Submitted to the Annals of Statistics

OPTIMAL ESTIMATION OF GAUSSIAN MIXTURES VIA DENOISED METHOD OF MOMENTS

By Yihong Wu[*] and Pengkun Yang[†]

Yale University[*] *and Princeton University*[†]

The Method of Moments [46] is one of the most widely used methods in statistics for parameter estimation, by means of solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or non-uniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analyses of the method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

This paper considers estimating a k-component Gaussian location mixture with a common (possibly unknown) variance parameter. To overcome the aforementioned theoretic and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space (via semidefinite programming) before solving the method of moments equations. Not only does this regularization ensures existence and uniqueness of solutions, it also yields fast solvers by means of Gauss quadrature. Furthermore, by proving new moment comparison theorems in the Wasserstein distance via polynomial interpolation and majorization techniques, we establish the statistical guarantees and adaptive optimality of the proposed procedure, as well as oracle inequality in misspecified models. These results can also be viewed as provable algorithms for Generalized Method of Moments [21] which involves non-convex optimization and lacks theoretical guarantees.

1. Introduction.

1.1. *Gaussian mixture model.* Consider a k-component Gaussian location mixture model, where each observation is distributed as

$$X \sim \sum_{i=1}^{k} w_i N(\mu_i, \sigma^2). \tag{1}$$

MSC 2010 subject classifications: Primary 62G05; secondary 62C20

Keywords and phrases: Gaussian mixture, finite mixture model, deconvolution, method of moments, Wasserstein distance, semidefinite programming, moment space, Gauss quadrature, minimax optimality

Here w_i is the mixing weight such that $w_i \geq 0$ and $\sum_i w_i = 1$, μ_i is the mean (center) of the i^{th} component, and σ is the common standard deviation. Equivalently, we can write the distribution of an observation X as a convolution

$$X \sim \nu * N(0, \sigma^2), \tag{2}$$

where $\nu = \sum_{i=1}^{k} w_i \delta_{\mu_i}$ denotes the *mixing distribution*. Thus, we can write $X = U + \sigma Z$, where $U \sim \nu$ is referred to as the latent variable, and Z is standard normal and independent of U.

Generally speaking, there are three formulations of learning mixture models:

- **Parameter estimation**: estimate the means μ_i's and the weights w_i's up to a global permutation, and possibly also σ^2.
- **Density estimation**: estimate the probability density function of the Gaussian mixture under certain loss such as L_2 or Hellinger distance. This task is further divided into the cases of *proper* and *improper* learning, depending on whether the estimator is required to be a k-Gaussian mixture or not; in the latter case, there is more flexibility in designing the estimator but less interpretability.
- **Clustering**: estimate the latent variable of each sample (i.e. U_i, if the ith sample is represented as $X_i = U_i + \sigma Z_i$) with a small misclassification rate.

It is clear that clustering necessarily relies on the separation between the clusters; however, as far as estimation is concerned, both parametric and non-parametric, no separation condition should be needed and one can obtain accurate estimates of the parameters even when clustering is impossible. Furthermore, one should be able to learn from the data the order of the mixture model, that is, the number of components. However, in the present literature, most of the estimation procedures with finite sample guarantees are either clustering-based, or rely on separation conditions in the analysis (e.g. [4, 42, 24]). Bridging this conceptual divide is one of the main motivations of the present paper.

Existing methodologies for mixture models are largely divided into likelihood-based and moment-based methods; see Section 1.5 for a detailed review. Among likelihood-based methods, the *Maximum Likelihood Estimate* (MLE) is not efficiently computable due to the non-convexity of the likelihood function. The most popular heuristic procedure to approximate the MLE is the *Expectation-Maximization* (EM) algorithm [11]; however, absent separation conditions, no theoretical guarantee is known in general. Moment-based methods include the classical *method of moments* [46] and many extensions [21, 2]; however, the usual method of moments suffers from many issues as elaborated in the next subsection. In the theoretical computer science literature, [27, 44, 22] proposed moment-based polynomial-time algorithms with provable guarantees; however, these methods are typically based on grid search and far from being practical. Finding theoretically sound, numerically stable, and computationally efficient version of the method of moments is a major objective of this paper.

1.2. *Failure of the classical method of moments.* The method of moments, commonly attributed to Pearson [46], produces an estimator by equating the population moments to the sample moments. While conceptually simple, this method suffers from the following problems, especially in the context of mixture models:

- *Solvability*: the method of moments entails solving a multivariate polynomial system, in which one frequently encounters non-existence or non-uniqueness of statistically meaningful solutions.
- *Computation*: solving moment equations can be computationally intensive. For instance, for k-component Gaussian mixture models, the system of moment equations consist of $2k-1$ polynomial equations with $2k-1$ variables.
- *Accuracy*: existing statistical literature on the method of moments [54, 21] either shows mere consistency under weak assumptions, or proves asymptotic normality assuming very strong regularity conditions (so that the delta method works), which generally do not hold in mixture models since the convergence rates can be slower than parametric. Some results on nonparametric rates are known (cf. [54, Theorem 5.52] and [34, Theorem 14.4]) but the conditions are extremely hard to verify.

To explain the failure of the vanilla method of moments in Gaussian mixture models, we analyze the following simple two-component example:

Example 1. Consider a Gaussian mixture model with two unit variance components: $X \sim w_1 N(\mu_1, 1) + w_2 N(\mu_2, 1)$. Since there are three parameters μ_1, μ_2 and $w_1 = 1 - w_2$, we use the first three moments and solve the following system of equations:

$$
\begin{aligned}
\mathbb{E}_n[X] &= \mathbb{E}[X] = w_1\mu_1 + w_2\mu_2, \\
\mathbb{E}_n[X^2] &= \mathbb{E}[X^2] = w_1\mu_1^2 + w_2\mu_2^2 + 1, \\
\mathbb{E}_n[X^3] &= \mathbb{E}[X^3] = w_1\mu_1^3 + w_2\mu_2^3 + 3(w_1\mu_1 + w_2\mu_2),
\end{aligned}
\tag{3}
$$

where $\mathbb{E}_n[X^i] \triangleq \frac{1}{n}\sum_{j=1}^n X_j^i$ denotes the i^{th} moment of the empirical distribution from n i.i.d. samples. The right-hand sides of (3) are related to the moments of the mixing distribution by a linear transformation, which allow us to equivalently rewrite the moment equations (3) as:

$$\begin{aligned} \mathbb{E}_n[X] &= \mathbb{E}[U] = w_1\mu_1 + w_2\mu_2, \\ \mathbb{E}_n[X^2 - 1] &= \mathbb{E}[U^2] = w_1\mu_1^2 + w_2\mu_2^2, \\ \mathbb{E}_n[X^3 - 3X] &= \mathbb{E}[U^3] = w_1\mu_1^3 + w_2\mu_2^3, \end{aligned} \tag{4}$$

where $U \sim w_1\delta_{\mu_1} + w_1\delta_{\mu_2}$. It turns out that with finitely many samples, there is always a non-zero chance that (4) has no solution; even with infinite samples, it is possible that the solution does not exist with constant probability. To see this, note that, from the first two equations of (4), the solution does not exist whenever

$$\mathbb{E}_n[X^2] - 1 < \mathbb{E}_n^2[X], \tag{5}$$

that is, the Cauchy-Schwarz inequality fails. Consider the case $\mu_1 = \mu_2 = 0$, i.e., $X \sim N(0,1)$. Then (5) is equivalent to

$$n(\mathbb{E}_n[X^2] - \mathbb{E}_n^2[X]) \le n,$$

where the left-hand side follows the χ^2-distribution with $n-1$ degrees of freedom. Thus, (5) occurs with probability approaching $\frac{1}{2}$ as n diverges, according to the central limit theorem.

In view of the above example, we note that the main issue with the classical method of moments is the following: although individually each moment estimate is accurate ($\sqrt{n}$-consistent), jointly they do not correspond to the moments of any distribution. Moment vectors satisfy many geometric constraints, e.g., the Cauchy-Schwarz and Hölder inequalities, and lie in a convex set known as the *moment space*. Thus for any model parameters, with finitely many samples the method of moments fails with nonzero probability whenever the noisy estimates escape the moment space; even with infinitely many samples, it also provably happens with constant probability when the order of the mixture model is strictly less than k, or equivalently, the population moments lie on the boundary of the moment space (see Lemma 39 of the supplement [57] for a justification).

1.3. *Main results.* In this paper, we propose the *denoised method of moments* (DMM), which consists of three main steps: (1) compute noisy estimates of moments, e.g., the unbiased estimates; (2) jointly denoise the moment estimates by projecting them onto the moment space; (3) execute the usual method of moments. It turns out that the extra step of projection resolves the three issues of the vanilla version of the method of moments identified in Section 1.2 simultaneously:

- *Solvability*: a unique statistically meaningful solution is guaranteed to exist by the classical theory of moments;
- *Computation*: the solution can be found through an efficient algorithm (Gauss quadrature) instead of invoking generic solvers of polynomial systems;
- *Accuracy*: the solution provably achieves the optimal rate of convergence, and automatically adaptive to the clustering structure of the population.

We emphasize that the denoising (projection) step is explicitly carried out via a convex optimization in Section 4.1, and implicitly used in analyzing Lindsay's algorithm [40] in Section 4.2, when the variance parameter is known and unknown, respectively.

Following the framework proposed in [7, 23], in this paper we consider the estimation of the mixing distribution, rather than estimating the parameters of each component. The main benefits of this formulation include the following:

- Assumption-free: to recover individual components it is necessary to impose certain assumptions to ensure identifiability, such as lower bounds on the mixing weights and separations between components, none of which is needed for estimating the mixing distribution. Furthermore, under the usual assumption such as separation conditions, statistical guarantees on estimating the mixing distribution can be naturally translated to those for estimating the individual parameters.
- Inference on the number of components: this formulation allows us to deal with misspecified models and estimate the order of the mixture model.

Equivalently, estimating the mixing distribution can be viewed as a deconvolution problem, where the goal is to recover the distribution ν based on observations drawn from the convolution (2).

In this framework, a meaningful and flexible loss function for estimating the mixing distribution is the 1-*Wasserstein distance* (see Section 1.4 for a justification in the context of mixture models), defined by

$$W_1(\nu,\nu') \triangleq \inf\{\mathbb{E}[\|X-Y\|] : X\sim\nu, Y\sim\nu'\}, \tag{6}$$

where the infimum is taken over all couplings, i.e., joint distributions of X and Y which are marginally distributed as ν and ν' respectively. In one dimension, the W_1 distance coincides with the L_1-distance between the cumulative distribution functions (CDFs) [55].

Next we present the theoretical results, which can be classified into two categories:

- To estimate the mixing distribution, our methodology produces moment-based estimators that are optimal in both worst-case (Theorem 1) and adaptive sense (Theorem 2), for both known and unknown σ.
- To estimate the mixture density, the same procedure produces a *proper* estimate that attains the optimal parametric rate (Theorem 3), despite the fact that the mixing distribution can only be estimated at a non-parametric rate. Moreover, the procedure is robust to model misspecification (Theorem 4).

