alpha_i \in \mathbb{R} \right\}. \tag{4.3}$$

An order-$\ell$ tensor $T$ is *symmetric* if $T_{j_1,\ldots,j_\ell} = T_{j_{\pi(1)},\ldots,j_{\pi(\ell)}}$ for all $j_1, \ldots, j_\ell \in [d]$ and all permutations $\pi$ on $[\ell]$. The Frobenius norm of a tensor $T$ is defined as $\|T\|_F \triangleq \sqrt{\langle T, T\rangle}$, where

the tensor inner product is defined as $\langle S, T\rangle = \sum_{j_1,\ldots,j_\ell\in[d]} S_{j_1,\ldots,j_\ell} T_{j_1,\ldots,j_\ell}$. The spectral norm (operator norm) of a tensor $T$ is defined as

$$\|T\| \triangleq \max\{\langle T, u_1 \otimes u_2 \otimes \cdots \otimes u_\ell\rangle : \|u_i\| = 1, i = 1, \ldots, \ell\}. \tag{4.4}$$

Denote the set of $d$-dimensional order-$\ell$ symmetric tensors by $\mathbb{S}_\ell(\mathbb{R}^d)$. For a symmetric tensor, the following result attributed to Banach ([8, 25]) is crucial for the present paper:

$$\|T\| = \max\{|\langle T, u^{\otimes\ell}\rangle| : \|u\| = 1\}. \tag{4.5}$$

For $T \in \mathbb{S}_\ell(\mathbb{R}^d)$, if $\mathrm{rank}_s(T) \le r$, then the spectral norm can be bounded by the Frobenius norm as follows [66]:[3]

$$\frac{1}{\sqrt{r^{\ell-1}}} \|T\|_F \le \|T\| \le \|T\|_F . \tag{4.6}$$

For any $d$-dimensional random vector $U$, its order-$\ell$ *moment tensor* is

$$M_\ell(U) \triangleq \mathbb{E}[\underbrace{U \otimes \cdots \otimes U}_{\ell \text{ times}}], \tag{4.7}$$

which, by definition, is a symmetric tensor; in particular, $M_1(U) = \mathbb{E}[U]$ and $M_2(U - \mathbb{E}[U])$ are the mean and the covariance matrix of $U$, respectively. Given a multi-index $\mathbf{j} = (j_1, \ldots, j_d) \in \mathbb{Z}_+^d$, the $\mathbf{j}$th (multivariate) moment of $U$

$$m_{\mathbf{j}}(U) = \mathbb{E}[U_{j_1} \cdots U_{j_d}] \tag{4.8}$$

is the $\mathbf{j}$th entry of the moment tensor $M_{|\mathbf{j}|}(U)$, with $|\mathbf{j}| \triangleq j_1 + \ldots j_d$. Since moments are functionals of the underlying distribution, we also use the notation $M_\ell(\Gamma) = M_\ell(U)$ where $U \sim \Gamma$. An important observation is that the moment of the projection of a random vector can be expressed in terms of the moment tensor as follows: for any $u \in \mathbb{R}^d$,

$$m_\ell(\langle X, u\rangle) = \mathbb{E}[\langle X, u\rangle^\ell] = \mathbb{E}[\langle X^{\otimes\ell}, u^{\otimes\ell}\rangle] = \langle M_\ell(X), u^{\otimes\ell}\rangle.$$

Consequently, the difference between two moment tensors measured in the spectral norm is equal to the maximal moment difference of their projections. Indeed, thanks to (4.5),

$$\|M_\ell(X) - M_\ell(Y)\| = \sup_{\|u\|=1} |m_\ell(\langle X, u\rangle) - m_\ell(\langle Y, u\rangle)|. \tag{4.9}$$

Furthermore, if $U$ is a discrete random variable with a few atoms, then its moment tensor has low rank. Specifically, if $U$ is distributed according to some $k$-atomic distribution $\Gamma = \sum_{i=1}^k w_i \delta_{\mu_i}$, then

$$M_\ell(\Gamma) = \sum_{i=1}^k w_i \mu_i^{\otimes\ell}, \tag{4.10}$$

whose symmetric rank is at most $k$.

The following result gives a characterization of statistical distances (squared Hellinger, KL, or $\chi^2$-divergence) between $k$-GMs in terms of the moment tensors up to *dimension-independent* constant factors. Note that the upper bound in one dimension has been established in [81] (by combining Lemma 9 and 10 therein).

[3] The weaker bound $\|T\| \ge r^{-\ell/2} \|T\|_F$, which suffices for the purpose of this paper, takes less effort to show. Indeed, in view of the Tucker decomposition (4.23), combining (4.4) with (4.26) yields that $\|T\| \ge \max_{\mathbf{j}\in[r]^\ell} |\alpha_{\mathbf{j}}| \ge r^{-\ell/2} \|\alpha\|_F = r^{-\ell/2} \|T\|_F$.

THEOREM 4.2 (Moment characterization of statistical distances). *For any pair of $k$-atomic distributions $\Gamma, \Gamma'$ supported on the ball $B(0,R)$ in $\mathbb{R}^d$, for any $D \in \{H^2, \mathsf{KL}, \chi^2\}$,*

$$(Ck)^{-4k} \max_{\ell \le 2k-1} \left\|M_\ell(\Gamma) - M_\ell(\Gamma')\right\|_{\mathrm{F}}^2 \le D(P_\Gamma, P_{\Gamma'}) \le Ce^{36k^2} \max_{\ell \le 2k-1} \left\|M_\ell(\Gamma) - M_\ell(\Gamma')\right\|_{\mathrm{F}}^2. \tag{4.11}$$

*where the constant $C$ may depend on $R$ but not $k$ or $d$.*

To prove Theorem 4.2 we need a few auxiliary lemmas. The following lemma bounds the difference of higher-order moment tensors of $k$-atomic distributions using those of the first $2k-1$ moment tensors. The one-dimensional version was shown in [81, Lemma 10] using polynomial interpolation techniques; however, it is hard to extend this proof to multiple dimensions as multivariate polynomial interpolation (on arbitrary points) is much less well-understood. Fortunately, this difficulty can be sidestepped by exploiting the relationship between moment tensor norms and projections in (4.9).

LEMMA 4.3. *Let $U, U'$ be $k$-atomic random variables in $\mathbb{R}^d$. Then for any $j \ge 2k$,*

$$\|M_j(U) - M_j(U')\| \le 3^j \max_{\ell \in [2k-1]} \|M_\ell(U) - M_\ell(U')\|.$$

PROOF.

$$\begin{aligned}
\|M_j(U) - M_j(U')\| &\overset{\text{(a)}}{=} \sup_{\|v\|=1} |m_j(\langle U, v\rangle) - m_j(\langle U', v\rangle)| \\
&\overset{\text{(b)}}{\le} 3^j \sup_{\|v\|=1} \max_{\ell \in [2k-1]} |m_\ell(\langle U, v\rangle) - m_\ell(\langle U', v\rangle)| \\
&\overset{\text{(c)}}{=} 3^j \max_{\ell \in [2k-1]} \|M_\ell(U) - M_\ell(U')\|,
\end{aligned}$$

where (a) and (c) follow from (4.9), and (b) follows from [81, Lemma 10]. □

The lower bound part of Theorem 4.2 can be reduced to the one-dimensional case, which is covered by the following lemma. The proof relies on Newton interpolating polynomials and is deferred till Section 3 of the supplement [22].

LEMMA 4.4. *Let $\gamma, \gamma'$ be $k$-atomic distributions supported on $[-R, R]$. Then for any $(2k-1)$-times differentiable test function $h$,*

$$H(\gamma * N(0,1), \gamma' * N(0,1)) \ge c \left| \int h d\gamma - \int h d\gamma' \right|, \tag{4.12}$$

*where $c$ is some constant depending only on $k$, $R$, and $\max_{0 \le i \le 2k-1} \|h^{(i)}\|_{L_\infty([-R,R])}$. In the particular case where $h(x) = x^i$ for $i \in [2k-1]$, $c \ge (Ck)^{-k}$ for a constant $C$ depending only on $R$.*

PROOF OF THEOREM 4.2. Since

$$H^2(P,Q) \le \mathsf{KL}(P\|Q) \le \chi^2(P\|Q), \tag{4.13}$$

(see, e.g., [74, Section 2.4.1]), it suffices to prove the lower bound for $H^2$ and the upper bound for $\chi^2$.

Let $U \sim \Gamma$ and $U' \sim \Gamma'$, $X \sim P_\Gamma = \Gamma * N(0, I_d)$ and $X' \sim P_{\Gamma'} = \Gamma' * N(0, I_d)$. Then $\langle \theta, X\rangle \sim P_{\Gamma_\theta}$ and $\langle \theta, X'\rangle \sim P_{\Gamma'_\theta}$. By the data processing inequality,

$$H(P_\Gamma, P_{\Gamma'}) \geq \sup_{\theta \in S^{d-1}} H(P_{\Gamma_\theta}, P_{\Gamma'_\theta}). \tag{4.14}$$

Applying Lemma 4.4 to all monomials of degree at most $2k-1$, we obtain

$$H(P_\Gamma, P_{\Gamma'}) \geq (Ck)^{-k} \sup_{\theta \in S^{d-1}} \max_{\ell \leq 2k-1} |m_\ell(\langle \theta, U\rangle) - m_\ell(\langle \theta, U'\rangle)| = (Ck)^{-k} \max_{\ell \leq 2k-1} \left\| M_\ell(U) - M_\ell(U') \right\|, \tag{4.15}$$

for some constant $C$, where the last equality is due to (4.9). Thus the desired lower bound for Hellinger follows from the tensor norm comparison in (4.6).