Throughout the paper, we assume that the number of components satisfies

$$k = O\left(\frac{\log n}{\log\log n}\right). \tag{7}$$

If the order of mixture is large, namely, $k \geq \Omega(\frac{\log n}{\log\log n})$, including continuous mixtures, then one can approximate it by a finite mixture with $O(\frac{\log n}{\log\log n})$ components and estimate the mixing distribution using the DMM estimator. Furthermore, this method is optimal (see Theorem 5 at the end of this subsection). Our main result is the following theorem:

THEOREM 1 (Optimal rates). *Suppose that $|\mu_i| \leq M$ for $M \geq 1$ and σ is bounded by a constant, and both k and M are given.*

- *If σ is known, then there exists an estimator $\hat{\nu}$ computable in $O(kn)$ time such that, with probability at least $1-\delta$,*

$$W_1(\nu,\hat{\nu}) \leq O\left(Mk^{1.5}\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{4k-2}}\right). \tag{8}$$

- *If σ is unknown, then there exists an estimator $(\hat{\nu},\hat{\sigma})$ computable in $O(kn)$ time such that, with probability at least $1-\delta$,*

$$W_1(\nu,\hat{\nu}) \leq O\left(Mk^{2}\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{4k}}\right), \tag{9}$$

and

$$|\sigma^2-\hat{\sigma}^2| \leq O\left(M^2k\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{2k}}\right). \tag{10}$$

For fixed for constant k, the above convergence rates are minimax optimal as shown in Section 6 of the supplement [57]; in the case of known σ, the optimality of (8) has been previously shown in [23], while the matching lower bounds for (9)–(10) are new.

Note that the results in Theorem 1 are proved under the worst-case scenario where the centers can be arbitrarily close, e.g., components completely overlap. It is reasonable to expect a faster convergence rate when the components are better separated, and, in fact, a parametric rate in the best-case scenario where the components are fully separated and weights are bounded away from zero. To capture the clustering structure of the mixture model, we introduce the following definition:

Definition 1. The Gaussian mixture (1) has k_0 (γ, ω)-separated clusters if there exists a partition $S_1, \ldots, S_{k_0}$ of $[k]$ such that

- $|\mu_i - \mu_{i'}| \geq \gamma$ for any $i \in S_\ell$ and $i' \in S_{\ell'}$ such that $\ell \neq \ell'$;
- $\sum_{i \in S_\ell} w_i \geq \omega$ for each ℓ.

In the absence of the minimal weight condition (i.e. $\omega = 0$), we say the Gaussian mixture has k_0 γ-separated clusters.

The next result shows that the DMM estimators attain the following adaptive rates:

Theorem 2 (Adaptive rate). *Under the conditions of Theorem 1, suppose there are k_0 (γ, ω)-separated clusters such that $\gamma\omega \geq C\epsilon$ for some absolute constant $C > 2$, where ϵ denotes the right-hand side of (8) and (9) when σ is known and unknown, respectively.*

- *If σ is known, then, with probability at least $1 - \delta$,*[1]

$$W_1(\nu, \hat{\nu}) \leq O_k\left(M\gamma^{-\frac{2k_0-2}{2(k-k_0)+1}} \left(\frac{n}{\log(k/\delta)} \right)^{-\frac{1}{4(k-k_0)+2}} \right). \tag{11}$$

- *If σ is unknown, then, with probability at least $1 - \delta$,*[2]

$$\sqrt{|\sigma^2 - \hat{\sigma}^2|},\ W_1(\nu, \hat{\nu}) \leq O_k\left(M\gamma^{-\frac{k_0-1}{k-k_0+1}} \left(\frac{n}{\log(k/\delta)} \right)^{-\frac{1}{4(k-k_0+1)}} \right). \tag{12}$$

[1] Here $O_k(\cdot)$ denotes a constant factor that depends on k only.

[2] Note that the estimation rate for the mean part ν is the square root of the rate for estimating the variance parameter σ^2. Intuitively, this phenomenon is due to the infinite divisibility of the Gaussian distribution: note that for the location mixture model $\nu * N(0, \sigma^2)$ with $\nu \sim N(0, \epsilon^2)$ and $\sigma^2 = 1$ has the same distribution as that of $\nu \sim \delta_0$ and $\sigma^2 = 1 + \epsilon^2$.

For fixed k, k_0 and γ, the rate in (11) is minimax optimal in view of the lower bounds in [23]; we also provide a simple proof in [57, Remark 4] by extending the lower bound argument in Section 6 of the supplement [57]. If σ is unknown, we do not have a matching lower bound for (12). In fact, in the fully-separated case ($k_0 = k$), (12) reduces to $n^{-\frac{1}{4}}$ while the parametric rate is clearly achievable. Let us emphasize that, for known σ, the rates (8) and (11) for fixed k, k_0 and γ have been previously obtained in [23] by means of the computationally expensive minimum distance estimator; for unknown σ, the results in (9), (10), and (12) are new.

Next we discuss the implication on density estimation (*proper* learning), where the goal is to estimate the density function of the Gaussian mixture by another k-Gaussian mixture density. Given that the estimated mixing distribution $\hat{\nu}$ from Theorem 1, a natural density estimate is the convolution $\hat{f} = \hat{\nu} * N(0, \sigma^2)$. Theorem 3 below shows that the density estimate $\hat{f}$ is $O(\frac{1}{\sqrt{n}})$-close to the true density f in the total variation distance $\mathsf{TV}(f, g) \triangleq \frac{1}{2}\|f - g\|_1$.

THEOREM 3 (Density estimation). *Under the conditions of Theorem 1, denote the density of the underlying model by $f = \nu * N(0, \sigma^2)$. If σ is given, then there exists an estimate $\hat{f}$ such that*

$$\mathsf{TV}(\hat{f}, f) \le O_k(\sqrt{\log(1/\delta)/n}),$$

with probability $1 - \delta$.

So far we have been focusing on well-specified models. In the case of misspecified models, the data need not be generated from a k-Gaussian mixture. In this case, the DMM procedure still reports a meaningful estimate that is close to the best k-Gaussian mixture fit of the unknown distribution. This is made precise by the next result of oracle inequality type. Analogous results hold for χ^2-divergence, Kullback-Leibler divergence, and Hellinger distance as well.

THEOREM 4 (Misspecified model). *Assume that $X_1, \ldots, X_n$ is independently drawn from a density f which is 1-subgaussian. Suppose there exists a k-component Gaussian location mixture g with a given variance σ^2 such that $\mathsf{TV}(f, g) \le \epsilon$. Then, there exists an estimate $\hat{f}$ such that*

$$\mathsf{TV}(\hat{f}, f) \le O_k\left(\epsilon\sqrt{\log(1/\epsilon)} + \sqrt{\log(1/\delta)/n}\right),$$

with probability $1 - \delta$.

To conclude this subsection, we present a result for estimating mixtures of an arbitrarily large order, including continuous mixtures, in the case of known variance. In this situation we apply the DMM method to produce a mixture of order $\min\{k, O(\frac{\log n}{\log\log n})\}$. The convergence rate is minimax optimal in view of the matching lower bound in Proposition 9 of the supplement [57].

THEOREM 5 (Higher-order mixture). *Suppose $|\mu_i| \leq M$ for $M \geq 1$ and σ is a bounded constant, where M, σ are given. Then there exists an estimate $\hat{\nu}$ such that, with probability at least $1-\delta$,*

$$W_1(\nu, \hat{\nu}) \leq O\left(M\left(\frac{\log\log n}{\log n} + \sqrt{\frac{\log(1/\delta)}{n^{1-c}}}\right)\right),$$

for some constant $c < 1$.

1.4. *Why Wasserstein distance?.* Throughout the paper we consider estimating the mixing distribution ν with respect to the Wasserstein distance. This is a natural criterion, which is not too stringent to yield trivial result (such as the Kolmogorov-Smirnov (KS) distance[3]) and, at the same time, strong enough to provide meaningful guarantees on the means and weights. In fact, the commonly used criterion $\min_\Pi \sum_i |\mu_i - \hat{\mu}_{\Pi(i)}|$ over all permutations Π is precisely (k times) the Wasserstein distance between two equally weighted distributions [55].

Furthermore, we can obtain statistical guarantees on the support sets and weights of the estimated mixing distribution under the usual assumptions in literature [8, 27, 22] that include separation between the means and lower bound on the weights. See Section 2.2 for a detailed discussion. We highlight the following result, phrased in terms of the parameter estimation error up to a permutation:

LEMMA 1. *Let $\nu = \sum_{i=1}^k w_i \delta_{\mu_i}$ and $\hat{\nu} = \sum_{i=1}^k \hat{w}_i \delta_{\hat{\mu}_i}$. Suppose that $W_1(\nu, \hat{\nu}) < \epsilon$. Let $\epsilon_1 = \min\{|\mu_i - \mu_j|, |\hat{\mu}_i - \hat{\mu}_j| : 1 \leq i < j \leq k\}$ and $\epsilon_2 = \min\{w_i, \hat{w}_i : i \in [k]\}$. If $\epsilon < \epsilon_1\epsilon_2/4$, then, there exists a permutation Π such that*

$$\|\mu - \Pi\hat{\mu}\|_\infty < \epsilon/\epsilon_2, \quad \|w - \Pi\hat{w}\|_\infty < 2\epsilon/\epsilon_1,$$

where $\mu = (\mu_1, \ldots, \mu_k)$, $w = (w_1, \ldots, w_k)$ denote the atoms and weights of ν, respectively, and $\hat{\mu}, \hat{w}$ denote those of $\hat{\nu}$,

[3] Consider two mixing distributions δ_0 and δ_ϵ with arbitrarily small ϵ, whose KS distance is always one.

1.5. *Related work.* There exist a vast literature on mixture models, in particular Gaussian mixtures, and the method of moments. For a comprehensive review see [41, 15]. Below we highlight a few existing results that are related to the present paper.

Likelihood-based methods. Maximum likelihood estimation (MLE) is one of the most useful method for parameter estimation. Under strong separation assumptions, MLE is consistent and asymptotically normal [48]; however, those assumptions are difficult to verify, and it is computationally hard to obtain the global maximizer due to the non-convexity of the likelihood function in the location parameters.

Expectation-Maximization (EM) [11] is an iterative algorithm that aims to approximate the MLE. It has been widely applied in Gaussian mixture models [48, 59] and more recently in high-dimensional settings [4]. In general, this method is only guaranteed to converge to a local maximizer of the likelihood function rather than the global MLE. In practice we need to employ heuristic choices of the initialization [29] and stopping criteria [50], as well as possibly data augmentation techniques [43, 47]. Furthermore, its slow convergence rate is widely observed in practice [48, 29]. Global convergence of the EM algorithm is recently analyzed by [58, 9] but only in the special case of two equally weighted components. Additionally, the EM algorithm accesses the entire data set in each iteration, which is particularly expensive for large sample size and high dimensions.

Lastly, we mention the nonparametric maximum likelihood estimation (NPMLE) in mixture models proposed by [30], where the maximization is taken over all mixing distributions which need not be k-atomic. This is an infinite-dimensional convex optimization problem, which has been studied in [36, 39, 41] and more recently in [32] on its computation based on discretization. One of the drawbacks of NPMLE is its lack of interpretability since the solution is a discrete distribution with at most n atoms cf. [32, Theorem 2]. Furthermore, few statistical guarantees in terms of convergence rate are available.

Moment-based methods. The simplest moment-based method is the method of moments (MM) introduced by Pearson [46]. The failure of the vanilla MM described in Section 1.2 has motivated various modifications including, notably, the *Generalized Method of Moments* (GMM) introduced by Hansen [21]. GMM is a widely used methodology for analyzing economic and financial data (cf. [20] for a thorough review). Instead of exactly solving the MM equations, GMM aims to minimize the sum of squared differences between the sample moments and the fitted moments. Despite its nice asymptotic properties [21], GMM involves a non-convex optimization problem which is computationally challenging to solve. In practice, heuristics such as gradient descent are used [6] which converge slowly and lack theoretical guarantees.

For Gaussian mixture models (and more generally finite mixture models), our results can be viewed as a solver for GMM which is provably exact and computationally efficient, improving over existing heuristic methods in terms of both speed and accuracy significantly; this is another algorithmic contribution of the present paper. The key is to switch the view from optimizing over k-atomic mixing distributions (which is non-convex) to moment space (which is convex and efficiently optimizable via SDP). We also note that minimizing the sum of squares in GMM is not crucial and minimizing any distance yields the same theoretical guarantee. We discuss the connections to GMM in details in Section 4.1.

There are a number of recent work in the theoretical computer science literature on provable results for moment-based estimators in Gaussian location-scale mixture models, see, e.g., [44, 27, 5, 22, 38]. For instance, the algorithm [44] is based on exhaustive search over the discretized parameter space such that the population moments is close to the empirical moments. In addition to being computationally expensive, this method achieves the estimation accuracy $n^{-C/k}$ for some constant C, which is suboptimal in view of Theorem 1. By carefully analyzing Pearson's method of moments equations [46], [22] showed that the optimal rate for two-component location-scale mixtures is $\Theta(n^{-1/12})$; however, this approach is difficult to generalize to higher order mixtures. Finally, for moment-based methods in multiple dimensions, such as spectral and tensor decomposition, we defer the discussion to Section 9.2 of the supplement [57].

Minimum distance estimators. In the case of known variance, the minimum distance estimator is studied by [10, 7, 23]. Specifically, the estimator is a k-atomic distribution $\hat{\nu}$ such that $\hat{\nu} * N(0, \sigma^2)$ is the closest to the empirical distribution of the samples in certain distance. The minimax optimal rate $O(n^{-\frac{1}{4k-2}})$ for estimating the mixing distribution under the Wasserstein distance is shown in [23] (which corrects the previous result in [7]), by bounding the W_1 distance between the mixing distributions in terms of the KS distance of the Gaussian mixtures [23, Lemma 4.5]. However, the minimum distance estimator is in general computationally expensive and suffers from the same non-convexity issue of the MLE. In contrast, denoised method of moments is efficiently computable and adaptively achieves the optimal rate of accuracy as given in Theorem 2. For arbitrary Gaussian location mixtures in one dimension, the minimum distance estimator was considered in [14] in the context of empirical Bayes. Under the assumptions of bounded first moment, it is shown in [14, Corollary 2] that the mixing distribution can be estimated at rate $O((\log n)^{-1/4})$ under the L_2-distance between the CDFs; this loss is, however, weaker than the W_1-distance (i.e. L_1 distance between the CDFs).

Density estimation. If the estimator is allowed to be any density (*improper* learning), it is known that as long as the mixing distribution has a bounded support, the rate of convergence is close to parametric regardless of the number of components. Specifically, the optimal squared L_2-risk is found to be $\Theta(\frac{\sqrt{\log n}}{n})$ [31], achieved by the kernel density estimator designed for analytic densities [26]. As mentioned before, *proper* density estimate (which is required to be a k-Gaussian mixture) is more desirable for the sake of interpretability; however, finding the k-Gaussian mixture that best approximates a given function such as a kernel density estimate can be computationally challenging due to, again, the non-convexity in the location parameters. In this regard, another contribution of Theorems 3 and 4 is the observation that proper and near optimal estimates/approximates can be found efficiently via the method of moments. Finally, we note that MLE for estimating the density of general Gaussian mixtures has been studied in [17, 18].

1.6. *Notations.* A discrete distribution supported on k atoms is called a k-atomic distribution. The expectation of a given function f under a distribution μ is denoted by $\mathbb{E}_\mu f = \mathbb{E}_\mu[f(X)] = \int f(x)\mu(\mathrm{d}x)$, and the subscript μ may be omitted if it is specified from the context. The empirical mean of f from n samples is denoted as $\mathbb{E}_n[f(X)] = \frac{1}{n}\sum_{i=1}^n f(X_i)$, where $X_1, \dots, X_n$ are i.i.d. copies of X. The r^{th} moment of a distribution μ is denoted by $m_r(\mu) \triangleq \mathbb{E}_\mu X^r$. The moment matrix associated with $m_0, m_1, \dots, m_{2r}$ is a Hankel matrix of order $r+1$:

$$\mathbf{M}_r = \begin{bmatrix} m_0 & m_1 & \cdots & m_r \\ m_1 & m_2 & \cdots & m_{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_r & m_{r+1} & \cdots & m_{2r} \end{bmatrix}. \tag{13}$$

For matrices $A \succeq B$ stands for $A-B$ being positive semidefinite. The interval $[x-a, x+a]$ is abbreviated as $[x \pm a]$. For any $x, y \in \mathbb{R}$, $x \wedge y \triangleq \min\{x, y\}$ and $(x)_+ \triangleq \max\{x, y\}$. For two vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, let $\langle x, y\rangle \triangleq \sum_i x_i y_i$. A distribution π is called σ-subgaussian if $\mathbb{E}_\pi[e^{tX}] \leq \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$. We use standard big-O notations, e.g., for two positive sequence $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ if $a_n \leq Cb_n$ for some constant $C > 0$; $a_n = \Omega(b_n)$ if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. We write $a_n = O_\beta(b_n)$ if C depends on another parameter β.

1.7. *Organization.* The paper is organized as follows. In Section 2 we provide some basic results of the theory of moments and the Wasserstein distance. In Section 3 we introduce the moment comparison theorems, which bound the Wasserstein distance between two discrete distributions in terms of the discrepancy of their moments. These are key results to prove the main theorems. In Section 4, we propose estimation algorithms and provide their statistical guarantees. We provide a proof of the moment comparison theorems in Section 5 (with two alternative proofs given in [57, Section 10]); in particular, Section 5.1 contains a brief discussion on polynomial interpolation and majorization, which play a crucial role in the proof. Matching minimax lower bounds, numerical experiments and comparison with other methods such as the EM algorithm, and extensions and open problems including location-scale mixtures, the multivariate case, and general finite mixtures are given in the supplemental article [57]. Auxiliary results are collected in [57, Appendix B].

2. Preliminaries.

2.1. *Moment space, SDP characterization, and Gauss quadrature.* The theory of moments plays a key role in the developments of analysis, probability, statistics, and optimization. See the classics [51, 28] and the recent monographs [37, 49] for a detailed treatment. Below, we briefly review a few basic facts that are related to this paper.

The r^{th} moment vector of a distribution π is a r-tuple $\mathbf{m}_r(\pi) = (m_1(\pi), \dots, m_r(\pi))$. The r^{th} moment space on $K \subseteq \mathbb{R}$ is defined as

$$\mathcal{M}_r(K) = \{\mathbf{m}_r(\pi) : \pi \text{ is supported on } K\},$$

which is the convex hull of $\{(x, x^2, \dots, x^r) : x \in K\}$. A valid moment vector satisfies many geometric constraints such as the Cauchy-Schwarz and Hölder inequalities. When $K = [a, b]$ is a compact interval, $\mathcal{M}_r([a, b])$ is completely described by (see [51, Theorem 3.1], and also [28, 37]) the following condition:

$$\begin{cases} \mathbf{M}_{0,r} \succeq 0, \quad (a+b)\mathbf{M}_{1,r-1} \succeq ab\mathbf{M}_{0,r-2} + \mathbf{M}_{2,r}, & r \text{ even}, \\ b\mathbf{M}_{0,r-1} \succeq \mathbf{M}_{1,r} \succeq a\mathbf{M}_{0,r-1}, & r \text{ odd}, \end{cases} \tag{14}$$

where $\mathbf{M}_{i,j}$ denotes the Hankel matrix with entries $m_i, m_{i+1}, \dots, m_j$:

$$\mathbf{M}_{i,j} = \begin{bmatrix} m_i & m_{i+1} & \cdots & m_{\frac{i+j}{2}} \\ m_{i+1} & m_{i+2} & \cdots & m_{\frac{i+j}{2}+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{\frac{i+j}{2}} & m_{\frac{i+j}{2}+1} & \cdots & m_j \end{bmatrix}.$$

EXAMPLE 2 (Moment spaces on $[0,1]$). For the first two moments, $\mathcal{M}_2([0,1])$ is simply described by $m_1 \geq m_2 \geq 0$ and $m_2 \geq m_1^2$. For $r=3$, according to (14), $\mathcal{M}_3([0,1])$ is described by

$$\begin{bmatrix} 1 & m_1 \\ m_1 & m_2 \end{bmatrix} \succeq \begin{bmatrix} m_1 & m_2 \\ m_2 & m_3 \end{bmatrix} \succeq 0.$$

Using Sylvester's criterion (see [25, Theorem 7.2.5]), they are equivalent to

$$\begin{aligned} &0 \leq m_1 \leq 1, \quad m_2 \geq m_3 \geq 0, \\ &m_1 m_3 \geq m_2^2, \quad (1-m_1)(m_2-m_3) \geq (m_1-m_2)^2. \end{aligned}$$

The necessity of the above inequalities is apparent: the first two follow from the support being $[0,1]$, and the last two follow from the Cauchy-Schwarz inequality. It turns out that they are also sufficient.

Moment matrices of discrete distributions satisfy more structural properties. For instances, the moment matrix of a k-atomic distribution of any order is of rank at most k, and is a deterministic function of $\mathbf{m}_{2k-1}$; the number of atoms can be characterized using the determinants of moment matrices (see [53, p. 362] or [40, Theorem 2A]) as follows:

THEOREM 6. *$(m_1,\ldots,m_{2r})$ are the first $2r$ moments of a distribution with exactly r points of support if and only if $\det(\mathbf{M}_{r-1})>0$ and $\det(\mathbf{M}_r)=0$.*

Next we discuss the closely related notion of *Gauss quadrature*, which is a discrete approximation for a given distribution in the sense of moments and plays an important role in the execution of the DMM estimator. Given π supported on an interval $[a,b] \subseteq \mathbb{R}$, a k-point Gauss quadrature is a k-atomic distribution $\pi_k = \sum_{i=1}^k w_i \delta_{x_i}$, also supported on $[a,b]$, such that, for any polynomial P of degree at most $2k-1$,

$$\mathbb{E}_\pi P = \mathbb{E}_{\pi_k} P = \sum_{i=1}^k w_i P(x_i). \tag{15}$$

Gauss quadrature is known to always exist and is uniquely determined by $\mathbf{m}_{2k-1}(\pi)$ (cf. e.g. [52, Section 3.6]), which shows that any valid moment vector of order $2k-1$ can be realized by a unique k-atomic distribution. A basic algorithm to compute Gauss quadrature is Algorithm 1 [19] and many algorithms with improved computational efficiency and numerical stability have been proposed; cf. [16, Chapter 3].

```
Algorithm 1 Quadrature rule
Input: a valid moment vector $(m_1, \dots, m_{2k-1})$.
Output: nodes $x = (x_1, \dots, x_k)$ and weights $w = (w_1, \dots, w_k)$.
  Define the following degree-$k$ polynomial $P$
      $P(x) = \det \begin{bmatrix} 1 & m_1 & \cdots & m_k \\ \vdots & \vdots & \ddots & \vdots \\ m_{k-1} & m_k & \cdots & m_{2k-1} \\ 1 & x & \cdots & x^k \end{bmatrix}.$
  Let the nodes $(x_1, \dots, x_k)$ be the roots of the polynomial $P$.
  Let the weights $w = (w_1, \dots, w_k)$ be
      $w = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_k^{k-1} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_{k-1} \end{bmatrix}.$
```

2.2. *Wasserstein distance.* A central quantity in the theory of optimal transportation, the Wasserstein distance is the minimum cost of mapping one distribution to another. In this paper, we will be mainly concerned with the 1-Wasserstein distance defined in (6), which can be equivalently expressed, through the Kantorovich duality [55], as

$$W_1(\nu, \nu') = \sup\{\mathbb{E}_\nu[\varphi] - \mathbb{E}_{\nu'}[\varphi] : \varphi \text{ is 1-Lipschitz}\}. \tag{16}$$

The optimal coupling in (6) has many equivalent characterization [55] but is often difficult to compute analytically in general. Nevertheless, the situation is especially simple for distributions on the real line, where the quantile coupling is known to be optimal and hence

$$W_1(\nu, \nu') = \int |F_\nu(t) - F_{\nu'}(t)| \mathrm{d}t, \tag{17}$$

where F_ν and $F_{\nu'}$ denote the CDFs of ν and ν', respectively. Both (16) and (17) provide convenient characterizations to bound the Wasserstein distance in Section 3.