To show the upper bound for $\chi^2$, we first reduce the dimension from $d$ to $2k$. Without loss of generality, assume that $d \geq 2k$ (for otherwise we can skip this step). Since both $U$ and $U'$ are $k$-atomic, the collection of atoms of $U$ and $U'$ lie in some subspace spanned by the orthonormal basis $\{v_1, \ldots, v_{2k}\}$. Let $V = [v_1, \ldots, v_{2k}]$ and let $V_\perp = [v_{2k+1}, \ldots, v_d]$ consist of orthonormal basis of the complement, so that $[V, V_\perp]$ is an orthogonal matrix. Write $X = U + Z$, where $Z \sim N(0, I_d)$ is independent of $U$. Then $V^\top X = V^\top U + V^\top Z \sim \nu * N(0, I_{2k}) = P_\nu$, where $\nu = \mathcal{L}(V^\top U)$ is a $k$-atomic distribution on $\mathbb{R}^{2k}$. Furthermore, $V_\perp^\top X = V_\perp^\top Z \sim N(0, I_{d-2k})$ and is independent of $V^\top X$. Similarly, $(V^\top X', V_\perp^\top X') \sim P_{\nu'} \otimes N(0, I_{d-2k})$, where $\nu' = \mathcal{L}(V^\top U')$. Therefore,

$$\begin{aligned}\chi^2(P_\Gamma \| P_{\Gamma'}) &= \chi^2(\mathcal{L}(V^\top X, V_\perp^\top X) \| \mathcal{L}(V^\top X', V_\perp^\top X')) = \chi^2(P_\nu \otimes N(0, I_{d-2k}) \| P_{\nu'} \otimes N(0, I_{d-2k})) \\ &= \chi^2(P_\nu \| P_{\nu'}).\end{aligned}$$

For notational convenience, let $B = V^\top U \sim \nu$ and $B' = V^\top U' \sim \nu'$.

To bound $\chi^2(P_\nu \| P_{\nu'})$, we first assume that $\mathbb{E}[B'] = 0$. For each multi-index $\mathbf{j} = (j_1, \ldots, j_d) \in \mathbb{Z}_+^{2k}$, define the $\mathbf{j}$th Hermite polynomial as

$$H_{\mathbf{j}}(x) = \prod_{i=1}^{2k} H_{j_i}(x_i), \quad x \in \mathbb{R}^{2k} \tag{4.16}$$

which is a degree-$|\mathbf{j}|$ polynomial in $x$. Furthermore, the following orthogonality property is inherited from that of univariate Hermite polynomials: for $Z \sim N(0, I_{2k})$,

$$\mathbb{E}[H_{\mathbf{j}}(Z) H_{\mathbf{j}'}(Z)] = \mathbf{j}! \mathbf{1}_{\{\mathbf{j} = \mathbf{j}'\}}. \tag{4.17}$$

Recall the exponential generating function of Hermite polynomials (see [1, 22.9.17]): for $x, b \in \mathbb{R}$, $\phi(x - b) = \phi(x) \sum_{j \geq 0} H_j(x) \frac{b^j}{j!}$. It is straightforward to obtain the multivariate extension of this result:

$$\phi_{2k}(x - b) = \phi_{2k}(x) \sum_{\mathbf{j} \in \mathbb{Z}_+^{2k}} \frac{H_{\mathbf{j}}(x)}{\mathbf{j}!} \prod_{i=1}^{2k} b_i^{j_i}, \quad x, b \in \mathbb{R}^{2k}.$$

Integrating $b$ over $B \sim \nu$, we obtain the following expansion of the density of $P_\nu$:

$$p_\nu(x) = \mathbb{E}[\phi_{2k}(x - B)] = \phi_{2k}(x) \sum_{\mathbf{j} \in \mathbb{Z}_+^{2k}} \frac{H_{\mathbf{j}}(x)}{\mathbf{j}!} \underbrace{\mathbb{E}\left[\prod_{i=1}^{2k} B_i^{j_i}\right]}_{m_{\mathbf{j}}(B)}.$$

Similarly, $p_{\nu'}(x) = \phi_{2k}(x) \sum_{\mathbf{j} \in \mathbb{Z}_+^{2k}} \frac{1}{\mathbf{j}!} m_{\mathbf{j}}(B') H_{\mathbf{j}}(x)$. Furthermore, by the assumption that $\mathbb{E}[B'] = 0$ and $\|B'\| \leq \|U'\| \leq R$ almost surely, Jensen's inequality yields

$$p_{\nu'}(x) = \phi_{2k}(x) \mathbb{E}[\exp(\langle B', x\rangle - \|B'\|^2/2)] \geq \phi_{2k}(x) \exp(-R^2/2).$$

Consequently,

$$\begin{aligned}
\chi^2(P_\nu\|P_{\nu'}) &\le e^{R^2/2}\int_{\mathbb{R}^{2k}} dx \frac{(p_\nu(x)-p_{\nu'}(x))^2}{\phi_{2k}(x)} \\
&\overset{\text{(a)}}{=} e^{\frac{R^2}{2}} \sum_{\mathbf{j}\in\mathbb{Z}_+^{2k}} \frac{(m_{\mathbf{j}}(B)-m_{\mathbf{j}}(B'))^2}{\mathbf{j}!} \\
&\overset{\text{(b)}}{\le} e^{\frac{R^2}{2}} \sum_{\ell\ge 1} \frac{\|M_\ell(B)-M_\ell(B')\|_{\mathrm{F}}^2}{\ell!}(2k)^\ell \\
&\overset{\text{(c)}}{\le} e^{\frac{R^2}{2}} e^{2k} \max_{\ell\in[2k-1]} \left\|M_\ell(B)-M_\ell(B')\right\|_{\mathrm{F}}^2 + e^{\frac{R^2}{2}} \sum_{\ell\ge 2k} \frac{(4k^2)^\ell}{\ell!} \left\|M_\ell(B)-M_\ell(B')\right\|^2 \\
&\overset{\text{(d)}}{\le} e^{\frac{R^2}{2}} \left( e^{2k} + \underbrace{\sum_{\ell\ge 2k} \frac{(36k^2)^\ell}{\ell!}}_{\le e^{36k^2}} \right) \max_{\ell\in[2k-1]} \left\|M_\ell(B)-M_\ell(B')\right\|_{\mathrm{F}}^2,
\end{aligned}$$

where (a) follows from the orthogonality relation (4.17); (b) is by the fact that $(|\mathbf{j}|)! \le \mathbf{j}!(2k)^{|\mathbf{j}|}$ for any $\mathbf{j}\in\mathbb{Z}_+^{2k}$; (c) follows from the tensor norm comparison inequality (4.6), since the symmetric rank of $M_\ell(B)-M_\ell(B')$ is at most $2k$ for all $\ell$; (d) follows from Lemma 4.3.

Finally, if $\mathbb{E}[B']\neq 0$, by the shift-invariance of $\chi^2$-divergence, applying the following simple lemma to $\mu=\mathbb{E}[B']$ (which satisfies $\|\mu\|\le R$) yields the desired upper bound. □

LEMMA 4.5. *For any random vectors $X$ and $Y$ and any deterministic $\mu\in\mathbb{R}^d$,*

$$\|M_\ell(X-\mu)-M_\ell(Y-\mu)\| \le \sum_{k=0}^{\ell}\binom{\ell}{k}\|M_k(X)-M_k(Y)\|\|\mu\|^{\ell-k}$$

PROOF. Using (4.9) and binomial expansion, we have:

$$\begin{aligned}
\|M_\ell(X-\mu)-M_\ell(Y-\mu)\| &= \sup_{\|u\|=1} |m_\ell(\langle X,u\rangle-\langle\mu,u\rangle)-m_\ell(\langle Y,u\rangle-\langle\mu,u\rangle)|. \\
&\le \sup_{\|u\|=1}\sum_{k=0}^{\ell}\binom{\ell}{k}|m_k(\langle X,u\rangle)-m_k(\langle Y,u\rangle)||\langle\mu,u\rangle|^{\ell-k} \\
&\le \sum_{k=0}^{\ell}\binom{\ell}{k}\|M_k(X)-M_k(Y)\|\|\mu\|^{\ell-k}
\end{aligned}$$

where in the step we used the Cauchy-Schwarz inequality. □

4.2. *Local entropy of Hellinger balls.* Before presenting the proof of Lemma 4.1, we discuss the connection and distinction between our approach and the existing literature on the metric entropy of mixture densities [32, 39, 40, 83, 58, 14, 70] and clarify the role of the moment tensors. Both the previous work and the current paper bound the statistical difference between mixtures in terms of moment differences (through either Taylor expansion or orthogonal expansion). For example, the seminal work [32] bounds the global entropy of nonparametric Gaussian mixtures in one dimension by first constructing a finite mixture approximation via moment matching, then discretizing the weights and atoms. The crucial

difference is that in the present paper we directly work with moment parametrization as opposed to the natural parametrization (atoms and weights). As mentioned in Section 1.1, to eliminate the unnecessary logarithmic factors and obtain the exact parametric rate in high dimensions, it is crucial to obtain a tight control of the *local* entropy as opposed to the global entropy, which relies on a good parametrization that bounds the Hellinger distance from both above and below – see (1.13). This *two-sided bound* is satisfied by the moment tensor reparametrization, thanks to Theorem 4.2, but not the natural parametrization. Therefore, to construct a good local covering, we do so in the moment space, by leveraging the low-rank structure of moment tensors.

PROOF OF LEMMA 4.1. Recall from Section 2 that $N(\epsilon, A, \rho)$ the $\epsilon$-covering number of the set $A$ with respect to the metric $\rho$, i.e., the minimum cardinality of an $\epsilon$-covering set $A_\epsilon$ such that, for any $v \in A$, there exists $\tilde{v} \in A_\epsilon$ with $\rho(v, \tilde{v}) < \epsilon$.

Let $\mathcal{M}_\epsilon = \{M(\Gamma) \colon P_\Gamma \in \mathcal{P}_\epsilon\}$, where $\mathcal{P}_\epsilon$ is the Hellinger neighborhood of $P_{\Gamma_0}$, $M(\Gamma) = (M_1(\Gamma), \ldots, M_{2k-1}(\Gamma))$ consists of the moment tensors of $\Gamma$ up to degree $2k-1$. Let $c_k' = \sqrt{c_k}$ and $C_k' = \sqrt{C_k}$ where $c_k \triangleq (Ck)^{-4k}$ and $C_k \triangleq C' e^{36k^2}$ are the constants from Theorem 4.2. To obtain a $\delta$-covering of $\mathcal{P}_\epsilon$, we first show that it suffices to construct a $\frac{\delta}{2C_k'}$-covering of the moment space $\mathcal{M}_\epsilon$ with respect to the distance $\rho(M, M') \triangleq \max_{\ell \le 2k-1} \|M_\ell - M_\ell'\|_F$ and thus

$$N(\delta, \mathcal{P}_\epsilon, H) \le N(\delta/(2C_k'), \mathcal{M}_\epsilon, \rho). \tag{4.18}$$

To this end, let $\mathcal{N}$ be the optimal $\frac{\delta}{2C_k'}$-covering of $\mathcal{M}_\epsilon$ with respect to $\rho$, and we show that $\mathcal{N}' = \{P_\Gamma : \Gamma = \operatorname{argmin}_{\Gamma' : P_{\Gamma'} \in \mathcal{P}_\epsilon} \rho(M(\Gamma'), M), M \in \mathcal{N}\}$ is a $\delta$-covering of $\mathcal{P}_\epsilon$. For any $P_\Gamma \in \mathcal{P}_\epsilon$, by the covering property of $\mathcal{N}$, there exists a tensor $M \in \mathcal{N}$ such that $\rho(M, M(\Gamma)) < \frac{\delta}{2C_k'}$. By the definition of $\mathcal{N}'$, there exists $P_{\tilde{\Gamma}} \in \mathcal{N}'$ such that $\rho(M(\tilde{\Gamma}), M) < \frac{\delta}{2C_k'}$. Therefore, $\rho(M(\tilde{\Gamma}), M(\Gamma)) < \frac{\delta}{C_k'}$ and thus $H(P_{\tilde{\Gamma}}, P_\Gamma) < \delta$ by Theorem 4.2.