As previously mentioned in Section 1.4, two discrete distributions close in the Wasserstein distance have similar support sets and weights. This is made precise by Lemma 2 and 3 next:

Lemma 2. *Suppose ν and ν' are discrete distributions supported on S and S', respectively. Let $\epsilon = \min\{\nu(x) : x \in S\} \wedge \min\{\nu'(x) : x \in S'\}$. Then,*

$$d_H(S, S') \le W_1(\nu, \nu')/\epsilon,$$

where d_H denotes the Hausdorff distance defined as

$$d_H(S,S') = \max\left\{\sup_{x\in S}\inf_{x'\in S'}|x-x'|, \sup_{x'\in S'}\inf_{x\in S}|x-x'|\right\}. \tag{18}$$

Lemma 3. *For any $\delta > 0$,*

$$\nu(x) - \nu'([x\pm\delta]) \le W_1(\nu,\nu')/\delta, \quad \nu'(x) - \nu([x\pm\delta]) \le W_1(\nu,\nu')/\delta.$$

3. Optimal transport and moment comparison theorems.

A discrete distribution with k atoms has $2k-1$ free parameters. Therefore it is reasonable to expect that it can be uniquely determined by its first $2k-1$ moments. Indeed, we have the following simple identifiability results for discrete distributions:

Lemma 4. *Let ν and ν' be distributions on the real line.*

1. *If ν and ν' are both k-atomic, then $\nu = \nu'$ if and only if $\mathbf{m}_{2k-1}(\nu) = \mathbf{m}_{2k-1}(\nu')$.*
2. *If ν is k-atomic, then $\nu = \nu'$ if and only if $\mathbf{m}_{2k}(\nu) = \mathbf{m}_{2k}(\nu')$.*

In the context of statistical estimation, we only have access to samples and noisy estimates of moments. To solve the inverse problems from moments to distributions, our theory relies on the following stable version of the identifiability in Lemma 4, which show that closeness of moments implies closeness of distributions in Wasserstein distance. In the sequel we refer to Propositions 1 and 2 as moment comparison theorems.

Proposition 1. *Let ν and ν' be k-atomic distributions supported on $[-1,1]$. If $|m_i(\nu) - m_i(\nu')| \le \delta$ for $i = 1,\ldots,2k-1$, then*

$$W_1(\nu,\nu') \le O\left(k\delta^{\frac{1}{2k-1}}\right).$$

Proposition 2. *Let ν be a k-atomic distribution supported on $[-1,1]$. If $|m_i(\nu) - m_i(\nu')| \le \delta$ for $i = 1,\ldots,2k$, then*

$$W_1(\nu,\nu') \le O\left(k\delta^{\frac{1}{2k}}\right).$$

Remark 1. The exponents in Proposition 1 and 2 are optimal. To see this, we first note that the number of moments needed for identifiability in Lemma 4 cannot be reduced:

1. Given any $2k$ distinct points, there exist two k-atomic distributions with disjoint support sets but identical first $2k-2$ moments (see Lemma 30 of the supplement [57]).
2. Given any continuous distribution, its k-point Gauss quadrature is k-atomic and have identical first $2k-1$ moments (see Section 2.1).

By the first observation, there exist two k-atomic distributions ν and ν' such that

$$m_i(\nu) = m_i(\nu'),\ i = 1, \ldots, 2k-2, \quad |m_{2k-1}(\nu) - m_{2k-1}(\nu')| = c_k, \quad W_1(\nu, \nu') = d_k,$$

where c_k and d_k are strictly positive constants that depend on k. Let $\tilde{\nu}$ and $\tilde{\nu}'$ denote the distributions of ϵX and $\epsilon X'$ such that $X \sim \nu$ and $X' \sim \nu'$, respectively. Then, we have

$$\max_{i \in [2k-1]} |m_i(\tilde{\nu}) - m_i(\tilde{\nu})| = \epsilon^{2k-1} c_k, \quad W_1(\tilde{\nu}, \tilde{\nu}') = \epsilon d_k.$$

This concludes the tightness of the exponent in Proposition 1. Similarly, the exponent in Proposition 2 is also tight using the second observation.

REMARK 2. Classical moments comparison theorems aim to show convergence of distributions by comparing a *growing* number of moments. For example, Chebyshev's theorem (see [12, Theorem 2]) states if $\mathbf{m}_r(\pi) = \mathbf{m}_r(N(0,1))$, then

$$\sup_{x \in \mathbb{R}} |F_\pi(x) - \Phi(x)| \le \sqrt{\frac{\pi}{2r}},$$

where F_π and Φ denote the CDFs of π and $N(0,1)$, respectively. For two compactly supported distributions, the above estimate can be sharpened to $O(\frac{\log r}{r})$ [35]. In contrast, in the context of estimating finite mixtures we are dealing with discrete mixing distributions, which can be identified by a *fixed* number of moments. However, with finitely many samples, it is impossible to exactly determine the moments, and measuring the error in the KS distance leads to triviality (see Section 1.4). It turns out that W_1-distance is a suitable metric for this purpose, and the closeness of moments does imply the closeness of distribution in the W_1 distance, which is the integrated difference (L_1-distance) between CDFs as opposed the uniform error (L_∞-distance). An upper bound on the W_1 distance is obtained in [33] (see also Lemma 24 of the supplement [57]) involving the differences of the first k moments and a $\Theta(\frac{1}{k})$ term that does not vanish for fixed k. The discrepancy between parameters of two Gaussian mixtures is obtained by comparing moments in [27, 44], which is not applicable for estimating the mixing distribution.

4. Estimators and statistical guarantees. In this section we introduce the DMM estimators and prove the statistical bounds announced in Section 1. To keep the presentation simple, we focus on estimators with expected risk guarantees. To obtain a high-probability bound, one can employ the usual technique of dividing the samples into batches, applying the unbiased moment estimator to each batch and taking the median, then finally executing the DMM method to estimate the mixing distribution.

The estimators considered in this section[4] are evaluated by numerical experiments in comparison with the EM algorithm and a popular implementation of GMM [6]. Overall the performance of moment-based estimators is on par with that of EM, but the running time of significantly shorter especially when the components are poorly separated. Compared to the existing heuristic solver of GMM [6], the DMM estimator (which exactly solves the GMM) is more accurate and achieves a speedup by orders of magnitude. Furthermore, consistent with the theory in Theorem 2, better estimation accuracy is achieved when the components are more separated. Due to page limit, the details are reported in Section 8 of the supplement [57].

4.1. *Known variance.* The denoised method of moments for estimating Gaussian location mixture models (2) with known variance parameter σ^2 consists of three main steps:

1. estimate $\mathbf{m}_{2k-1}(\nu)$ by $\tilde{m} = (\tilde{m}_1, \dots, \tilde{m}_{2k-1})$ (using Hermite polynomials);
2. denoise $\tilde{m}$ by its projection $\hat{m}$ onto the moment space (semidefinite programming);
3. find a k-atomic distribution $\hat{\nu}$ such that $\mathbf{m}_{2k-1}(\hat{\nu}) = \hat{m}$ (Gauss quadrature).

The complete algorithm is summarized in Algorithm 2.

We estimate the moments of the mixing distribution in lines 1 to 4. The unique unbiased estimators for the polynomials of the mean parameter in a Gaussian location model are Hermite polynomials

$$H_r(x) = r! \sum_{j=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^j}{j!(r-2j)!} x^{r-2j}, \tag{20}$$

such that $\mathbb{E} H_r(X) = \mu^r$ when $X \sim N(\mu, 1)$. Thus, if we define

$$\gamma_r(x, \sigma) = \sigma^r H_r(x/\sigma) = r! \sum_{j=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^j}{j!(r-2j)!} \sigma^{2j} x^{r-2j}, \tag{21}$$

[4] The implementations are available at https://github.com/Albuso0/mixture.

Algorithm 2 Denoised method of moments (DMM) with known variance.

```
Input: $n$ independent samples $X_1, \ldots, X_n$, order $k$, variance $\sigma^2$, interval $I = [a, b]$.
Output: estimated mixing distribution.
for $r = 1$ to $2k - 1$ do
    $\hat{\gamma}_r = \frac{1}{n} \sum_i X_i^r$
    $\tilde{m}_r = r! \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^i}{i!(r-2i)!} \hat{\gamma}_{r-2i} \sigma^{2i}$
end for
Let $\hat{m}$ be the optimal solution of the following:
    (19)    $\min\{\|\tilde{m} - \hat{m}\| : \hat{m} \text{ satisfies } (14)\}$,
where $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_{2k-1})$.
Report the outcome of the Gauss quadrature (Algorithm 1) with input $\hat{m}$.
```

then $\mathbb{E}\gamma_r(X, \sigma) = \mu^r$ when $X \sim N(\mu, \sigma^2)$. Hence, by linearity, $\tilde{m}_r$ is an unbiased estimate of $m_r(\nu)$. The variance of $\tilde{m}_r$ is bounded by the following lemma:

LEMMA 5. *If* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \nu * N(0, \sigma^2)$ *and* ν *is supported on* $[-M, M]$, *then*

$$\mathsf{var}[\tilde{m}_r] \leq \frac{1}{n}(O(M + \sigma\sqrt{r}))^{2r}.$$

As observed in Section 1.2, the major reason for the failure of the usual method of moments is that the unbiased estimate $\tilde{m}$ needs not constitute a legitimate moment sequence, despite the consistency of each individual $\tilde{m}_i$. To resolve this issue, we project $\tilde{m}$ to the moment space using (19). As explained in Section 2.1, (14) consists of positive semidefinite constraints, and thus the optimal solution of (19) can be obtained by semidefinite programming (SDP).[5] In fact, it suffices to solve a *feasibility* program and find any valid moment vector $\hat{m}$ that is within the desired $\frac{1}{\sqrt{n}}$ statistical accuracy.

Now that $\hat{m}$ is indeed a valid moment sequence, we use the Gauss quadrature introduced in Section 2.1 (see Algorithm 1 in Section 2.1) to find the unique k-atomic distribution $\hat{\nu}$ such that $\mathbf{m}_{2k-1}(\hat{\nu}) = \hat{m}$. Using Algorithm 2, $\tilde{m}$ is computed in $O(kn)$ time, the semidefinite programming is solvable in $O(k^{6.5})$ time using the interior-point method (see [56]), and the Gauss quadrature can be evaluated in $O(k^3)$ time [19]. In view of the global assumption (7), Algorithm 2 can be executed in $O(kn)$ time.

We now prove the statistical guarantee (8) for the DMM estimator previously announced in Theorem 1:

[5] The formulation (19) with Euclidean norm can already be implemented in popular modeling languages for convex optimization problem such as CVXPY [13]. A standard form of SDP is given in Appendix A of the supplement [57].

Proof. By scaling it suffices consider $M = 1$. We use Algorithm 2 with Euclidean norm in (19). Using the variance of $\tilde{m}$ in Lemma 5 and Chebyshev inequality yield that, for each $r = 1, \ldots, 2k-1$, with probability $1 - \frac{1}{8k}$,

$$|\tilde{m}_r - m_r(\nu)| \leq \sqrt{k/n}(c\sqrt{r})^r, \tag{22}$$

for some absolute constant c. By the union bound, with probability $3/4$, (22) holds simultaneously for every $r = 1, \ldots, 2k-1$, and thus

$$\|\tilde{m} - \mathbf{m}_{2k-1}(\nu)\|_2 \leq \epsilon, \quad \epsilon \triangleq \frac{(\sqrt{ck})^{2k+1}}{\sqrt{n}}.$$

Since $\mathbf{m}_{2k-1}(\nu)$ satisfies (14) and thus is one feasible solution for (19), we have $\|\tilde{m} - \hat{m}\|_2 \leq \epsilon$. Note that $\hat{m} = \mathbf{m}_{2k-1}(\hat{\nu})$. Hence, by triangle inequality, we obtain the following statistical accuracy:

$$\|\mathbf{m}_{2k-1}(\hat{\nu}) - \mathbf{m}_{2k-1}(\nu)\|_2 \leq \epsilon, \tag{23}$$

Applying Proposition 1 yields that, with probability $3/4$,

$$W_1(\hat{\nu}, \nu) \leq O\left(k^{1.5} n^{-\frac{1}{4k-2}}\right).$$

The confidence $1 - \delta$ in (8) can be obtained by the usual "median trick": divide the samples into $T = \log \frac{2k}{\delta}$ batches, apply Algorithm 2 to each batch of n/T samples, and take $\tilde{m}_r$ to be the median of these estimates. Then Hoeffding's inequality and the union bound imply that, with probability $1 - \delta$,

$$|\tilde{m}_r - m_r(\nu)| \leq \sqrt{\frac{\log(2k/\delta)}{n}}(c\sqrt{r})^r, \quad \forall\, r = 1, \ldots, 2k-1, \tag{24}$$

and the desired (8) follows. □

To conclude this subsection, we discuss the connection to the Generalized Method of Moments (GMM). Instead of solving the moment equations, GMM aims to minimize the difference between estimated and fitted moments:

$$Q(\theta) = (\hat{m} - m(\theta))^\top W(\hat{m} - m(\theta)), \tag{25}$$

where $\hat{m}$ is the estimated moment, θ is the model parameter, and W is a positive semidefinite weighting matrix. The minimizer of $Q(\theta)$ serves as the GMM estimate for the unknown model parameter θ_0. In general the objective function Q is non-convex in θ, notably under the Gaussian mixture model with θ corresponding to the unknown means and weights, which is hard to optimize. Note that (19) with the Euclidean norm is *equivalent* to GMM with the identity weighting matrix. Therefore Algorithm 2 is an exact solver for GMM in the Gaussian location mixture model.

In theory, the optimal weighting matrix W^* that minimizes the asymptotic variance is the inverse of $\lim_{n\to\infty} \mathsf{cov}[\sqrt{n}(\hat{m} - m(\theta_0))]$, which depends the unknown model parameters θ_0. Thus, a popular approach is a two-step estimator [20]:

1. a suboptimal weighting matrix, e.g., identify matrix, is used in the GMM to obtain a consistent estimate of θ_0 and hence a consistent estimate $\hat{W}$ for W^*;
2. θ_0 is re-estimated using the weighting matrix $\hat{W}$.

The above two-step approach can be similarly implemented in the denoised method of moments.

4.2. *Unknown variance.* When the variance parameter σ^2 is unknown, unbiased estimator for the moments of the mixing distribution no longer exists (see Lemma 31 of the supplement [57]). It is not difficult to consistently estimate the variance,[6] then plug into the DMM estimator in Section 4.1 to obtain a consistent estimate of the mixing distribution ν; however, the convergence rate is far from optimal. In fact, to achieve the optimal rate in Theorem 1, it is crucial to simultaneously estimate both the means and the variance parameters. To this end, again we take a moment-based approach. The following result provides a guarantee for any joint estimate of both the mixing distribution and the variance parameter in terms of the moments accuracy.

PROPOSITION 3. *Let*

$$\pi = \nu * N(0,\sigma^2), \quad \hat{\pi} = \hat{\nu} * N(0,\hat{\sigma}^2),$$

where $\nu, \hat{\nu}$ are k-atomic distributions supported on $[-M, M]$, and $\sigma, \hat{\sigma}$ are bounded by a constant. If $|m_r(\pi) - m_r(\hat{\pi})| \le \epsilon$ for $r = 1, \ldots, 2k$, then

$$|\sigma^2 - \hat{\sigma}^2| \le O(M^2 \epsilon^{\frac{1}{k}}), \quad W_1(\nu, \hat{\nu}) \le O(M k^{1.5} \epsilon^{\frac{1}{2k}}).$$

To apply Proposition 3, we can solve the method of moments equations, namely, find a k-atomic distribution $\hat{\nu}$ and $\hat{\sigma}^2$ such that

$$\mathbb{E}_n[X^r] = \mathbb{E}_{\hat{\pi}}[X^r], \qquad r = 1, \ldots, 2k \tag{26}$$

[6] For instance, the simple estimator $\hat{\sigma} = \frac{\max_i X_i}{\sqrt{2\log n}}$ satisfies $|\sigma - \hat{\sigma}| = O_P(\log n)^{-\frac{1}{2}}$.

where $\hat{\pi} = \hat{\mu} * N(0, \hat{\sigma}^2)$ is the fitted Gaussian mixture. Here both the number of equations and the number of variables are equal to $2k$. Suppose (26) has a solution $(\hat{\mu}, \hat{\sigma})$. Then applying Proposition 3 with $\delta = O_k(\frac{1}{\sqrt{n}})$ achieves the rate $O_k(n^{-1/(4k)})$ in Theorem 1, which is minimax optimal (see Section 6 of the supplement [57]). In sharp contrast to the case of known σ, where we have shown in Section 1.2 that the vanilla method of moments equation can have no solution unless we denoise by projection to the moment space, here with one extra scale parameter σ, one can show that (26) has a solution with probability one![7] Furthermore, an efficient method of finding *a* solution to (26) is due to Lindsay [40] and summarized in Algorithm 3. Here, the sample moments can be computed in $O(kn)$ time, and the smallest non-negative root of the polynomial of degree $k(k+1)$ can be found in $O(k^2)$ time using Newton's method (see [3]). So overall Lindsay's estimator can be evaluated in $O(kn)$ time.

Algorithm 3 Lindsay's estimator for normal mixtures with an unknown common variance

```
Input: $n$ samples $X_1, \dots, X_n$.
Output: estimated mixing distribution $\hat{\nu}$, and estimated variance $\hat{\sigma}^2$.
for $r = 1$ to $2k$ do
    $\hat{\gamma}_r = \frac{1}{n}\sum_i X_i^r$
    $\hat{m}_r(\sigma) = r! \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^i}{i!(r-2i)!} \hat{\gamma}_{r-2i}\sigma^{2i}$
end for
Let $\hat{d}_k(\sigma)$ be the determinant of the matrix $\{\hat{m}_{i+j}(\sigma)\}_{i,j=0}^{k}$.
Let $\hat{\sigma}$ be the smallest positive root of $\hat{d}_k(\sigma) = 0$.
for $r = 1$ to $2k$ do
    $\hat{m}_r = \hat{m}_r(\hat{\sigma})$
end for
Let $\hat{\nu}$ be the outcome of the Gauss quadrature (Algorithm 1) with input $\hat{m}_1, \dots, \hat{m}_{2k-1}$
Report $\hat{\nu}$ and $\hat{\sigma}^2$.
```

In [40] the consistency of this estimator was proved under the extra condition that $\hat{\sigma}$ (which is a random variable) as a root of d_k has multiplicity one. It is unclear whether this condition is guaranteed to hold. We will show that, unconditionally, Lindsay's estimator is not only consistent, but in fact achieves the minimax optimal rate (9) and (10) previously announced in

[7] It is possible that the equation (26) has no solution, for instance, when $k = 2, n = 7$ and the empirical distribution is $\pi_7 = \frac{1}{7}\delta_{-\sqrt{7}} + \frac{1}{7}\delta_{\sqrt{7}} + \frac{5}{7}\delta_0$. The first four empirical moments are $\mathbf{m}_4(\pi_7) = (0, 2, 0, 14)$, which cannot be realized by any two-component Gaussian mixture (1). Indeed, suppose $\hat{\pi} = w_1 N(\mu_1, \sigma^2) + (1 - w_1) N(\mu_2, \sigma^2)$ is a solution to (26). Eliminating variables leads to the contradiction that $2\mu_1^4 + 2 = 0$. Assuringly, as we will show later in Lemma 7, such cases occur with probability zero.

Theorem 1. We start by proving that Lindsay's algorithm produces an estimator $\hat{\sigma}$ so that the corresponding the moment estimates lie in the moment space with probability one. In this sense, although no explicit projection is involved, the noisy estimates are *implicitly* denoised.

We first describe the intuition of the choice of $\hat{\sigma}$ in Lindsay's algorithm, i.e., line 6 of Algorithm 3. Let $X \sim \nu * N(0, \sigma^2)$. For any $\sigma' \leq \sigma$, we have

$$\mathbb{E}[\gamma_j(X, \sigma')] = m_j(\nu * N(0, \sigma^2 - \sigma'^2)).$$

Let $d_k(\sigma')$ denote the determinant of the moment matrix $\{\mathbb{E}[\gamma_{i+j}(X, \sigma')]\}_{i,j=0}^k$, which is an even polynomial in σ' of degree $k(k+1)$. According to Theorem 6, $d_k(\sigma') > 0$ when $0 \leq \sigma' < \sigma$ and becomes zero at $\sigma' = \sigma$, and thus σ is characterized by the smallest positive zero of d_k. In lines $5-6$, d_k is estimated by $\hat{d}_k$ using the empirical moments, and σ is estimated by the smallest positive zero of $\hat{d}_k$. We first note that $\hat{d}_k$ indeed has a positive zero:

LEMMA 6. *Assume $n > k$ and the mixture distribution has a density. Then, almost surely, $\hat{d}_k$ has a positive root within $(0, s]$, where $s^2 \triangleq \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}_n[X])^2$ denotes the sample variance.*

The next result shows that, with the above choice of $\hat{\sigma}$, the moment estimates $\hat{m}_j = \mathbb{E}_n[\gamma_j(X, \hat{\sigma})]$ for $j = 1, \ldots, 2k$ given in line 8 are implicitly denoised and lie in the moment space with probability one. Thus (26) has a solution, and the estimated mixing distribution $\hat{\nu}$ can be found by the Gauss quadrature. This result was previously shown in [40] assuming that $\hat{\sigma}$ is of multiplicity one. In contrast, Lemma 7 only requires that $n \geq 2k-1$ and the mixture distribution has a density.