Next we bound the right side of (4.18). Since $\Gamma_0, \Gamma$ are both $k$-atomic, it follows from Theorem 4.2 that

$$\mathcal{M}_\epsilon \subseteq M(\Gamma_0) + \{\Delta : \|\Delta_\ell\|_F \le \epsilon / c_k', \operatorname{rank}_s(\Delta_\ell) \le 2k, \forall \ell \le 2k-1\}, \tag{4.19}$$

where $\Delta = (\Delta_1, \ldots, \Delta_{2k-1})$ and $\Delta_\ell \in \mathbb{S}_\ell(\mathbb{R}^d)$. Let $\mathcal{D}_\ell = \{\Delta_\ell \in \mathbb{S}_\ell(\mathbb{R}^d) : \|\Delta_\ell\|_F \le \epsilon/c_k', \operatorname{rank}_s(\Delta_\ell) \le 2k\}$, and $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_{2k-1}$ be their Cartesian product. By monotonicity,

$$N(\delta/(2C_k'), \mathcal{M}_\epsilon, \rho) \le N(\delta/(2C_k'), \mathcal{D}, \rho) \le \prod_{\ell=1}^{2k-1} N(\delta/(2C_k'), \mathcal{D}_\ell, \|\cdot\|_F). \tag{4.20}$$

By Lemma 4.6 next,

$$N(\delta/(2C_k'), \mathcal{M}_\epsilon, \rho) \le \prod_{\ell=1}^{2k-1} \left(\frac{cC_k' \ell \epsilon}{2c_k' \delta}\right)^{2dk} \left(\frac{cC_k' \epsilon}{2c_k' \delta}\right)^{(2k)^\ell}, \tag{4.21}$$

for some absolute constant $c$. So we obtain, for constants $\tilde{C}, \tilde{C}'$ that does not depend on $d$ or $k$, that

$$\begin{aligned} N(\delta/(2C_k'), \mathcal{M}_\epsilon, \rho) &\le \left(\frac{\tilde{C} k^{4k} e^{36k^2} k\epsilon}{\delta}\right)^{4dk^2} \left(\frac{\tilde{C} k^{4k} e^{36k^2} \epsilon}{\delta}\right)^{(2k)^{2k}} \\ &\le \left(\frac{\tilde{C}' k^{4k+1} e^{36k^2} \epsilon}{\delta}\right)^{4dk^2 + (2k)^{2k}}. \end{aligned}$$

□

LEMMA 4.6. *Let* $\mathcal{T} = \{T \in \mathbb{S}_\ell(\mathbb{R}^d) : \|T\|_F \le \epsilon, \mathrm{rank}_s(T) \le r\}$. *Then, for any* $\delta \le \epsilon/2$,

$$N(\delta, \mathcal{T}, \|\cdot\|_F) \le \left(\frac{c\ell\epsilon}{\delta}\right)^{dr} \left(\frac{c\epsilon}{\delta}\right)^{r^\ell}, \tag{4.22}$$

*for some absolute constant* $c$.

PROOF. For any $T \in \mathcal{T}$, $\mathrm{rank}_s(T) \le r$. Thus $T = \sum_{i=1}^r a_i v_i^{\otimes \ell}$ for some $a_i \in \mathbb{R}$ and $v_i \in S^{d-1}$. Furthermore, $\|T\|_F \le \epsilon$. Ideally, if the coefficients satisfied $|a_i| \le \epsilon$ for all $i$, then we could cover the $r$-dimensional $\epsilon$-hypercube with an $\frac{\epsilon}{2}$-covering, which, combined with a $\frac{1}{2}$-covering of the unit sphere that covers the unit vectors $v_i$'s, constitutes a desired covering for the tensor. Unfortunately the coefficients $a_i$'s need not be small due to the possible cancellation between the rank-one components (consider the counterexample of $0 = v^{\otimes\ell} - v^{\otimes\ell}$). Next, to construct the desired covering we turn to the Tucker decomposition of the tensor $T$.

Let $u = (u_1, \ldots, u_r)$ be an orthonormal basis for the subspace spanned by $(v_1, \ldots, v_r)$. In particular, let $v_i = \sum_{j=1}^r b_{ij} u_j$. Then

$$T = \sum_{\mathbf{j}=(j_1,\ldots,j_\ell)\in[r]^\ell} \alpha_{\mathbf{j}} \underbrace{u_{j_1} \otimes \cdots \otimes u_{j_\ell}}_{\triangleq u_{\mathbf{j}}}, \tag{4.23}$$

where $\alpha_{\mathbf{j}} = \sum_{i=1}^r a_i b_{ij_1} \cdots b_{ij_\ell}$. In tensor notation, $T$ admits the following *Tucker decomposition*

$$T = \alpha \times_1 U \cdots \times_\ell U \tag{4.24}$$

where the symmetric tensor $\alpha = (\alpha_{\mathbf{j}}) \in \mathbb{S}_\ell(\mathbb{R}^r)$ is called the core tensor and $U$ is a $r \times d$ matrix whose rows are given by $u_1, \ldots, u_r$.

Due to the orthonormality of $(u_1, \ldots, u_r)$, we have for any $\mathbf{j}, \mathbf{j}' \in [r]^\ell$,

$$\langle u_{\mathbf{j}}, u_{\mathbf{j}'} \rangle = \prod_{i=1}^\ell \langle u_{j_i}, u_{j_i'} \rangle = \mathbf{1}_{\{\mathbf{j}=\mathbf{j}'\}}. \tag{4.25}$$

Hence we conclude from (4.23) that

$$\|\alpha\|_F = \|T\|_F. \tag{4.26}$$

In particular $\|\alpha\|_F \le \epsilon$. Therefore,

$$\mathcal{T} \subseteq \mathcal{T}' \triangleq \left\{ T = \sum_{\mathbf{j}\in[r]^\ell} \alpha_{\mathbf{j}} u_{j_1} \otimes \cdots \otimes u_{j_\ell} : \|\alpha\|_F \le \epsilon, \langle u_i, u_j \rangle = \mathbf{1}_{\{i=j\}} \right\}. \tag{4.27}$$

Let $\tilde{A}$ be a $\frac{\delta}{2}$-covering of $\{\alpha \in \mathbb{S}_\ell(\mathbb{R}^r) : \|\alpha\|_F \le \epsilon\}$ under $\|\cdot\|_F$ of size $(\frac{c\epsilon}{\delta})^{r^\ell}$ for some absolute constant $c$; let $\tilde{B}$ be a $\frac{\delta}{2\ell\epsilon}$-covering of $\{(u_1, \ldots, u_r) : \langle u_i, u_j \rangle = \mathbf{1}_{\{i=j\}}\}$ under the maximum of column norms of size $(\frac{c\ell\epsilon}{\delta})^{dr}$. Let $\tilde{\mathcal{T}}' = \{\sum_{\mathbf{j}\in[r]^\ell} \tilde{\alpha}_{\mathbf{j}} \tilde{u}_{j_1} \otimes \cdots \otimes \tilde{u}_{j_\ell} : \tilde{\alpha} \in \tilde{A}, \tilde{u} \in \tilde{B}\}$. Next we verify the covering property.

For any $T \in \mathcal{T}'$, there exists $\tilde{T} \in \tilde{\mathcal{T}}'$ such that $\|\alpha - \tilde{\alpha}\|_F \le \frac{\delta}{2}$ and $\max_{i\le r} \|u_i - \tilde{u}_i\| \le \frac{\delta}{2\ell\epsilon}$. Then, by the triangle inequality,

$$\left\|T - \tilde{T}\right\|_F \le \sum_{\mathbf{j}} |\alpha_{\mathbf{j}}| \, \|u_{j_1} \otimes \cdots \otimes u_{j_\ell} - \tilde{u}_{j_1} \otimes \cdots \otimes \tilde{u}_{j_\ell}\|_F + \left\| \sum_{\mathbf{j}} (\alpha_{\mathbf{j}} - \tilde{\alpha}_{\mathbf{j}}) \tilde{u}_{j_1} \otimes \cdots \otimes \tilde{u}_{j_\ell} \right\|_F. \tag{4.28}$$

The second term is at most $\|\alpha - \tilde{\alpha}\|_F \le \delta/2$. For the first term, it follows from the triangle inequality that

(4.29)
$$\|u_{j_1} \otimes \cdots \otimes u_{j_\ell} - \tilde{u}_{j_1} \otimes \cdots \otimes \tilde{u}_{j_\ell}\|_F \le \sum_{i=1}^{\ell} \|u_{j_1} \otimes \cdots \otimes (u_{j_i} - \tilde{u}_{j_i}) \otimes \cdots \otimes \tilde{u}_{j_\ell}\|_F \le \frac{\delta}{2\epsilon}.$$

Therefore, the first term is at most $\frac{\delta}{2\epsilon}\|\alpha\|_F \le \delta/2$. □

4.3. *Efficient proper density estimation.* To remedy the computational intractability of the Le Cam-Birgé estimator, in this subsection we adapt the procedure for mixing distribution estimation in Section 3 for density estimation. Let $\hat{\Gamma}$ be the estimated mixing distribution as defined in (3.4), with the following modifications:

- The grid size in (3.13) is adjusted to to $\epsilon_n = n^{-1/2}$. As such, in the set of $k$-atomic candidate distributions in (3.15), $\mathcal{W}$ denotes an $(\epsilon_n, \|\cdot\|_1)$-covering of the probability simplex $\Delta^{k-1}$, and $\mathcal{A}$ denotes an $(\epsilon_n, \|\cdot\|_2)$-covering the $k$-dimensional ball $\{x \in \mathbb{R}^k : \|x\|_2 \le R\}$.
- In the determination of the best mixing distribution in (3.18), instead of comparing the Wasserstein distance, we directly compare the projected moments on the directions over an $(\epsilon_n, \|\cdot\|_2)$-covering $\mathcal{N}$ of $S^{k-1}$:

$$\hat{\gamma} = \arg\min_{\gamma' \in \mathcal{S}} \max_{\theta \in \mathcal{N}} \max_{r \in [2k-1]} |m_r(\gamma'_\theta) - m_r(\widehat{\gamma}_\theta)|,$$

  where $\gamma'_\theta$ denotes the projection of $\gamma'$ onto the direction $\theta$ (recall (2.1)).

We then report $P_{\hat{\Gamma}} = \hat{\Gamma} * N(0, I_d)$ as a proper density estimate. By a similar analysis to Remark 1, using those finer grids, the run time of the procedure increases to $n^{O(k)}$. The next theorem provides a theoretical guarantee for this estimator in general $k$-GM model.

THEOREM 4.7. *There exists an absolute constant $C$ such that, with probability $1 - \delta$,*

$$H(P_\Gamma, P_{\hat{\Gamma}}) \le C\sqrt{k}\left(\frac{d + k^2 \log(k/\delta)}{n}\right)^{1/4} + \frac{(e^{Ck}\log(1/\delta))^{\frac{2k-1}{2}}}{\sqrt{n}}.$$

PROOF. Recall the notation $\Sigma, \hat{\Sigma}, \hat{V}, \hat{H}, \gamma$ defined in Section 3.3. By the triangle inequality,

$$H(P_\Gamma, P_{\hat{\Gamma}}) \le H(P_\Gamma, P_{\Gamma_{\hat{H}}}) + H(P_{\Gamma_{\hat{H}}}, P_{\hat{\Gamma}}) \le H(P_\Gamma, P_{\Gamma_{\hat{H}}}) + H(P_\gamma, P_{\hat{\gamma}}). \tag{4.30}$$

For the first term, we have

$$H^2(P_\Gamma, P_{\Gamma_{\hat{H}}}) \le \mathsf{KL}(P_\Gamma, P_{\Gamma_{\hat{H}}}) \overset{\text{(a)}}{\le} \frac{1}{2} W_2^2(\Gamma, \Gamma_{\hat{H}}) \overset{\text{(b)}}{\le} k\|\Sigma - \hat{\Sigma}\|_2, \tag{4.31}$$

where (a) follows from the convexity of the KL divergence (see [65, Remark 5]); (b) applies Lemma 3.5.

Next we analyze the second term in the right-hand side of (4.30) conditioning on $\hat{V}$. By Theorem 4.2, (4.6), and (4.9), the Hellinger distance is upper bounded by the difference between the projected moments:

$$H(P_\gamma, P_{\hat{\gamma}}) \le e^{Ck^2} \sup_{\theta \in S^{k-1}} \max_{r \in [2k-1]} |m_r(\gamma_\theta) - m_r(\hat{\gamma}_\theta)|. \tag{4.32}$$

It follows from Lemma 1.5 of the supplement [22] that

$$\min_{\gamma' \in \mathcal{S}} \max_{\theta \in S^{k-1}} \max_{r \in [2k-1]} |m_r(\gamma'_\theta) - m_r(\gamma_\theta)| \le \frac{2kR^{2k-1}}{\sqrt{n}}. \tag{4.33}$$

By (2.5) and (2.7) of the supplement [22], with probability $1-\delta/2$,

$$\max_{\theta\in\mathcal{N}}\max_{r\in[2k-1]}|m_r(\widehat{\gamma}_\theta)-m_r(\gamma_\theta)|\le\frac{(Ck^2\log(k/\delta))^{\frac{2k-1}{2}}}{\sqrt{n}}, \tag{4.34}$$

for an absolute constant $C$. Therefore, by (4.33) and (4.34),

$$\min_{\gamma'\in\mathcal{S}}\max_{\theta\in\mathcal{N}}\max_{r\in[2k-1]}|m_r(\gamma'_\theta)-m_r(\widehat{\gamma}_\theta)|\le\frac{(C'k^2\log(k/\delta))^{\frac{2k-1}{2}}+(C'k)^{4k}}{\sqrt{n}}, \tag{4.35}$$

for an absolute constant $C'\ge C$. Note that the minimizer of (4.35) is our estimator $\hat{\gamma}$. Consequently, combining (4.34) and (4.35), we obtain

$$\max_{\theta\in\mathcal{N}}\max_{r\in[2k-1]}|m_r(\hat{\gamma}_\theta)-m_r(\gamma_\theta)|\le\frac{2\left((C'k^2\log(k/\delta))^{\frac{2k-1}{2}}+(C'k)^{4k}\right)}{\sqrt{n}}.$$

Then it follows from Lemma 1.6 of the supplement [22] that

$$\sup_{\theta\in S^{k-1}}\max_{r\in[2k-1]}|m_r(\gamma_\theta)-m_r(\hat{\gamma}_\theta)|\le\frac{(C''k^2\log(k/\delta))^{\frac{2k-1}{2}}+(C''k)^{4k}}{\sqrt{n}},$$

for an absolute constant $C''$. Applying (4.32) yields that, with probability $1-\delta/2$,

$$H(P_\gamma,P_{\hat{\gamma}})\le\frac{(e^{C'''k}\log(1/\delta))^{\frac{2k-1}{2}}}{\sqrt{n}}, \tag{4.36}$$

for an absolute constant $C'''$. We conclude the theorem by applying Lemma 3.6, (4.31), and (4.36) to (4.30). □

Compared with the optimal parametric rate $O_k(\sqrt{d/n})$ in Theorem 1.2, the rate $O_k((d/n)^{1/4})$ in Theorem 4.7 is suboptimal. It turns out that, for the special case of 2-GMs, we can achieve the optimal rate using the same procedure with an extra centering step. Specifically, using the first half of observations $\{X_1,\dots,X_n\}$, we compute the sample mean and covariance matrix by

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^n X_i,\qquad S=\frac{1}{n}\sum_{i=1}^n(X_i-\hat{\mu})(X_i-\hat{\mu})^\top-I_d.$$

Let $\hat{u}\in\mathbb{R}^d$ be the top eigenvector of the sample covariance matrix $S$. Then we center and project the second half of the observations by $x_i=\langle\hat{u},X_{i+n}-\hat{\mu}\rangle$ for $i\in[n]$, which reduces the problem to one dimension. Then we apply the one-dimensional DMM algorithm with $x_1,\dots,x_n$ and obtain $\hat{\gamma}=\sum_{i=1}^2\hat{w}_i\delta_{\hat{\theta}_i}$. Finally, we report $P_{\hat{\Gamma}}$ with the mixing distribution

$$\hat{\Gamma}=\sum_{i=1}^2\hat{w}_i\delta_{\hat{\theta}_i\hat{u}+\hat{\mu}}.$$

The next result shows the optimality of $P_{\hat{\Gamma}}$.

THEOREM 4.8. *With probability at least* $1-\delta$,

$$H(P_\Gamma,P_{\hat{\Gamma}})\lesssim\sqrt{\frac{d+\log(1/\delta)}{n}}.$$

PROOF. Let $\Gamma = w_1\delta_{\mu_1} + w_2\delta_{\mu_2}$, where $\|\mu_i\| \le R, i = 1, 2$. Denote the population mean and covariance matrix by $\mu = \mathbb{E}_\Gamma[U]$ and $\Xi = \mathbb{E}_\Gamma(U-\mu)(U-\mu)^\top$, respectively. We have the following standard results for the sample mean and covariance matrix (see, e.g., [57, Eq. (1.1)] and Lemma 3.6):

$$\|\hat{\mu} - \mu\|_2,\ \|S - \Xi\|_2 \le C(R)\sqrt{\frac{d + \log(1/\delta)}{n}}, \tag{4.37}$$

with probability at least $1 - \delta/2$. Let $\Gamma' = \sum_{i=1}^2 w_i \delta_{\langle \hat{u}, \mu_i - \mu\rangle \hat{u} + \mu}$, and $\tilde{\Gamma} = \sum_{i=1}^2 w_i \delta_{\langle \hat{u}, \mu_i - \hat{\mu}\rangle \hat{u} + \hat{\mu}}$. By the triangle inequality,

$$H(P_\Gamma, P_{\hat{\Gamma}}) \le H(P_\Gamma, P_{\Gamma'}) + H(P_{\Gamma'}, P_{\tilde{\Gamma}}) + H(P_{\tilde{\Gamma}}, P_{\hat{\Gamma}}). \tag{4.38}$$

We upper bound three terms separately conditioning on $\hat{\mu}$ and $\hat{u}$. For the second term of (4.38), applying the convexity of the squared Hellinger distance yields that

$$\begin{aligned} H^2(P_{\tilde{\Gamma}}, P_{\Gamma'}) &\le \sum_{i=1}^2 w_i H^2\left(N(\hat{u}\hat{u}^\top(\mu_i - \mu) + \mu, I_d), N(\hat{u}\hat{u}^\top(\mu_i - \hat{\mu}) + \hat{\mu}, I_d)\right) \\ &= \sum_{i=1}^2 w_i \left(2 - 2e^{-\frac{\|(I - \hat{u}\hat{u}^\top)(\mu - \hat{\mu})\|_2^2}{8}}\right) \le \|\mu - \hat{\mu}\|_2^2, \end{aligned} \tag{4.39}$$

where we used $e^x \ge 1 + x$. For the third term (4.38), note that conditioned on the $(\hat{u}, \hat{\mu})$, $x_i \overset{\text{i.i.d.}}{\sim} P_{\tilde{\gamma}}$ where $\tilde{\gamma} = \sum_{i=1}^2 w_i \delta_{\langle \hat{u}, \mu_i - \hat{\mu}\rangle}$. Note that $\hat{\Gamma}$ and $\tilde{\Gamma}$ are supported on the same affine subspace $\{\theta\hat{u} + \hat{\mu} : \theta \in \mathbb{R}\}$. Thus

$$H(P_{\hat{\Gamma}}, P_{\tilde{\Gamma}}) = H(P_{\hat{\gamma}}, P_{\tilde{\gamma}}) \lesssim \sqrt{\frac{\log(1/\delta)}{n}}, \tag{4.40}$$

with probability at least $1 - \delta/2$, where the inequality follows from the statistical guarantee of the DMM algorithm in [81, Theorem 3].