LEMMA 7. *Assume $n \geq 2k-1$ and the mixture distribution has a density. Then, almost surely, there exists a k-atomic distribution $\hat{\nu}$ such that $m_j(\hat{\nu}) = \hat{m}_j$ for $j \leq 2k$, where $\hat{m}_j$ is from Algorithm 3.*

With the above analysis, we now prove the statistical guarantee (9) and (10) for Lindsay's algorithm announced in Theorem 1:

PROOF. It suffices to consider $M = 1$. Let $\hat{\pi} = \hat{\nu} * N(0, \hat{\sigma}^2)$ and $\pi = \nu * N(0, \sigma^2)$ denote the estimated mixture distribution and the ground truth, respectively. Let $\hat{m}_r = \mathbb{E}_n[X^r]$ and $m_r = m_r(\pi)$. The variance of $\hat{m}_r$ is upper bounded by

$$\mathsf{var}[\hat{m}_r] = \frac{1}{n}\mathsf{var}[X_1^r] \leq \frac{1}{n}\mathbb{E}[X^{2r}] \leq \frac{(\sqrt{cr})^{2r}}{n},$$

for some absolute constant c. Using Chebyshev inequality, for each $r = 1, \ldots, 2k$, with probability $1 - \frac{1}{8k}$, we have,

$$|\hat{m}_r - m_r| \leq (\sqrt{cr})^r \sqrt{k/n}. \tag{27}$$

By the union bound, with probability $3/4$, the above holds holds simultaneously for every $r = 1, \ldots, 2k$. It follows from Lemma 6 and 7 that (26) holds with probability one. Therefore,

$$|m_r(\hat{\pi}) - m_r(\pi)| \leq (\sqrt{cr})^r \sqrt{k/n}, \quad r = 1, \ldots, 2k.$$

for some absolute constant c. In the following, the error of variance estimate is denoted by $\tau^2 = |\sigma^2 - \hat{\sigma}^2|$.

- If $\sigma \leq \hat{\sigma}$, let $\nu' = \hat{\nu} * N(0, \tau^2)$. Using $\mathbb{E}_\pi[\gamma_r(X, \sigma)] = m_r(\nu)$ and $\mathbb{E}_{\hat{\pi}}[\gamma_r(X, \sigma)] = m_r(\nu')$, where γ_r is the Hermite polynomial (21), we obtain that (see Lemma 27 of the supplement [57])

  $$|m_r(\nu') - m_r(\nu)| \leq (\sqrt{c'k})^{2k} \sqrt{k/n}, \quad r = 1, \ldots, 2k, \tag{28}$$

  for an absolute constant c'. Applying Proposition 3 yields that,

  $$|\sigma^2 - \hat{\sigma}^2| \leq O(kn^{-\frac{1}{2k}}), \quad W_1(\nu, \hat{\nu}) \leq O(k^2 n^{-\frac{1}{4k}}).$$

- If $\sigma \geq \hat{\sigma}$, let $\nu' = \nu * N(0, \tau^2)$. Similar to (28), we have

  $$|m_r(\hat{\nu}) - m_r(\nu')| \leq (\sqrt{c'k})^{2k} \sqrt{k/n} \triangleq \epsilon, \quad r = 1, \ldots, 2k.$$

  To apply Proposition 3, we also need to ensure that $\hat{\nu}$ has a bounded support, which is not obvious. To circumvent this issue, we apply a truncation argument thanks to the following tail probability bound for $\hat{\nu}$ (see Lemma 16 of the supplement [57]):

  $$\mathbb{P}[|\hat{U}| \geq \sqrt{c_0 k}] \leq \epsilon(\sqrt{c_1 k}/t)^{2k}, \quad \hat{U} \sim \hat{\nu}, \tag{29}$$

  for absolute constants c and c'. To this end, consider $\tilde{U} = \hat{U}\mathbf{1}_{\{|\hat{U}| \leq \sqrt{c_0 k}\}} \sim \tilde{\nu}$. Note that $\tilde{U}$ is k-atomic supported on $[-\sqrt{c_0 k}, \sqrt{c_0 k}]$, we have $W_1(\nu, \hat{\nu}) \leq \epsilon e^{O(k)}$ and $|m_r(\tilde{\nu}) - m_r(\hat{\nu})| \leq k\epsilon(c_1 k)^k$ for $r = 1, \ldots, 2k$. Using the triangle inequality yields that

  $$|m_r(\tilde{\nu}) - m_r(\nu')| \leq \epsilon + k\epsilon(c_1 k)^k.$$

  Now we apply Proposition 3 with $\tilde{\nu}$ and $\nu * N(0, \tau^2)$ where both $\tilde{\nu}$ and ν are k-atomic supported on $[-\sqrt{c_0 k}, \sqrt{c_0 k}]$. In the case $\tilde{\nu}$ is discrete, the dependence on k in Proposition 3 can be improved (by improving [57, (64)] in the proof) and we obtain that

$$|\sigma^2 - \hat{\sigma}^2| \le O(kn^{-\frac{1}{2k}}), \quad W_1(\nu, \tilde{\nu}) \le O(k^2 n^{-\frac{1}{4k}}).$$

Using $k \le O(\frac{\log n}{\log\log n})$, we also obtain $W_1(\nu, \hat{\nu}) \le O(k^2 n^{-\frac{1}{2k}})$ by the triangle inequality.

To obtain a confidence $1-\delta$ in (9) and (10), we can replace the empirical moments $\hat{m}_r$ by the median of $T = \log\frac{2k}{\delta}$ independent estimates similar to (24). □

4.3. *Adaptive rates.* In sections 4.1 and 4.2, we proved the statistical guarantees of our estimators under the worst-case scenario where the means can be arbitrarily close. Under separation conditions on the means (see Definition 1), our estimators automatically achieve a strictly better accuracy than the one claimed in Theorem 1. The goal in this subsection is to show those adaptive results. The key is the following adaptive version of the moment comparison theorems (cf. Propositions 1 and 2):

PROPOSITION 4. *Suppose both ν and ν' are supported on a set of ℓ atoms in $[-1,1]$, and each atom is at least γ away from all but at most ℓ' other atoms. Let $\delta = \max_{i\in[\ell-1]}|m_i(\nu) - m_i(\nu')|$. Then,*

$$W_1(\nu, \nu') \le \ell \left(\frac{\ell 4^{\ell-1}\delta}{\gamma^{\ell-\ell'-1}}\right)^{\frac{1}{\ell'}}.$$

PROPOSITION 5. *Suppose ν is supported on k atoms in $[-1,1]$ and any $t\in\mathbb{R}$ is at least γ away from all but k' atoms. Let $\delta = \max_{i\in[2k]}|m_i(\nu) - m_i(\nu')|$. Then,*

$$W_1(\nu, \nu') \le 8k \left(\frac{k 4^{2k}\delta}{\gamma^{2(k-k')}}\right)^{\frac{1}{2k'}}.$$

The adaptive result (11) in the known variance parameter case is obtained using Proposition 4 in place of Proposition 1. To deal with unknown variance parameter case, using Proposition 5, we first show the following adaptive version of Proposition 3:

PROPOSITION 6. *Under the conditions of Proposition 3, if both Gaussian mixtures both have k_0 γ-separated clusters in the sense of Definition 1, then,*

$$\sqrt{|\sigma^2 - \hat{\sigma}^2|},\ W_1(\nu, \hat{\nu}) \le O_k\left(\left(\frac{\epsilon}{\gamma^{2(k_0-1)}}\right)^{\frac{1}{2(k-k_0+1)}}\right).$$

Using these propositions, we now prove the adaptive rate of the denoised method of moments previously announced in Theorem 2:

PROOF OF THEOREM 2. By scaling it suffices to consider $M = 1$. Recall that the Gaussian mixture is assumed to have k_0 (γ, ω)-separated clusters in the sense of Definition 1, that is, there exists a partition $S_1, \ldots, S_{k_0}$ of $[k]$ such that $|\mu_i - \mu_{i'}| \geq \gamma$ for any $i \in S_\ell$ and $i' \in S_{\ell'}$ such that $\ell \neq \ell'$, and $\sum_{i \in S_\ell} w_i \geq \omega$ for each ℓ.

Let $\hat{\nu}$ be the estimated mixing distribution which satisfies $W_1(\nu, \hat{\nu}) \leq \epsilon$ by Theorem 1. Since $\gamma\omega \geq C\epsilon$ by assumption, for each S_ℓ, there exists $i \in S_\ell$ such that μ_i is within distance $c\gamma$, where $c = 1/C$, to some atom of $\hat{\nu}$. Therefore, the estimated mixing distribution $\hat{\nu}$ has k_0 $(1-2c)\gamma$-separated clusters. Denote the union of the support sets of ν and $\hat{\nu}$ by $\mathcal{S}$.

- When σ is known, each atom in $\mathcal{S}$ is $\Omega(\gamma)$ away from at least $2(k_0 - 1)$ other atoms. Then (11) follows from Proposition 4 with $\ell = 2k$ and $\ell' = (2k-1) - 2(k_0 - 1)$.
- When σ is unknown, (12) follows from a similar proof of (9) and (10) with Proposition 3 replaced by Proposition 6. □

5. Proof of moments comparison theorems.

We begin by briefly reviewing some background on polynomial interpolation, which plays a key role in the proofs.

5.1. *Polynomial interpolation, majorization, and the Neville diagram.*

Given a function f and a set of distinct points (commonly referred to as *nodes*) $\{x_0, \ldots, x_k\}$, there exists a unique polynomial P of degree k that coincides with f on every node. The interpolating polynomial P can be expressed in the *Lagrange form* as

$$P(x) = \sum_{i=0}^{k} f(x_i) \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}, \tag{30}$$

and, alternatively, in the *Newton form* as

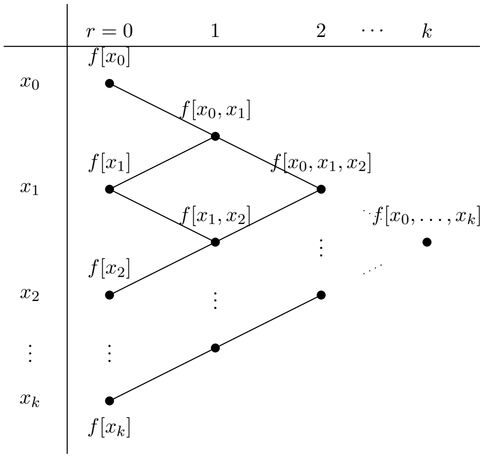
$$P(x) = a_0 + a_1(x - x_0) + \cdots + a_k(x - x_0)\cdots(x - x_{k-1}). \tag{31}$$

Let us pause to emphasize that, in numerical analysis, typically the Newton form is introduced for computational considerations so that one does not need to recompute all coefficients when an extra node is introduce [52]. Here for our theoretical analysis the Newton form turns out to be crucial, which offers better bound on the coefficients of the interpolating polynomials.

The coefficients in (31) can be successively calculated using $a_0 = f(x_0)$, $a_0 + a_1(x_1 - x_0) = f(x_1)$, etc. In general, they coincide with the divided differences $a_r = f[x_0, \ldots, x_r]$ that are recursively defined as

$$f[x_i] = f(x_i) \quad f[x_i, \ldots, x_{i+r}] = \frac{f[x_{i+1}, \ldots, x_{i+r}] - f[x_i, \ldots, x_{i+r-1}]}{x_{i+r} - x_i}. \tag{32}$$

The above recursion can be calculated by the following *Neville's diagram* (cf. [52, Section 2.1.2]):



In Neville's diagram, the r^{th} order divided differences are computed in the r^{th} column, and are determined by the previous column and the nodes. The coefficients in (31) are found in the top diagonal. In this paper Neville's diagram will be used to bound the coefficients in Newton formula (31); cf. Lemma 25 of the supplement [57].