It remains to upper bound the first term of (4.38). Denote the centered version of $\Gamma, \Gamma'$ by $\pi, \pi'$. Using $\mu = w_1\mu_1 + w_2\mu_2$ and $w_1 + w_2 = 1$, we may write $\pi = w_1\delta_{\lambda w_2 u} + w_2\delta_{-\lambda w_1 u}$, $\pi' = w_1\delta_{\lambda w_2 \hat{u}\hat{u}^\top u} + w_2\delta_{-\lambda w_1 \hat{u}\hat{u}^\top u}$, where $\lambda \triangleq \|\mu_1 - \mu_2\|_2$, and $u \triangleq \frac{\mu_1 - \mu_2}{\|\mu_1 - \mu_2\|_2}$. Then

$$H(P_\Gamma, P_{\Gamma'}) = H(P_\pi, P_{\pi'}). \tag{4.41}$$

By Theorem 4.2, it is equivalent to upper bound the difference between the first three moment tensors. Both $\pi$ and $\pi'$ have zero mean. Using $w_2^2 - w_1^2 = w_2 - w_1$, their covariance matrices and the third-order moment tensors are found to be

$$\Xi = \mathbb{E}_\pi[UU^\top] = \lambda^2 w_1 w_2 uu^\top, \qquad T = \mathbb{E}_\pi[U^{\otimes 3}] = \lambda^3 w_1 w_2 (w_1 - w_2) u^{\otimes 3},$$

$$\Xi' = \mathbb{E}_{\pi'}[UU^\top] = \lambda^2 w_1 w_2 \langle u, \hat{u}\rangle^2 \hat{u}\hat{u}^\top, \qquad T' = \mathbb{E}_{\pi'}[U^{\otimes 3}] = \lambda^3 w_1 w_2 (w_1 - w_2) \langle u, \hat{u}\rangle^3 \hat{u}^{\otimes 3}.$$

Applying Theorem 4.2 and Lemma 4.9 below yields that $H(P_\pi, P_{\pi'}) \lesssim \|S - \Xi\|$. The proof is completed by combining (4.37) – (4.41). □

LEMMA 4.9.

$$\|\Xi - \tilde{\Xi}\|_F + \|T - T'\|_F \lesssim \|S - \Xi\|.$$

PROOF. Let $\sigma = \lambda^2 w_1 w_2$ and $\cos\theta = \langle u, \hat{u}\rangle$. Since $\lambda \lesssim 1$, we have

$$\begin{aligned} \|\Xi - \tilde{\Xi}\|_F^2 &= \sigma^2(1 - \cos^4\theta) \lesssim \sigma^2 \sin^2\theta, \\ \|T - T'\|_F^2 &= \lambda^2\sigma^2(w_1 - w_2)^2(1 - \cos^6\theta) \lesssim \sigma^2 \sin^2\theta. \end{aligned}$$

Since $\Xi = \sigma uu^\top$, by the Davis-Kahan theorem [19], $\sin\theta \lesssim \frac{\|S - \Xi\|}{\sigma}$, completing the proof. □

4.4. *Connection to mixing distribution estimation.* The next result shows that optimal estimation of the mixing distribution can be reduced to that of the mixture density, both statistically and computationally, provided that the density estimate is proper (a valid $k$-GM). Note that this does not mean an optimal density estimate $P_{\hat{\Gamma}}$ automatically yields an optimal estimator of the mixing distribution $\hat{\Gamma}$ for Theorem 1.1. Instead, we rely on an intermediate step that allows us to estimate the appropriate subspace and then perform density estimation in this low-dimensional space.

THEOREM 4.10. *Suppose that for each $d \in \mathbb{N}$, there exists a proper density estimator $\hat{P} = \hat{P}(X_1, \ldots, X_n)$, such that for every $\Gamma \in \mathcal{G}_{k,d}$ and $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\Gamma$,*

$$\mathbb{E}H(\hat{P}, P_\Gamma) \le c_k (d/n)^{1/2}, \tag{4.42}$$

*for some constant $c_k$. Then there is an estimator $\hat{\Gamma}$ of the mixing distribution $\Gamma$ and a positive constant $C$ such that*

$$\mathbb{E}W_1(\hat{\Gamma}, \Gamma) \le \left( (Ck)^{k/2} \sqrt{c_k} \left(\frac{d}{n}\right)^{1/4} + Ck^5 c_k^{\frac{1}{2k-1}} \left(\frac{1}{n}\right)^{\frac{1}{4k-2}} \right). \tag{4.43}$$

PROOF OF THEOREM 4.10. We first construct the estimator $\hat{\Gamma}$ using $X_1, \ldots, X_{2n} \overset{\text{i.i.d.}}{\sim} P_\Gamma$. Let $\hat{P} \in \mathcal{P}_{k,d}$ be the proper mixture density estimator from $\{X_i\}_{i \le n}$ satisfying

$$\mathbb{E}H(\hat{P}, P_\Gamma) \le c_k \sqrt{d/n}, \tag{4.44}$$

for a positive constant $c_k$, as guaranteed by (4.42). Since $\hat{P}$ is a proper estimator, it can be written $\hat{P} = \hat{\Gamma}' * N(0, I_d)$ for some $\hat{\Gamma}' \in \mathcal{G}_{k,d}$.

Let $\hat{V} \in \mathbb{R}^{d \times k}$ be a matrix whose columns form an orthonormal basis for the space spanned by the atoms of $\hat{\Gamma}'$, $\hat{H} = \hat{V}\hat{V}^\top$, and $\gamma = \Gamma_{\hat{V}}$. Note that conditioned on $\hat{V}$, $\{\hat{V}^\top X_i\}_{i=n+1,\ldots,2n}$ is an i.i.d. sample drawn from the $k$-GM $P_\gamma$. Invoking (4.42) again, there exists an estimator $\hat{\gamma} = \sum_{j=1}^k \hat{w}_j \delta_{\hat{\psi}_j} \in \mathcal{G}_{k,k}$ such that

$$\mathbb{E}H(P_{\hat{\gamma}}, P_\gamma) \le c_k \sqrt{k/n}. \tag{4.45}$$

We will show that $\hat{\Gamma} \triangleq \hat{\gamma}_{\hat{V}^\top} = \sum_{j=1}^k \hat{w}_j \delta_{\hat{V}\hat{\psi}_j}$ achieves the desired rate (4.43). Recall from (3.19) the risk decomposition:

$$W_1(\Gamma, \hat{\Gamma}) \le W_1(\Gamma, \Gamma_{\hat{H}}) + W_1(\gamma, \hat{\gamma}). \tag{4.46}$$

Let $\Sigma = \mathbb{E}_{U \sim \Gamma}[UU^\top]$ and $\hat{\Sigma} = \mathbb{E}_{U \sim \hat{\Gamma}'}[UU^\top]$ whose ranks are at most $k$. Then $\hat{H}$ is the projection matrix onto the space spanned by the top $k$ eigenvectors of $\hat{\Sigma}$. It follows from Lemma 3.5 and the Cauchy-Schwarz inequality that $W_1(\Gamma, \Gamma_{\hat{H}}) \le \sqrt{2k\|\Sigma - \hat{\Sigma}\|_2}$. By Lemma 4.4 and the data processing inequality of the Hellinger distance,

$$\|\Sigma - \hat{\Sigma}\|_2 = \sup_{\theta \in S^{d-1}} |m_2(\Gamma_\theta) - m_2(\hat{\Gamma}'_\theta)| \le C_k \sup_{\theta \in S^{d-1}} H(P_{\Gamma_\theta}, P_{\hat{\Gamma}'_\theta}) \le C_k H(P_\Gamma, P_{\hat{\Gamma}'}),$$

where $C_k = (Ck)^k$ for a constant $C$. Therefore, by (4.44), we obtain that

$$\mathbb{E}W_1(\Gamma, \Gamma_{\hat{H}}) \le \sqrt{2kC_k c_k \left(\frac{d}{n}\right)^{1/2}}. \tag{4.47}$$

We condition on $\hat{V}$ to analyze the second term on the right-hand side of (4.46). By Lemma 3.1 and Lemma 1.1 of the supplement [22], there is a constant $C'$ such that

$$W_1(\gamma,\hat{\gamma}) \le k^{5/2} \sup_{\theta\in S^{k-1}} W_1(\gamma_\theta,\hat{\gamma}_\theta) \le C' k^{7/2} \sup_{\theta\in S^{k-1}, r\in[2k-1]} |m_r(\gamma_\theta) - m_r(\hat{\gamma}_\theta)|^{\frac{1}{2k-1}}.$$

Again, by Lemma 4.4 and the data processing inequality, for any $\theta \in S^{k-1}$ and $r \in [2k-1]$,

$$|m_r(\gamma_\theta) - m_r(\hat{\gamma}_\theta)| \le C_k H\left(P_{\hat{\gamma}_\theta}, P_{\gamma_\theta}\right) \le C_k H\left(P_{\hat{\gamma}}, P_{\gamma}\right).$$

Therefore, by (4.45), we obtain that

$$\mathbb{E} W_1(\gamma,\hat{\gamma}) \le C' k^{7/2} \left( C_k c_k \left(\frac{k}{n}\right)^{1/2} \right)^{\frac{1}{2k-1}}. \tag{4.48}$$

The conclusion follows by applying (4.47) and (4.48) in (4.46). □

At the crux of the above proof is the following key inequality for $k$-GMs in $d$ dimensions:

$$W_1(\gamma,\hat{\gamma}) \lesssim_{k,d} H(P_\gamma, P_{\hat{\gamma}})^{1/(2k-1)}, \tag{4.49}$$

which we apply after a dimension reduction step. The proof of (4.49) relies on two crucial facts: for one-dimensional $k$-atomic distributions $\gamma, \hat{\gamma}$,

$$W_1(\gamma,\hat{\gamma}) \lesssim_k \max_{\ell\in[2k-1]} |m_\ell(\gamma) - m_\ell(\hat{\gamma})|^{1/(2k-1)}, \tag{4.50}$$

and

$$\max_{\ell\in[2k-1]} |m_\ell(\gamma) - m_\ell(\hat{\gamma})| \lesssim_k H(P_\gamma, P_{\hat{\gamma}}). \tag{4.51}$$

Then (4.49) immediately follows from Lemma 3.1 and (4.9).

Relationships similar to (4.49) are found elsewhere in the literature on mixture models, e.g., [15, 39, 40, 38], where they are commonly used to translate a density estimation guarantee into one for mixing distributions. For example, [39, Proposition 2.2(b)] showed the non-uniform bound $W_r^r(\gamma,\hat{\gamma}) \le C(\gamma) H(P_\gamma, P_{\hat{\gamma}})$, where $r$ is a parameter that depends on the level of overfitting in the model; see [39, 40] for more results for other models such as location-scale Gaussian mixtures. For uniform bound similar to (4.49), [38, Theorem 6.3] showed $W_{2k-1}^{2k-1}(\gamma,\hat{\gamma}) \lesssim \|p_\gamma - p_{\hat{\gamma}}\|_\infty$ in one dimension.