Interpolating polynomials are the main tool to prove moment comparison theorems in Section 3. Specifically, we will interpolate step functions by polynomials in order to bound the difference of two CDFs via their moment difference. Therefore, it is crucial to have a good control over the coefficients of the interpolating polynomial. To this end, it turns out the Newton form is more convenient to use than the Lagrange form because the former takes into account the cancellation between each term in the polynomial. Indeed, in the Lagrange form (30), if two nodes are very close, then the individual terms can be arbitrarily large, even if f itself is a smooth function. In contrast, each term of (31) is stable when f is smooth since divided differences are closely related to derivatives. The following example illustrates this point:

EXAMPLE 3 (Lagrange versus Newton form). Given three points $x_1 = 0, x_2 = \epsilon, x_3 = 1$ with $f(x_1) = 1, f(x_2) = 1 + \epsilon, f(x_3) = 2$, the interpolating polynomial is $P(x) = x + 1$. The next equation gives the interpolating polynomial in Lagrange's and Newton's form respectively.

$$\begin{aligned} \text{Lagrange: } P(x) &= \frac{(x-\epsilon)(x-1)}{\epsilon} + (1+\epsilon)\frac{x(x-1)}{\epsilon(\epsilon-1)} + 2\frac{x(x-\epsilon)}{1-\epsilon}; \\ \text{Newton: } P(x) &= 1 + x + 0. \end{aligned}$$

The coefficients in the Newton form are bounded, while those in the Lagrange form blow up as $\epsilon \to 0$.

Polynomial interpolation can be generalized to interpolate the value of derivatives, known as the *Hermite interpolation*. Formally, given a function f and distinct nodes $x_0 < x_1 < \ldots < x_m$, there exists a unique polynomial P of degree k satisfying $P^{(j)}(x_i) = f^{(j)}(x_i)$ for $i = 0, \ldots, m$ and $j = 0, \ldots, k_i - 1$, where $k + 1 = \sum_{i=0}^{m} k_i$. Analogous to the Lagrange formula (30), P can be explicitly constructed with the help of the generalized Lagrange polynomials, and an explicit formula is given in [52, pp. 52–53]. The Newton form (31) can also be extended by using generalized divided differences, which, for repeated nodes, is defined as the value of the derivative:

$$f[x_i, \ldots, x_{i+r}] \triangleq \frac{f^{(r)}(x_0)}{r!}, \quad x_i = x_{i+1} = \ldots = x_{i+r}, \tag{33}$$

To this end, we define an expanded set of nodes by repeating each x_i for k_i times:

$$\underbrace{x_0 = \ldots = x_0}_{k_0} < \underbrace{x_1 = \ldots = x_1}_{k_1} < \ldots < \underbrace{x_m = \ldots = x_m}_{k_m}. \tag{34}$$

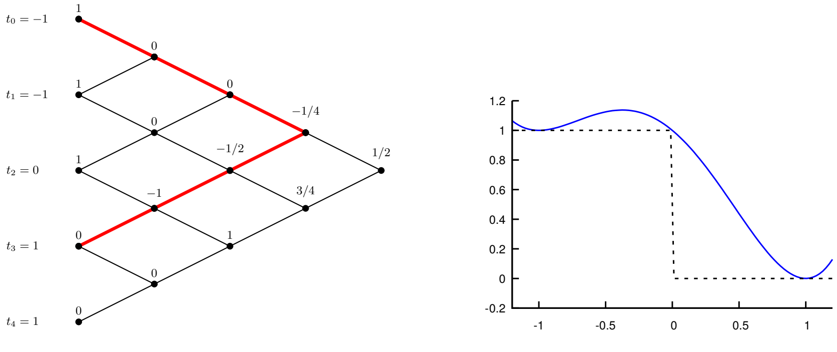
The Hermite interpolating polynomial is obtained by (31) using this new set of nodes and generalized divided differences, which can also be calculated from the Neville's diagram verbatim by replacing divided differences by derivatives whenever encountering repeated nodes. Below we give an example using Hermite interpolation to construct polynomial majorant, which will be used to prove moment comparison theorems in Section 3.

Example 4 (Hermite interpolation and polynomial majorization). Let $f(x) = \mathbf{1}_{\{x \le 0\}}$. We want to find a polynomial majorant $P \ge f$ such that $P(x) = f(x)$ on $x = \pm 1$. To this end we interpolate the values of f on $\{-1, 0, 1\}$ with the following constraints:

x	-1	0	1
$P(x)$	1	1	0
$P'(x)$	0	any	0

The resulting polynomial P has degree four and majorizes f [1, p. 65]. To see this, we note that $P'(\xi) = 0$ for some $\xi \in (-1, 0)$ by Rolle's theorem. Since $P'(-1) = P'(1) = 0$, P has no other stationary point than $-1, \xi, 1$, and thus decreases monotonically in $(\xi, 1)$. Hence, $-1, 1$ are the only local minimum points of P, and thus $P \geq f$ everywhere. The polynomial P is shown in Fig. 1(b).

To explicitly construct the polynomial, we expand the set of nodes to $-1, -1, 0, 1, 1$ according to (34). Applying Newton formula (31) with generalized divided differences from the Neville's diagram Fig. 1(a), we obtain that $P(x) = 1 - \frac{1}{4}x(x+1)^2 + \frac{1}{2}x(x+1)^2(x-1)$.



(a) Neville's diagram.

(b) Hermite interpolation.

Fig 1. *Neville's diagram and Hermite interpolation. In (a), values are recursively calculated from left to right. For example, the red thick line shows that* $f[-1, -1, 0, 1]$ *is obtained by* $\frac{-1/2-0}{1-(-1)} = -1/4$.

5.2. *Proofs of Propositions 1 and 2.* In this subsection we prove Propositions 1 and 2. As a warm-up, we start by proving Lemma 4, with the purpose of introducing the apparatus of interpolating polynomials. Throughout this section, we use

$$F_\pi(x) \triangleq \pi((-\infty, x]).$$

to denote the CDF of a distribution π.

Proof of Lemma 4. We only need to prove the "if" part.

1. Denote the union of the support sets of ν and ν' by S. Here S is of size at most $2k$. For any $t \in \mathbb{R}$, there exists a polynomial P of degree at most $2k-1$ to interpolate $x \mapsto \mathbf{1}_{\{x \leq t\}}$ on S. Since $m_i(\nu) = m_i(\nu')$ for $i = 1, ..., 2k-1$, we have

$$F_\nu(t) = \mathbb{E}_\nu[\mathbf{1}_{\{X \leq t\}}] = \mathbb{E}_\nu[P(X)] = \mathbb{E}_{\nu'}[P(X)] = \mathbb{E}_{\nu'}[\mathbf{1}_{\{X \leq t\}}] = F_{\nu'}(t).$$

2. Denote the support set of ν by $S' = \{x_1, \ldots, x_k\}$. Let $Q(x) = \prod_i (x - x_i)^2$, a non-negative polynomial of degree $2k$. Since $m_i(\nu) = m_i(\nu')$ for $i = 1, ..., 2k$, we have

$$\mathbb{E}_{\nu'}[Q(X)] = \mathbb{E}_{\nu}[Q(X)] = 0.$$

Therefore, ν' is also supported on S' and thus is k-atomic. The conclusion follows from the first case of Lemma 4. □

Next we prove Proposition 7, which is slightly stronger than Proposition 1. We provide three proofs: the first two are based on the primal (coupling) formulation of W_1 distance (17), and the third proof uses the dual formulation (16). Specifically,

- The first proof uses polynomials to interpolate step functions, whose expected values are the CDFs. The closeness of moments imply the closeness of distribution functions and thus, by (17), a small Wasserstein distance. Similar idea applies to the proof of Proposition 2 later.
- The second proof finds a polynomial that preserves the sign of the difference between two CDFs, and then relate the Wasserstein distance to the integral of that polynomial. Related idea has been used in [44, Lemma 20] which finds a polynomial that preserves the sign of the difference between two Gaussian mixture densities.
- The third proof uses polynomials to approximate 1-Lipschitz functions, whose expected values are related to the Wasserstein distance via the dual formulation (16).

The first proof is presented below, and the other two proofs are given in Section 10.1 of the supplement [57].

Proposition 7. *Let ν and ν' be discrete distributions supported on a total of ℓ atoms in $[-1, 1]$. If*

$$|m_i(\nu) - m_i(\nu')| \le \delta, \quad i = 1, \ldots, \ell - 1, \tag{35}$$

then

$$W_1(\nu, \nu') \le O\left(\ell \delta^{\frac{1}{\ell - 1}}\right).$$

First proof of Proposition 7. Suppose ν and ν' are supported on

$$S = \{t_1, \ldots, t_\ell\}, \quad t_1 < t_2 < \cdots < t_\ell. \tag{36}$$

Then, using the integral representation (17), the W_1 distance reduces to

$$W_1(\nu,\nu') = \sum_{r=1}^{\ell-1} |F_\nu(t_r) - F_{\nu'}(t_r)| \cdot |t_{r+1} - t_r|. \tag{37}$$

For each r, let $f_r(x) = \mathbf{1}_{\{x \le t_r\}}$, and P_r be the unique polynomial of degree $\ell - 1$ to interpolate f_r on S. In this way we have $f_r = P_r$ almost surely under both ν and ν', and thus

$$|F_\nu(t_r) - F_{\nu'}(t_r)| = |\mathbb{E}_\nu f_r - \mathbb{E}_{\nu'} f_r| = |\mathbb{E}_\nu P_r - \mathbb{E}_{\nu'} P_r|. \tag{38}$$

P_r can expressed using Newton formula (31) as

$$P_r(x) = 1 + \sum_{i=r+1}^{\ell} f_r[t_1, \ldots, t_i] g_{i-1}(x), \tag{39}$$

where $g_r(x) = \prod_{j=1}^r (x - t_j)$ and we used $f_r[t_1, \ldots, t_i] = 0$ for $i = 1, \ldots, r$. In (39), the absolute values of divided differences are obtained in Lemma 25 of the supplement [57]:

$$|f_r[t_1, \ldots, t_i]| \le \frac{\binom{i-2}{r-1}}{(t_{r+1} - t_r)^{i-1}}. \tag{40}$$

In the summation of (39), let $g_{i-1}(x) = \sum_{j=0}^{i-1} a_j x^j$. Since $|t_j| \le 1$ for every j, we have $\sum_{j=0}^{i-1} |a_j| \le 2^{i-1}$ (see Lemma 26 of the supplement [57]). Applying (35) yields that

$$|\mathbb{E}_\nu[g_{i-1}] - \mathbb{E}_{\nu'}[g_{i-1}]| \le \sum_{j=1}^{i-1} |a_j| \delta \le 2^{i-1}\delta. \tag{41}$$

Then we obtain from (38) and (39) that

$$|F_\nu(t_r) - F_{\nu'}(t_r)| \le \sum_{i=r+1}^{\ell} \frac{\binom{i-2}{r-1} 2^{i-1}\delta}{(t_{r+1} - t_r)^{i-1}} \le \frac{\ell 4^{\ell-1}\delta}{(t_{r+1} - t_r)^{\ell-1}}. \tag{42}$$

Also, $|F_\nu(t_r) - F_{\nu'}(t_r)| \le 1$ trivially. Therefore,

$$W_1(\nu,\nu') \le \sum_{r=1}^{\ell-1} \left(\frac{\ell 4^{\ell-1}\delta}{(t_{r+1} - t_r)^{\ell-1}} \wedge 1 \right) \cdot |t_{r+1} - t_r| \le 4e(\ell - 1)\delta^{\frac{1}{\ell-1}}, \tag{43}$$

where we used $\max\{\frac{\alpha}{x^{\ell-2}} \wedge x : x > 0\} = \alpha^{\frac{1}{\ell-1}}$ and $x^{\frac{1}{x-1}} \le e$ for $x \ge 1$. □

The proof of Proposition 2 uses a similar idea as the first proof of Proposition 7 to approximate step functions for all values of ν and ν'; however, this is clearly impossible for non-discrete ν'. For this reason, we turn from interpolation to majorization. A classical method to bound a distribution function by moments is to construct two polynomials that majorizes and minorizes a step function, respectively. Then the expectations of these two polynomials provide a sandwich bound for the distribution function. This idea is used, for example, in the proof of Chebyshev-Markov-Stieltjes inequality (cf. [1, Theorem 2.5.4]).