Conversely, distances between mixtures can also be bounded by transportation distances between mixing distributions, e.g., the middle inequality in (4.31) for the KL divergence. Total variation inequality of the form $\mathsf{TV}(F*\gamma, F*\hat{\gamma}) \lesssim W_1(\gamma,\hat{\gamma})$ for arbitrary $\gamma, \gamma'$ are shown in [65, Proposition 7] or [14, Proposition 5.3], provided that $F$ has bounded density. See also [40, Theorem 3.2(c)] and [39, Proposition 2.2(a)] for results along this line.

**5. Numerical studies.** This section presents numerical experiments comparing the estimator (3.4) to the classical EM algorithm. The EM algorithm is guaranteed only to converge to a local optimum (and very slowly without separation conditions) [44], and its performance depends heavily on the initialization chosen. It moreover takes a pass at the entire sample on each iteration.

As opposed to the worst-case scenario considered in the minimax analysis, in the experiments we consider fixed mixing distribution $\Gamma$ that does not depend on $n$. The empirical results demonstrate that in certain cases, the proposed algorithm, while not scalable to large $k$, can be a good alternative to EM in terms of both accuracy and speed. However, we reiterate that this algorithm, which is meant to show that the minimax rate can be achieved in

time polynomial in $n$, is not practical even for modest values of $k$. The EM algorithm, while hampered by its repeated accessing of all data points and the possibility of being derailed by spurious local maxima, scales much better in $k$.

All simulations are run in Python. The DMM algorithm relies on the CVXPY [21] and CVXOPT [5] packages; see Section 6 of [81] for more details on the implementation of DMM. We also use the Python Optimal Transport package [24] to compute the $d$-dimensional 1-Wasserstein distance.

In all experiments, we set $\sigma = 1, d = 100$, with $n$ ranging from $10,000$ to $200,000$ in increments of $10,000$. For each model and each value of $n$, we run 10 repeated experiments; we plot the mean error and standard deviation of the error in the figures. We initialize EM randomly, and our stopping criterion for the EM algorithm is either after 1000 iterations or once the relative change in log likelihood is below $10^{-6}$. For the dimension reduction step in the computation of (3.4), we first center our data then project onto the top $k-1$ singular vectors. Thus when $k = 2$, we project onto a one-dimensional subspace and only run DMM once, so the grid search of Algorithm 1 is never invoked. While sample splitting is used for the sake of analyzing the estimator (3.4), in the actual experiments we forgo this step.

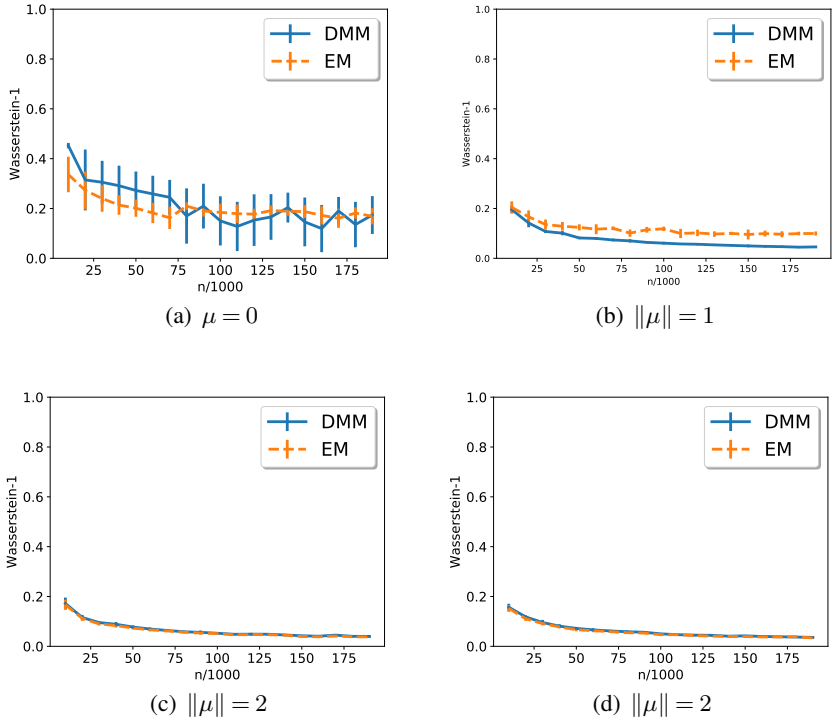
When $k = 3$, the data are projected to a 2-dimensional subspace after centering. In this case, we need to choose $\mathcal{W}, \mathcal{N}$, the $\epsilon_{n,k}$-nets on the simplex $\Delta^{k-1}$ and on the unit sphere $S^{k-2}$, respectively. Here $\mathcal{W}$ is chosen by discretizing the probabilities and $\mathcal{N}$ is formed by by gridding the angles $\alpha \in [-\pi, \pi]$ and using the points $(\cos\alpha, \sin\alpha)$. Note that here, $|\mathcal{W}| \le (C_1/\epsilon_{n,k})^{k-1}, |\mathcal{N}| \le (C_2/\epsilon_{n,k})^{k-2}$. For example, when $n = 10000$, $1/\epsilon_{n,k} \approx 3$. In the experiments in Fig. 2, we choose $C_1 = 1, C_2 = 2$.

In Fig. 1, we compare the performance on the symmetric 2-GM, where the sample is drawn from the distribution $\frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d)$. For Fig. 1(a), $\mu = 0$, i.e., the components completely overlap. And for Fig. 1(b) and Fig. 1(c), $\mu$ is uniformly drawn from the sphere of radius 1 and 2, respectively. In Fig. 1(d), the model is $P_\Gamma = \frac{1}{4}N(\mu, I_d) + \frac{3}{4}N(-\mu, I_d)$ where $\mu$ is drawn from the sphere of radius 2. Our algorithm and EM perform similarly for the model with overlapping components; our algorithm is more accurate than EM in the model where $\|\mu\|_2 = 1$, but EM improves as the model components become more separated. There is little difference in the performance of either algorithm in the uneven weights scenario.

In Fig. 2, we compare the performance on the 3-GM model $\frac{1}{3}N(\mu, I_d) + \frac{1}{3}N(0, I_d) + \frac{1}{3}N(-\mu, I_d)$ for different values of separation $\|\mu\|$. In these experiments, we see the opposite phenomenon in terms of the relative performance of our algorithm and EM: the former improves more as the centers become more separated. This seems to be because in, for instance, the case where $\mu = 0$, the error in each coordinate for DMM is fairly high, and this is compounded when we select the two-coordinate final distribution. The performance of our algorithm improves rapidly here because as the model becomes more separated, the errors in each coordinate become very small. Note that since we have made the model more difficult to learn by adding a center at 0, the errors are higher than for the $k = 2$ example in every experiment for both algorithms.

In Fig. 3, we provide further experiments to explore the adaptivity of the estimator produced by the algorithm in Section 3. The settings are the same as in the previous experiments except we choose a finer grid with parameter $C_1 = 2$ instead of $C_1 = 1$, for otherwise the quantization error of the weights is too large.

In Fig. 3(a), we let the true model be exactly as in Fig. 1(c), but we run the algorithm from Section 3 using $k = 3$. As in Fig. 2, DMM seems to improve more rapidly than EM as $n$ increases. But here, DMM has higher error than EM for small $n$. In Fig. 3(b), we let $k = 3$ and create a model without the symmetry structures of the models in previous experiments by drawing the atoms uniformly from the unit sphere and the weights from a Dirichlet$(1, 1, 1)$ distribution. This model is more difficult to learn for both DMM and EM, but DMM still outperforms EM in terms of accuracy.




(a) $\mu = 0$ (b) $\|\mu\| = 1$

(c) $\|\mu\| = 2$ (d) $\|\mu\| = 2$

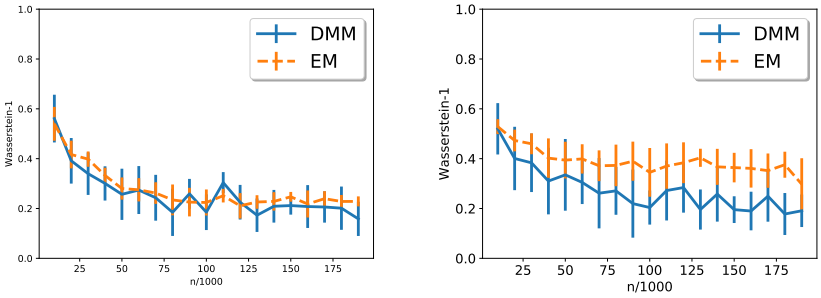
FIG 1. *In the first three figures,* $P_\Gamma = \frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d)$*, for increasing values of* $\|\mu\|_2$*. In the final figure,* $P_\Gamma = \frac{1}{4}N(\mu, I_d) + \frac{3}{4}N(-\mu, I_d)$ *where* $\|\mu\|_2 = 2$*.*

Finally, we provide a table of the average running time (in seconds) for each experiment. As expected, in the experiments in Fig. 1, the DMM algorithm is faster than EM. In the experiments in Fig. 2, DMM manages to run faster on average than EM in the two more separated models, Fig. 2(b) and Fig. 2(c). Also unsurprisingly, DMM is slower in the Fig. 2 experiments than those in Fig. 1, because grid search is invoked in the former. In the over-fitted case in Fig. 3(a), DMM is much slower on average than EM, and in fact is slower on average than DMM in any other experiment setup. In Fig. 3(b), where the model does not have special structure, DMM nonetheless runs on average in time faster than for some of the symmetric models in Fig. 2, and moreover again improves on the average run time of EM.

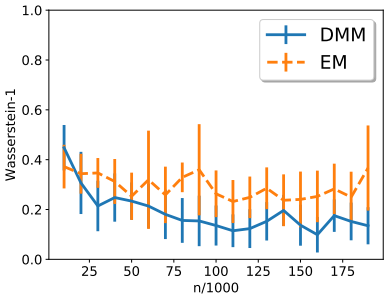
| Experiment | DMM | EM |
|---|---|---|
| Fig. 1(a) | 0.114407 | 0.678521 |
| Fig. 1(b) | 0.121561 | 1.163886 |
| Fig. 1(c) | 0.206713 | 0.573640 |
| Fig. 1(d) | 0.221704 | 0.818138 |
| Fig. 2(a) | 1.118308 | 0.985668 |
| Fig. 2(b) | 2.257503 | 2.582501 |
| Fig. 2(c) | 2.179928 | 3.576998 |
| Fig. 3(a) | 3.840112 | 1.350464 |
| Fig. 3(b) | 1.907299 | 2.546508 |

TABLE 1

*Running time comparison (in seconds).*

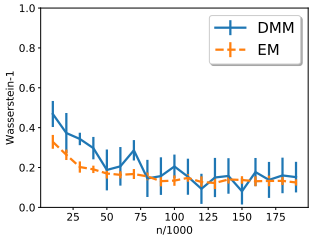



(a) $\mu = 0$ (b) $\|\mu\| = 1$

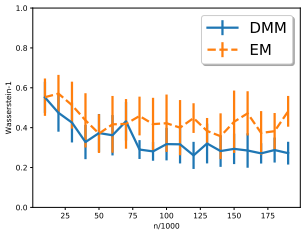



(c) $\|\mu\| = 2$

FIG 2. $P_\Gamma = \frac{1}{3}N(\mu, I_d) + \frac{1}{3}N(0, I_d) + \frac{1}{3}N(-\mu, I_d)$ *for increasing values of* $\|\mu\|_2$.




(a) $P_\Gamma = \frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d)$ with $\|\mu\|_2 = 2$.




(b) $P_\Gamma = \sum_{j=1}^{3} w_j N(\mu_j, I_d)$ with $\mu_j$'s drawn uniformly from the unit sphere and $(w_1, w_2, w_3) \sim$ Dirichlet$(1, 1, 1)$.

FIG 3. *Adaptivity and asymmetry.*

**6. Discussion.** In this paper we focused on the Gaussian location mixture model (1.1) in high dimensions, where the variance parameter $\sigma^2$ and the number of components $k$ are known, and the centers lie in a ball of bounded radius. Below we discuss weakening these assumptions and other open problems.

*Unbounded centers.* While the assumption of bounded support is necessary for estimating the mixing distribution (otherwise the worst-case $W_1$-loss is infinity), it is not needed for density estimation [2, 52, 7]. In fact, [2] first uses a crude clustering procedure to partition the sample into clusters whose means are close to each other, then zooms into each cluster to perform density estimation. For the lower bound, the worst case occurs when each cluster

is equally weighted and highly separated, so that the effective sample size for each component is $n/k$, leading to the lower bound of $\Omega(\frac{kd}{n})$. On the other hand, the density estimation guarantee for NPMLE in [32, 83, 70] relies on assumptions of either compact support or tail bound on the mixing distribution.

*Location-scale mixtures.* We have assumed that the covariance of our mixture is known and common across components. There is a large body of work studying general location-scale Gaussian mixtures, see, e.g., [60, 39, 7]. The introduction of the scale parameters makes the problem significantly more difficult. For parameter estimation, the optimal rate remains unknown even in one dimension except for $k = 2$ [35]. In the special case where all components share the same unknown variance $\sigma^2$, the optimal rate in one dimension is shown in [81] to be $n^{-1/(4k)}$ for estimating the mixing distribution and $n^{-1/(2k)}$ for $\sigma^2$, achieved by Lindsay's estimator [54]. Modifying the procedure in Section 3 by replacing the DMM subroutine with Lindsay's estimator, this result can be extended to high dimensions as follows (see Section 4 of the supplemental material [22] for details), provided that the unknown covariance matrix is isotropic; otherwise the optimal rate is open.

THEOREM 6.1 (Unknown common variance). *Assume the setting of Theorem 1.1, where $P_\Gamma = \Gamma * N(0, \sigma^2 I_d)$ for some unknown $\sigma$ bounded by some absolute constant $C$ and $\Gamma \in \mathcal{G}_{k,d}$. Given $n$ i.i.d. observations from $P_\Gamma$, there exists an estimator $(\hat{\Gamma}, \hat{\sigma})$ such that*

$$\mathbb{E}W_1(\hat{\Gamma}, \Gamma) \lesssim_k \left(\frac{d}{n}\right)^{1/4} \wedge 1 + n^{-1/(4k)}, \quad \mathbb{E}|\hat{\sigma}^2 - \sigma^2| \lesssim_k n^{-1/(2k)}. \tag{6.1}$$

*Furthermore, both rates are minimax optimal.*

*Number of components.* This work assumes that the parameter $k$ is known and fixed. Since the centers are allowed to overlap arbitrarily, $k$ is effectively an upper bound on the number of components. If $k$ is allowed to depend on $n$, the optimal $W_1$-rate is shown in [81, Theorem 5] to be $\Theta(\frac{\log\log n}{\log n})$ provided $k = \Omega(\frac{\log n}{\log\log n})$, including nonparametric mixtures. Extending this result to the high-dimensional setting of Theorem 1.1 is an interesting future direction.

The problem of selecting the mixture order $k$ has been extensively studied. For instance, many authors have considered likelihood-ratio based tests; however, standard asymptotics for such tests may not hold [37]. Various workarounds have been considered, including method of moments [54, 17], tests inspired by the EM algorithm [53], quadratic approximation of the log-likelihood ratio [55], and penalized likelihood [26]. A common practical method is to infer $k$ from an eigengap in the sample covariance matrix. In our setting, this technique is not viable even if the model centers are separated, since the atoms may all lie on a low-dimensional subspace. However, under separation assumptions we may infer a good value of $k$ from the estimated mixing distribution $\hat{\Gamma}$ of our algorithm.

*Efficient algorithms for density estimation.* As mentioned in Section 1.2, for the high-dimensional $k$-GM model, achieving the optimal rate $O_k(\sqrt{d/n})$ with a proper density estimate in polynomial time is unresolved except for the special case of $k = 2$. Such a procedure, as described in Section 4.3, is of method-of-moments type (involving the first three moments); nevertheless, thanks to the observation that the one-dimensional subspace of spanned by the centers of a zero-mean 2-GM can be extracted from the covariance matrix, we can reduce the problem to one dimension by projection, thereby sidestepping third-order tensor decomposition which poses computational difficulty. Unfortunately, this observation breaks down for $k$-GM with $k \geq 3$, as covariance alone does not provide enough information for learning the subspace accurately. For this reason it is unclear whether the algorithm in Section 4.3 is capable to achieve the optimal rate of $\sqrt{d/n}$ and so far we can only prove a rate of $(d/n)^{1/4}$ in Theorem 4.7. Closing this computational gap (or proving its impossibility) is a challenging open question.

*Analysis of the MLE.* A natural approach to any estimation problem is the maximum likelihood estimator, which, for the $k$-GM model (1.6), is defined as $\hat{\Gamma}_{\mathrm{MLE}} = \operatorname{argmax}_{\Gamma \in \mathcal{G}_{k,d}} \sum_{i=1}^{n} \log p_{\Gamma}(X_i)$. Although this non-convex optimization is difficult to solve in high dimensions, it is of interest to understand the statistical performance of the MLE and whether it can achieve the optimal rate of density estimation in Theorem 1.2.

A rate of convergence for the MLE is typically found by bounding the *bracketing entropy* of the class of square-root densities; see, e.g., [77, 75]. Given a function class $\mathcal{F}$ of real-valued functions on $\mathbb{R}^d$, its $\epsilon$-bracketing number $N_{[]}(\epsilon)$ is defined as the minimum number of brackets (pairs of functions which differ by $\epsilon$ in $L_2$-norm), such that each $f \in \mathcal{F}$ is sandwiched between one of such brackets. Suppose that the class $\mathcal{F}$ is parametrized by $\theta$ in some $D$-dimensional space $\Theta$. For such parametric problems, it is reasonable to expect that the bracketing number of $\mathcal{F}$ behaves similarly to the covering number of $\Theta$ as $\left(\frac{1}{\epsilon}\right)^{O(D)}$ (see, for instance, the discussion on [75, p. 122]). Such bounds for Gaussian mixtures were obtained in [58]. For example, for $d$-dimensional $k$-GMs, [58, Proposition B.4] yields the following bound for the global bracketing entropy:

$$\log N_{[]}(\epsilon) \lesssim kd \log \frac{Cd}{\epsilon}. \tag{6.2}$$

Using standard result based on bracketing entropy integral (c.f. e.g. [75, Theorem 7.4]), this result leads to the following high-probability bound for the MLE $\hat{\Gamma}_{\mathrm{ML}}$:

$$H(P_{\hat{\Gamma}_{\mathrm{ML}}}, P_{\Gamma}) \leq C\sqrt{\frac{dk \log(dn)}{n}}, \tag{6.3}$$

which has the correct dependency on $k$, but is suboptimal by a logarithmic factor compared to Theorem 1.2. It is for this reason that we turn to the Le Cam-Birgé estimator, which relies on bounding the local Hellinger entropy without brackets, in proving Theorem 1.2. Obtaining a local version of the bracketing entropy bound in (6.2) and determining the optimality (without the undesirable log factors) of the MLE for high-dimensional GM model remains open.

*Adaptivity.* The rate in Theorem 1.1 is optimal in the worst-case scenario where the centers of the Gaussian mixture can overlap. To go beyond this pessimistic result, in one dimension, [38] showed that when the atoms of $\Gamma$ form $k_0$ well-separated (by a constant) clusters (see [81, Definition 1] for a precise definition), the optimal rate is $n^{-1/(4(k-k_0)+2)}$, interpolating the rate $n^{-1/(4k-2)}$ in the worst case ($k_0 = 1$) and the parametric rate $n^{-1/2}$ in the best case ($k_0 = k$). Furthermore, this can be achieved adaptively by either the minimum distance estimator [38, Theorem 3.3] or the DMM algorithm [81, Theorem 2].

In high dimensions, it is unclear how to extend the adaptive framework in [38]. For the procedure considered in Section 3, by Lemma 3.5, the projection $\hat{V}$ obtained from PCA preserves the separation of the atoms of $\Gamma$. Therefore, in the special case of $k = 2$, if we first center the data so that the projection $\gamma$ in (3.2) is one-dimensional, then the adaptive guarantee of the DMM algorithm allows us to adapt to the clustering structure of the original high-dimensional mixture; however, if $k > 2$, Algorithm 1 must be invoked to learn the multivariate $\gamma$, and it does not seem possible to obtain an adaptive version of Lemma 3.2, since some of the projections may have poor separation, e.g. when all the atoms are aligned with the first coordinate vector.

## SUPPLEMENTARY MATERIAL

**Supplement to "Optimal estimation of high-dimensional Gaussian location mixtures" (DOI: ).**

In this supplemental material, we present the proofs of several remaining technical lemmas.