Proof of Proposition 2. Suppose ν is supported on $x_1 < x_2 < \ldots < x_k$. Fix $t \in \mathbb{R}$ and let $f_t(x) = \mathbf{1}_{\{x \leq t\}}$. Suppose $x_m < t < x_{m+1}$. We construct polynomial majorant and minorant using Hermite interpolation. To this end, let P_t and Q_t be the unique degree-$2k$ polynomials to interpolate f_t with the following:

	x_1	$\ldots$	x_m	t	x_{m+1}	$\ldots$	x_k
P	1	$\ldots$	1	1	0	$\ldots$	0
P'	0	$\ldots$	0	any	0	$\ldots$	0
Q	1	$\ldots$	1	0	0	$\ldots$	0
Q'	0	$\ldots$	0	any	0	$\ldots$	0

As a consequence of Rolle's theorem, $P_t \geq f_t \geq Q_t$ (cf. [1, p. 65], and an illustration in Fig. 2): Using Lagrange formula of Hermite interpolation [52, pp. 52–53], P_t and Q_t differ by

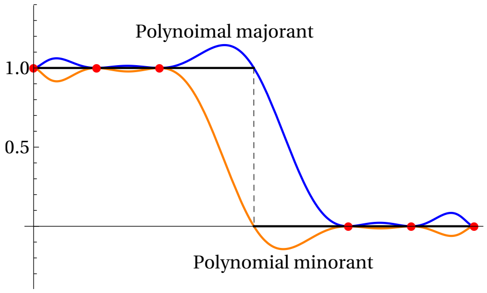


Fig 2. *Polynomial majorant P_t and minorant Q_t that coincide with the step function on 6 red points. The polynomials are of degree 12, obtained by Hermite interpolation.*

$$P_t(x) - Q_t(x) = R_t(x) \triangleq \prod_i \left(\frac{x - x_i}{t - x_i} \right)^2 .$$

The sandwich bound for f_t yields a sandwich bound for the CDFs:

$$\begin{aligned}&\mathbb{E}_{\nu'}[Q_t] \le F_{\nu'}(t) \le \mathbb{E}_{\nu'}[P_t] = \mathbb{E}_{\nu'}[Q_t] + \mathbb{E}_{\nu'}[R_t],\\ &\mathbb{E}_{\nu}[Q_t] \le F_{\nu}(t) \le \mathbb{E}_{\nu}[P_t] = \mathbb{E}_{\nu}[Q_t].\end{aligned}$$

Then the CDFs differ by

$$\begin{aligned}|F_\nu(t) - F_{\nu'}(t)| &\le (f(t) + g(t)) \wedge 1 \le f(t) \wedge 1 + g(t) \wedge 1,\\ f(t) &\triangleq |\mathbb{E}_{\nu'}[Q_t] - \mathbb{E}_\nu[Q_t]|, \quad g(t) \triangleq \mathbb{E}_{\nu'}[R_t].\end{aligned} \tag{44}$$

The conclusion will be obtained from the integral of CDF difference using (17). Since R_t is almost surely zero under ν, we also have $g(t) = |\mathbb{E}_{\nu'}[R_t] - \mathbb{E}_\nu[R_t]|$. Similar to (41), we obtain that

$$g(t) = |\mathbb{E}_{\nu'}[R_t] - \mathbb{E}_\nu[R_t]| \le \frac{2^{2k}\delta}{\prod_{i=1}^k (t - x_i)^2}.$$

Hence,

$$\int (g(t) \wedge 1)\mathrm{d}t \le \int \left(\frac{2^{2k}\delta}{\prod_{i=1}^k (t - x_i)^2} \wedge 1\right) \mathrm{d}t \le 16k\delta^{\frac{1}{2k}}, \tag{45}$$

where the last inequality is proved in Lemma 29 of the supplement [57].

Next we analyze $f(t)$. The polynomial Q_t (and also P_t) can be expressed using Newton formula (31) as

$$Q_t(x) = 1 + \sum_{i=2m+1}^{2k+1} f_t[t_1, \ldots, t_i] g_{i-1}(x), \tag{46}$$

where $t_1, \ldots, t_{2k+1}$ denotes the expanded sequence

$$x_1, x_1, \ldots, x_m, x_m, t, x_{m+1}, x_{m+1}, \ldots, x_k, x_k$$

obtained by (34), $g_r(x) = \prod_{j=1}^r (x - t_j)$, and we used $f_t[t_1, \ldots, t_i] = 0$ for $i = 1, \ldots, 2m$. In (46), the absolute values of divided differences are obtained in Lemma 25 of the supplement [57]:

$$f_t[t_1, \ldots, t_i] \le \frac{\binom{i-2}{2m-1}}{(t - x_m)^{i-1}}.$$

Using (46), and applying the upper bound for $|\mathbb{E}_\nu[g_{i-1}] - \mathbb{E}_{\nu'}[g_{i-1}]|$ in (41), we obtain that, for $x_m < t < x_{m+1}, m \ge 1$,

$$f(t) = |\mathbb{E}_{\nu'}[Q_t] - \mathbb{E}_\nu[Q_t]| \le \sum_{i=2m+1}^{2k+1} \frac{\binom{i-2}{2m-1} 2^{i-1}\delta}{(t - x_m)^{i-1}} \le \frac{k4^{2k}\delta}{(t - x_m)^{2k}}.$$

If $t < x_1$, then $Q_t = 0$ and thus $f(t) = 0$. Then, analogous to (45), we obtain that

$$\int (f(t) \wedge 1)\mathrm{d}t \le 16k\delta^{\frac{1}{2k}}. \tag{47}$$

Using (45) and (47), the conclusion follows by applying (44) to the integral representation of Wasserstein distance (17). □

Acknowledgment. We are grateful to Philippe Rigollet for bringing [23] to our attention and Harry Zhou for pointing out [7]. We thank Roger Koenker for discussions on NPMLE and sharing his experimental results. We also thank Sivaraman Balakrishnan for helpful comments on [4, 45].

SUPPLEMENTARY MATERIAL

Supplementary material for "Optimal estimation of Gaussian mixtures via denoised method of moments"
(; .pdf). Due to space constraints, additional results are given in the supplementary document [57], which contains minimax lower bounds, extensions to unbounded means, multiple dimensions, and Gaussian scale mixtures, numerical experiments, discussion on open problems, and all proofs and technical results omitted from the main article.

References.

[1] Akhiezer, N. I. (1965). *The classical moment problem: and some related questions in analysis* **5**. Oliver & Boyd.

[2] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research* **15** 2773–2832.

[3] Atkinson, K. E. (2008). *An introduction to numerical analysis*. John Wiley & Sons.

[4] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* **45** 77–120.

[5] Belkin, M. and Sinha, K. (2010). Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on* 103–112. IEEE.

[6] Chaussé, P. (2010). Computing Generalized Method of Moments and Generalized Empirical Likelihood with R. *Journal of Statistical Software* **34** 1–35.

[7] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics* 221–233.

[8] Dasgupta, S. (1999). Learning mixtures of Gaussians. In *Foundations of computer science, 1999. 40th annual symposium on* 634–644. IEEE.

[9] Daskalakis, C., Tzamos, C. and Zampetakis, M. (2017). Ten Steps of EM Suffice for Mixtures of Two Gaussians. In *Conference on Learning Theory* 704–710.

[10] Deely, J. and Kruse, R. (1968). Construction of sequences estimating the mixing distribution. *The Annals of Mathematical Statistics* **39** 286–288.

[11] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.

[12] Diaconis, P. (1987). Application of the method of moments in probability and statistics. In *Moments in mathematics*, **37** 125–139. Amer. Math. Soc.: Providence, RI.

[13] Diamond, S. and Boyd, S. (2016). CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research* **17** 1–5.

[14] Edelman, D. (1988). Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics* **16** 1609–1622.

[15] Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* Springer Science & Business Media.

[16] Gautschi, W. (2004). *Orthogonal polynomials: computation and approximation.* Oxford University Press on Demand.

[17] Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Annals of Statistics* **28** 1105–1127.

[18] 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.

[19] Golub, G. H. and Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of computation* **23** 221–230.

[20] Hall, A. R. (2005). *Generalized method of moments.* Oxford University Press.

[21] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.

[22] 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.

[23] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics* **46** 2844–2870.

[24] 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* 1021–1034. ACM.

[25] Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*, 2 ed. Cambridge University Press.

[26] Ibragimov, I. (2001). Estimation of analytic functions. *Lecture Notes-Monograph Series* 359–383.

[27] 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. ACM.

[28] Karlin, S. and Shapley, L. S. (1953). *Geometry of moment spaces* **12**. American Mathematical Soc.

[29] Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis* **41** 577–590.

[30] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* 887–906.

[31] Kim, A. K. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli* **20** 1802–1818.

[32] Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical As-*

sociation **109** 674–685.

[33] Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. *The Annals of Statistics* **45** 2218–2247.

[34] Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

[35] Krawtchouk, M. (1932). Sur le problème de moments. In *ICM Proceedings* 127–128. Available at https://www.mathunion.org/fileadmin/ICM/Proceedings/ICM1932.2/ICM1932.2.ocr.pdf.

[36] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73** 805–811.

[37] Lasserre, J. B. (2009). *Moments, positive polynomials and their applications* **1**. World Scientific.

[38] Li, J. and Schmidt, L. (2017). Robust and proper learning for mixtures of Gaussians via systems of polynomial inequalities. In *Conference on Learning Theory* 1302–1382.

[39] Lindsay, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work* 95–109. Springer.

[40] Lindsay, B. G. (1989). Moment matrices: applications in mixtures. *The Annals of Statistics* 722–740.

[41] Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics* i–163. JSTOR.

[42] Lu, Y. and Zhou, H. H. (2016). Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. *arXiv preprint arXiv:1612.02099*.

[43] Meng, X.-L. and Van Dyk, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 511–567.

[44] Moitra, A. and Valiant, G. (2010). Settling the polynomial learnability of mixtures of Gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on* 93–102. IEEE.

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

[46] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185** 71–110.

[47] Pilla, R. S. and Lindsay, B. G. (2001). Alternative EM methods for nonparametric finite mixture models. *Biometrika* **88** 535–550.

[48] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review* **26** 195–239.

[49] Schmüdgen, K. (2017). *The moment problem*. Springer.

[50] Seidel, W., Mosler, K. and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* **52** 481–487.

[51] Shohat, J. A. and Tamarkin, J. D. (1943). *The problem of moments* **1**. American Mathematical Soc.

[52] Stoer, J. and Bulirsch, R. (2002). *Introduction to Numerical Analysis*, 3rd ed. Springer-Verlag, New York, NY.

[53] Uspensky, J. V. (1937). *Introduction to mathematical probability*. McGraw-Hill.

[54] Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press, Cambridge, United Kingdom.

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

[56] Wolkowicz, H., Saigal, R. and Vandenberghe, L. (2012). *Handbook of semidefi-*

nite programming: theory, algorithms, and applications **27**. Springer Science & Business Media.

[57] Wu, Y. and Yang, P. (2019). Supplement to "Optimal estimation of Gaussian mixtures via denoised method of moments".

[58] Xu, J., Hsu, D. J. and Maleki, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems* 2676–2684.

[59] Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation* **8** 129–151.

Department of Statistics and Data Science
Yale University
New Haven, CT 06511
E-mail: yihong.wu@yale.edu

Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
E-mail: pengkuny@princeton.edu