**Funding.** Y. Wu is supported in part by the NSF Grants CCF-1900507, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. P. Yang was supported in part by the National Science Foundation of China (NSFC) Grant 12101353. H. H. Zhou was supported in part by the National Science Foundation (NSF) Grant DMS 2112918.

## REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation.

[2] ACHARYA, J., JAFARPOUR, A., ORLITKSY, A. and SURESH, A. T. (2014). Near-Optimal Sample Estimators for Spherical Gaussian Mixtures. In *Neural Information Processing Systems* 1395–1403. Curran Associates, Inc.

[3] ACHLIOPTAS, D. and MCSHERRY, F. (2005). On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory* 458–469. Springer.

[4] ANANDKUMAR, A., HSU, D. J. and KAKADE, S. M. (2012). A Method of Moments for Mixture Models and Hidden Markov Models. In *Conference on Learning Theory*.

[5] ANDERSEN, M. S., DAHL, J. and VANDENBERGHE, L. (2013). CVXOPT: A Python package for convex optimization.

[6] ARORA, S. and KANNAN, R. (2005). Learning mixtures of separated nonspherical Gaussians. *Annals of Applied Probability* **15** 69-92.

[7] ASHTIANI, H., BEN-DAVID, S., HARVEY, N. J. A., LIAW, C., MEHRABIAN, A. and PLAN, Y. (2018). Near-optimal Sample Complexity Bounds for Robust Learning of Gaussian Mixtures via Compression Schemes. *Neural Information Processing Systems*.

[8] BANACH, S. (1938). Über homogene Polynome in ($L^2$). *Studia Mathematica* **7** 36-44.

[9] BANDEIRA, A. S., RIGOLLET, P. and WEED, J. (2017). Optimal rates of estimation for multi-reference alignment. *arXiv preprint arXiv:1702.08546*.

[10] BAYRAKTAR, E. and GUO, G. (2021). Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability* **26** 1 – 13. https://doi.org/10.1214/21-ECP383

[11] BELKIN, M. and SINHA, K. (2010). Toward learning Gaussian mixtures with arbitrary separation. In *23rd Annual Conference on Learning Theory (COLT)* 407–419.

[12] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. für Wahrscheinlichkeitstheorie und Verw. Geb.* **65** 181–237.

[13] BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability theory and related fields* **71** 271–291.

[14] BONTEMPS, D. and GADAT, S. (2014). Bayesian methods for the Shape Invariant Model. *Electron. J. Statist.* **8** 1522–1568. https://doi.org/10.1214/14-EJS933

[15] CHEN, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics* **23** 221-233.

[16] COMON, P., GOLUB, G., LIM, L.-H. and MOURRAIN, B. (2008). Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications* **30** 1254-1279.

[17] DACUNHA-CASTELLE, D. and GASSIAT, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3** 279–299.

[18] DASGUPTA, S. (1999). Learning Mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science* 634. IEEE Computer Society, USA.

[19] DAVIS, C. and KAHAN, W. M. (1970). The Rotation of Eigenvectors by a Peturbation. *SIAM Journal of Numerical Analysis* **7**.

[20] DESHPANDE, I., HU, Y.-T., SUN, R., PYRROS, A., SIDDIQUI, N., KOYEJO, S., ZHAO, Z., FORSYTH, D. and SCHWING, A. (2019). Max-Sliced Wasserstein Distance and its use for GANs. *arXiv:1904.05877*.

[21] DIAMOND, S. and BOYD, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* **17** 1-5.

[22] DOSS, N., WU, Y., YANG, P. and ZHOU, H. H. (2021). Supplement to "Optimal estimation of high-dimensional Gaussian location mixtures".

[23] FELDMAN, J., SERVEDIO, R. A. and O'DONNELL, R. (2006). PAC Learning Axis-Aligned Mixtures of Gaussians with No Separation Assumption. In *Learning Theory* (G. LUGOSI and H. U. SIMON, eds.) 20-34. Springer Berlin Heidelberg, Berlin, Heidelberg.

[24] FLAMARY, R. and COURTY, N. (2017). POT: Python Optimal Transport library.

[25] FRIEDLAND, S. and LIM, L.-H. (2018). Nuclear norm of higher-order tensors. *Mathematics of Computation* **87** 1255–1281.

[26] GASSIAT, E. and VAN HANDEL, R. (2012). Consistent order estimation and minimal penalties. *IEEE Transactions on Information Theory* **59** 1115–1128.
[27] GASSIAT, E. and VAN HANDEL, R. (2014). The local geometry of finite mixtures. *Transactions of the American Mathematical Society* **366** 1047–1072.
[28] GENOVESE, C. R. and WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Annals of Statistics* **28** 1105–1127.
[29] GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27** 143 – 158. https://doi.org/10.1214/aos/1018031105
[30] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500 – 531. https://doi.org/10.1214/aos/1016218228
[31] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35** 697 – 723. https://doi.org/10.1214/009053606000001271
[32] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics* 1233–1263.
[33] GUHA, A., HO, N. and NGUYEN, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* **27** 2159 – 2188. https://doi.org/10.3150/20-BEJ1275
[34] HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.
[35] HARDT, M. and PRICE, E. (2015). Tight Bounds for Learning a Mixture of Two Gaussians. In *Proceedings of the forty-seventh annual ACM symposium on theory of computing* 753-760.
[36] HARDT, M. and PRICE, E. (2015). Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing* 753–760. ACM.
[37] HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer* **2** 807-810.
[38] HEINRICH, P. and KAHN, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics* **46** 2844–2870.
[39] HO, N. and NGUYEN, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics* **44** 2726–2755. https://doi.org/10.1214/16-AOS1444
[40] HO, N. and NGUYEN, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics* **10** 271-307.
[41] HOPKINS, S. B. and LI, J. (2018). Mixture Models, Robustness, and Sum of Squares Proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. *STOC 2018* 1021–1034. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3188745.3188748
[42] HSU, D. and KAKADE, S. M. (2013). Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. *Fourth Innovations in Theoretical Computer Science*.
[43] IBRAGIMOV, I. (2001). *Estimation of analytic functions*. In *State of the art in probability and statistics*. *Lecture Notes–Monograph Series* **36** 359–383. Institute of Mathematical Statistics, Beachwood, OH. https://doi.org/10.1214/lnms/1215090078
[44] JIN, C., ZHANG, Y., BALAKRISHNAN, S., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. *Advances in neural information processing systems* **29**.
[45] KALAI, A. T., MOITRA, A. and VALIANT, G. (2010). Efficiently learning mixtures of two Gaussians. In *Proceedings of the forty-second ACM symposium on theory of computing* 553-562.
[46] KANNAN, R., SALMASIAN, H. and VEMPALA, S. (2005). The spectral method for general mixture models. In *18th Annual Conference on Learning Theory (COLT)* 444–457.
[47] KIM, A. K. H. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli* **20** 1802-1818.
[48] KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association* **109** 674–685.
[49] KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.
[50] KRUSKAL, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications* **18** 95–138.
[51] LE CAM, L. (1973). Convergence of Estimates Under Dimensionality Restrictions. *The Annals of Statistics* **1** 38 – 53.
[52] LI, J. and SCHMIDT, L. (2017). Robust and Proper Learning for Mixtures of Gaussians via Systems of Polynomial Inequalities. In *Proceedings of the 2017 Conference on Learning Theory* (S. KALE and O. SHAMIR, eds.). *Proceedings of Machine Learning Research* **65** 1302–1382. PMLR, Amsterdam, Netherlands.
[53] LI, P. and CHEN, J. (2010). Testing the Order of a Finite Mixture. *Journal of the American Statistical Association* **105**.

[54] LINDSAY, B. G. (1989). Moment matrices: applications in mixtures. *Annals of Statistics* **17** 722-740.

[55] LIU, X. and SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics* **31** 807-832.

[56] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2019). Optimality of Spectral Clustering for Gaussian Mixture Model. *to appear in The Annals of Statistics*. arXiv:1911.00538.

[57] LUGOSI, G. and MENDELSON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics* **47** 783–794.

[58] MAUGIS, C. and MICHEL, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: Probability and Statistics* **15** 41–68.

[59] MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman-Yor Process Mixtures for the Number of Components. *Journal of Machine Learning Research* **15** 3333-3370.

[60] MOITRA, A. and VALIANT, G. (2010). Settling the Polynomial Learnability of Mixtures of Gaussians. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science* 93-102.

[61] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **41** 370–400.

[62] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **41** 370 – 400. https://doi.org/10.1214/12-AOS1065

[63] NILES-WEED, J. and RIGOLLET, P. (2019). Estimation of Wasserstein distances in the Spiked Transport Model. *arXiv:1909.07513*.

[64] PATY, F.-P. and CUTURI, M. (2019). Subspace Robust Wasserstein Distances. In *Proceedings of the 36th International Conference on Machine Learning* **97** 5072–5081. PMLR.

[65] POLYANSKIY, Y. and WU, Y. (2015). Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory* **62** 35–55.

[66] QI, L. (2011). The best rank-one approximation ratio of a tensor space. *SIAM journal on matrix analysis and applications* **32** 430-442.

[67] RABIN, J., PEYRÉ, G., DELON, J. and BERNOT, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision* 435–446. Springer.

[68] ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 689-710. https://doi.org/10.1111/j.1467-9868.2011.00781.x

[69] RUDELSON, M. and VERSHYNIN, R. (2009). Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics* **62** 1707–1739.

[70] SAHA, S. and GUNTUBOYINA, A. (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *Annals of Statistics* **48** 738–762.

[71] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623?640. https://doi.org/10.1093/biomet/ast015

[72] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics* **29** 687 – 714. https://doi.org/10.1214/aos/1009210686

[73] SHOHAT, J. A. and TAMARKIN, J. D. (1943). *The problem of moments* **1**. American Mathematical Soc.

[74] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY.

[75] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[76] VAN DER VAART, A. (2002). The statistical work of Lucien Le Cam. *The Annals of Statistics* 631–682.

[77] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag New York, Inc.

[78] VEMPALA, S. and WANG, G. (2004). A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci* 2004.

[79] VILLANI, C. (2003). *Topics in optimal transportation*. American Mathematical Society, Providence, RI.

[80] WU, Y. (2017). Lecture notes on information-theoretic methods for high-dimensional statistics. http://www.stat.yale.edu/~yw562/teaching/598/it-stats.pdf.

[81] WU, Y. and YANG, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics* **48** 1981–2007.

[82] WU, Y. and ZHOU, H. H. (2021). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning* **4** 143–220. arxiv preprint arXiv:1908.10935.

[83] ZHANG, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica* **19** 1297-1318.